ELSEVIER
ACADEMIC
PRESS

ENCYCLOPEDIA OF HUMAN NUTRITION

EDITED BY
BENJAMIN CABALLERO
LINDSAY ALLEN
ANDREW PRENTICE

SECOND EDITION

ENCYCLOPEDIA OF HUMAN NUTRITION

SECOND EDITION

ENCYCLOPEDIA OF HUMAN NUTRITION

SECOND EDITION

Editor-in-Chief
BENJAMIN CABALLERO

Editors
LINDSAY ALLEN
ANDREW PRENTICE

ELSEVIER
ACADEMIC
PRESS

Amsterdam Boston Heidelberg London New York Oxford
Paris San Diego San Francisco Singapore Sydney Tokyo

Elsevier Ltd., The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK

The following articles are US Government works in the public domain and not subject to copyright:

CAROTENOIDS/Chemistry, Sources and Physiology

FOOD FORTIFICATION/Developed Countries

FRUCTOSE

LEGUMES

TEA

TUBERCULOSIS/Nutrition and Susceptibility

TUBERCULOSIS/Nutritional Management

VEGETARIAN DIETS

Second edition 2005

Library of Congress Control Number: 2004113614

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

ISBN 0-12-150110-8 (set)

This book is printed on acid-free paper
Printed and bound in Spain

EDITORIAL ADVISORY BOARD

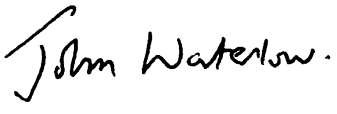
FOREWORD

Why an encyclopedia? The original Greek word means 'the circle of arts and sciences essential for a liberal education', and such a book was intended to embrace all knowledge. That was the aim of the famous Encyclopedie produced by Diderot and d'Alembert in the middle of the 18th century, which contributed so much to what has been called the Enlightenment. It is recorded that after all the authors had corrected the proofs of their contributions, the printer secretly cut out whatever he thought might give offence to the king, mutilated most of the best articles and burnt the manuscripts! Later, and less controversially, the word 'encyclopedia' came to be used for an exhaustive repertory of information on some particular department of knowledge. It is in this class that the present work falls.

In recent years the scope of Human Nutrition as a scientific discipline has expanded enormously. I used to think of it as an applied subject, relying on the basic sciences of physiology and biochemistry in much the same way that engineering relies on physics. That traditional relationship remains and is fundamental, but the field is now much wider. At one end of the spectrum epidemiological studies and the techniques on which they depend have played a major part in establishing the relationships between diet, nutritional status and health, and there is greater recognition of the importance of social factors. At the other end of the spectrum we are becoming increasingly aware of the genetic determinants of ways in which the body handles food and is able to resist adverse influences of the environment. Nutritionists are thus beginning to explore the mechanisms by which nutrients influence the expression of genes in the knowledge that nutrients are among the most powerful of all influences on gene expression. This has brought nutrition to the centre of the new 'post-genome' challenge of understanding the effects on human health of gene-environment interactions.

In parallel with this widening of the subject there has been an increase in opportunities for training and research in nutrition, with new departments and new courses being developed in universities, medical schools and schools of public health, along with a greater involvement of schoolchildren and their teachers. Public interest in nutrition is intense and needs to be guided by sound science. Governments are realizing more and more the role that nutrition plays in the prevention of disease and the maintenance of good health, and the need to develop a nutrition policy that is integrated with policies for food production.

The first edition of the Encyclopaedia of Human Nutrition established it as one of the major reference works in our discipline. The second edition has been completely revised to take account of new knowledge in our rapidly advancing field. This new edition is as comprehensive as the present state of knowledge allows, but is not overly technical and is well supplied with suggestions for further reading. All the articles have been carefully reviewed and although some of the subjects are controversial and sensitive, the publishers have not exerted the kind of political censorship that so infuriated Diderot.



J.C. Waterlow
Emeritus Professor of Human Nutrition
London School of Hygiene and Tropical Medicine
February 2005

INTRODUCTION

The science of human nutrition and its applications to health promotion continue to gain momentum. In the relatively short time since the release of the first edition of this Encyclopedia, a few landmark discoveries have had a dramatic multiplying effect over nutrition science: the mapping of the human genome, the links between molecular bioenergetics and lifespan, the influence of nutrients on viral mutation, to name a few.

But perhaps the strongest evidence of the importance of nutrition for human health comes from the fact that almost 60% of the diseases that kill humans are related to diet and lifestyle (including smoking and physical activity). These are all modifiable risk factors. As individuals and organizations intensify their efforts to reduce disease risks, the need for multidisciplinary work becomes more apparent. Today, an effective research or program team is likely to include several professionals from fields other than nutrition. For both nutrition and non-nutrition scientists, keeping up to date on the concepts and interrelationships between nutrient needs, dietary intake and health outcomes is essential. The new edition of the Encyclopedia of Human Nutrition hopes to address these needs. While rigorously scientific and up to date, EHN provides concise and easily understandable summaries on a wide variety of topics. The nutrition scientist will find that the Encyclopedia is an effective tool to "fill the void" of information in areas beyond his/her field of expertise. Professionals from other fields will appreciate the ease of alphabetical listing of topics, and the presentation of information in a rigorous but concise way, with generous aid from graphs and diagrams.

For a work that involved more than 340 authors requires, coordination and attention to detail is critical. The editors were fortunate to have the support of an excellent team from Elsevier's Major Reference Works division. Sara Gorman and Paula O'Connell initiated the project, and Tracey Mills and Samuel Coleman saw it to its successful completion.

We trust that this Encyclopedia will be a useful addition to the knowledge base of professionals involved in research, patient care, and health promotion around the globe.

Benjamin Caballero, Lindsay Allen and Andrew Prentice
Editors
April 2005

GUIDE TO USE OF THE ENCYCLOPEDIA

Structure of the Encyclopedia

The material in the Encyclopedia is arranged as a series of entries in alphabetical order. Most entries consist of several articles that deal with various aspects of a topic and are arranged in a logical sequence within an entry. Some entries comprise a single article.

To help you realize the full potential of the material in the Encyclopedia we have provided three features to help you find the topic of your choice: a Contents List, Cross-References and an Index.

1. Contents List

Your first point of reference will probably be the contents list. The complete contents lists, which appears at the front of each volume will provide you with both the volume number and the page number of the entry. On the opening page of an entry a contents list is provided so that the full details of the articles within the entry are immediately available.

Alternatively you may choose to browse through a volume using the alphabetical order of the entries as your guide. To assist you in identifying your location within the Encyclopedia a running headline indicates the current entry and the current article within that entry.

You will find 'dummy entries' where obvious synonyms exist for entries or where we have grouped together related topics. Dummy entries appear in both the contents lists and the body of the text.

Example
If you were attempting to locate material on food intake measurement via the contents list:

FOOD INTAKE *see* DIETARY INTAKE MEASUREMENT: Methodology; Validation. DIETARY SURVEYS. MEAL SIZE AND FREQUENCY

The dummy entry directs you to the Methodology article, in The Dietary Intake Measurement entry. At the appropriate location in the contents list, the page numbers for articles under Dietary Intake Measurement are given.

If you were trying to locate the material by browsing through the text and you looked up Food intake then the following information would be provided in the dummy entry:

Food Intake *see* **Dietary Intake Measurement**: Methodology; Validation. **Dietary Surveys. Meal Size and Frequency**

Alternatively, if you were looking up Dietary Intake Measurement the following information would be provided:

DIETARY INTAKE MEASUREMENT

Contents

Methodology

Validation

2. Cross-References

All of the articles in the Encyclopedia have been extensively cross-referenced.

The cross-references, which appear at the end of an article, serve three different functions. For example, at the end of the ADOLESCENTS/Nutritional Problems article, cross-references are used:

i. To indicate if a topic is discussed in greater detail elsewhere.

> *See also*: **Adolescents**: Nutritional Requirements of Adolescents. **Anemia**: Iron-Deficiency Anemia. **Calcium**: Physiology. **Eating Disorders**: Anorexia Nervosa; Bulimia Nervosa; Binge Eating. **Folic Acid**: Physiology, Dietary Sources, and Requirements. **Iron**: Physiology, Dietary Sources, and Requirements. **Obesity**: Definition, Aetiology, and Assessment. **Osteoporosis**: Nutritional Factors. **Zinc**: Physiology.

ii. To draw the reader's attention to parallel discussions in other articles.

> *See also*: **Adolescents**: Nutritional Requirements of Adolescents. **Anemia**: Iron-Deficiency Anemia. **Calcium**: Physiology. **Eating Disorders**: Anorexia Nervosa; Bulimia Nervosa; Binge Eating. **Folic Acid**: Physiology, Dietary Sources, and Requirements. **Iron**: Physiology, Dietary Sources, and Requirements. **Obesity**: Definition, Aetiology, and Assessment. **Osteoporosis**: Nutritional Factors **Zinc**: Physiology.

iii. To indicate material that broadens the discussion.

> *See also*: **Adolescents**: Nutritional Requirements of Adolescents. **Anemia**: Iron-Deficiency Anemia. **Calcium**: Physiology. **Eating Disorders**: Anorexia Nervosa; Bulimia Nervosa; Binge Eating. **Follic Acid**: Physiology, Dietary Sources, and Requirements. **Iron**: Physiology, Dietary Sources, and Requirements. **Obesity**: Definition, Aetiology, and Assessment. **Osteoporosis**: Nutritional Factors. **Zinc**: Physiology.

3. Index

The index will provide you with the page number where the material is located, and the index entries differentiate between material that is a whole article, is part of an article or is data presented in a figure or table. Detailed notes are provided on the opening page of the index.

4. Contributors

A full list of contributors appears at the beginning of each volume.

CONTRIBUTORS

E Abalos
Centro Rosarino de Estudios Perinatales
Rosario, Argentina

A Abi-Hanna
Johns Hopkins School of Medicine
Baltimore, MD, USA

L S Adair
University of North Carolina
Chapel Hill, NC, USA

A Ahmed
Obetech Obesity Research Center
Richmond, VA, USA

B Ahrén
Lund University
Lund, Sweden

J Akré
World Health Organization, Geneva, Switzerland

A J Alberg
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD, USA

L H Allen
University of California at Davis
Davis, CA, USA

D Anderson
University of Bradford
Bradford, UK

J J B Anderson
University of North Carolina
Chapel Hill, NC, USA

R A Anderson
US Department of Agriculture
Beltsville, MD, USA

L J Appel
Johns Hopkins University
Baltimore, MD, USA

A Ariño
University of Zaragoza
Zaragoza, Spain

M J Arnaud
Nestle S.A.
Vevey, Switzerland

E W Askew
University of Utah
Salt Lake City, UT, USA

R L Atkinson
Obetech Obesity Research Center
Richmond, VA, USA

S A Atkinson
McMaster University
Hamilton, ON, Canada

L S A Augustin
University of Toronto
Toronto, ON, Canada

D J Baer
US Department of Agriculture
Beltsville, MD, USA

A Baqui
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD, USA

Y Barnett
Nottingham Trent University
Nottingham, UK

G E Bartley
Agricultural Research Service
Albany, CA, USA

C J Bates
MRC Human Nutrition Research
Cambridge, UK

J A Beltrán
University of Zaragoza
Zaragoza, Spain

A E Bender
Leatherhead, UK

D A Bender
University College London
London, UK

I F F Benzie
The Hong Kong Polytechnic University
Hong Kong SAR, China

C D Berdanier
University of Georgia
Athens, GA, USA

R Bhatia
United Nations World Food Programme
Rome, Italy

Z A Bhutta
The Aga Khan University
Karachi, Pakistan

J E Bines
University of Melbourne
Melbourne, VIC, Australia

J Binkley
Vanderbilt Center for Human Nutrition
Nashville, TN, USA

R Black
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD, USA

J E Blundell
University of Leeds
Leeds, UK

A T Borchers
University of California at Davis
Davis, CA, USA

C Boreham
University of Ulster at Jordanstown
Jordanstown, UK

F Branca
Istituto Nazionale di Ricerca per gli Alimenti e la Nutrizione
Rome, Italy

J Brand-Miller
University of Sydney
Sydney, NSW, Australia

A Briend
Institut de Recherche pour le Développement
Paris, France

P Browne
St James's Hospital
Dublin, Ireland

I A Brownlee
University of Newcastle
Newcastle-upon-Tyne, UK

H Brunner
Centre Hospitalier Universitaire Vaudois
Lausanne, Switzerland

A J Buckley
University of Cambridge
Cambridge, UK

H H Butchko
Exponent, Inc.
Wood Dale, IL, USA

J Buttriss
British Nutrition Foundation
London, UK

B Caballero
Johns Hopkins Bloomberg School of Public Health and
Johns Hopkins University
Baltimore, MD, USA

E A Carrey
Institute of Child Health
London, UK

A Cassidy
School of Medicine
University of East Anglia
Norwich, UK

G E Caughey
Royal Adelaide Hospital
Adelaide, SA, Australia

J P Cegielski
Centers for Disease Control and Prevention
Atlanta, GA, USA

C M Champagne
Pennington Biomedical Research Center
Baton Rouge, LA, USA

S C Chen
US Department of Agriculture
Beltsville, MD, USA

L Cheskin
Johns Hopkins University
Baltimore, MD, USA

S Chung
Columbia University
New York, NY, USA

L G Cleland
Royal Adelaide Hospital
Adelaide, SA, Australia

L Cobiac
CSIRO Health Sciences and Nutrition
Adelaide, SA, Australia

G A Colditz
Harvard Medical School
Boston, MA, USA

T J Cole
Institute of Child Health
London, UK

L A Coleman
Marshfield Clinic Research Foundation
Marshfield, WI, USA

S Collier
Children's Hospital, Boston, Harvard Medical School,
and Harvard School of Public Health
Boston, MA, USA

M Collins
Muckamore Abbey Hospital
Antrim, UK

K G Conner
Johns Hopkins Hospital
Baltimore, MD, USA

K C Costas
Children's Hospital Boston
Boston, MA, USA

R C Cottrell
The Sugar Bureau
London, UK

W A Coward
MRC Human Nutrition Research
Cambridge, UK

J M Cox
Johns Hopkins Hospital
Baltimore, MD, USA

S Cox
London School of Hygiene and Tropical Medicine
London, UK

P D'Acapito
Istituto Nazionale di Ricerca per gli Alimenti e la Nutrizione
Rome, Italy

S Daniell
Vanderbilt Center for Human Nutrition
Nashville, TN, USA

O Dary
The MOST Project
Arlington, VA, USA

T J David
University of Manchester
Manchester, UK

C P G M de Groot
Wageningen University
Wageningen, The Netherlands

M de Onis
World Health Organization
Geneva, Switzerland

M C de Souza
Universidad de Mogi das Cruzes
São Paulo, Brazil

R de Souza
University of Toronto
Toronto, ON, Canada

C H C Dejong
University Hospital Maastricht
Maastricht, The Netherlands

L Demeshlaira
Emory University
Atlanta, GA, USA

K G Dewey
University of California at Davis
Davis, CA, USA

H L Dewraj
The Aga Khan University
Karachi, Pakistan

C Doherty
MRC Keneba
The Gambia

C M Donangelo
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil

A Dornhorst
Imperial College at Hammersmith Hospital
London, UK

E Dowler
University of Warwick
Coventry, UK

J Dowsett
St Vincent's University Hospital
Dublin, Ireland

A K Draper
University of Westminster
London, UK

M L Dreyfuss
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD, USA

R D'Souza
Queen Mary's, University of London
London, UK

C Duggan
Harvard Medical School
Boston, MA, USA

A G Dulloo
University of Fribourg
Fribourg, Switzerland

E B Duly
Ulster Hospital
Belfast, UK

J L Dupont
Florida State University
Tallahassee, FL, USA

J Dwyer
Tufts University
Boston, MA, USA

J Eaton–Evans
University of Ulster
Coleraine, UK

C A Edwards
University of Glasgow
Glasgow, UK

M Elia
University of Southampton
Southampton, UK

P W Emery
King's College London
London, UK

J L Ensunsa
University of California at Davis
Davis, CA, USA

C Feillet-Coudray
National Institute for Agricultural Research
Clermont-Ferrand, France

J D Fernstrom
University of Pittsburgh
Pittsburgh, PA, USA

M H Fernstrom
University of Pittsburgh
Pittsburgh, PA, USA

F Fidanza
University of Rome Tor Vergata
Rome, Italy

P Fieldhouse
The University of Manitoba
Winnipeg, MB, Canada

N Finer
Luton and Dunstable Hospital NHS Trust
Luton, UK

J Fiore
University of Westminster
London, UK

H C Freake
University of Connecticut
Storrs, CT, USA

J Freitas
Tufts University
Boston, MA, USA

R E Frisch
Harvard Center for Population and Development Studies
Cambridge, MA, USA

G Frost
Imperial College at Hammersmith Hospital
London, UK

G Frühbeck
Universidad de Navarra
Pamplona, Spain

D Gallagher
Columbia University
New York, NY, USA

L Galland
Applied Nutrition Inc.
New York, NY, USA

C Geissler
King's College London
London, UK

M E Gershwin
University of California at Davis
Davis, CA, USA

H Ghattas
London School of Hygiene and Tropical Medicine
London, UK

E L Gibson
University College London
London, UK

T P Gill
University of Sydney
Sydney, NSW, Australia

W Gilmore
University of Ulster
Coleraine, UK

G R Goldberg
MRC Human Nutrition Research
Cambridge, UK

J Gómez-Ambrosi
Universidad de Navarra
Pamplona, Spain

J M Graham
University of California at Davis
Davis, CA, USA

J Gray
Guildford, UK

J P Greaves
London, UK

M W Green
Aston University
Birmingham, UK

R Green
University of California
Davis, CA, USA

R F Grimble
University of Southampton
Southampton, UK

M Grønbæk
National Institute of Public Health
Copenhagen, Denmark

J D Groopman
Johns Hopkins University
Baltimore MD, USA

S M Grundy
University of Texas Southwestern Medical Center
Dallas, TX, USA

M A Grusak
Baylor College of Medicine
Houston, TX, USA

M Gueimonde
University of Turku
Turku, Finland

C S Gulotta
Johns Hopkins University and Kennedy Krieger Institute
Baltimore, MD, USA

P Haggarty
Rowett Research Institute
Aberdeen, UK

J C G Halford
University of Liverpool
Liverpool, UK

C H Halsted
University of California at Davis
Davis, CA, USA

J Hampsey
Johns Hopkins School of Medicine
Baltimore, MD, USA

E D Harris
Texas A&M University
College Station, TX, USA

Z L Harris
Johns Hopkins Hospital and School of Medicine
Baltimore, MD, USA

P J Havel
University of California at Davis
Davis, CA, USA

W W Hay Jr
University of Colorado Health Sciences Center
Aurora, CO, USA

R G Heine
University of Melbourne
Melbourne, VIC, Australia

R Heinzen
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD, USA

A Herrera
University of Zaragoza
Zaragoza, Spain

B S Hetzel
Women's and Children's Hospital
North Adelaide, SA, Australia

A J Hill
University of Leeds
Leeds, UK

S A Hill
Southampton General Hospital
Southampton, UK

G A Hitman
Queen Mary's, University of London
London, UK

J M Hodgson
University of Western Australia
Perth, WA, Australia

M F Holick
Boston University Medical Center
Boston, MA, USA

C Hotz
National Institute of Public Health
Morelos, Mexico

R Houston
Emory University
Atlanta, GA, USA

H-Y Huang
Johns Hopkins University
Baltimore, MD, USA

J R Hunt
USDA-ARS Grand Forks Human Nutrition Research Center
Grand Forks, ND, USA

R Hunter
King's College London
London, UK

P Hyland
Nottingham Trent University
Nottingham, UK

B K Ishida
Agricultural Research Service
Albany, CA, USA

J Jacquet
University of Geneva
Geneva, Switzerland

M J James
Royal Adelaide Hospital
Adelaide, SA, Australia

W P T James
International Association for the Study of Obesity/
International Obesity Task Force Offices
London, UK

A G Jardine
University of Glasgow
Glasgow, UK

S A Jebb
MRC Human Nutrition Research
Cambridge, UK

K N Jeejeebhoy
University of Toronto
Toronto, ON, Canada

D J A Jenkins
University of Toronto
Toronto, ON, Canada

G L Jensen
Vanderbilt Center for Human Nutrition
Nashville, TN, USA

I T Johnson
Institute of Food Research
Norwich, UK

P A Judd
University of Central Lancashire
Preston, UK

M A Kalarchian
University of Pittsburgh
Pittsburgh, PA, USA

R M Katz
Johns Hopkins University School of Medicine and Mount Washington Pediatric Hospital
Baltimore, MD, USA

C L Keen
University of California at Davis
Davis, CA, USA

N L Keim
US Department of Agriculture
Davis, CA, USA

E Kelly
Harvard Medical School
Boston, MA, USA

C W C Kendall
University of Toronto
Toronto, ON, Canada

T W Kensler
Johns Hopkins University
Baltimore, MD, USA

J E Kerstetter
University of Connecticut
Storrs, CT, USA

M Kiely
University College Cork
Cork, Ireland

P Kirk
University of Ulster
Coleraine, UK

S F L Kirk
University of Leeds
Leeds, UK

P N Kirke
The Health Research Board
Dublin, Ireland

G L Klein
University of Texas Medical Branch at Galveston
Galveston TX, USA

R D W Klemm
Johns Hopkins University
Baltimore, MD, USA

D M Klurfeld
US Department of Agriculture
Beltville, MD, USA

P G Kopelman
Queen Mary's, University of London
London, UK

J Krick
Kennedy–Krieger Institute
Baltimore, MD, USA

D Kritchevsky
Wistar Institute
Philadelphia, PA, USA

R Lang
University of Teeside
Middlesbrough, UK

A Laurentin
Universidad Central de Venezuela
Caracas, Venezuela

A Laverty
Muckamore Abbey Hospital
Antrim, UK

M Lawson
Institute of Child Health
London, UK

F E Leahy
University of Auckland
Auckland, New Zealand

A R Leeds
King's College London
London, UK

J Leiper
University of Aberdeen
Aberdeen, UK

M D Levine
University of Pittsburgh
Pittsburgh, PA, USA

A H Lichtenstein
Tufts University
Boston MA, USA

E Lin
Emory University
Atlanta, GA, USA

L Lissner
Sahlgrenska Academy at Göteborg University
Göteborg, Sweden

C Lo
Children's Hospital, Boston, Harvard Medical School, and Harvard School of Public Health
Boston, MA, USA

P A Lofgren
Oak Park, IL, USA

B Lönnerdal
University of California at Davis
Davis, CA, USA

M J Luetkemeier
Alma College
Alma, MI, USA

Y C Luiking
University Hospital Maastricht
Maastricht, The Netherlands

P G Lunn
University of Cambridge
Cambridge, UK

C K Lutter
Pan American Health Organization
Washington, DC, USA

A MacDonald
The Children's Hospital
Birmingham, UK

A Maqbool
The Children's Hospital of Philadelphia
Philadelphia, PA, USA

M D Marcus
University of Pittsburgh
Pittsburgh, PA, USA

E Marietta
The Mayo Clinic College of Medicine
Rochester, MN, USA

P B Mark
University of Glasgow
Glasgow, UK

V Marks
University of Surrey
Guildford, UK

D L Marsden
Children's Hospital Boston
Boston, MA, USA

R J Maughan
Loughborough University
Loughborough, UK

K C McCowen
Beth Israel Deaconess Medical Center and Harvard Medical School
Boston, MA, USA

S S McDonald
Raleigh, NC, USA

S McLaren
London South Bank University
London, UK

J L McManaman
University of Colorado
Denver, CO, USA

D N McMurray
Texas A&M University
College Station, TX, USA

D J McNamara
Egg Nutrition Center
Washington, DC, USA

J McPartlin
Trinity College
Dublin, Ireland

R P Mensink
Maastricht University
Maastricht, The Netherlands

M Merialdi
World Health Organization
Geneva, Switzerland

A R Michell
St Bartholomew's Hospital
London, UK

J W Miller
UC Davis Medical Center
Sacramento, CA, USA

P Miller
Kennedy–Krieger Institute
Baltimore, MD, USA

D J Millward
University of Surrey
Guildford, UK

D M Mock
University of Arkansas for Medical Sciences
Little Rock, AR, USA

N Moore
John Hopkins School of Medicine
Baltimore, MD, USA

J O Mora
The MOST Project
Arlington, VA, USA

T Morgan
University of Melbourne
Melbourne, VIC, Australia

T A Mori
University of Western Australia
Perth, WA, Australia

J E Morley
St Louis University
St Louis, MO, USA

P A Morrissey
University College Cork
Cork, Ireland

M H Murphy
University of Ulster at Jordanstown
Jordanstown, UK

S P Murphy
University of Hawaii
Honolulu, HI, USA

J Murray
The Mayo Clinic College of Medicine
Rochester, MN, USA

R Nalubola
Center for Food Safety and Applied Nutrition,
US Food and Drug Administration, MD, USA

J L Napoli
University of California
Berkeley, CA, USA

V Nehra
The Mayo Clinic College of Medicine
Rochester, MN, USA

B Nejadnik
Johns Hopkins University
Baltimore, MD, USA

M Nelson
King's College London
London, UK

P Nestel
International Food Policy Research Institute
Washington, DC, USA

L M Neufeld
National Institute of Public Health
Cuernavaca, Mexico

M C Neville
University of Colorado
Denver, CO, USA

F Nielsen
Grand Forks Human Nutrition Research Center
Grand Forks, ND, USA

N Noah
London School of Hygiene and Tropical Medicine
London, UK

K O O'Brien
Johns Hopkins University
Baltimore, MD, USA

S H Oh
Johns Hopkins General Clinical Research Center
Baltimore, MD, USA

J M Ordovas
Tufts University
Boston, MA, USA

S E Ozanne
University of Cambridge
Cambridge, UK

D M Paige
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD, USA

J P Pearson
University of Newcastle
Newcastle-upon-Tyne, UK

S S Percival
University of Florida
Gainesville, FL, USA

T Peters
King's College Hospital
London, UK

B J Petersen
Exponent, Inc.
Washington DC, USA

J C Phillips
BIBRA International Ltd
Carshalton, UK

M F Picciano
National Institutes of Health
Bethesda, MD, USA

A Pietrobelli
Verona University Medical School
Verona, Italy

S Pin
Johns Hopkins Hospital and School of Medicine
Baltimore, MD, USA

B M Popkin
University of North Carolina
Chapel Hill, NC, USA

E M E Poskitt
London School of Hygiene and Tropical Medicine
London, UK

A D Postle
University of Southampton
Southampton, UK

J Powell-Tuck
Queen Mary's, University of London
London, UK

V Preedy
King's College London
London, UK

N D Priest
Middlesex University
London, UK

R Rajendram
King's College London
London, UK

A Raman
University of Wisconsin–Madison
Madison, WI, USA

H A Raynor
Brown University
Providence, RI, USA

Y Rayssiguier
National Institute for Agricultural Research
Clermont-Ferrand, France

L N Richardson
United Nations World Food Programme
Rome, Italy

F J Rohr
Children's Hospital Boston
Boston, MA, USA

A R Rolla
Harvard Medical School
Boston, MA, USA

P Roncalés
University of Zaragoza
Zaragoza, Spain

A C Ross
The Pennsylvania State University
University Park, PA, USA

R Roubenoff
Millennium Pharmaceuticals, Inc.
Cambridge, MA, USA and Tufts University
Boston, MA, USA

D Rumsey
University of Sheffield
Sheffield, UK

C H S Ruxton
Nutrition Communications
Cupar, UK

J M Saavedra
John Hopkins School of Medicine
Baltimore, MD, USA

J E Sable
University of California at Davis
Davis, CA, USA

M J Sadler
MJSR Associates
Ashford, UK

N R Sahyoun
University of Maryland
College Park, MD, USA

S Salminen
University of Turku
Turku, Finland

M Saltmarsh
Alton, UK

J M Samet
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD, USA

C P Sánchez-Castillo
National Institute of Medical Sciences and Nutrition
Salvador Zubirán, Tlalpan, Mexico

M Santosham
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD, USA

C D Saudek
Johns Hopkins School of Medicine
Baltimore, MD, USA

A O Scheimann
Johns Hopkins School of Medicine
Baltimore, MD, USA

B Schneeman
University of California at Davis
Davis, CA, USA

D A Schoeller
University of Wisconsin–Madison
Madison, WI, USA

L Schuberth
Kennedy Krieger Institute
Baltimore, MD, USA

K J Schulze
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD, USA

Y Schutz
University of Lausanne
Lausanne, Switzerland

K B Schwarz
Johns Hopkins School of Medicine
Baltimore, MD, USA

J M Scott
Trinity College Dublin
Dublin, Ireland

C Shaw
Royal Marsden NHS Foundation Trust
London, UK

J Shedlock
Johns Hopkins Hospital and School of Medicine
Baltimore, MD, USA

S M Shirreffs
Loughborough University
Loughborough, UK

R Shrimpton
Institute of Child Health
London, UK

H A Simmonds
Guy's Hospital
London, UK

A P Simopoulos
The Center for Genetics, Nutrition and Health
Washington, DC, USA

R J Smith
Brown Medical School
Providence, RI, USA

P B Soeters
University Hospital Maastricht
Maastricht, The Netherlands

N Solomons
Center for Studies of Sensory Impairment, Aging and Metabolism (CeSSIAM)
Guatemala City, Guatemala

J A Solon
MRC Laboratories Gambia
Banjul, The Gambia

K Srinath Reddy
All India Institute of Medical Sciences
New Delhi, India

S Stanner
British Nutrition Foundation
London, UK

J Stevens
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA

J J Strain
University of Ulster
Coleraine, UK

R J Stratton
University of Southampton
Southampton, UK

R J Stubbs
The Rowett Research Institute
Aberdeen, UK

C L Stylianopoulos
Johns Hopkins University
Baltimore, MD, USA

A W Subudhi
University of Colorado at Colorado
Colorado Springs, CO, USA

J Sudagani
Queen Mary's, University of London
London, UK

S A Tanumihardjo
University of Wisconsin-Madison
Madison, WI, USA

J A Tayek
Harbor–UCLA Medical Center
Torrance, CA, USA

E H M Temme
University of Leuven
Leuven, Belgium

H S Thesmar
Egg Nutrition Center
Washington, DC, USA

B M Thomson
Rowett Research Institute
Aberdeen, UK

D I Thurnham
University of Ulster
Coleraine, UK

L Tolentino
National Institute of Public Health
Cuernavaca, Mexico

D L Topping
CSIRO Health Sciences and Nutrition
Adelaide, SA, Australia

B Torun
Center for Research and Teaching in Latin America (CIDAL)
Guatemala City, Guatemala

M G Traber
Oregon State University
Corvallis, OR, USA

T R Trinick
Ulster Hospital
Belfast, UK

K P Truesdale
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA

N M F Trugo
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil

P M Tsai
Harvard Medical School
Boston, MA, USA

K L Tucker
Tufts University
Boston, MA, USA

O Tully
St Vincent's University Hospital
Dublin, Ireland

E C Uchegbu
Royal Hallamshire Hospital
Sheffield, UK

M C G van de Poll
University Hospital Maastricht
Maastricht, The Netherlands

W A van Staveren
Wageningen University
Wageningen, The Netherlands

J Villar
World Health Organization
Geneva, Switzerland

M L Wahlqvist
Monash University
Victoria, VIC, Australia

A F Walker
The University of Reading
Reading, UK

P A Watkins
Kennedy Krieger Institute and Johns Hopkins University School of Medicine
Baltimore, MD, USA

A A Welch
University of Cambridge
Cambridge, UK

R W Welch
University of Ulster
Coleraine, UK

K P West Jr
Johns Hopkins University
Baltimore, MD, USA

S Whybrow
The Rowett Research Institute
Aberdeen, UK

D H Williamson
Radcliffe Infirmary
Oxford, UK

M-M G Wilson
St Louis University
St Louis, MO, USA

R R Wing
Brown University
Providence, RI, USA

C K Winter
University of California at Davis
Davis, CA, USA

H Wiseman
King's College London
London, UK

M Wolraich
Vanderbilt University
Nashville, TN, USA

R J Wood
Tufts University
Boston, MA, USA

X Xu
Johns Hopkins Hospital and School of Medicine
Baltimore, MD, USA

Z Yang
University of Wisconsin-Madison
Madison, WI, USA

A A Yates
ENVIRON Health Sciences
Arlington, VA, USA

S H Zeisel
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA

X Zhu
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA

S Zidenberg-Cherr
University of California at Davis
Davis, CA, USA

T R Ziegler
Emory University
Atlanta, GA, USA

CONTENTS

VOLUME 1

B

C

D

VOLUME 2

E

F

G

H

VOLUME 3

I

K

L

M

N

O

P

VOLUME 4

R

S

T

U

V

W

Y

Z

A

Acids *see* **Electrolytes**: Acid-Base Balance

ADIPOSE TISSUE

G Frühbeck and J Gómez-Ambrosi, Universidad de Navarra, Pamplona, Spain

Introduction

The role of white adipose tissue (WAT) in storing and releasing lipids for oxidation by skeletal muscle and other tissues became so firmly established decades ago that a persistent lack of interest hindered the study of the extraordinarily dynamic behavior of adipocytes. However, disentangling the neuroendocrine systems that regulate energy homeostasis and adiposity has jumped to a first-priority challenge, with the recognition of obesity as one of the major public health problems. Strictly speaking, obesity is not defined as an excess of body weight but as an increased adipose tissue accretion, to the extent that health may be adversely affected. Therefore, in the last decades, adipose tissue has become the research focus of biomedical scientists for epidemiological, pathophysiological, and molecular reasons. Although the primary role of adipocytes is to store triglycerides during periods of caloric excess and to mobilize this reserve when expenditure exceeds intake, it is now widely recognized that adipose tissue lies at the heart of a complex network that participates in the regulation of a variety of quite diverse biological functions (**Figure 1**).

Development

Adipose tissue develops extensively in homeotherms with the proportion to body weight varying greatly among species. Adipocytes differentiate from stellate or fusiform precursor cells of mesenchymal origin. There are two processes of adipose tissue formation. In the primary fat formation, which takes place relatively early (in human fetuses the first traces of a fat organ are detectable between the 14th and 16th weeks of prenatal life), gland-like aggregations of epitheloid precursor cells, called lipoblasts or preadipocytes, are laid down in specific locations and accumulate multiple lipid droplets becoming brown adipocytes. The secondary fat formation takes place later in fetal life (after the 23rd week of gestation) as well as in the early postnatal period, whereby the differentiation of other fusiform precursor cells that accumulate lipid to ultimately coalesce into a single large drop per cell leads to the dissemination of fat depots formed by unilocular white adipocytes in many areas of connective tissue. Adipose tissue may be partitioned by connective tissue septa into lobules. The number of fat lobules remains constant, while in the subsequent developmental phases the lobules continuously increase in size. At the sites of early fat development, a multilocular morphology of adipocytes predominates, reflecting the early developmental stage. Microscopic studies have shown that the second trimester may be a critical period for the development of obesity in later life. At the beginning of the third trimester, adipocytes are present in the main fat depots but are still relatively small. During embryonic development it is important to emphasize the temporospacial tight coordination of angiogenesis with the formation of fat cell clusters. At birth, body fat has been reported to

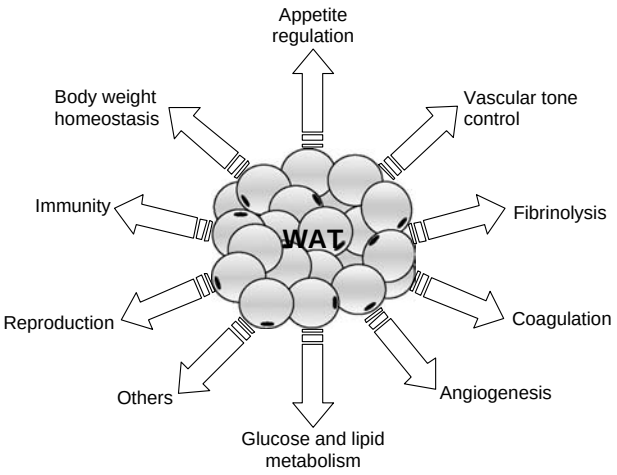


Figure 1 Dynamic view of white adipose tissue based on the pleiotropic effects on quite diverse physiological functions.

account for approximately 16% of total body weight (with brown fat constituting 2–5%) with an increase in body fat of around 0.7–2.8 kg during the first year of life.

Adipogenesis, i.e., the development of adipose tissue, varies according to sex and age. Furthermore, the existence of sensitive periods for changes in adipose tissue cellularity throughout life has been postulated. In this regard, two peaks of accelerated adipose mass enlargement have been established, namely after birth and between 9 and 13 years of age. The capacity for cell proliferation and differentiation is highest during the first year of life, while it is less pronounced in the years before puberty. Thereafter, the rate of cell proliferation slows down during adolescence and, in weight stable individuals, remains fairly constant throughout adulthood. In case of a maintained positive energy balance adipose mass expansion takes places initially by an enlargement of the existing fat cells. The perpetuation of this situation ends up in severe obesity where the total fat cell number can be easily trebled. Childhood-onset obesity is characterized by a combination of fat cell hyperplasia and hypertrophy, whereas in adult-onset obesity a hypertrophic growth predominates. However, it has been recently shown that adult humans are capable of new adipocyte formation, with fat tissue containing a significant proportion of cells with the ability to undergo differentiation. Interestingly, the hyperplasic growth of fat cells in adults does not take place until the existing adipocytes reach a critical cell size.

Initially, excess energy storage starts as hypertrophic obesity resulting from the accumulation of excess lipid in a normal number of unilocular adipose cells. In this case, adipocytes may be four times their normal size. If the positive energy balance is maintained, a hyperplasic or hypercellular obesity characterized by a greater than normal number of cells is developed. Recent observations regarding the occurrence of apoptosis in WAT have changed the traditional belief that acquisition of fat cells is irreversible. The adipose lineage originates from multipotent mesenchymal stem cells that develop into adipoblasts (**Figure 2**). Commitment of these adipoblasts gives rise to preadipose cells (preadipocytes), which are cells that have expressed early but not late markers and have yet to accumulate triacylglycerol stores (**Figure 3**). Multipotent stem cells and adipoblasts, which are found during embryonic development, are still present postnatally. The relationship between brown and white fat during development has not been completely solved. Brown adipocytes can be detected among all white fat depots in variable amounts depending on species, localization, and environmental temperature. The transformation of characteristic brown adipocytes into white fat cells can take place rapidly in numerous species and depots during postnatal development.

The morphological and functional changes that take place in the course of adipogenesis represent a shift in transcription factor expression and activity leading from a primitive, multipotent state to a final phenotype characterized by alterations in cell shape and lipid accumulation. Various redundant signaling pathways and transcription factors directly influence fat cell development by converging in the upregulation of PPARγ, which embodies a common and essential regulator of adipogenesis as well as of adipocyte hypertrophy. Among the broad panoply of transcription factors, C/EBPs and the basic helix-loop-helix family (ADD1/SREBP-1c) also stand out together with their link with the existing nutritional status. The transcriptional repression of adipogenesis includes both active and passive mechanisms. The former directly interferes with the transcriptional machinery, while the latter is based on the binding of negative regulators to yield inactive forms of known activators.

Hormones, cytokines, growth factors, and nutrients influence the dynamic changes related to adipose tissue mass as well as its pattern of distribution (**Figure 4**). The responsiveness of fat cells to neurohumoral signals may vary according to peculiarities in the adipose lineage stage at the moment of exposure. Moreover, the simultaneous presence of some adipogenic factors at specific threshold concentrations may be a necessary requirement to trigger terminal differentiation.

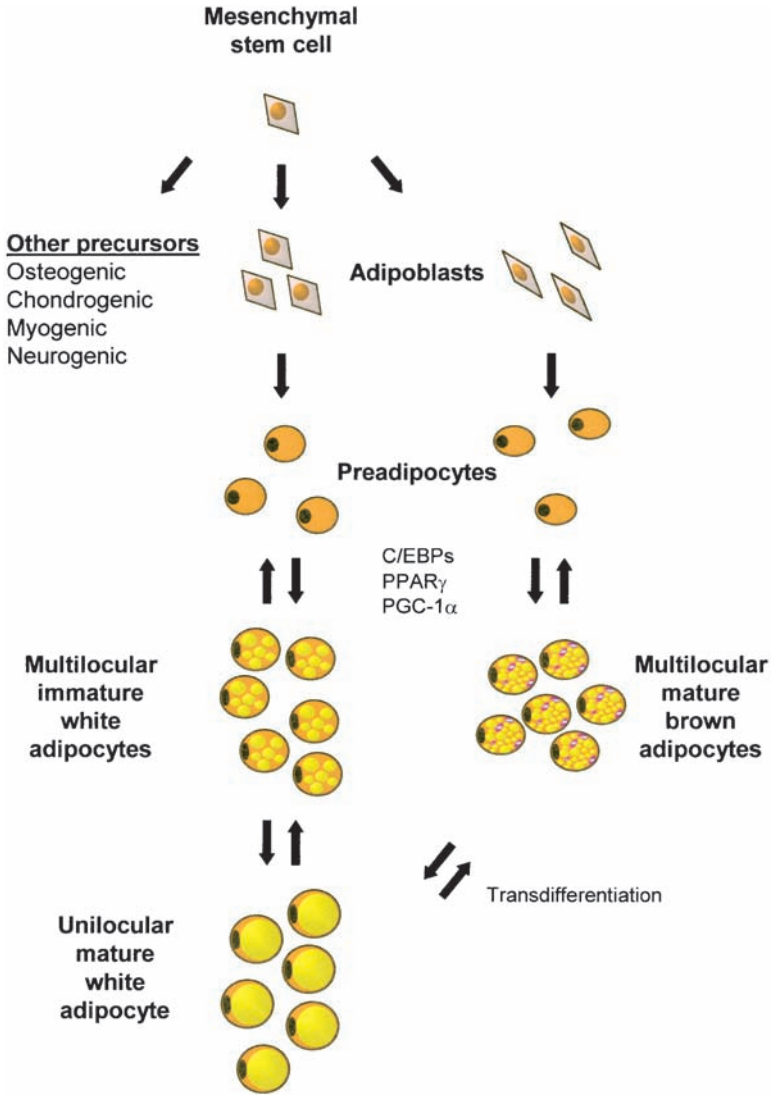


Figure 2 Schematic diagram of the histogenesis of white and brown adipocytes. C/EBPs, CCAAT/enhancer binding proteins; PGC-1α, peroxisome proliferator-activated receptor-γ coactivator-1; PPARγ, peroxisome proliferator-activated receptor-γ.

Structure

Adipose tissue is a special loose connective tissue dominated by adipocytes. The name of these cells is based on the presence of a large lipid droplet with 'adipo' derived from the Latin *adeps* meaning 'pertaining to fat.' In adipose tissue, fat cells are individually held in place by delicate reticular fibers clustering in lobular masses bounded by fibrous septa surrounded by a rich capillary network. In adults, adipocytes may comprise around 90% of adipose mass accounting only for roughly 25% of the total cell population. Thus, adipose tissue itself is composed not only of adipocytes, but also other cell types called the stroma-vascular fraction, comprising blood cells, endothelial cells, pericytes, and adipose precursor cells among others (**Figure 5**); these account for the remaining 75% of the total cell population, representing a wide range of targets for extensive autocrine-paracrine cross-talk.

Adipocytes, which are typically spherical and vary enormously in size (20–200 μm in diameter, with variable volumes ranging from a few picoliters to about 3 nanoliters), are embedded in a connective tissue matrix and are uniquely adapted to store and release energy. Surplus energy is assimilated by adipocytes and stored as lipid droplets. The stored fat is composed mainly of triacylglycerols (about 95% of the total lipid content comprised principally of oleic and palmitic acids) and to a smaller degree of diacylglycerols, phospholipids, unesterified fatty acids, and cholesterol. To accommodate the lipids adipocytes are capable of changing their

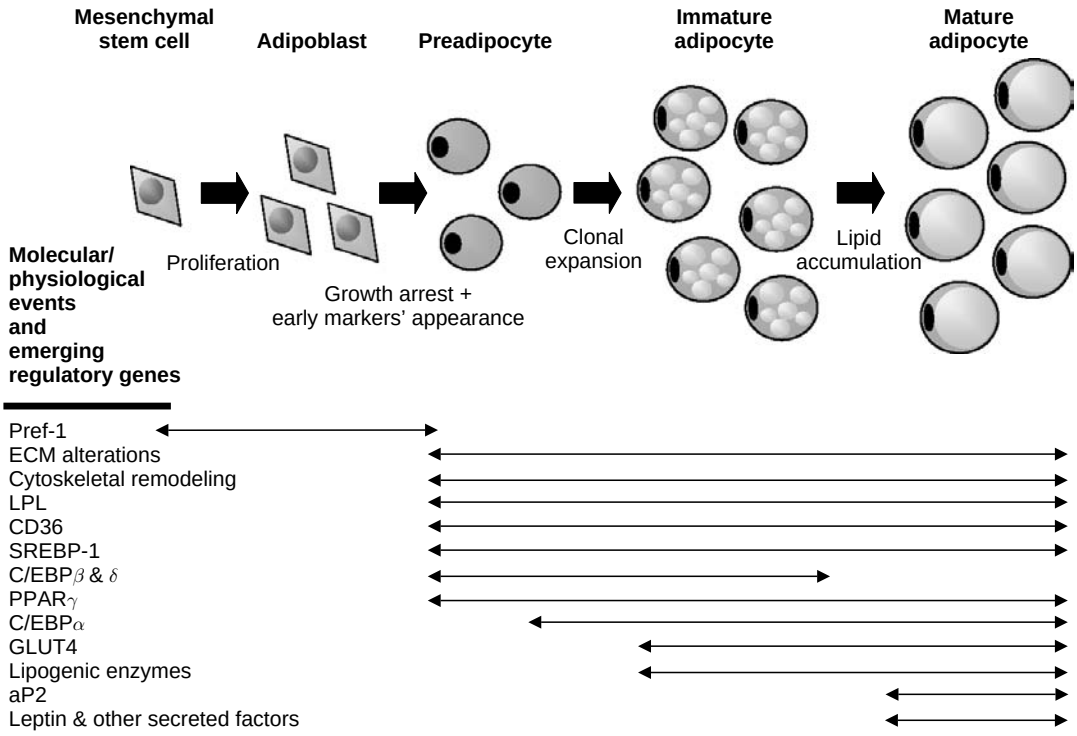


Figure 3 Multistep process of adipogenesis together with events and participating regulatory elements. aP2, adipocyte fatty acid binding protein; C/EBPα, CCAAT/enhancer binding protein α; C/EBPβ & δ, CCAAT/enhancer binding protein β & δ; CD36, fatty acid translocase; ECM, extracellular matrix; GLUT4, glucose transporter type 4; LPL, lipoprotein lipase; PPARγ, peroxisome proliferator-activated receptor-γ; Pref-1, preadipocyte factor-1; SREBP-1, sterol regulatory element binding protein-1.

diameter 20-fold and their volumes by several thousand-fold. However, fat cells do not increase in size indefinitely. Once a maximum capacity is attained, which in humans averages 1000 picoliters, the formation of new adipocytes from the precursor pool takes place.

Histologically, the interior of adipocytes appears unstained since the techniques of standard tissue

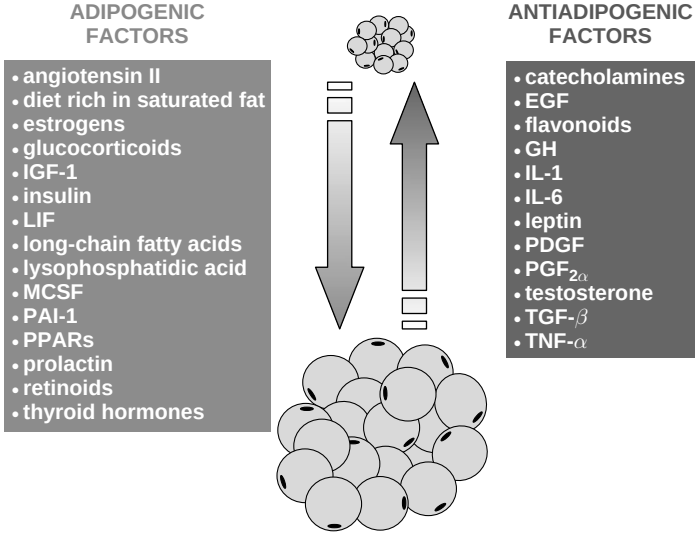


Figure 4 Factors exerting a direct effect on adipose mass. EGF, epidermal growth factor; GH, growth hormone; IGF-1, insulin-like growth factor-1; IL-1, interleukin-1; IL-6, interleukin-6; LIF, leukemia inhibitory factor; MCSF, macrophage colony stimulating factor; PAI-1, plasminogen activator inhibitor-1; PDGF, platelet-derived growth factor; $PGF_{2\alpha}$, prostaglandin $F_{2\alpha}$; PPARs, peroxisome proliferator-activated receptors; TGF-β, transforming growth factor-β; TNF-α, tumor necrosis factor-α.

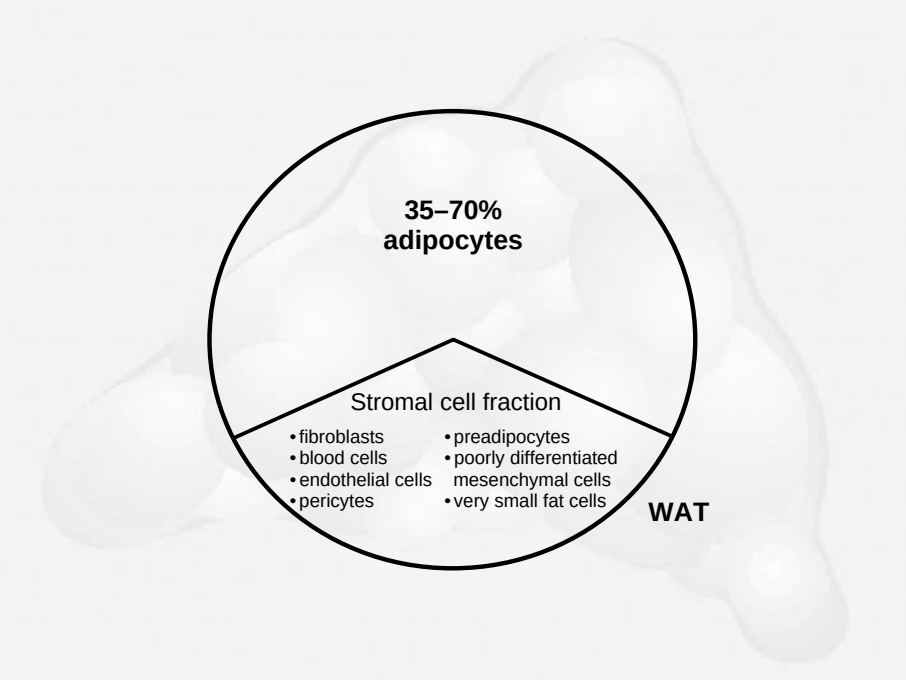
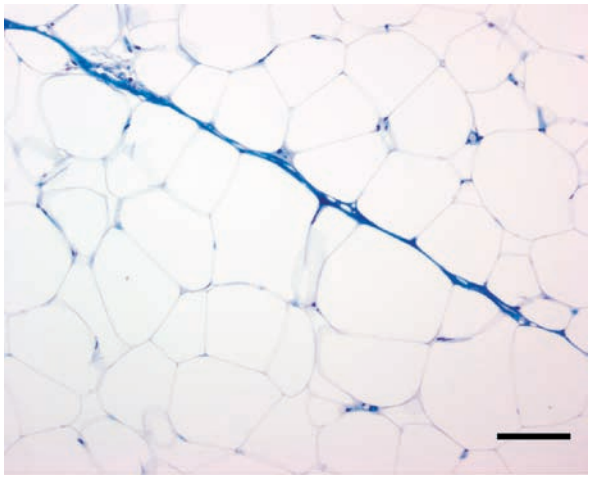


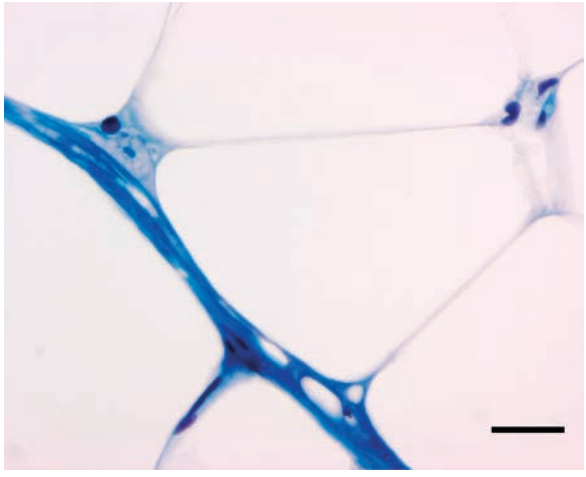
Figure 5 Schematic representation of cell types present in adipose tissue. WAT, white adipose tissue.

preparation dissolve out the lipids, leaving a thin rim of eosinophilic cytoplasm that typically loses its round shape during tissue processing, thus contributing to the sponge-like appearance of WAT in routine preparations for light microscopy (**Figure 6** and **Figure 7**). Owing to the fact that about 90% of the cell volume is a lipid droplet, the small dark nucleus becomes a flattened semilunar structure pushed against the edge of the cell and the thin cytoplasmic rim is also pushed to the periphery of the adipocytes. Mature white adipose cells contain a single large lipid droplet and are described as unilocular. However, developing white adipocytes are transiently multilocular containing multiple lipid droplets before these finally coalesce into a single large drop (**Figure 8**). The nucleus is round or oval in young fat cells, but is cup-shaped and peripherally displaced in mature adipocytes. The cytoplasm is stretched to form a thin sheath around the fat globule, although a relatively large volume is concentrated around the nucleus. A thin external lamina called basal lamina surrounds the cell. The smooth cell membrane shows no microvilli but has abundant smooth micropinocytotic invaginations that often fuse to form small vacuoles appearing as rosette-like configurations (**Figure 9**). Mitochondria are few in number with loosely arranged membranous cristae. The Golgi zone is small and the cytoplasm is filled with free ribosomes, but contains only a limited number of short profiles of the granular endoplasmic reticulum. Occasional lysosomes can be found. The coalescent lipid droplets contain a mixture of neutral fats, triglycerides, fatty acids, phospholipids, and cholesterol. A thin interface membrane separates the lipid droplet from the cytoplasmic matrix. Peripheral to this membrane is a system of parallel meridional thin filaments. Because of the size of these cells, relative to the thickness of the section, the nucleus (accounting for only one-fortieth of the cell volume) may not always be present in the section. Unilocular adipocytes usually appear in clumps near blood vessels, which is reasonable since the source and dispersion of material stored in fat cells depends on transportation by the vascular system.

Brown fat is a specialized type of adipose tissue that plays an important role in body temperature regulation. In the newborn brown fat is well developed in the neck and interscapular region. It has a limited distribution in childhood, and occurs only to a small degree in adult humans, while it is present in significant amounts in rodents and hibernating animals. The brown color is derived from a rich vascular network and abundant mitochondria and lysosomes. The individual multilocular adipocytes are frothy appearing cells due to the fact that the lipid, which does not coalesce as readily as in white fat cells and is normally stored in multiple small droplets, has been leached out during tissue

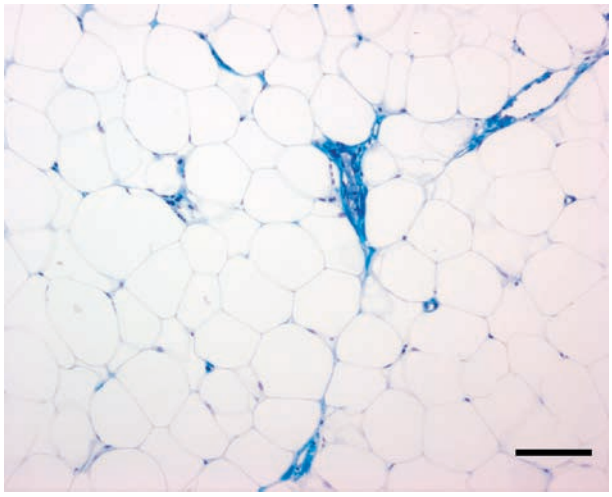


(A)

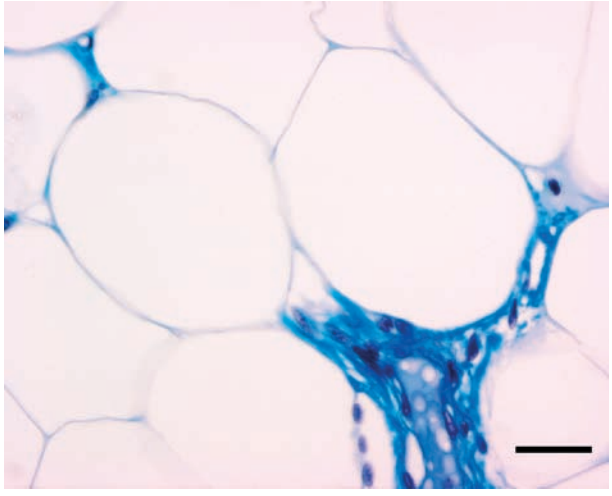


(B)

Figure 6 (A) Human subcutaneous white adipose tissue with Masson trichrome staining (10×; bar = 100 μm). (B) Same tissue at a higher magnification (40×; bar = 25 μm). (Courtesy of Dr. M A Burrell and M Archanco, University of Navarra, Spain.)



(A)



(B)

Figure 7 (A) Human omental white adipose tissue with Masson trichrome staining (10×; bar = 100 μm). (B) Same tissue at a higher magnification (40×; bar = 25 μm). (Courtesy of Dr. M A Burrell and M Archanco, University of Navarra, Spain.)

processing (**Figure 10**). The spherical nuclei are centrally or eccentrically located within the cell. Compared to the unilocular white adipocytes, the cytoplasm of the multilocular brown fat cell is relatively abundant and strongly stained because of the numerous mitochondria present. The mitochondria are involved in the oxidation of the stored lipid, but because they exhibit a reduced potential to carry out oxidative phosphorylation, the energy produced is released in the form of heat due to the uncoupling activity of UCP and not captured in adenosine triphosphate (ATP). Therefore, brown adipose tissue is extremely well vascularized so that the blood is warmed when it passes through the active tissue.

Distribution

White adipose tissue may represent the largest endocrine tissue of the whole organism, especially in overweight and obese patients. The anatomical distribution of individual fat pads dispersed throughout the whole body and not connected to each other contradicts the classic organ-specific localization. WAT exhibits clear, regional differences in its sites of predilection (**Table 1**). The hypodermal region invariably contains fat, except in a few places such as the eyelids and the scrotum. Adipocytes also accumulate around organs like the kidneys and adrenals, in the coronary sulcus of the heart, in bone marrow, mesentery, and omentum. Unilocular fat is

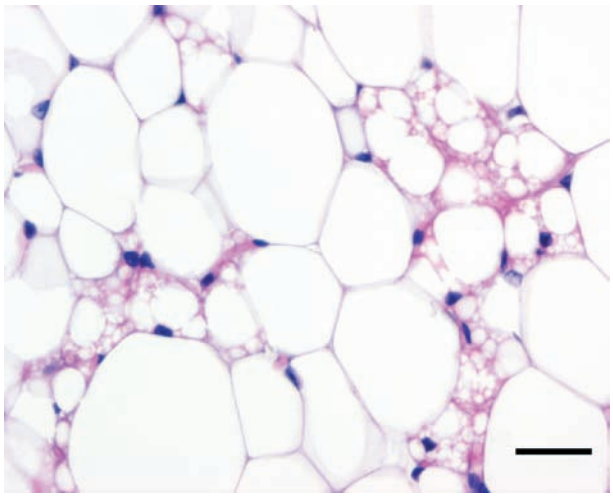
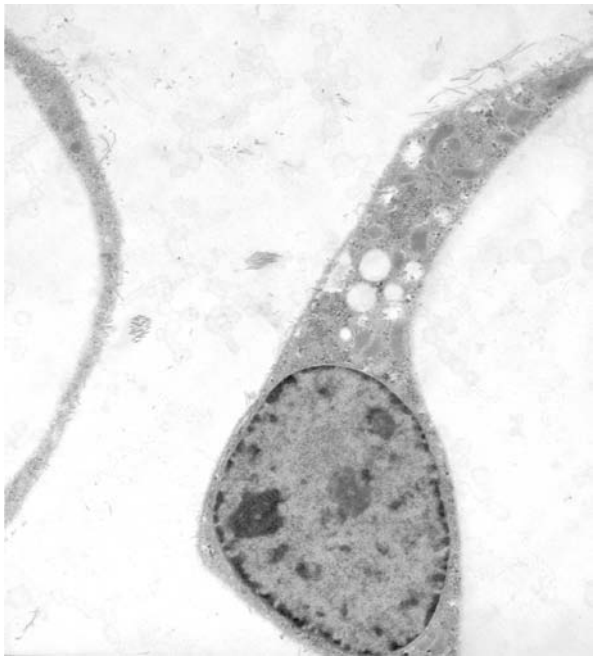


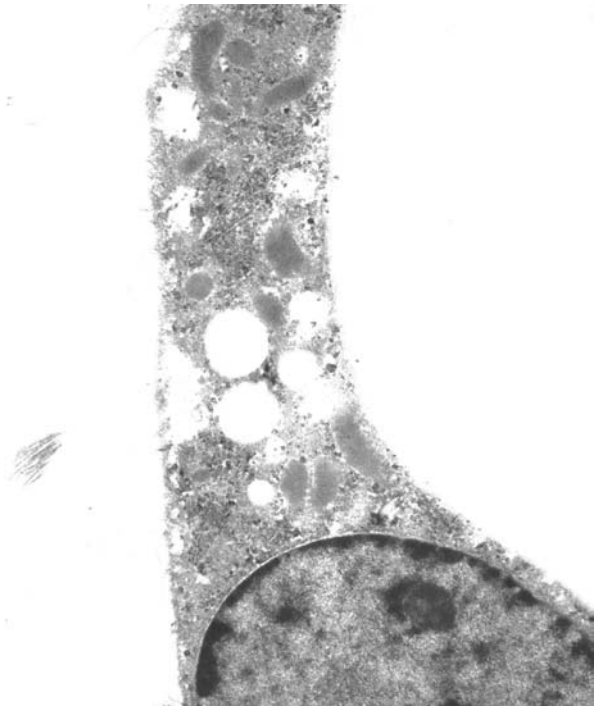
Figure 8 Paraffin section of rat abdominal white adipose tissue with a hematoxylin and eosin stain showing the simultaneous presence of uni- and multilocular adipocytes (40×; bar = 25 μm). (Courtesy of Dr. M A Burrell and M Archanco, University of Navarra, Spain.)

widely distributed in the subcutaneous tissue of humans but exhibits quantitative regional differences that are influenced by age and sex. In infants and young children there is a continuous subcutaneous fat layer, the panniculus adiposus, over the whole body. This layer thins out in some areas in adults but persists and grows thicker in certain other regions. The sites differ in their distribution among sexes, being responsible for the characteristic body form of males and females, termed android and ginecoid fat distribution. In males, the main regions include the nape of the neck, the subcutaneous area over the deltoid and triceps muscles, and the lumbosacral region. In females, subcutaneous fat is most abundant in the buttocks, epitrochanteric region, anterior and lateral aspects of the thighs, as well as the breasts. Additionally, extensive fat depots are found in the omentum, mesenteries, and the retroperitoneal area of both sexes. In well-nourished, sedentary individuals, the fat distribution persists and becomes more obvious with advancing age with males tending to deposit more fat in the visceral compartment. Depot-specific differences may be related not only to the metabolism of fat cells but also to their capacity to form new adipocytes. Additionally, regional differences may result from variations in hormone receptor distribution as well as from specific local environmental characteristics as a consequence of differences in innervation and vascularization.

Regional distribution of body fat is known to be an important indicator for metabolic and cardiovascular alterations in some individuals.

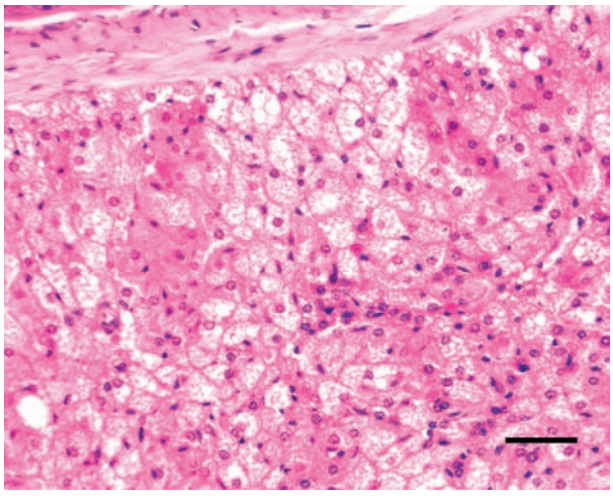


(A)

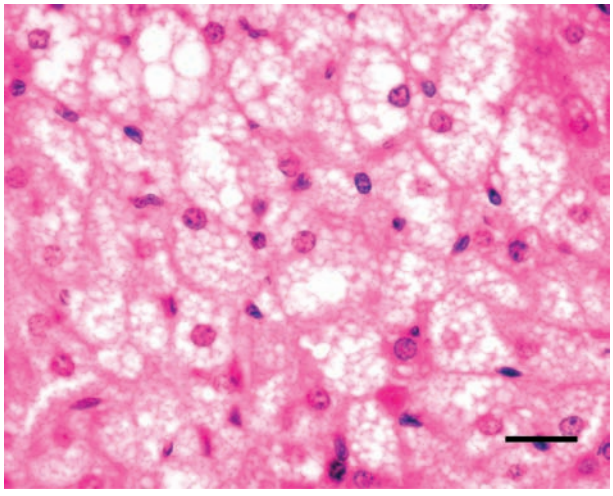


(B)

Figure 9 (A) Transmission electron micrographs with the characteristically displaced nucleus to one side and slightly flattened by the accumulated lipid. The cytoplasm of the fat cell is reduced to a thin rim around the lipid droplet (7725×). (B) The cytoplasm contains several small lipid droplets that have not yet coalesced. A few filamentous mitochondria, occasional cisternae of endoplasmic reticulum, and a moderate number of free ribosomes are usually visible (15 000×). (Courtesy of Dr. M A Burrell and M Archanco, University of Navarra, Spain.)



(A)



(B)

Figure 10 (A) Paraffin section of rat brown adipose tissue with a hematoxylin and eosin stain (20×; bar = 50 μm). (B) Same tissue at a higher magnification (40×; bar = 25 μm). (Courtesy of Dr. M A Burrell and M Archanco, University of Navarra.)

Table 1 Distribution of main human adipose tissue depots

Subcutaneous (approx. 80%; deep + superficial layers)
Truncal
– Cervical
– Dorsal
– Lumbar
Abdominal
Gluteofemoral
Mammary
Visceral (approx. 20%; thoracic-abdominal-pelvic)
Intrathoracic (extra-intrapericardial)
Intra-abdominopelvic
– Intraperitoneal
Omental (greater and lesser omentum)
Mesenteric (epiplon, small intestine, colon, rectum)
Umbilical
– Extraperitoneal
Peripancreatic (infiltrated with brown adipocytes)
Perirenal (infiltrated with brown adipocytes)
– Intrapelvic
Gonadal (parametrial, retrouterine, retropubic)
Urogenital (paravesical, para-retrorectal)
Intraparenchymatous (physiologically or pathologically)
Inter-intramuscular and perimuscular (inside the muscle fascia)
Perivascular
Paraosseal (interface between bone and muscle)
Ectopic (steatosis, intramyocardial, lypodystrophy, etc.)

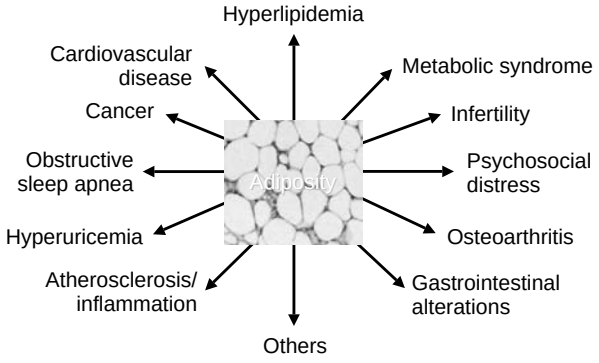


Figure 11 Main comorbidities associated with increased adiposity.

The observation that the topographic distribution of adipose tissue is relevant to understanding the relation of obesity to disturbances in glucose and lipid metabolism was formulated before the 1950s. Since then numerous prospective studies have revealed that android or male-type obesity correlates more often with an elevated mortality and risk for the development of diabetes mellitus type 2, dyslipidemia, hypertension, and atherosclerosis than gynoid or female-type obesity. Obesity has been reported to cause or exacerbate a large number of health problems with a known impact on both life expectancy and quality of life. In this respect, the association of increased adiposity is accompanied by important pathophysiological alterations, which lead to the development of a wide range of comorbidities (**Figure 11**).

Function

Although many cell types contain small reserves of carbohydrate and lipid, the adipose tissue is the body's most capacious energy reservoir. Because of the high energy content per unit weight of fat as well as its hydrophobicity, the storage of energy in the form of triglycerides is a highly efficient biochemical phenomenon (1 g of adipose tissue contains around 800 mg triacylglycerol and only about 100 mg of

water). It represents quantitatively the most variable component of the organism, ranging from a few per cent of body weight in top athletes to more than half of the total body weight in severely obese patients. The normal range is about 10–20% body fat for males and around 20–30% for females, accounting approximately for a 2-month energy reserve. During pregnancy most species accrue additional reserves of adipose tissue to help support the development of the fetus and to further facilitate the lactation period.

Energy balance regulation is an extremely complex process composed of multiple interacting homeostatic and behavioral pathways aimed at maintaining constant energy stores. It is now evident that body weight control is achieved through highly orchestrated interactions between nutrient selection, organoleptic influences, and neuroendocrine responses to diet as well as being influenced by genetic and environmental factors. The concept that circulating signals generated in proportion to body fat stores influence appetite and energy expenditure in a coordinated manner to regulate body weight was proposed almost 50 years ago. According to this model, changes in energy balance sufficient to alter body fat stores are signaled via one or more circulating factors acting in the brain to elicit compensatory changes in order to match energy intake to energy expenditure. This was formulated as the 'lipostatic theory' assuming that as adipose tissue mass enlarges, a factor that acts as a sensing hormone or 'lipostat' in a negative feedback control from adipose tissue to hypothalamic receptors informs the brain about the abundance of body fat, thereby allowing feeding behavior, metabolism, and endocrine physiology to be coupled to the nutritional state of the organism. The existing body of evidence gathered in the last decades through targeted expression or knockout of specific genes involved in different steps of the pathways controlling food intake, body weight, adiposity, or fat distribution has clearly contributed to unraveling the underlying mechanisms of energy homeostasis. The findings have fostered the notion of a far more complex system than previously thought, involving the integration of a plethora of factors.

The identification of adipose tissue as a multifunctional organ as opposed to a passive organ for the storage of excess energy in the form of fat has been brought about by the emerging body of evidence gathered during the last few decades. This pleiotropic nature is based on the ability of fat cells to secrete a large number of hormones, growth factors, enzymes, cytokines, complement factors, and matrix proteins, collectively termed adipokines or adipocytokines (**Table 2**, **Figure 12**), at the same time as expressing receptors for most of these factors (**Table 3**), which warrants extensive cross-talk at a local and systemic level in response to specific external stimuli or metabolic changes. The vast majority of adipocyte-derived factors have been shown to be dysregulated in alterations accompanied by changes

Table 2 Relevant factors secreted by adipose tissue into the bloodstream

Molecule	*Function/effect*
Adiponectin/ACRP30/AdipoQ/ apM1/GBP28	Plays a protective role in the pathogenesis of type 2 diabetes and cardiovascular diseases
Adipsin	Possible link between the complement pathway and adipose tissue metabolism
Angiotensinogen	Precursor of angiotensin II; regulator of blood pressure and electrolyte homeostasis
ASP	Influences the rate of triacylglycerol synthesis in adipose tissue
FFA	Oxidized in tissues to produce local energy. Serve as a substrate for triglyceride and structural molecules synthesis. Involved in the development of insulin resistance
Glycerol	Structural component of the major classes of biological lipids and gluconeogenic precursor
IGF-I	Stimulates proliferation of a wide variety of cells and mediates many of the effects of growth hormone
IL-6	Implicated in host defense, glucose and lipid metabolism, and regulation of body weight
Leptin	Signals to the brain about body fat stores. Regulation of appetite and energy expenditure. Wide variety of physiological functions
NO	Important regulator of vascular tone. Pleiotropic involvement in pathophysiological conditions
PAI-1	Potent inhibitor of the fibrinolytic system
PGI_2 & $PGF_{2\alpha}$	Implicated in regulatory functions such as inflammation and blood clotting, ovulation, menstruation, and acid secretion
Resistin	Putative role in insulin resistance May participate in inflammation
TNF-α	Interferes with insulin receptor signaling and is a possible cause of the development of insulin resistance in obesity
VEGF	Stimulation of angiogenesis

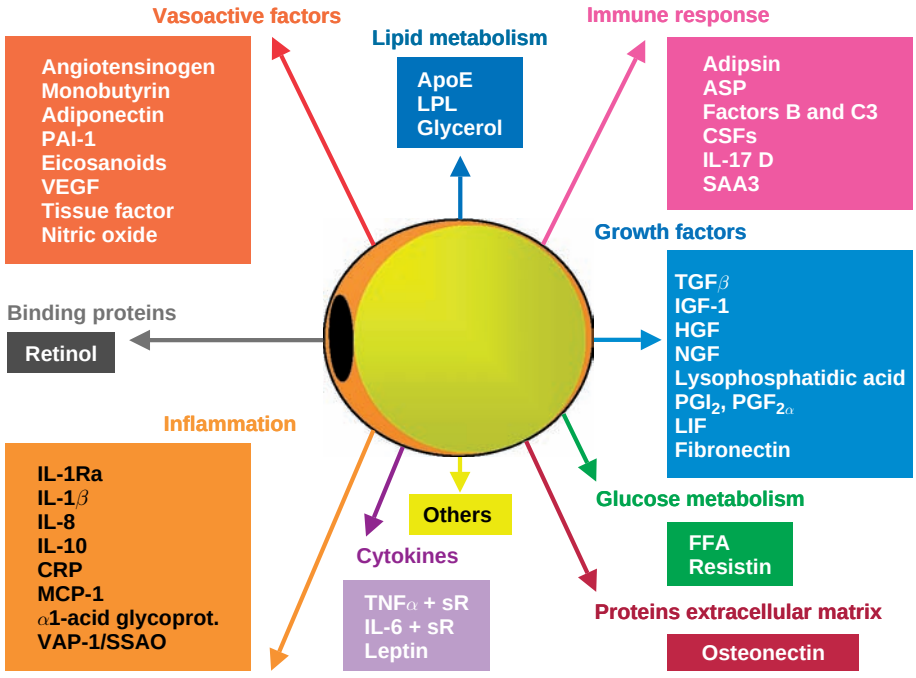


Figure 12 Factors secreted by white adipose tissue, which underlie the multifunctional nature of this endocrine organ. Although due to their pleiotropic effects some of the elements might be included in more than one physiological role, they have been included only under one function for simplicity reasons. apoE, apolipoprotein E; ASP, acylation-stimulating protein; CRP, C-reactive protein; CSFs, colony-stimulating factors; FFA, free fatty acids; HGF, hepatocyte growth factor; IGF-1, insulin-like growth factor-1; IL-10, interleukin-10; IL-17 D, interleukin-17 D; IL-1Ra, interleukin-1 receptor antagonist; IL-1β, interleukin-1β; IL-6, interleukin, 6; IL-8, interleukin-8; LIF, leukemia inhibitory factor; LPL, lipoprotein lipase; MCP-1, monocyte chemoattractant protein-1; NGF, nerve growth factor; PAI-1, plasminogen activator inhibitor -1; $PGF_{2\alpha}$, prostaglandin $F_{2\alpha}$; PGI_2, prostacyclin; SAA3, serum amyloid A3; sR, soluble receptor; TGF-β, transforming growth factor-β; TNF-α, tumor necrosis factor-α; VAP-1/SSAO, vascular adhesion protein-1/semicarbazide-sensitive amine oxidase; VEGF, vascular endothelial growth factor.

in adipose tissue mass such as overfeeding and lipodystrophy, thus providing evidence for their implication in the etiopathology and comorbidities asssociated with obesity and cachexia.

WAT is actively involved in cell function regulation through a complex network of endocrine, paracrine, and autocrine signals that influence the response of many tissues, including hypothalamus, pancreas, liver, skeletal muscle, kidneys, endothelium, and immune system, among others. Adipose tissue serves the functions of being a store for reserve energy, insulation against heat loss through the skin, and a protective padding of certain organs. A rapid turnover of stored fat can take place, and with only a few exceptions (orbit, major joints as well as palm and foot sole), the adipose tissue can be used up almost completely during starvation. Adipocytes are uniquely equipped to participate in the regulation of other functions such as reproduction, immune response, blood pressure control, coagulation, fibrinolysis, and angiogenesis, among others. This multifunctional nature is based on the existence of the full complement of enzymes, regulatory proteins, hormones, cytokines, and receptors needed to carry out an extensive cross-talk at both a local and systemic level in response to specific external stimuli or neuroendocrine changes. This secretory nature has prompted the view of WAT as an extremely active endocrine tissue. Interestingly, the high number and ample spectrum of genes found to be expressed in WAT together with the changes observed in samples of obese patients substantiates the view of an extraordinarily active and plastic tissue. The complex and complementary nature of the expression profile observed in adipose tissue from obese organisms reflects a plethora of adaptive changes affecting crucial physiological functions that may need to be further explored through genomic and proteomic approaches.

The endocrine activity of WAT was postulated almost 20 years ago when the tissue's ability for steroid hormone interconversion was alluded to. In recent years, especially since the discovery of leptin, the list of adipocyte-derived factors has been increasing at a phenomenal pace. Another way of addressing the production of adipose-derived factors is by focusing on the function they are implicated in (**Figure 12**). One of the best known

Table 3 Main receptors expressed by adipose tissue

Receptor	*Main effect of receptor activation on adipocyte metabolism*
Hormone-cytokine receptors	
Adenosine	Inhibition of lipolysis
Adiponectin (AdipoR1 & AdipoR2)	Regulation of insulin sensitivity and fatty acid oxidation
Angiotensin II	Increase of lipogenesis Stimulation of prostacyclin production by mature fat cells. Interaction with insulin in regulation of adipocyte metabolism
GH	Induction of leptin and IGF-I expression. Stimulation of lipolysis
IGF-I & -II	Inhibition of lipolysis. Stimulation of glucose transport and oxidation
IL-6	LPL activity inhibition. Induction of lipolysis
Insulin	Inhibition of lipolysis and stimulation of lipogenesis. Induction of glucose uptake and oxidation. Stimulation of leptin expression
Leptin (OB-R)	Stimulation of lipolysis. Autocrine regulation of leptin expression
NPY-Y1 & Y5	Inhibition of lipolysis. Induction of leptin expression
Prostaglandin	Strong antilypolitic effects (PGE_2). Modulation of preadipocyte differentiation ($PGF_{2\alpha}$ and PGI_2)
TGF-β	Potent inhibition of adipocyte differentiation
TNF-α	Stimulation of lipolysis. Regulation of leptin secretion. Potent inhibition of adipocyte differentiation. Involvement in development of insulin resistance
VEGF	Stimulation of angiogenesis
Catecholamine-nervous system receptors	
Muscarinic	Inhibition of lipolysis
Nicotinic	Stimulation of lipolysis
α_1-AR	Induction of inositol phosphate production and PKC activation
α_2-AR	Inhibition of lipolysis. Regulation of preadipocyte growth
β_1-, β_2- & β_3-AR	Stimulation of lipolysis. Induction of thermogenesis. Reduction of leptin mRNA levels
Nuclear receptors	
Androgen	Control of adipose tissue development (antiadipogenic signals). Modulation of leptin expression
Estrogen	Control of adipose tissue development (proadipogenic signals). Modulation of leptin expression
Glucocorticoids	Stimulation of adipocyte differentiation
PPARδ	Regulation of fat metabolism. Plays a central role in fatty acid-controlled differentiation of preadipose cells
PPARγ	Induction of adipocyte differentiation and insulin sensitivity
RAR/RXR	Regulation of adipocyte differentiation
T_3	Stimulation of lipolysis. Regulation of leptin secretion. Induction of adipocyte differentiation. Regulation of insulin effects
Lipoprotein receptors	
HDL	Clearance and metabolism of HDL
LDL	Stimulation of cholesterol uptake
VLDL	Binding and internalization of VLDL particles. Involvement in lipid accumulation

Abbreviations: ACRP30, adipocyte complement-related protein of 30 kDa; apM1, adipose most abundant gene transcript 1; ASP, acylation-stimulating protein; FFA, free fatty acids; GBP28, gelatin-binding protein 28; GH, growth hormone; HDL, high density lipoprotein; IGF, insulin-like growth factor; IL-6, interleukin 6; LDL, low density lipoprotein; LPL, lipoprotein lipase; NO, nitric oxide; NPY-Y1 & -Y5, neuropeptide receptors Y-1 & -5; OB-R, leptin receptor; PAI-1, plasminogen activator inhibitor -1; PGE_2, prostaglandin E_2; $PGF_{2\alpha}$, prostaglandin $F_{2\alpha}$; PGI_2, prostacyclin; PPAR, peroxisome proliferator-activated receptor; RAR, retinoic acid receptor; RXR, retinoid x receptor; T3, triiodothyronine; TGF-β, transforming growth factor-β; TNF-α, tumor necrosis factor-α; VEGF, vascular endothelial growth factor; VLDL, very low-density lipoprotein; α_1- & α_2-AR, α_1- & α_2-adrenergic receptors; β_1-, β_2- & β_3-AR, β_1-, β_2- & β_3 adrenergic receptors.

aspects of WAT physiology relates to the synthesis of products involved in lipid metabolism such as perilipin, adipocyte lipid-binding protein (ALBP, FABP4, or aP2), CETP (cholesteryl ester transfer protein), and retinol binding protein (RBP). Adipose tissue has also been identified as a source of production of factors with immunological properties participating in immunity and stress responses, as is the case for ASP (acylation-simulating protein) and metallothionein. More recently, the pivotal role of adipocyte-derived factors in cardiovascular function control such as angiotensinogen, adiponectin, peroxisome proliferator-activated receptor γ angiopoietin related protein/fasting-induced adipose factor (PGAR/FIAF), and C-reactive protein (CRP) has been established. A further subsection of proteins produced by adipose tissue concerns other factors with an autocrine-paracrine function like

PPAR-γ (peroxisome proliferator-activated receptor), IGF-1, monobutyrin, and the UCPs.

It is generally assumed that under normal physiological circumstances adult humans are practically devoid of functional brown adipose tissue. As is the case in other larger mammals the functional capacity of brown adipose tissue decreases because of the relatively higher ratio between heat production from basal metabolism and the smaller surface area encountered in adult animals. In addition, clothing and indoor life have reduced the need for adaptive nonshivering thermogenesis. However, it has been recently shown that human WAT can be infiltrated with brown adipocytes expressing UCP-1.

Regulation of Metabolism

The control of fat storage and mobilization has been marked by the identification of a number of regulatory mechanisms in the last few decades. Isotopic tracer studies have clearly shown that lipids are continuously being mobilized and renewed even in individuals in energy balance. Fatty acid esterification and triglyceride hydrolysis take place continuously. The half-life of depot lipids in rodents is about 8 days, meaning that almost 10% of the fatty acid stored in adipose tissue is replaced daily by new fatty acids. The balance between lipid loss and accretion determines the net outcome on energy homeostasis.

The synthesis of triglycerides, also termed lipogenesis, requires a supply of fatty acids and glycerol. The main sources of fatty acids are the liver and the small intestine. Fatty acids are esterified with glycerol phosphate in the liver to produce triglycerides. Since triglycerides are bulky polar molecules that do not cross cell membranes well, they must be hydrolyzed to fatty acids and glycerol before entering fat cells. Serum very low-density lipoproteins (VLDLs) are the major form in which triacylglycerols are carried from the liver to WAT. Short-chain fatty acids (16 carbons or less) can be absorbed from the gastrointestinal tract and carried in chylomicra directly to the adipocyte. Inside fat cells, glycerol is mainly synthesized from glucose. In WAT, fatty acids can be synthesized from several precursors, such as glucose, lactate, and certain amino acids, with glucose being quantitatively the most important in humans. In the case of glucose, GLUT4, the principal glucose transporter of adipocytes, controls the entry of the substrate into the adipocyte. Insulin is known to stimulate glucose transport by promoting GLUT4 recruitment as well as increasing its activity. Inside the adipocyte, glucose is initially phosphorylated and then metabolized both in the cytosol and in the mitochondria to produce cytosolic acetyl-CoA with the flux being influenced by phosphofructokinase and pyruvate dehydrogenase. Glycerol does not readily enter the adipocyte, but the membrane-permeable fatty acids do. Once inside the fat cells, fatty acids are re-esterified with glycerol phosphate to yield triglycerides. Lipogenesis is favored by insulin, which activates pyruvate kinase, pyruvate dehydrogenase, acetyl-CoA carboxylase, and glycerol phosphate acyltransferase. When excess nutrients are available insulin decreases acetyl-CoA entry into the tricarboxylic acid cycle while directing it towards fat synthesis. This insulin effect is antagonized by growth hormone. The gut hormones glucagon-like peptide 1 and gastric inhibitory peptide also increase fatty acid synthesis, while glucagon and catecholamines inactivate acetyl-CoA carboxylase, thus decreasing the rate of fatty acid synthesis.

The release of glycerol and free fatty acids by lipolysis plays a critical role in the ability of the organism to provide energy from triglyceride stores. In this sense, the processes of lipolysis and lipogenesis are crucial for the attainment of body weight control. For this purpose adipocytes are equipped with a well-developed enzymatic machinery, together with a number of nonsecreted proteins and binding factors directly involved in the regulation of lipid metabolism. The hydrolysis of triglycerides from circulating VLDL and chylomicrons is catalyzed by lipoprotein lipase (LPL). This rate-limiting step plays an important role in directing fat partitioning. Although LPL controls fatty acid entry into adipocytes, fat mass has been shown to be preserved by endogenous synthesis. From observations made in patients with total LPL deficiency it can also be concluded that fat deposition can take place in the absence of LPL. A further key enzyme catalyzing a rate-limiting step of lipolysis is HSL (hormome sensitive lipase), which cleaves triacylglycerol to yield glycerol and fatty acids. Some fatty acids are re-esterified, so that the fatty acid: glycerol ratio leaving the cell is usually less than the theoretical 3:1. Increased concentrations of cAMP activate HSL as well as promote its movement from the cytosol to the lipid droplet surface. Catecholamines and glucagon are known inducers of the lipolytic activity, while the stimulation of lipolysis is attenuated by adenosine and protaglandin E_2. Interestingly, HSL deficiency leads to male sterility and adipocyte hypertrophy, but not to obesity, with an unaltered basal lipolytic activity suggesting that other lipases may also play a relevant role in fat mobilization.

The lipid droplets contained in adipocytes are coated by structural proteins, such as perilipin, that stabilize the single fat drops and prevent triglyceride

hydrolysis in the basal state. The phosphorylation of perilipin following adrenergic stimulation or other hormonal inputs induces a structural change of the lipid droplet that allows the hydrolysis of triglycerides. After hormonal stimulation, HSL and perilipin are phosphorylated and HSL translocates to the lipid droplet. ALBP, also termed aP2, then binds to the N-terminal region of HSL, preventing fatty acid inhibition of the enzyme's hydrolytic activity.

The function of CETP is to promote the exchange of cholesterol esters of triglycerides between plasma lipoproteins. Fasting, high-cholesterol diets as well as insulin stimulate CETP synthesis and secretion in WAT. In plasma, CETP participates in the modulation of reverse cholesterol transport by facilitating the transfer of cholesterol esters from high-density lipoprotein (HDL) to triglyceride-rich apoB-containing lipoproteins. VLDLs, in particular, are converted to low-density lipoproteins (LDLs), which are subjected to hepatic clearance by the apoB/E receptor system. Adipose tissue probably represents one of the major sources of CETP in humans. Therefore, WAT represents a cholesterol storage organ, whereby peripheral cholesterol is taken up by HDL particles, acting as cholesterol efflux acceptors, and is returned for hepatic excretion. In obesity, the activity and protein mass of circulating CETP is increased showing a negative correlation with HDL concentrations at the same time as a positive correlation with fasting glycemia and insulinemia suggesting a potential link with insulin resistance.

Synthesis and secretion of RBP by adipocytes is induced by retinoic acid and shows that WAT plays an important role in retinoid storage and metabolism. In fact, RBP mRNA is one of the most abundant transcripts present in both rodent and human adipose tissue. Hepatic and renal tissues have been regarded as the main sites of RBP production, while the quantitative and physiological significance of the WAT contribution remains to be fully elucidated.

The processes participating in the control of energy balance, as well as the intermediary lipid and carbohydrate metabolism, are intricately linked by neurohumoral mediators. The coordination of the implicated molecular and biochemical pathways underlies, at least in part, the large number of intracellular and secreted proteins produced by WAT with autocrine, paracrine, and endocrine effects. The finding that WAT secretes a plethora of pleiotropic adipokines at the same time as expressing receptors for a huge range of compounds has led to the development of new insights into the functions of adipose tissue at both the basic and clinical level. At this early juncture in the course of adipose tissue research, much has been discovered. However, a great deal more remains to be learned about its physiology and clinical relevance. Given the adipocyte's versatile and ever-expanding list of secretory proteins, additional and unexpected discoveries are sure to emerge. The growth, cellular composition, and gene expression pattern of adipose tissue is under the regulation of a large selection of central mechanisms and local effectors. The exact nature and control of this complex cross-talk has not been fully elucidated and represents an exciting research topic.

Abbreviations

ACRP30/apM1/ GBP28	adipocyte complement-related protein of 30 kDa/adipose most abundant gene transcript 1/gelatin-binding protein 28
ADD1/SREBP-1C	adipocyte determination and differentiation factor-1/sterol regulatory element binding protein-1c
ALBP/FABP4/aP2	adipocyte fatty acid binding protein
apoE	apolipoprotein E
ASP	acylation-stimulating protein
ATP	adenosine triphosphate
cAMP	cyclic adenosin monophosphate
CD36	fatty acid translocase
C/EBPs	CCAAT/enhancer binding proteins
CETP	cholesteryl ester transfer protein
CRP	C-reactive protein
CSF	colony-stimulating factor
ECM	extracellular matrix
EGF	epidermal growth factor
FFA	free fatty acids
FGF	fibroblast growth factor
GH	growth hormone
GLP-1	glucagon-like peptide-1
GLUT4	glucose transporter type 4
HDL	high density lipoprotein
HGF	hepatocyte growth factor
HSL	hormone-sensitive lipase
IGF	insulin-like growth factor
IL	interleukin
IL-1Ra	interleukin-1 receptor antagonist
LDL	low density lipoprotein
LIF	leukemia inhibitory factor
LPL	lipoprotein lipase
MCP-1	monocyte chemoattractant protein-1
MCSF	macrophage colony stimulating factor
MIF	macrophage migration inhibitory factor
MIP-1α	macrophage inflammatory protein-1α

NGF	nerve growth factor
NO	nitric oxide
NPY-Y1 & -Y5	neuropeptide receptors Y-1 & -5
OB-R	leptin receptor
PAI-1	plasminogen activator inhibitor-1
PDGF	platelet-derived growth factor
PGAR/FIAF	peroxisome proliferator-activated receptor angiopoietin related protein/fasting-induced adipose factor
PGC-1α	peroxisome proliferator-activated receptor-γ coactivator-1α
PGE_2	prostaglandin E_2
$PGF_{2\alpha}$	prostaglandin $F_{2\alpha}$
PGI_2	prostacyclin
PPAR	peroxisome proliferator-activated receptor
Pref-1	preadipocyte factor-1
RAR	retinoic acid receptor
RBP	retinol binding protein
RXR	retinoid x receptor
SAA3	serum amyloid A3
T_3	triiodothyronine
TGF-β	transforming growth factor-β
TNF-α	tumor necrosis factor-α
UCP	uncoupling protein
VAP-1/SSAO	vascular adhesion protein-1/ semicarbazide-sensitive amine oxidase
VEGF	vascular endothelial growth factor
VLDL	very low density lipoprotein
WAT	white adipose tissue
α_1- & α_2-AR	α_1- & α_2-adrenergic receptors
β_1-, β_2- & β_3-AR	β_1-, β_2- & β_3 adrenergic receptors

See also: **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Hypertension**: Etiology. **Lipids**: Chemistry and Classification; Composition and Role of Phospholipids. **Lipoproteins**. **Obesity**: Definition, Etiology and Assessment; Fat Distribution; Childhood Obesity; Complications; Prevention; Treatment. **Pregnancy**: Safe Diet for Pregnancy.

Further Reading

Ailhaud G and Hauner H (2004) Development of white adipose tissue. In: Bray GA and Bouchard C (eds.) *Handbook of Obesity. Etiology and Pathophysiology*, 2nd edn, pp. 481–514. New York: Marcel Dekker, Inc.

Frayn KN, Karpe F, Fielding BA, Macdonald IA, and Coppack SW (2003) Integrative physiology of human adipose tissue. *International Journal of Obesity* **27**: 875–888.

Fried SK and Ross RR (2004) Biology of visceral adipose tissue. In: Bray GA and Bouchard C (eds.) *Handbook of Obesity. Etiology and Pathophysiology*, 2nd edn, pp. 589–614. New York: Marcel Dekker, Inc.

Frühbeck G (2004) The adipose tissue as a source of vasoactive factors. *Current Medicinal Chemistry (Cardiovascular & Hematological Agents)* **2**: 197–208.

Frühbeck G and Gómez-Ambrosi J (2003) Control of body weight: a physiologic and transgenic perspective. *Diabetologia* **46**: 143–172.

Frühbeck G, Gómez-Ambrosi J, Muruzábal FJ, and Burrell MA (2001) The adipocyte: a model for integration of endocrine and metabolic signaling in energy metabolism regulation. *American Journal of Physiology* **280**: E827–E847.

Gómez-Ambrosi J, Catalán V, Diez-Caballero A, Martínez-Cruz A, Gil MJ, García-Foncillas J, Cienfuegos JA, Salvador J, Mato JM, and Frühbeck G (2004) Gene expression profile of omental adipose tissue in human obesity. *The FASEB Journal* **18**: 215–217.

Lafontan M and Berlan M (2003) Do regional differences in adipocyte biology provide new pathophysiological insights? *Trends in Pharmacological Sciences* **24**: 276–283.

Langin D and Lafontan M (2000) Millennium fat-cell lipolysis reveals unsuspected novel tracks. *Hormone and Metabolic Research* **32**: 443–452.

Pond CM (1999) Physiological specialisation of adipose tissue. *Progress in Lipid Research* **38**: 225–248.

Rosen ED, Walkey CJ, Puigserver P, and Spiegelman BM (2000) Transcriptional regulation of adipogenesis. *Genes and Development* **14**: 1293–1307.

Shen W, Wang Z, Punyanita M, Lei J, Sinav A, Kral JG, Imielinska C, Ross R, and Heymsfield SB (2003) Adipose quantification by imaging methods: a proposed classification. *Obesity Research* **11**: 5–16.

Trayhurn P and Beattie JH (2001) Physiological role of adipose tissue: white adipose tissue as an endocrine and secretory organ. *Proceedings of the Nutrition Society* **60**: 329–339.

Unger RH (2003) The physiology of cellular liporegulation. *Annual Review of Physiology* **65**: 333–347.

Wajchenberg BL (2000) Subcutaneous and visceral adipose tissue: their relation to the metabolic syndrome. *Endocrine Reviews* **21**: 697–738.

ADOLESCENTS

Contents
Nutritional Requirements
Nutritional Problems

Nutritional Requirements

C H S Ruxton, Nutrition Communications, Cupar, UK
J Fiore, University of Westminster, London, UK

Introduction

Adolescence is the period of transition between childhood and adulthood. This reflects not only the physical and emotional changes experienced by the adolescent, but the development of dietary behaviors. Whereas younger children are characterized by their resistance to new experiences, the adolescent may use food to assert their independence, not always in a beneficial way. This section will cover development in adolescence and highlight nutrients that are important during this time. Information on adolescent energy and nutrient intakes from a broad range of countries will be presented. The findings will be put in context with dietary recommendations.

Physical Changes During Adolescence

Adolescence is generally assumed to be the period of human development from 10 to 18 years of age, a time during which rapid growth and physical maturity take place.

Growth

During prepubescent childhood, the growth of boys and girls follows a similar trajectory, although boys may be slightly taller and heavier than girls. Around the 9th year, the pubertal growth spurt, which can last up to 3.5 years, will occur in girls with boys beginning 2 years later. Girls reach their full height approximately 2 years before boys and are, therefore, the taller of the two sexes for a period of time. Current UK standards for height and weight during adolescence are presented in **Table 1**.

Maximum height velocity is generally seen in the year preceding menarche for girls and at around 14 years for boys. On average, weight velocity peaks at 12.9 years for girls and 14.3 years for boys. Annual growth rates during adolescence can be as much as 9 cm/8.8 kg in girls and 10.3 cm/9.8 kg in boys. Energy and protein intakes per kilogram body weight have been observed to peak during maximal growth, suggesting increased requirements during adolescence. Undernutrition in this crucial window of development can result in a slow height increment, lower peak bone mass, and delayed puberty. On the other hand, overnutrition is not without its risks. It is believed that obesity in young girls can bring about an early menarche, which then increases the risk of breast cancer in later adulthood. Menarche is deemed precocious if it occurs before the age of eight. Rising childhood obesity levels in Western countries have resulted in a rise in the proportion of girls displaying precocious menarche.

Table 1 Percentiles for height, weight, and body mass index

Age (years)	*Height (cm)*			*Weight (kg)*			*Body mass index*		
	3rd	*50th*	*97th*	*3rd*	*50th*	*97th*	*2nd*	*50th*	*99.6th*
(a) Boys									
11	130.8	143.2	155.8	26.1	34.5	50.9	14	17	26
16	158.9	173.0	187.4	44.9	60.2	83.2	16	20	30
18	163.3	176.4	189.7	52.0	66.2	87.9	17	21	32
(b) Girls									
11	130.9	143.8	156.9	26.0	35.9	53.6	14	17	27
16	151.6	163.0	174.6	42.8	55.3	74.1	16	20	31
18	152.3	163.6	175.0	44.7	57.2	76.3	17	21	32

It is not fully known when growth ceases. Certainly, height gains of up to 2 cm can still occur between 17 and 28 years. Important nutrients for growth include protein, iron, calcium, vitamin C, vitamin D, and zinc. Calcium, in particular, has a key role in bone development, and huge increments in bone density are seen during adolescence under the influence of sex hormones. Bone density peaks in the early twenties and a low bone density at this time is related to increased osteoporosis risk in later life, especially for women. Studies have suggested that body mass index in adolescence is the best predictor of adult bone density, explaining why children who experience anorexia nervosa are likely to have a higher risk of osteoporosis.

Adipose stores

There are few differences in body fat between boys and girls in the prepubertal stage. However, during puberty, girls develop adipose tissue at a greater rate than boys, laying down stores in the breast and hip regions. The pattern for boys is rather different and tends towards a more central deposition. Methods for estimating fatness in adolescents include weight for height, body mass index (weight in kilograms/height in meters^2), skinfold thickness measures, bioelectrical impedance analysis, densitometry, magnetic resonance imaging, dual energy X-ray absorptiometry, and computer tomography. Waist circumference is gaining popularity as a useful proxy of fatness in the field. Many researchers argue that it is a better predictor than body mass index (BMI) of the central adipose stores, which place the individual most at risk from later obesity, diabetes, and coronary heart disease.

Current UK standards for BMI and waist circumference are outlined in **Table 1**. The 90th percentile is viewed as the lower cut-off point for classification of overweight and can identify those at risk of chronic disease. In a Norwegian longitudinal survey, adolescents with a mean baseline BMI above the 95th centile increased their risk of early mortality by 80–100% compared with adolescents whose mean baseline BMI was between 25th and 75th centiles. Despite this intriguing data, it is notoriously difficult to establish which adolescents will persist with an excess body weight into adulthood. This is partly because adolescents have yet to reach their full height and partly because the etiology of obesity is related to lifestyle factors that may change with time. Attempts to track fatness from childhood to adulthood have produced contradictory results, with some authors claiming that certain ages, such as 7 years and adolescence, are 'risk' points for the development of later obesity and others finding that only the adiposity of older adolescents tracks to adulthood. Thus, there is no guarantee that the overweight adolescent will remain so in later life.

Sexual Development

In girls, the onset of menarche at around 13 years is triggered by the attainment of a specific level of body fat, with taller, heavier girls more likely to experience an early menarche. Vigorous exercise, e.g., gymnastics and endurance running, can delay the menarche, due both to the physiological effects of regular training and the depletion of body fat. Iron becomes more important for girls as menstrual periods become regular and heavier, and there is evidence that the iron status of many girls may be inadequate. Low iron status in this age group is, in part, due to higher requirements, but it is also linked to nutritional practices such as missing breakfast, avoiding red meat, and dieting.

Dietary Recommendations

There are, of course, a variety of national recommendations for nutritional intake, which, for adolescents, are normally based on a combination of deficiency studies and extrapolations from adult studies. In the UK, US, and Canada, guidelines have evolved from a simple recommended dietary intake (RDI) to a more complex bell-shaped distribution with a mean representing the intake likely to satisfy the needs of 50% of the population. The upper extreme, at the 97.5th centile, represents the intake likely to meet the needs of the majority of the population, while the lower extreme, at the 2.5th centile, represents the lowest acceptable intake. Current UK reference nutrient intakes (RNIs), presented in **Table 2**, cover a range of nutrients from fats and sugars to the main micronutrients. Dietary guidelines are an important reference point for nutrition scientists and dietitians, but it must also be borne in mind that they relate to the average needs of populations, rather than individuals.

Instead of numerical recommendations, many nations have adopted more conceptual ways of representing the ideal diet. This makes sense as recommended nutrient intakes are poorly understood by the public and need to be put into context by health professionals. Communication tools such as the plate model, pyramid system, food groups, and traffic light system can help to get healthy eating messages across to adolescents.

Dietary Intakes

There is a lay belief that most adolescents have a nutritionally inadequate diet yet, despite reported

Table 2 UK Dietary guidelines for adolescents

(a) Dietary reference values macronutrients

Age group (years)	*Sex*	*Energy (MJ)*	*Protein (g)*	*NSP (g)*	*Fat (% energy)*	*Starch/intrinsic sugars (% energy)*	*Nonmilk extrinsic sugars (% energy)*
11–14	M	9.27	42.1	18	35	39	11
	F	7.92	41.2	18	35	39	11
15–18	M	11.51	55.2	18	35	39	11
	F	8.83	45.0	18	35	39	11

(b) Reference nutrient intakes vitamins and minerals

Age group (years)	*Sex*	*Vit. B_2 (mg)*	*Vit. B_2 (mg)*	*Niacin (mg)*	*Vit. B_6 (mg)*	*Vit. B_{12} (μg)*	*Folate (μg)*	*Vit. C (mg)*	*Vit. A (μg)*	*Ca (mg)*	*Fe (mg)*	*Zn (mg)*
11–14	M	0.9	1.2	15	1.2	1.2	200	35	600	1000	11.3	9.0
	F	0.7	1.1	12	1.0	1.2	200	35	600	800	14.8	9.0
15–18	M	1.1	1.3	18	1.5	1.5	200	40	700	1000	11.3	9.5
	F	0.8	1.1	14	1.2	1.5	200	40	600	800	14.8	7.0

NSP, nonstarch polysaccharide.

low intakes of some micronutrients in surveys, there is little evidence of widespread clinical deficiencies, or indications that adolescents are failing to achieve appropriate heights and weights. Iron is the exception, where mean intakes are low and clinical markers suggest deficiency in some age groups. There is justifiable concern about the general healthiness of diets eaten by 'at risk' subgroups such as dieters, smokers, strict vegetarians, and adolescents who drink excess amounts of alcohol.

Dietary surveys

Mean daily intakes of energy and selected micronutrients from a selection of major international surveys of adolescents are presented in **Table 3**. Caution should be exercised when interpreting data from dietary surveys because under-reporting of energy is widespread in adolescent and adult populations. Selective under-reporting, often focused on energy-dense or high-fat foods, can partially explain low reported intakes of energy and certain micronutrients. It is also complex to make comparisons between the data from different countries given the range of dietary assessment methods used. There is normally a trade-off between sample size and methodology, which sees the larger surveys favoring less precise methods such as 24-h recalls or food frequency questionnaires in order to make data collection more economical. The results of the most recent UK National Diet and Nutrition Survey (NDNS) of 2672 young people aged 4–18 years (adolescent values given in **Table 4**) will be discussed in detail as this represents a survey with particularly strong dietary methodology (i.e., 7-day weighed inventory).

Energy and Protein

Despite mean height and weight data, which are consistent with expected results, energy intakes in UK adolescents remain below estimated average requirements (EARs). Mean energy intakes for boys and girls were 77–89% of EARs; a similar finding to that demonstrated by surveys of younger children and adults. Girls aged 15–18 years had the lowest energy intakes as a proportion of EARs and, apart from under-reporting, this could be due to smoking, slimming, or indeed lower than anticipated energy expenditure. It is well documented that physical activity is particularly low in adolescent girls. Indeed, the NDNS reported that 60% of girls (and 40% of boys) failed to perform the recommended amount of 1 h moderate physical activity per day. Popular sources of energy in the UK adolescent diet included cereal products (one third of energy), savory snacks, potatoes, meat/meat products, white bread, milk/dairy products, biscuits/cakes, spreading fats, and confectionery. Soft drinks contributed on average 6% of energy intakes.

Figure 1 gives a comparison of energy intakes across a range of countries; mainly in Europe. The values represent the mean of reported energy intakes for children aged 9–18 years in these countries, with the majority of surveys focusing on intakes of 11–18 year olds. It is interesting that a large number of countries display similar results (around 10 000 kJ day^{-1}), with a handful of countries, namely Germany, Greece, Portugal, Sweden, and the UK displaying intakes closer to 8000 kJ. For these countries, under-reporting, lower energy requirements, or conscious energy restriction prompted by weight concerns could be reasons for the apparent low intakes.

Table 3 Key international surveys of adolescent dietary intakes

Country	*Sex (age in years)*	*Energy (mJ)*	*Energy (kcal)*	*Protein (% energy)*	*CHO (% energy)*	*Sugars (g)*	*Fat (% energy)*	*Fe (mg)*	*Ca (mg)*	*Vit. A (μg)*	*Vit. B_1 (mg)*	*Vit. B_2 (mg)*	*Vit. B_6 (μg)*	*Vit. B_{12} (μg)*	*Niacin (mg)*	*Folate (μg)*	*Vit. C (mg)*
Australia 24HR 1995	M (12–15)	11.59	2777	15.1	50.9	33.5	24.7	16.1	1093	1296	2.4	3.0	–	–	46.0	271	121
	M (16–18)	13.53	3233	15.4	49.6	32.9	24.5	17.9	1280	1186	2.3	3.0	–	–	53.5	313	154
	F (12–15)	8.53	2038	8.5	51.1	33.1	25.6	11.0	784	1130	1.5	2.0	–	–	33.4	206	124
	F (16–18)	8.69	2076	8.7	50.1	32.1	24.0	11.1	801	877	1.5	1.8	–	–	35.3	217	126
Austria 7dUR, 24HR 1991, 2002	M (11–14)	9.49	2268	13.2	48.2	–	35.2	13.0	903	–	1.4	1.6	1.5	5.7	–	229	113
	M (15–18)	11.65	2784	12.9	50.0	–	37.2	15.4	1002	–	1.4	1.7	1.5	–	–	247	140
	F (10–14)	9.49	2268	12.6	49.4	–	35.8	10.2	834	–	1.1	1.4	1.3	5.0	–	217	132
	F (15–18)	8.49	2029	12.7	49.5	–	33.5	13.4	784	–	1.0	1.2	1.2	4.0	–	201	99
Belgium 3dUR, FFQ 1991, 1995	M (11–12)	11.49	2746	11.6	–	–	–	–	–	–	–	–	–	–	–	–	–
	M (12–18)	13.06	3122	13.0	48.6	149	37.2	13.4	913	–	1.5	1.7	1.6	–	–	–	83
	F (11–12)	11.72	2802	11.6	–	–	–	–	–	–	1.0	–	–	–	–	–	–
	F (1–18)	9.44	2256	14.9	48.8	112	36.7	8.2	805	–	1.2	1.3	1.2	–	–	–	78
Canada 24HR, 1993	M (13–15)	9.71	2321	15.0	51.0	–	34.0	15.8	1299	1191	1.7	2.2	1.6	5.1	–	205	110
	F (13–15)	7.09	1695	15.0	54.0	–	32.0	10.5	954	892	1.2	1.6	1.1	3.4	–	155	118
Denmark 7dUR 1995	M (11–14)	10.90	2605	–	51.0	–	35.0	–	1286	–	1.5	2.2	1.7	6.7	27.0	304	79
	M (15–18)	12.15	2903	14.0	–	–	35.0	–	1362	–	1.5	2.3	–	7.1	30.0	295	80
	F (11–14)	8.70	2079	–	51.0	–	34.0	–	1061	–	1.1	1.7	1.4	5.1	23.0	238	72
	F (15–18)	9.70	2318	14.0	–	–	34.0	–	1121	–	1.2	1.8	1.5	5.5	23.0	266	79
Finland 24HR, 4dUR, 3DUR 1996–97	M (12–13)	10.2	2437	–	–	–	–	–	1230	–	1.8	2.2	–	–	28.0	–	81
	M (12–18)	–	–	–	47.0	–	40.0	19.7	–	–	–	–	–	–	17.0	–	–
	M (15–16)	11.8	2820	15.0	–	–	–	–	–	–	–	–	–	–	–	–	–
	F (12–13)	8.5	2031	–	–	–	–	–	–	–	1.7	–	–	–	–	–	–
	F (12–18)	–	–	–	50.0	–	37.0	13.5	–	–	–	–	–	–	–	–	–
	F (15–16)	8.6	2055	14.0	–	–	–	–	–	–	1.9	–	–	–	–	–	–
France DH, 1dWR 1988, 1993–94	M (10–13)	–	–	–	47.8	142.5	–	–	1250	–	1.2	2.1	1.7	11.0	–	–	88
	M (11–14)	10.83	2587	15.4	–	–	36.5	12.6	835	–	1.4	1.8	1.8	5.6	–	253	91
	M (11–18)	–	–	15.7	–	–	–	–	–	–	1.0	2.2	–	–	–	–	–
	M (13–18)	12.10	2892	14.9	48.8	126.8	36.0	12.5	1300	–	1.4	–	2.0	7.0	–	–	127
	F (10–13)	–	–	–	47.7	113.3	–	–	1100	–	1.0	1.8	1.5	7.5	–	–	99
	F (11–14)	8.84	2112	15.9	–	–	–	11.4	835	–	–	1.8	1.8	5.6	17.0	253	91
	F (11–18)	–	–	16.1	–	–	–	–	–	–	1.3	–	–	–	–	–	–
	F (13–18)	9.16	2188	16.1	45.7	98.2	–	10.4	1100	–	–	1.7	1.4	7.0	–	–	112

Continued

Table 3 Continued

Country	*Sex (age in years)*	*Energy (mJ)*	*Energy (kcal)*	*Protein (% energy)*	*CHO (% energy)*	*Sugars (g)*	*Fat (% energy)*	*Fe (mg)*	*Ca (mg)*	*Vit. A (μg)*	*Vit. B_1 (mg)*	*Vit. B_2 (mg)*	*Vit. B_6 (μg)*	*Vit. B_{12} (μg)*	*Niacin (mg)*	*Folate (μg)*	*Vit. C (mg)*
Germany DH, 3-d/7-d Recall, 1dWR 1985–95, 1998	M (10–12)	9.08	2170	12.9	46.0	–	38.0	12.4	795	–	1.1	1.2	1.3	4.4	24.0	221	87
	M (13–14)	10.41	2487	13.2	45.6	–	37.5	14.3	893	–	1.3	1.5	1.4	5.5	28.6	245	98
	M (15–18)	11.14	2661	13.2	46.9	–	38.3	14.8	902	–	1.4	1.7	1.6	5.7	30.4	263	97
	F (10–12)	7.78	1860	12.9	47.6	–	36.4	11.1	681	–	1.0	1.2	1.1	4.0	20.4	203	87
	F (13–14)	8.49	2028	12.6	45.1	–	39.2	12.2	754	–	1.1	1.1	1.2	4.2	24.1	210	98
	F (15–18)	8.59	2052	13.1	46.5	–	37.0	12.3	728	–	1.0	1.3	1.5	4.4	24.6	216	92
Greece 1dWR, 24HR 1993–94, 1999	M (10–11)	–	–	15.7	44.0	–		11.0	963	–	–	–	–	–	–	–	119
	M (12–14)	8.90	2126	15.6	45.0	97	40.0	13.5	1011	–	1.7	2.1	1.8	4.7	18.0	226	112
	M (14–16)	9.00	2151	15.0	47.2	–	–	13.8	871	–	2.5	2.0	1.9	4.3	19.2	251	123
	F (10–11)	–	–	15.6	44.0	–		10.0	851	–	–	–	–	–	–	–	108
	F (12–14)	9.70	2318	14.7	48.0	88	41.8	10.1	748	–	1.4	1.6	–	4.1	–	212	108
	F (14–16)	7.08	1692	14.5	46.0	–	–	9.4	771	–	2.4	1.5	1.3	3.2	13.2	217	118
Ireland DH 1988	M (11–14)	11.3	2700	–	50.3	–	36.3	14.7	1208	–	1.8	2.5	2.2	4.9	40.2	246	76
	M (15–17)	14.0	3346	14.2	49.3	–	36.0	19.3	1549	–	2.2	3.1	2.6	7.2	51.7	306	95
	F (11–14)	9.10	2174	–	50.2	–	36.0	–	962	–	–	1.9	1.7	3.9	32.0	198	76
	F (12–15)	–	–					12.4	–	–	1.4	–	–	–	–	–	–
	F (15–17)	8.90	2127	13.9	48.9	–	37.1	11.6	950	–	1.3	1.8	1.6	4.0	32.0	182	79
Japan UR n/a	M (15–19)	10.6	2545	14.4	51.9	–	28.3	8.6	633	978	1.2	1.4	1.3	8.1	16.6	303	89
	F (15–19)	8.0	1918	15.1	50.4	–	29.7	7.4	516	875	0.9	1.2	1.1	6.4	13.2	268	91
Netherlands 2dUR 1997–98	M (13–16)	10.9	2605	13.1	51.2	188	35.5	10.9	1045	778	1.2	1.6	1.6	3.9	–	–	79
	M (16–19)	11.6	2772	13.3	49.5	184	35.4	11.5	1095	972	1.3	1.6	1.8	4.4	–	–	71
	F (13–16)	8.7	2079	13.7	50.3	146	35.9	9.0	904	724	1.0	1.4	1.3	3.4	–	–	81
	F (16–19)	9.1	2175	13.4	50.3	152	35.5	9.9	908	754	1.2	1.4	1.4	3.4	–	–	81
New Zealand 24HR, FFQ 1997	M (15–18)	12.4	2963	15.0	49.0	82	35.0	15.2	957	505	1.8	2.1	1.8	4.9	43.0	280	155
	F (15–18)	8.86	2117	14.0	51.0	69	34.0	10.4	783	342	1.3	1.5	1.1	3.2	28.0	203	120
Norway 1dWR, FFQ n/a	M (11–14)	–	–	–	–	–	31.1	–	–	–	–	–	–	–	–	–	–
	M (13–14)	15.0	3585	–	–	–	–	–	1625	–	2.1	2.8	–	–	–	–	110
	M (13–15)	–	–	13.4	–	–	–	–	–	–	–	–	–	–	–	–	–
	F (11–14)	10.9	2605	–	54.9	–	28.9	–	–	–	–	–	–	–	–	–	–
	F (13–14)	–	–	–	–	–	–	–	1142	–	1.6	2.1	–	–	–	–	104
	F (13–15)	–	–	13.7	–	–	–	–	–	–	–	–	–	–	–	–	–

Continued

Table 3 Continued

Country	*Sex (age in years)*	*Energy (mJ)*	*Energy (kcal)*	*Protein (% energy)*	*CHO (% energy)*	*Sugars (g)*	*Fat (% energy)*	*Fe (mg)*	*Ca (mg)*	*Vit. A (μg)*	*Vit. B_1 (mg)*	*Vit. B_2 (mg)*	*Vit. B_6 (μg)*	*Vit. B_{12} (μg)*	*Niacin (mg)*	*Folate (μg)*	*Vit. C (mg)*
Portugal 24HR 1995	M (12–18)	8.86	2117	–	49.1	–	–	–	890	–	–	–	–	–	–	–	–
	M (13–17)	9.41	2248	17.6	–	–	–	–	–	–	–	–	–	–	–	–	77
	F (12–18)	9.40	2248	–	53.4	–	33.3	–	853	–	–	–	–	–	–	–	–
	F (13–17)	8.14	1945	17.8	–	–	–	–	–	–	–	–	–	–	–	–	99
Spain 24HR, FFQ 1989–92	M (10–12)	11.50	2747	15.4	51.8	–	40.3	11.3	713	749	1.4	1.4	–	4.7	28.0	128	71
	M (13–15)	12.54	2997	17.8	47.7	–	40.1	16.5	746	691	2.1	1.8	–	7.2	40.0	159	68
	F (10–12)	10.89	2602	13.5	43.9	–	40.8	–	666	1088	1.3	1.3	–	7.2	25.0	138	96
	F (13–15)	1052	2514	16.1	42.0	–	42.1	13.2	653	982	1.9	1.6	–	9.6	36.0	168	84
	F (17–18)	–	–	–	–	–	–	13.3	–	–	–	–	–	–	–	–	–
Sweden 7dUR 1989–90, 1993–94	M (13–14)	–	–	–	–	–	–	17.4	1279	–	–	–	–	–	–	–	–
	M (14–16)	8.90	2127	–	52.6	–	32.1	18.2	1406	–	1.8	2.4	2.0	6.6	33.5	178	68
	M (17–18)	10.50	2509	14.7	–	–	–	–	1472	–	1.8	2.8	2.2	8.7	36	138	77
	F (13–14)	–	–	–	49.4	–	–	–	1061	–	–	–	–	–	–	–	–
	F (14–15)	7.21	1722	–	–	–	–	13.4	1046	–	1.4	1.8	1.5	4.9	24.9	144	68
	F (17–18)	7.88	1884	14.2	54.1	–	–	13.3	966	–	1.2	1.8	1.5	5.5	23.0	105	77
Switzerland 7dUR 1994–95	M (11–12)	–	–	13.3	46.1	–	–	–	–	–	–	–	–	–	–	–	–
	M (13–14)	11.98	2863	–	–	–	40.1	16.0	1311	–	1.5	2.2	–	–	–	–	185
	M (15–18)	12.56	3001	–	–	–	35.0	–	1157	–	1.3	1.8	–	–	–	–	163
	F (11–12)	–	–	–	49.4	–	–	–	–	–	–	–	–	–	–	–	–
	F (13–14)	7.90	1887	–	–	–	37.4	9.3	819	–	–	1.3	–	–	–	–	110
	F (15–18)	8.12	1939	–	–	–	35.8	–	832	–	1.5	1.3	–	–	–	–	146
Turkey 24HR 2003	M (11–14)	9.92	2372	15.0	50.9	–	34.1	13.3	1030	1151	1.2	2.0	1.7	4.0	13.1	179	127
	F (11–14)	9.41	2250	14.6	48.2	–	37.2	11.8	1060	1386	1.1	2.0	1.7	3.9	12.6	163	135
USA 24HR 1999–2000	M (12–19)	11.24	2686	13.9	54.2	–	32.0	18.3	1081	–	–	–	–	–	–	421	–
	F (12–19)	8.34	1993	13.4	55.5	–	31.1	13.4	793	–	–	–	–	–	–	323	–

24HR refers to '24 hour' recall. WR, weighed record; FFQ, food intake questionnaire; UR, unweighed record; DH, diet history.
Vitamin A = micrograms retinol equivalent.
Dates of actual surveys are given where available. Data from more than one survey are presented for some countries.

Table 4 Average daily dietary intakes of UK adolescents from the National Diet and Nutrition Survey (2000)

Sex (age in years) Sample size	*Energy (MJ)*	*Protein (% energy)*	*CHO (% energy)*	*NMES (% energy)*	*Fat (% energy)*	*NSP (g)*	*Fe (mg)*	*Ca (mg)*	*Vit. A (μg)*	*Vit. B_1 (mg)*	*Vit. B_2 (mg)*	*Vit. B_6 (μg)*	*Vit. B_{12} (μg)*	*Niacin (mg)*	*Folate (μg)*	*Vit. C (mg)*
M (11–14)	8.28	13.1	51.7	16.9	35.2	11.6	10.8	799	577	1.71	1.74	2.2	4.5	30	247	78.4
N=234	89%	152%					96%	80%	96%	190%	145%	183%	375%	200%	124%	224%
	EAR	RNI					RNI	RNI	RNI	RNI	RNI	RNI	RNI	RNI	RNI	RNI
M (15–18)	9.69	13.9	50.5	15.8	35.9	13.3	12.6	878	628	1.93	1.95	2.7	5.0	36.8	309	86.5
N=179	83%	139%					112%	88%	90%	175%	150%	180%	333%	204%	154%	216%
	EAR	RNI					RNI	RNI	RNI	RNI	RNI	RNI	RNI	RNI	RNI	RNI
F (11–14)	7.03	12.7	51.2	16.2	36.1	10.2	9.1	641	482	1.42	1.35	1.9	3.3	24.8	210	73.7
N=238	89%	128%					61%	80%	80%	203%	123%	190%	275%	207%	105%	210%
	EAR	RNI					RNI	RNI	RNI	RNI	RNI	RNI	RNI	RNI	RNI	RNI
F (15–18)	6.82	13.9	50.6	15.3	35.9	10.6	8.9	653	562	1.41	1.34	2.0	3.4	25.6	215	81.2
N=210	77%	121%					60%	82%	94%	176%	122%	167%	227%	183%	108%	203%
	EAR	RNI					RNI	RNI	RNI	RNI	RNI	RNI	RNI	RNI	RNI	RNI

Study conducted January to December 1997 with a sample size of 2672.
EAR, estimated average requirement; RNI, reference nutrient intake; NMES, Nonmilk extrinsic sugars (similar to added sugars); NSP, nonstarch polysaccharide.

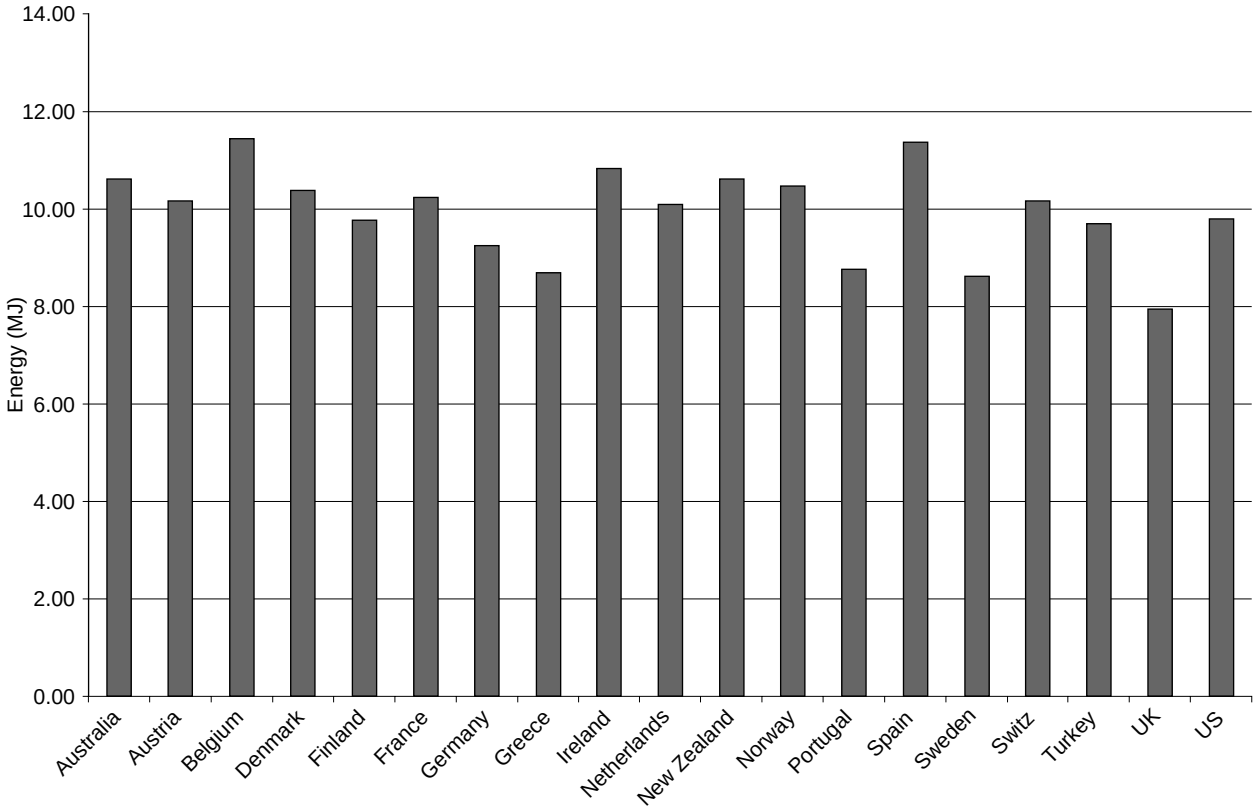


Figure 1 Reported energy intakes (kJ) for adolescents in a selection of countries.

In the NDNS, mean protein intakes were considerably in excess of requirements, as assessed by RNI, for all ages and both sexes. The main sources were meat and meat products (which contributed 30% of overall protein), cereals, bread, and milk products. It is believed that protein requirements in adolescents are between 0.8 and 1.0 g per kg body mass, although this fails to take into account any additional needs related to regular exercise (which are likely to be minor for most sports and be covered by normal protein intakes). As a proportion of energy, protein intakes were higher in Southern European countries, Australia, and New Zealand compared with intakes in the US and Northern European countries.

Fat

Mean total fat intake as a proportion of energy in the NDNS was around 35%, corresponding to the UK dietary reference value (DRV). This is lower than the intakes (38–40% energy from fat) found in previous studies. However, intakes of saturated fat, at 14% energy, still exceeded the DRV of 11% energy. Of more concern was the subgroup of adolescents in the highest percentile of intakes who consumed around 17% energy from saturated fat. This emphasizes the view that, although mean intakes may look acceptable when compared with dietary guidelines, there may be 'at risk' groups whose dietary habits predispose them to a greater risk of chronic disease. Main sources of saturated fat in the adolescent diet included meat and meat products (around 20%), savory snacks, and fried foods. In most other countries, fat intakes were 36–38% energy with the highest fat intake reported in Finland, Greece, Belgium, Germany, Switzerland, and Spain at around 38% energy. In the US, where the dietary guideline is 30%, intakes were around 32% energy from fat.

Carbohydrates

Average total carbohydrate intake in the NDNS was close to the DRV of 50% energy. The main sources were cereals, bread, savory snacks, vegetables, and potatoes. Fiber intakes, expressed as nonstarch polysaccharide (NSP), were 10–13 g day^{-1}, which approached 70% of the adult guideline. Vegetables, potatoes, and savory snacks together contributed 40% of NSP. Interesting, there was no clear relationship between NSP and bowel movements, although it was noted that adolescents who experienced less than one bowel

movement per day tended to have NSP intakes at the lowest end of the distribution spectrum. The mean intake of nonmilk extrinsic sugars (a proxy for added sugars) was 16% of energy, around 4 percentage points higher than the DRV of 11% food energy. Key sources were soft drinks (providing 42% of sugars), sugar preserves, and confectionery, particularly chocolate. Children from lower income households tended to have lower intakes of total carbohydrate, nonmilk extrinsic sugars, and NSP compared with children from higher income households.

Recommendations to reduce fat are often accompanied by those urging a decrease in added sugars due to concerns about obesity, dental health, and micronutrient dilution. However, an inverse relationship between fat and sugars is evident in the majority of dietary surveys, suggesting that concurrent reductions in fat and sugar may neither be realistic nor totally beneficial. A previous survey found a difference of 4% energy from fat between children in the lowest and highest thirds of sugar intake. Observational studies, including the latest NDNS, have also found an inverse relationship between body mass index and sugar intake. Explanations for this include self-imposed sugar restrictions amongst heavier people, and food choices in favor of higher sugar, low-fat foods, which could be less obesigenic. With respect to the potential impact of added sugars on micronutrient dilution, studies in the UK, Germany, and the US have found that a broad range of sugar intake is consistent with adequate micronutrient intakes. This may be partly due to fortification of sugar-containing foods, e.g., breakfast cereals. Lower levels of vitamins and minerals tend to be seen only at the upper and lower extremes of sugar consumption, suggesting that these diets lack variety.

Micronutrients

Main sources of micronutrients are breakfast cereals, milk, bread, chips/potatoes, and eggs. Surveys that report comparisons between intakes and recommendations have found satisfactory intakes for most micronutrients when means are considered. Intakes of vitamins B_1, B_2, B_6, B_{12} C, and niacin greatly exceeded RNIs in the NDNS, perhaps reflecting high protein intakes and the fortification of popular foods such as breakfast cereals, bread, and beverages. Even folate, a problem nutrient in earlier studies, was consumed at an acceptable level.

Nutrient intakes that remain at lower than expected levels were iron and zinc for both sexes, and calcium and vitamin A for girls. Mean iron intake was particularly low in 11–18-year-old girls at 60% of the RNI (see **Table 4**). Mean iron intakes often fail to meet recommended levels in the majority of studies reported, particularly in women and girls. This may reflect avoidance of iron-containing foods, e.g., red meat, for reasons of perceived health, food safety, or dislike. Iron status is also hampered by absorption rates, which can be as low as 10%. It is important to reverse this trend as increasing numbers of young girls are now demonstrating clinical evidence of poor iron status, e.g., more than a quarter of 15–18-year-old girls in the NDNS. A New Zealand survey reported that 4–6% of adolescents were anemic. Good sources of iron are meat/meat products, breakfast cereals, bread, chips/potatoes, chocolate, and crisps. Around 25% of iron intakes are from fortified foods, which supply non-heme iron. The latter four food groups are not particularly rich in iron but, nevertheless, contribute over 10% due to the significant amounts eaten.

Poor intakes of calcium are of concern due to the rising incidence of osteoporosis in later life, especially amongst women. While average calcium intakes were around 80% of the RNI in the NDNS, there was a considerable proportion of adolescents with intakes below the lower RNI (the bottom end of the acceptable spectrum). In 11–14-year-old children, 12% of boys and 24% of girls fell into this category, while in 15–18 year olds, the figures were 9% and 19%, respectively. Good sources of calcium are milk, cheese, yogurt, tinned fish, and, in many countries, fortified grain products. Concern has been expressed that the rise in soft drink consumption has displaced milk from the diets of adolescents and this could be contributing to the low calcium intakes found in many surveys. Fluid milk consumption has fallen dramatically over the last decade in Western countries and this is due to a range of factors including preference for other beverages, dieters' concerns about calories, and attitude of adolescents towards milk. It should not be forgotten that physical activity is an important aspect in the prevention of osteoporosis. Some life-style practices, such as smoking and drinking alcohol, are related to a higher requirement for micronutrients, suggesting that specific groups of adolescents may be more at risk from a poor nutrient status.

Impact of Lifestyle on Dietary Intakes

Young people consume particular foods and diets for a variety of reasons, often completely unrelated to their nutritional content. These can include:

slimming or weight control (whether justified or not); peer group pressure to consume certain foods or brands; the development of personal ideology, such as the use of vegetarian diets; following a specific diet to enhance sporting prowess; or even convenience. Energy and nutrient intakes are influenced by specialized eating patterns, thus it is important to consider life-style choices when interpreting dietary survey data.

Breakfast Consumption

Breakfast is identified in many studies as a nutrient-dense, low-fat meal, yet is often omitted by adolescents. Around 10% of younger children miss breakfast, rising to 20% as adulthood is approached. Boys omit breakfast less than girls and favor cereals rather than bread or a cooked breakfast. Data on breakfast habits have revealed higher intakes of sugars, fiber, and micronutrients, such as folate, niacin, iron, calcium and zinc, amongst high breakfast cereal consumers. Fat intakes, as a proportion of energy, are inversely related to breakfast cereal intake, probably due to the higher carbohydrate intakes of breakfast consumers. Previous surveys of adolescents have found an inverse relationship between breakfast cereal consumption and body mass index.

Consumption of School Lunches

Although the popularity of school lunches has diminished over the last 10 years, they are still eaten regularly by almost 40% of children, particularly those from lower socioeconomic groups. School lunches have been found to contribute 30–40% of total energy and are often criticized for containing a high proportion of fat and low levels of key micronutrients such as vitamin C and calcium. Older children often prefer to eat lunch at cafes and take-aways rather than consider school meals and this practice has been found to relate to lower nutrient-dense diets, particularly in the case of iron. Initiatives have been taken forward in many schools to improve the quality and perception of school meals including action groups involving pupils, caterers, and teachers. There have also been efforts at government level to integrate the production of school meals with classroom-based topics around nutrition, health, and life style. It is too early to say whether these efforts have had a significant impact on the nutrition of adolescents.

Snacking and Soft Drink Consumption

There has been a general shift over the last decade towards fewer meals eaten at home and more eaten in restaurants and cafes combined with an increase in snacking. Snacks, including soft drinks, now contribute a significant proportion of the daily energy intake of adolescents. Concerns about the possible impact of snacks on measures of overweight and nutrient composition have not been borne out by the evidence, although it is acknowledged that data collection in this area is complicated by the myriad of definitions for 'snack.' A number of observational studies have found that frequent snackers have similar nutrient intakes to those who snack infrequently. With respect to body size, snacking tends to relate to a lower body mass index rather than one that is high. Intervention studies also provide valuable evidence on the effects of snacking. A study in adults, which attempted to increase consumption of snacks to around 25% of daily energy using a variety of low- and high-fat products, found that the subjects compensated for the additional energy by reducing the amount eaten at meals. While these data suggest that snacking is more benign than was previously thought, it is important to emphasize the concept of balance. Common snack foods amongst adolescents are potato crisps, carbonated drinks, biscuits, and confectionery. While these foods certainly have a role in creating variety and enjoyment in the diet, no one would argue that they should represent the primary sources of energy for young people. In the case of soft drinks, evidence from short-term intervention studies suggests that higher intakes (in excess of two cans per day) are linked with higher energy intakes and lower intakes of micronutrients. Yet most epidemiological studies show an inverse correlation between sugar consumption (a proxy for soft drink consumption) and mean body mass index. Further work is needed to determine optimal cut-offs for soft drink intakes, particularly for adolescents who tend to be major consumers.

Smoking

The proportion of adolescent smokers rises with age and is between 8% and 20% with an average exposure, in older children, of around 40 cigarettes per week. Since the 1980s, smoking has decreased in adolescent boys but not in girls. Smokers tend to have different dietary habits from nonsmokers and this is reflected in their nutrient intakes. Studies have found that smokers consume less dairy foods, wholemeal bread, fruit and breakfast cereals, and more coffee, alcohol and chips. Smokers' diets tend to be lower in fiber, vitamin B_1, and vitamin C compared with nonsmokers. In a study of 18 year olds, male smokers had higher percentage energy from fat and lower intakes of sugars and iron. Contrary to

evidence from adult surveys, smoking has not been found to relate to body size in adolescents, although the opposite is believed to be true for teenage girls who use smoking as a misguided means to control energy intake. As would be expected, dietary restraint is more common amongst female smokers.

Consumption of Alcohol

In the NDNS, alcohol was consumed by 10% of 11–14 year olds and 37–46% of 15–18 year olds with older boys most likely to drink alcohol. Other European surveys have found higher proportions (60–90% in 14–18-year-old males), while US surveys have found similar proportions to the UK. The average contribution of alcohol to energy intakes in the NDNS was just over 1%, with higher contributions reported by Danish and Irish studies (around 2–5% energy). Excess alcohol intake can increase micronutrient requirements but few younger adolescents fall into this category. However, binge drinking in the 15–18-year-old age group is a concern. One US study found that 20% of adolescents could be classed as problem drinkers, while 7% could be classed as alcoholics. Regular moderate consumption of alcohol can contribute to obesity since the energy provided by alcoholic drinks rarely displaces energy from other food sources. This is likely to increase overall daily energy intakes and could lead to a positive energy balance.

Other Factors that Impact on Dietary Intakes

Comparisons between boys and girls often reveal differences in dietary patterns, yet these are seldom consistent between surveys. On the whole, boys eat more meat and dairy products, while girls favor fruit, salad vegetables, and artificially sweetened drinks. The dietary practices of girls are more likely to be influenced by a desire to limit energy intakes. Lower intakes of dairy products, meat, and breakfast cereals seen in older adolescent girls explain their typically poor intakes of iron and calcium.

Differences in diet are sometimes seen between children from different social classes or income groups. In the NDNS, children from a lower socioeconomic background were less likely to consume low-fat dairy foods, fruit juice, salad vegetables, high-fiber cereals, fruit juices, and fruit than children from a higher socioeconomic background. This impacted on mean daily nutrient intakes with lower socioeconomic children consuming less protein, total sugars, total carbohydrate, and fiber. There was a similar trend for micronutrients, particularly vitamin C. Some surveys have found higher fat intakes in children from lower socioeconomic backgrounds. Such a dietary pattern, characterized by lower than optimal levels of protective nutrients, combined with a higher prevalence of smoking, may partly explain the higher burden of chronic disease experienced by people from lower socioeconomic groups.

Promoting Optimal Diets

The findings of the studies shown in **Tables 3** and **4** reveal that most adolescents in the developed world are likely to be receiving adequate energy and protein to support growth. The intakes of micronutrients found in subgroups of the population may not be high enough to ensure optimal health but it is difficult to interpret the effects of these without appropriate biochemical data. For iron, there is good evidence of clinical deficiency in low iron consumers, particularly girls but for other nutrients, biochemical evidence is scarce. Longitudinal studies that attempt to link early diet with the incidence of later disease are a valuable tool and seem to suggest that high intakes of fruit, vegetables, folate, and *n*-3 polyunsaturated fatty acids (present in oily fish) are dietary indicators that relate to important aspects of health later in life. Despite these scientific findings, health messages relating to fruit and vegetables seem to have fallen on deaf ears. The NDNS showed that 70% of children had eaten no citrus fruit during the week of the dietary survey. Around 60% had eaten no green leafy vegetables or tomatoes, valuable sources of vitamins and minerals.

Since energy intake is the main predictor of micronutrient intakes, it makes sense to ensure that adolescents avoid restricting energy. Yet this finding needs to be considered against a background of rising obesity in the adolescent population. There is strong evidence that adolescence is the time when substantial reductions in physical activity are seen and such a trend, combined with lower energy intakes, could result in larger numbers of children failing to meet their individual nutrient requirements.

The key to tackling this lies as much with physical activity as it does with dietary intervention. Energy intakes need to be maintained at a level suitable for optimal micronutrient uptake while, at the same time, energy expenditure should be increased to ensure energy balance. A wide range of foods encompassing the main food groups will ensure a nutrient-dense diet. Special conditions in adolescence, such as pregnancy, lactation, and sports training, may increase requirements above normal and merit manipulation of the diet to

favor food groups known to be important sources of certain nutrients.

Conclusions

Diets of adolescents in developed countries meet the macronutrient requirements of the majority of individuals resulting in appropriate rates of growth. While fat intakes, as a proportion of energy, have continued to decline towards dietary guidelines, concern remains over the intakes of iron, calcium, zinc, and vitamin A in many subgroups of adolescents, particularly older girls. Maintaining adequate energy intakes and encouraging consumption of fruit, vegetables, lean meat, and oily fish may be a key route to achieving an optimal intake of micronutrients. Present recommendations for adolescents include a continuing reduction in dietary fat to help prevent later diseases of affluence. This should be combined with encouragement to increase physical activity in order to address the rising incidence of obesity in most developed countries.

See also: **Adolescents**: Nutritional Problems. **Alcohol**: Absorption, Metabolism and Physiological Effects; Disease Risk and Beneficial Effects; Effects of Consumption on Diet and Nutritional Status. **Calcium**. **Dietary Surveys**. **Osteoporosis**.

Further Reading

Alexy U, Sichert-Hellert W, and Kersting M (2003) Associations between intake of added sugars and intakes of nutrients and food groups in the diets of German children and adolescents. *British Journal of Nutrition* **90**: 441–447.

Cruz JA (2000) Dietary habits and nutritional status in adolescents over Europe–Southern Europe. *European Journal of Clinical Nutrition* **54**(supplement 1): S29–S35.

Deckelbaum RJ and Williams CL (2001) Childhood obesity: the health issue. *Obesity Research* **9**(supplement 4): 239S–243S.

Frary CD, Johnson RK, and Wang MQ (2004) Children and adolescents' choices of foods and beverages high in added sugars are associated with intakes of key nutrients and food groups. *Journal of Adolescent Health* **34**: 56–63.

Gregory JR, Lowe S, Bates CJ *et al.* (2000) *National Diet and Nutrition Survey: Young People Aged 4 to 18 Years*. London: The Stationery Office.

Lambert J, Agostoni C, Elmadfa I *et al.* (2004) Diet intake and nutritional status of children and adolescents in Europe. *British Journal of Nutrition* **92**(supplement 2): S147–S211.

Ruxton CHS, Storer H, Thomas B, and Talbot D (2000) Teenagers and young adults. In: Thomas B (ed.) *Manual of Dietetic Practice*, 2nd ed, pp. 256–262. UK: Blackwells: Oxford.

Serra-Majem L (2001) Vitamin and mineral intakes in European children. Is food fortification needed? *Public Health Nutrition* **4**: 101–107.

Nutritional Problems

C Lo, Childrens' Hospital Boston, Harvard Medical School and Harvard School of Public Health, Boston, MA, USA

Introduction: Normal Adolescent Growth and Diets

Adolescence is a unique time of rapid growth, with half of eventual adult weight and 45% of peak bone mass accumulated during adolescence. Adolescence is a time when peak physical muscular development and exercise performance is reached. However, adolescent diets are often notorious for their reliance on snacks and 'junk foods' that are high in calories, sugar, salt, and saturated fat, which could provide extra energy for high-activity demands of teenagers, but often risk becoming part of bad habits leading to obesity and increased risk of atherosclerotic heart disease in later life. Although most studies have been on older subjects, it is now clear that many Western diseases, especially heart disease, stroke, diabetes, hypertension, and many cancers, are diet related, and that diets high in saturated fat and low in fruits, vegetables, and fiber may increase risks of heart disease.

Indeed, autopsy reports of atherosclerotic plaques already present in adolescents who died accidentally suggests that prevention of heart disease should start quite early in life. Epidemiologic evidence from large cohort studies have concluded that a striking 80% reduction of heart disease and diabetes might be achieved in those with diets lower in saturated and trans fat and higher in fruits, vegetables, folate, fiber, and *n*-3 fish oils. Other factors include regular exercise, moderate alcohol use, and avoidance of obesity and smoking.

Nutrient Requirements

About every 10 years, the Institute of Medicine convenes several committees of nutrition scientists to review the scientific literature and recommend levels of daily dietary nutrients that would keep 95% of the population from developing deficiencies.

In the past, the dietary reference intakes (DRIs) or recommended dietary allowances (RDAs) concentrated on ensuring that nutrient deficiencies were minimized by specifying lower limits of intakes. However, it is now clear that many Western diets provide too much of some nutrients such as total calories, simple carbohydrates, saturated fats, and salt. Therefore, recent editions of DRIs (see **Table 1** to **5**) have

Table 1 Recommended dietary allowances and adequate intakes

Life stage group	*Vitamin A* ($\mu g\,day^{-1}$)	*Vitamin C* ($mg\,day^{-1}$)	*Vitamin D* ($\mu g\,day^{-1}$)	*Vitamin E* ($\mu g\,day^{-1}$)	*Vitamin K* ($\mu g\,day^{-1}$)	*Thiamin* ($mg\,day^{-1}$)	*Riboflavin* (mg/day^{-1})	*Niacin* ($\mu g\,day^{-1}$)	*Vitamin B_6* ($mg\,day^{-1}$)	*Folate* ($\mu g\,day^{-1}$)	*Vitamin B_{12}* ($\mu g\,day^{-1}$)	*Pantothenic Acid* ($mg\,day^{-1}$)	*Biotin* ($\mu g\,day^{-1}$)	*Choline* ($\mu g\,day^{-1}$)
Males														
9–13 years	**600**	**45**	5*	**11**	60*	**0.9**	**0.9**	**12**	**1.0**	**300**	**1.8**	4*	20*	375*
14–18 years	**900**	**75**	5*	**15**	75*	**1.2**	**1.3**	**16**	**1.3**	**400**	**2.4**	5*	25*	550*
19–30 years	**900**	**90**	5*	**15**	120*	**1.2**	**1.3**	**16**	**1.3**	**400**	**2.4**	5*	30*	550*
Females														
9–13 years	**600**	**45**	5*	**11**	60*	**0.9**	**0.9**	**12**	**1.0**	**300**	**1.8**	4*	20*	375*
14–18 years	**700**	**65**	5*	**15**	75*	**1.0**	**1.0**	**14**	**1.2**	**400**	**2.4**	5*	25*	400*
19–30 years	**700**	**75**	5*	**15**	90*	**1.1**	**1.1**	**14**	**1.3**	**400**	**2.4**	5*	30*	425*

This table (taken from the DRI reports, see http://www.nap.edu) presents recommended dietary allowances (RDAs) in **bold type** and adequate intakes (AIs) in ordinary type followed by an asterisk (*). RDAs and AIs may both be used as goals for individual intake. RDAs are set to meet the needs of almost all (97–98%) individuals in a group. For healthy breast-fed infants, the AI is the mean intake. The AI for other life-stage and gender groups is believed to cover needs of all individuals in the group, but lack of data or uncertainty in the data prevent being able to specify with confidence the percentage of individuals covered by this intake.

Table 2 Recommended dietary allowances and adequate intakes

Life stage group	*Calcium (mg day^{-1})*	*Chromium (μg day^{-1})*	*Copper (μg day^{-1})*	*Fluoride (mg day^{-1})*	*Iodine (μg day^{-1})*	*Iron (mg day^{-1})*	*Magnesium (mg day^{-1})*	*Manganese (mg day^{-1})*	*Molybdenum (μg day^{-1})*	*Phosphorus (mg day^{-1})*	*Selenium (μg day^{-1})*	*Zinc (mg day^{-1})*
Males												
9–13 years	1300*	25*	**700**	2*	**120**	**8**	**240**	1.9*	**34**	**1250**	**40**	**8**
14–18 years	1300*	35*	**890**	3*	**150**	**11**	**410**	2.2*	**43**	**1250**	**55**	**11**
19–30 years	1000*	35*	**900**	4*	**150**	**8**	**400**	2.3*	**45**	**700**	**55**	**11**
Females												
9–13 years	1300*	21*	**700**	2*	**120**	**8**	**240**	1.6*	**34**	**1250**	**40**	**8**
14–18 years	1300*	24*	**890**	3*	**150**	**15**	**360**	1.6*	**43**	**1250**	**55**	**9**
19–30 years	1000*	25*	**900**	3*	**150**	**18**	**310**	1.8*	**45**	**700**	**55**	**8**

This table presents recommended dietary allowances (RDAs) in **bold type** and adequate intakes (AIs) in ordinary type followed by an asterisk (*). RDAs and AIs may both be used as goals for individual intake. RDAs are set to meet the needs of almost all (97–98%) individuals in a group. For healthy breast-fed infants, the AI is the mean intake. The AI for other life-stage and gender groups is believed to cover needs of all individuals in the group, but lack of data or uncertainty in the data prevent being able to specify with confidence the percentage of individuals covered by this intake.

Sources: *Dietary Reference Intakes for Calcium, Phosphorous, Magnesium, Vitamin D, and Fluoride* (1997); *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin, and Choline* (1998); *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids* (2000); and *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc* (2001). These reports may be accessed via http://www.nap.edu

Table 3 Dietary reference intakes (DRIs): tolerable upper intake levels (UL)[a], vitamins

Life stage group	*Vitamin A* ($\mu g\ day^{-1}$)	*Vitamin C* ($mg\ day^{-1}$)	*Vitamin D* ($\mu g\ day^{-1}$)	*Vitamin E* ($\mu g\ day^{-1}$)	*Vitamin K*	*Thiamin*	*Riboflavin*	*Niacin* ($\mu g\ day^{-1}$)	*Vitamin B_6* ($mg\ day^{-1}$)	*Folate* ($\mu g\ day^{-1}$)	*Vitamin B_{12}*	*Pantothenic acid*	*Biotin*	*Choline* ($g\ day^{-1}$)	*Carotenoids*
Males, females															
9–13 years	1700	1200	50	600	ND	ND	ND	20	60	600	ND	ND	ND	2.0	ND
14–18 years	2800	1800	50	800	ND	ND	ND	30	80	800	ND	ND	ND	3.0	ND
19–70 years	3000	2000	50	1000	ND	ND	ND	35	100	1000	ND	ND	ND	3.5	ND

[a]UL = The maximum level of daily nutrient intake that is likely to pose no risk of adverse effects. Unless otherwise specified, the UL represents total intake from food, water, and supplements. Owing to lack of suitable data, ULs could not be established for vitamin K, thiamin, riboflavin, vitamin B_{12}, pantothenic acid, biotin, or carotenoids. In the absence of ULs, extra caution may be warranted in consuming levels above recommended intakes.

Sources: *Dietary Reference Intakes for Calcium, Phosphorous, Magnesium, Vitamin D, and Fluoride* (1997); *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin, and Choline* (1998); *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids* (2000); and *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc* (2001). These reports may be accessed via http://www.nap.edu

Table 4 Dietary reference intakes (DRIs): tolerable upper intake levels (UL)[a], Elements

Life stage group	*Arsenic*	*Boron* ($mg\,day^{-1}$)	*Calcium* ($g\,day^{-1}$)	*Chromium*	*Copper* ($\mu g\,day^{-1}$)	*Fluoride* ($mg\,day^{-1}$)	*Iodine* ($\mu g\,day^{-1}$)	*Iron* ($mg\,day^{-1}$)	*Magnesium* ($\mu g\,day^{-1}$)	*Manganese* ($mg\,day^{-1}$)	*Molybdenum* ($\mu g\,day^{-1}$)	*Nickel* ($mg\,day^{-1}$)	*Phosphorus* ($g\,day^{-1}$)	*Selenium* ($\mu g\,day^{-1}$)	*Silicon*	*Vanadium* ($\mu g\,day^{-1}$)	*Zinc* ($mg\,day^{-1}$)
Males, females																	
9–13 years	ND	11	2.5	ND	5000	10	600	40	350	6	1100	0.6	4	280	ND	ND	23
14–18 years	ND	17	2.5	ND	8000	10	900	45	350	9	1700	1.0	4	400	ND	ND	34
19–50 years	ND	20	2.5	ND	10 000	10	1100	45	350	11	2000	1.0	4	400	ND	1.8	40

[a]UL = The maximum level of daily nutrient intake that is likely to pose no risk of adverse effects. Unless otherwise specified, the UL represents total intake from food, water, and supplements. Owing to lack of suitable data, ULs could not be established for arsenic, chromium, and silicon. In the absence of ULs, extra caution may be warranted in consuming levels above recommended intakes.

Sources: *Dietary Reference Intakes for Calcium, Phosphorous, Magnesium, Vitamin D, and Fluoride* (1997); *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin, and Choline* (1998); *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids* (2000); *and Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc* (2001). These reports may be accessed via http://www.nap.edu

Table 5 Dietary reference intakes (DRIs): estimated average requirements

Life stage group	*Vit A* ($\mu g\,day^{-1}$)[a]	*Vit C* ($mg\,day^{-1}$)	*Vit E* ($\mu g\,day^{-1}$)[b]	*Thiamin* ($mg\,day^{-1}$)	*Riboflavin* ($mg\,day^{-1}$)	*Niacin* ($\mu g\,day^{-1}$)[c]	*Vit B_6* ($mg\,day^{-1}$)	*Folate* ($\mu g\,day^{-1}$)[d]	*Vit B_{12}* ($\mu g\,day^{-1}$)	*Copper* ($\mu g\,day^{-1}$)	*Iodine* ($\mu g\,day^{-1}$)	*Iron* ($mg\,day^{-1}$)	*Magnesium* ($mg\,day^{-1}$)	*Molybdenum* ($\mu g\,day^{-1}$)	*Phosphorus* ($mg\,day^{-1}$)	*Selenium* ($\mu g\,day^{-1}$)	*Zinc* ($mg\,day^{-1}$)
Males																	
9–13 years	445	39	9	0.7	0.8	9	0.8	250	1.5	540	73	5.9	200	26	1055	35	7.0
14–18 years	630	63	12	1.0	1.1	12	1.1	330	2.0	685	95	7.7	340	33	1055	45	8.5
19–30 years	625	75	12	1.0	1.1	12	1.1	320	2.0	700	95	6	330	34	580	45	9.4
Females																	
9–13 years	420	39	9	0.7	0.8	9	0.8	250	1.5	540	73	5.7	200	26	1055	35	7.0
14–18 years	485	56	12	0.9	0.9	11	1.0	330	2.0	685	95	7.9	300	33	1055	45	7.3
19–30 years	500	60	12	0.9	0.9	11	1.1	320	2.0	700	95	8.1	255	34	580	45	6.8

[a]As retinol activity equivalents (RAEs). 1 RAE = 1 Tg retinol, 12 Tg ϑ-carotene, 24 Tg l-carotene, or 24 Tg ϑ-cryptoxanthin. The RAE for dietary provitamin A carotenoids is twofold greater than retinol equivalents (RE), whereas the RAE for preformed vitamin A is the same as RE.

[b]As α-tocopherol. α-Tocopherol includes *RRR*-α-tocopherol, the only form of α-tocopherol that occurs naturally in foods, and the *2R*-stereoisomeric forms of α-tocopherol (*RRR*-, *RSR*-, *RRS*-, and *RSS*-α-tocopherol) that occur in fortified foods and supplements. It does not include the *2S*-stereoisomeric forms of α-tocopherol (*SRR*-, *SSR*-, *SRS*-, and *SSS*-α-tocopherol), also found in fortified foods and supplements.

[c]As niacin equivalents (NE). 1 mg of niacin = 60 mg of tryptophan.

[d]As dietary folate equivalents (DFE). 1 DFE = 1 μg food folate = 0.6 μg of folic acid from fortified food or as a supplement consumed with food = 0.5 μg of a supplement taken on an empty stomach.

This table presents estimated average requirements (EARs), which serve two purposes: for assessing adequacy of population intakes, and as the basis for calculating recommended dietary allowances (RDAs) for individuals for those nutrients. EARs have not been established for vitamin D, vitamin K, pantothenic acid, biotin, choline, calcium, chromium, fluoride, manganese, or other nutrients not yet evaluated via the DRI process.

specified estimated average requirements (EARs), adequate intakes (AIs), and upper limits (ULs).

Obesity

Obesity has recently become an epidemic in the US, with 31% of American adults classified as obese (body mass index $>30\,kg\,m^{-2}$) and 68% classified as overweight (body mass index $>25\,kg\,m^{-2}$) in 2000. The prevalence of obesity in childhood tripled from 5% in 1980 to 15% in 2000 according to National Health and Nutrition Examination Surveys (NHANES). There is every indication that the developed countries of Western Europe are not far behind. Indeed, obesity is becoming a worldwide problem, rapidly increasing in many developing countries including China and India, and overtaking undernutrition as the major nutritional problem.

Although obesity affects children in all socioeconomic classes, it is more prevalent in those of lower socioeconomic status in the US and developed countries, whereas it tends to affect the well-off in developing countries. This suggests that food insecurity and poor food choices are more the problem than lack of availability because of poverty. Although only 30% of obesity begins in adolescence, some estimate that 80% of obese adolescents will become obese adults, and obese adolescents are at much more risk for diabetes and major medical complications later in life. Since long-term weight loss is usually very difficult to achieve and is often unsuccessful despite widespread attempts at dieting, efforts to prevent obesity in early life are important.

Ultimately, weight gain results from dietary energy intake exceeding metabolic basal needs and activity. Only rarely is this due to some identifiable disorder of basal metabolic requirements such as hypothyroidism. However, it is difficult to measure either dietary intake or activity with enough accuracy to detect the relatively small mismatch necessary to add weight. For example, a small increase in dietary intake of $200\,kcal\,day^{-1}$, without a corresponding increase in activity could theoretically result in a weight gain of 8 kg over the course of a year.

Although the heritability of obesity has been estimated to be on the order of 60–80% on the basis of twin studies and family histories, the genetics of obesity are complex and just beginning to be understood. Adult weight is much more reflective of biological parents rather than adoptive parents in twin studies. Known genetic syndromes producing obesity in humans are rare (on the order of 1–2% of obese patients) but should be considered, such as trisomy 21 (Down's syndrome), Prader-Willi, Bardet-Biedl and Beckwith-Wiedemann syndromes, hypothyroidism, and polycystic ovary syndrome.

The adipose fat cell is not only a passive storage site but an endocrinologically active secretor of many substances like leptin, adiponectin, and cytokines, which participate in an inflammatory response and may mediate a host of adverse consequences, including insulin resistance and diabetes. Obesity is related to an increased risk of developing type 2 insulin-resistance diabetes mellitus, hyperlipidemia, heart disease, obstructive sleep apnea, asthma and other respiratory problems, back pain and orthopedic problems, fatty liver (nonalcoholic steato-hepatitis or NASH), gallstones, and depression. The increasing incidence of type 2 diabetes in obese adolescents is already being noticed, with estimates of 200 000 diabetics under age 20 years in the US predicted to rise to a lifetime risk of developing diabetes of 33–39% for those born in the year 2000.

The rapid increase in obesity has made standards based on population percentiles meaningless as medical obesity involved more than just the top 5% of weight-for-age. Instead of just relying on cross-sectional height- and weight-for-age graphs (see **Figures 1** and **2**), there has developed a need for a more valid indicator of obesity. The body mass index (BMI) charts recently released by the Centers for Disease Control allow for tracking of BMI standards for adolescents, who should have a BMI lower than the $20–25\,kg\,m^{-2}$ expected for adults. Although long-term validation data is not as available as in adults, in adolescents obesity is considered above the 95th percentile for age, with risk for obesity defined as above 85th percentile for age.

Body mass index is defined as weight (in kilograms) divided by height (in meters) squared, and is considered the best anthropometric surrogate for body composition (see **Figures 3** and **4**). Waist size may be an easier measurement to follow in adults, and particularly identifies central adiposity. Measurements by tape and caliper of mid-arm circumference and triceps skinfolds have a fairly good correlation (0.7–0.8) with more expensive research methods of underwater weighing and dual-energy X-ray absorptiometry (DEXA), and can be made even more accurate by including biceps, subscapular, and suprailiac skinfold measurements. Bioelectric impedance measures the difference in resistance between adipose and lean body tissue, but can be affected by fluid shifts especially in ill patients.

Physical examination should include blood pressure measurement because of the high percentage of comorbidity of the metabolic syndrome (obesity, hypertension, dyslipidemia, and/or diabetes).

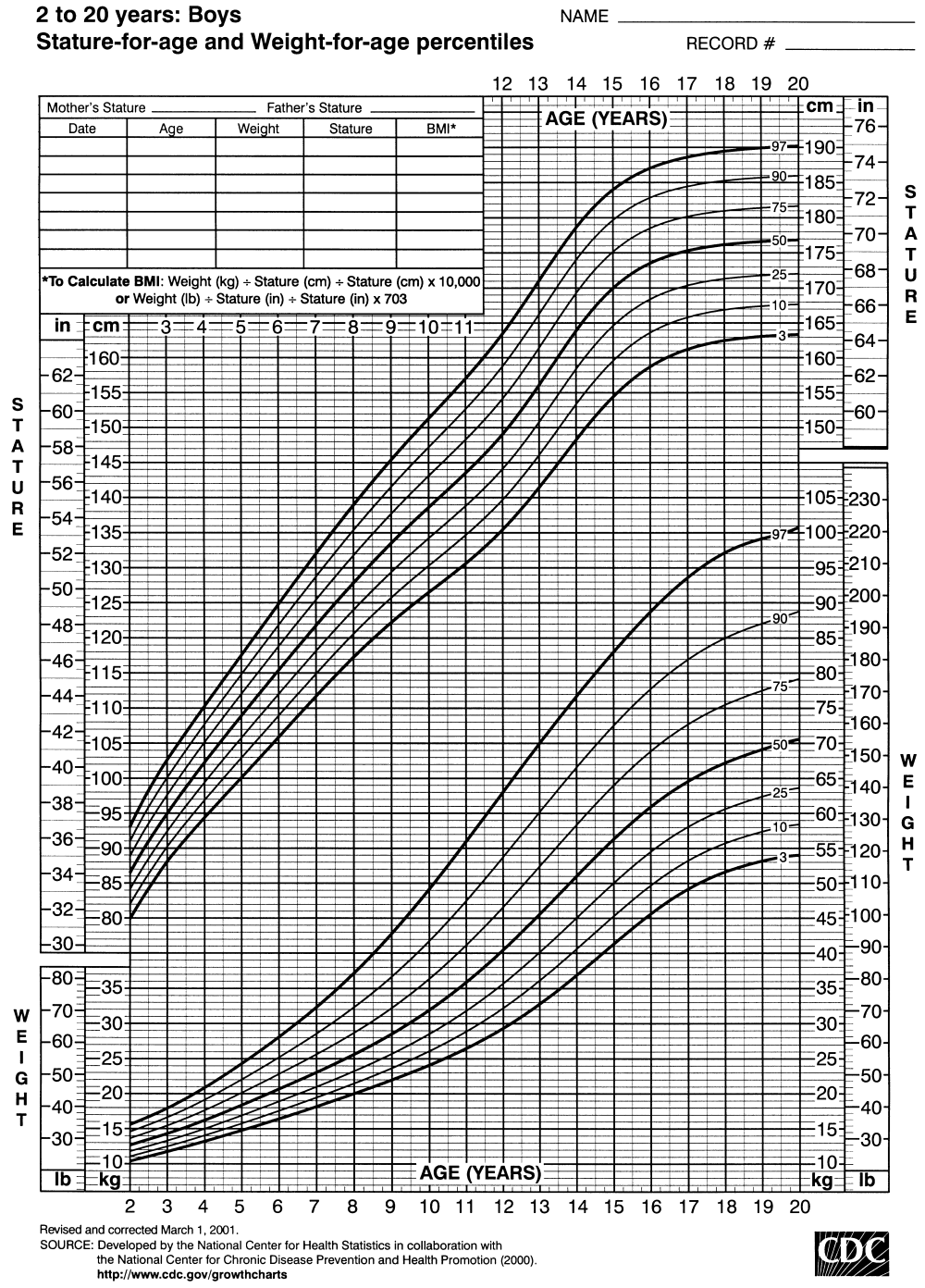


Figure 1 Weight-for-age percentiles: boys, 2–20 years. (Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion 2000: http://www.cdc.gov/growthcharts)

The metabolic syndrome is defined as three or more of the following: abdominal obesity (waist circumference greater than 40 inches (100 cm) in men or 35 inches (90 cm) in women), fasting hypertriglyceridemia ($<150\,mg\,dl^{-1}$), high fasting glucose greater than $110\,mg\,dl^{-1}$, low high-density cholesterol ($<40\,mg\,dl^{-1}$), and high blood pressure (>135/85 mm Hg). So far, it is mostly seen in later life (>40% of those over 60), but is increasingly seen at younger ages (7% of 20–29 years old). Acanthosis nigricans is a skin hyperpigmentation, chiefly around the neck, seen in about 20% of obese patients, especially African-Americans, which reflects insulin resistance and this finding should

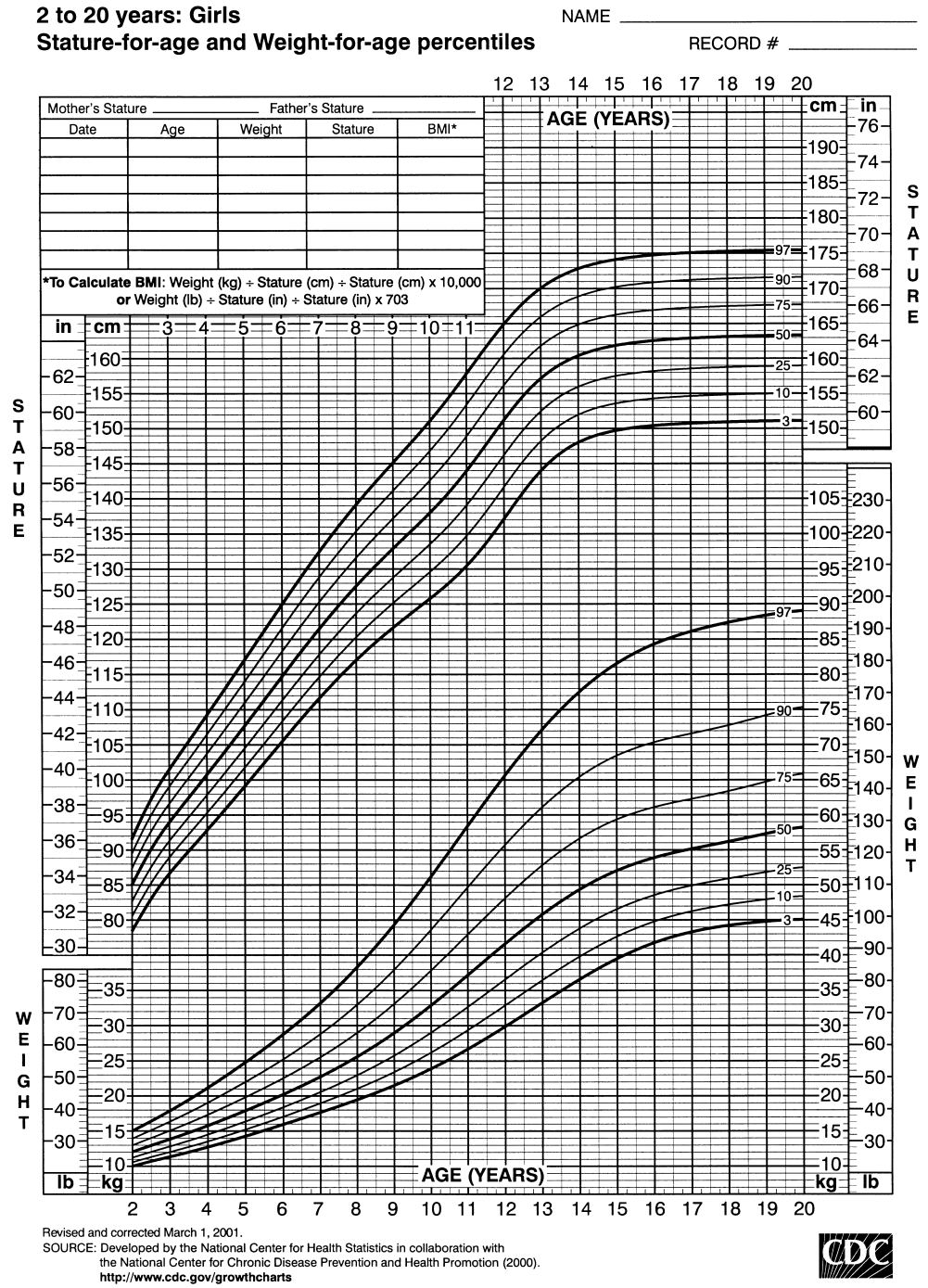


Figure 2 Weight-for-age percentiles: girls, 2–20 years. (Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion 2000: http://www.cdc.gov/growthcharts)

provoke screening tests for type 2 diabetes. Laboratory screening tests might include thyroid-stimulating hormone for hypothyroidism, fasting glucose, insulin, and glycosylated hemoglobin (HbA1C) for type 2 diabetes.

Diet histories and diet recalls are particularly important in nutritional assessments, but quantitative calorie counts are particularly unreliable in obese patients because of widespread conscious and subconscious underreporting of 20% or more. Regular meetings with a dietician should involve counseling on healthy eating choices. The recommendations regarding daily activity should include hours of television watching per day or per week because this is

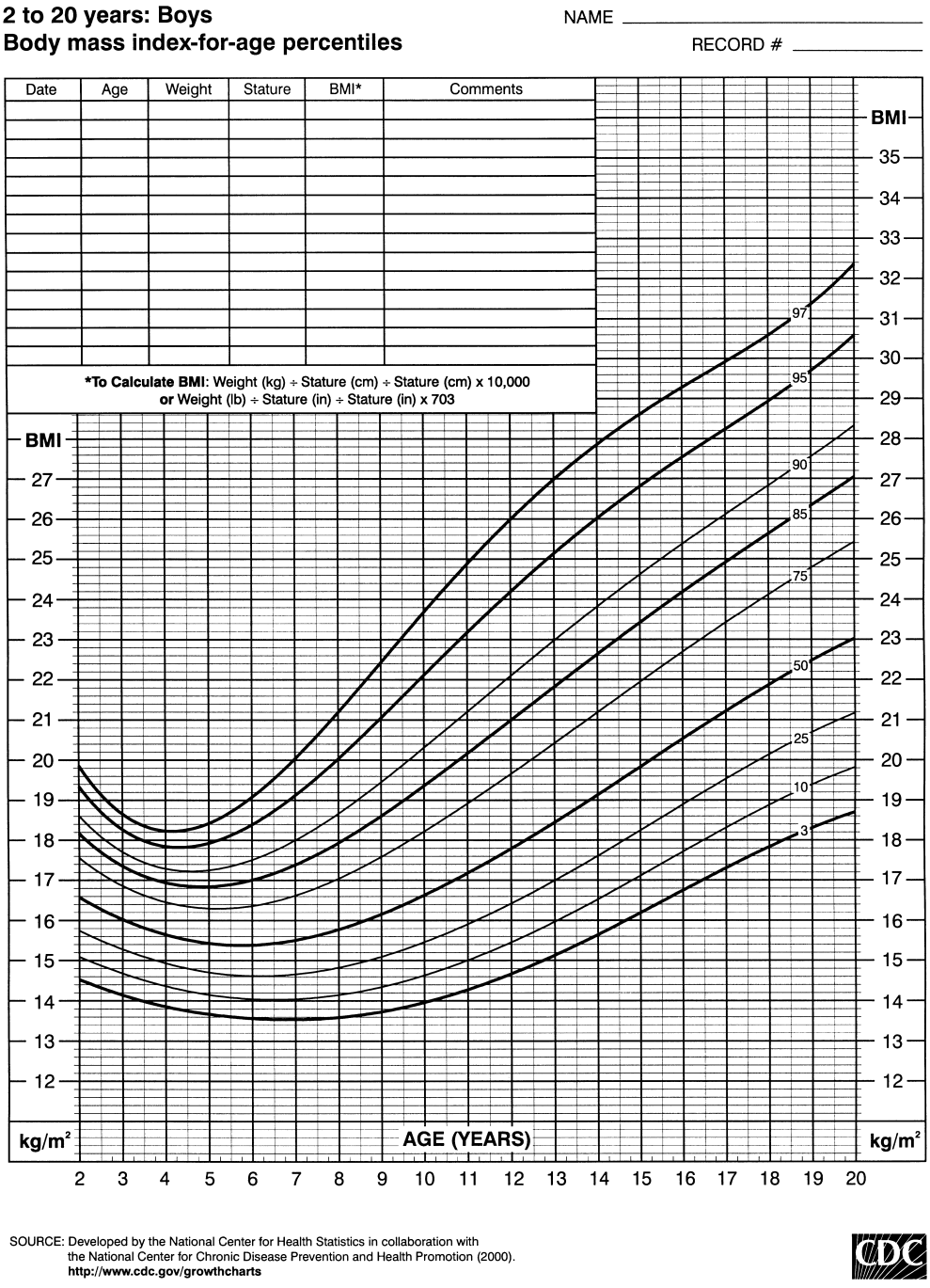


Figure 3 Body mass index-for-age percentiles: boys, 2–20 years. (Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion 2000: http://www.cdc.gov/growthcharts)

well correlated with obesity, not only because of decreased activity but also because of the influence of commercial snack food advertising.

Treatment should ideally involve a multidisciplinary team with a dietician, social worker, physical therapist, and physician, concentrating on lifestyle modification, moderate caloric restriction and regular exercise, with frequent follow-up and compliance being a good indicator of likelihood of success. Recent success with low-carbohydrate diets rather than the traditional low-fat diet advice suggests the importance of the role of satiety in maintaining caloric restriction. Most commercial diet plans promise short-term weight loss, but very few long-term studies have shown these to keep weight off for more than 6–12 months. As adolescents naturally

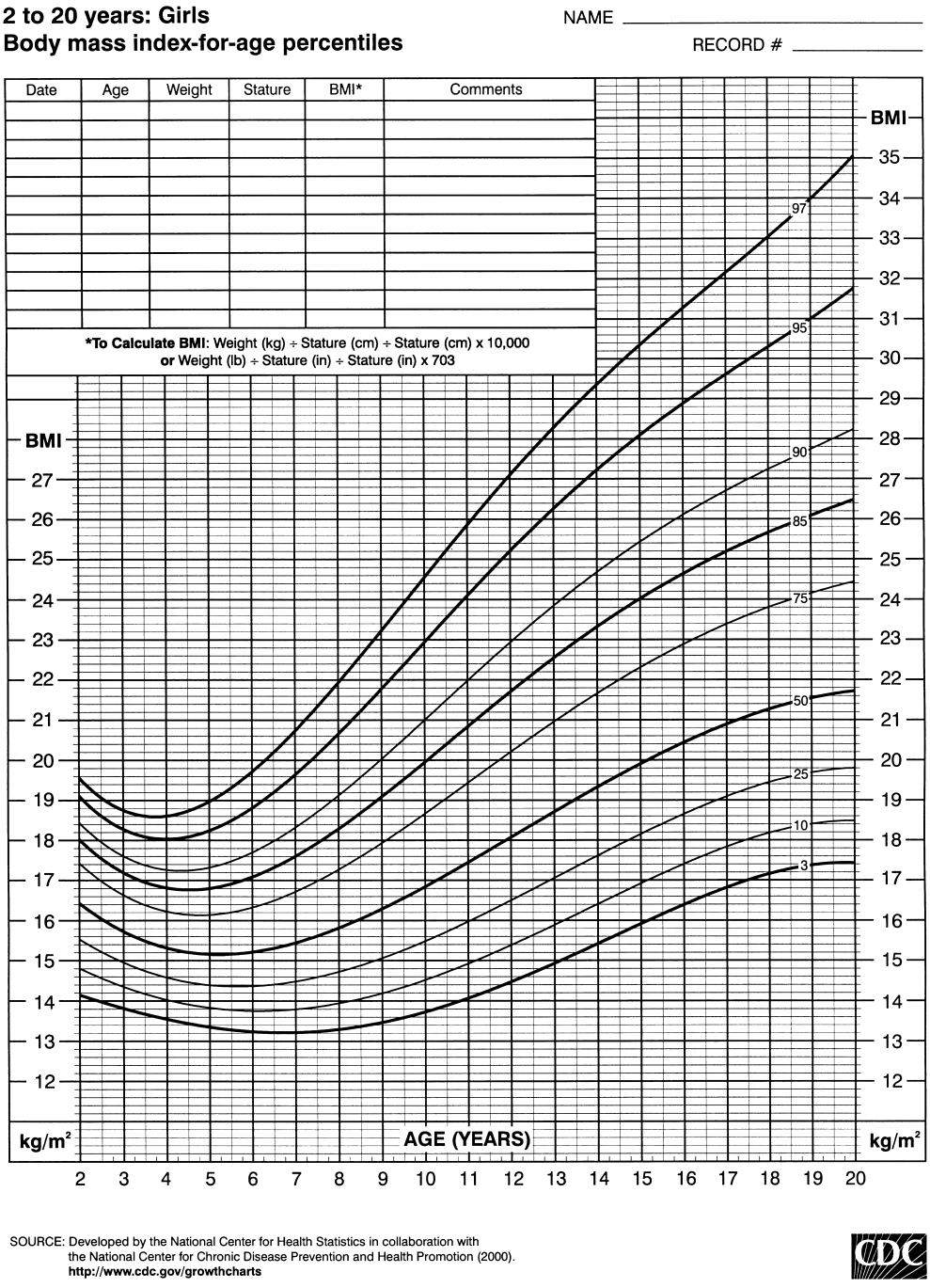


Figure 4 Body mass index-for-age percentiles: girls, 2–20 years. (Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion 2000: http://www.cdc.gov/growthcharts)

gain weight with height as they progress through puberty, it is probably more important that they learn healthy eating and activity habits over the long term rather than losing weight quickly only to gain it back within a few months.

Medications such as phenteramine-fenfluramine and stimulants have gained recent notoriety with unforeseen side effects. Possible treatment with leptin and other hormones or antagonists has much future promise, but so far has been effective only in rare patients with specific defects. Surgical gastroplasty has proven the most successful long-term therapy for massively obese adults, possibly because of suppression of ghrelin, increased satiety, and reduced hunger, but morbidity and mortality is variable and the option of major surgery should be

carefully considered only as a last resort before offering it to any adolescents.

Eating Disorders

Eating disorders affect 3–5 million in the US; 86% are diagnosed before the age of 20 and up to 11% of high-school students are affected. More than 90% are female, 95% Caucasian, and 75% have an onset in adolescence. Eating disorders are probably the most frequent causes of undernutrition in adolescents in developed countries, but only a relatively small percentage meet the full Diagnostic and Statistical Manual (DSM) IV criteria for anorexia nervosa (see **Table 6**), while most cases fall into the more general category eating disorder NOS (not otherwise specified). Bulimia, binge eating, and/or purging are probably much more common than full-blown anorexia nervosa, with some estimates of up to 20–30% of college women in the US, and often occur surreptitiously without telltale weight loss. Lifetime prevalence estimates range from 0.5% to 3% for anorexia nervosa and 1–19% for bulimia. So far eating disorders are considered rare in developing countries, but prevalence often increases dramatically when Western influences such as television advertising are introduced, as was the experience in the South Pacific Islands.

The pathophysiology of anorexia nervosa is not well understood, and there is probably a combination of environmental and psychological factors with a biochemical imbalance of neurotransmitters, especially serotonin and its precursor 5-hyroxyindole acetic acid, which tends to be reduced. There is a substantial biologic predisposition to run in families with heritability in twin studies of 35–90%.

Table 6 DSM-IV criteria for anorexia nervosa

A. Refusal to maintain body weight at or above a minimally normal weight for age and height (e.g., weight loss leading to maintenance of body weight less than 85% of that expected or failure to make expected weight gain during period of growth, leading to body weight less than 85% of that expected)
B. Intense fear of gaining weight or becoming fat, even though underweight
C. Disturbance in the way in which one's body weight or shape is experienced; undue influence of body weight or shape on self-evaluation, or denial of the seriousness of the current low body weight
D. In postmenarchal females, amenorrhea, that is, the absence of at least three consecutive menstrual cycles

Specify types

Restricting type: during the episode of anorexia nervosa, the person does not regularly engage in binge eating or purging behavior (i.e., self-induced vomiting or the misuse of laxatives or diuretics)

Binge-eating-purging type: during the episode of anorexia nervosa, the person has regularly engaged in binge eating or purging behavior (i.e., self-induced vomiting or the misuse of laxatives or diuretics)

Eating disorders should be suspected in any adolescent below normal weight ranges or with recent weight loss, but other medical conditions such as intestinal malabsorption, inflammatory bowel disease, and malignancy should also be considered. It is important to realize that most height and weight charts represent cross-sectional population norms, which may not be as sensitive as longitudinal tracking or height velocity of individuals, since puberty occurs at different ages. For example, a 12-year-old who does not gain weight for 6 months may just be entering puberty, or might be severely affected by growth failure due to a malignancy or inflammatory bowel disease.

Physical signs and symptoms of inadequate caloric intake may include amenorrhea, cold hands and feet, dry skin and hair, constipation, headaches, fainting, dizziness, lethargy, hypothermia, bradycardia, orthostatic hypotension, and edema. There is no specific laboratory diagnosis, but there are often endocrine and electrolyte abnormalities especially hypokalemia, hypophosphatemia, and hypochloremic metabolic alkalosis from vomiting, which often require careful supplementation.

Treatment may be very difficult and prolonged, often involving behavior therapy and occasionally long inpatient stays in a locked unit with threats of forced nasogastric feeding to maintain weight. There is a high risk of refeeding syndrome with edema, possible arrhythmias, and sudden death from electrolyte abnormalities, so protocols have been developed to provide a slow increase of calories, supplemented by adequate amounts of phosphorus and potassium. The anorexic patient's persistent distorted view of body image reality is very resistant to casual counseling.

The consequences of anorexia nervosa can be quite severe and include menstrual dysfunction, cardiovascular disease, arrhythmias, anemia, liver disease, swollen joints, endocrinopathies, cerebral atrophy, and even sudden death. There is a significant bone loss or osteopenia associated with amenorrhea and lack of estrogen stimulation, which is not completely reversed even with hormone replacement. Anorexia nervosa is well associated with other psychiatric diagnoses such as depression, anxiety, personality disorders, obsessive-compulsive disorder, and substance abuse, and psychiatric problems often continue to remain an issue even when normal weight is maintained. Prognosis is relatively poor compared to other adolescent medical illnesses,

with 33% persistence at 5 years and 17% at 11 years. Six per cent die within 5 years and 8.3% by 11 years.

Other Nutritional Diseases

In many countries of the world, HIV infection and acquired immunodeficiency syndrome (AIDS) has become one of the leading causes of undernutrition and cachexia, especially in younger patients. Indeed, many of the syndromes and consequences of protein-energy malnutrition are also seen in AIDS cachexia, such as frequent respiratory and other infections, diarrhea, malabsorption, and rashes. Weight loss is an AIDS-defining symptom, and weight loss of a third of usual weight usually signifies terminal illness. Fortunately, new generations of protease inhibitors and other medications have dramatically slowed the progression of HIV infection in many patients, as well as reducing the vertical transmission rate. Indeed, some studies have suggested that multivitamin supplementation of pregnant mothers may itself reduce vertical transmission rates in developing countries where antivirals are difficult to obtain. Proper attention to nutrition, with early enteral energy and micronutrient supplementation, is an important part of care, which is best instituted long before weight loss becomes manifest.

Specific Nutrients

Calcium

Calcium is the major component of bone, providing structural skeletal support to the human body (see 00033). The approximately 2–3 kg of bone calcium in each person also provides a storage reservoir for the small percentage of ionized calcium that allows muscle to contract, nerves to communicate, enzymes to function, and cells to react. The body has developed several hormonal mechanisms, including vitamin D, parathyroid hormone, and calcitonin, to protect the small amount of ionized calcium in the blood from changing drastically. Tight control of blood calcium levels is needed because unduly low blood calcium might result in uncontrolled tetanic muscle contractions and seizures, while high blood calcium levels may cause kidney stones and muscle calcifications. To increase blood calcium levels, vitamin D and its metabolites increase calcium absorption from the intestinal tract, parathyroid hormone increases calcium reabsorption from the kidney, and both increase resorption of calcium from the bone.

During the early years of life, calcium is deposited in the bone as it grows, but after about the 3rd decade, there is a steady decline in bone calcium. This is especially marked after menopause in women, when estrogen declines, and often leads to bone loss (osteopenia) to below a threshold that predisposes women in particular to fractures (osteoporosis). Osteoporosis is not just a disease of the elderly, and may occur in much younger patients, especially athletic young women, those with anorexia nervosa, those on steroids and other medications, and in anyone on prolonged bed rest, including astronauts experiencing long periods of weightlessness.

Dietary calcium is often seen as the most limiting factor in the development of peak bone mass, and strategies to increase dietary calcium have been promoted. Other factors in the development of bone mineral include height, weight, racial background and inheritance, gender, activity, vitamin D deficiency, parathyroid hormone deficiency, vitamin A, vitamin K, growth hormone, calcium, phosphorus, and magnesium. Phosphorus, the other major component of bone mineral, is relatively common in the diet.

In the 1997 DRIs, AIs of calcium were raised from 800 to 1300 mg in 9–18 year olds. Only a small percentage of the population takes in the RDA for calcium. The estimated average calcium intake in American women is only about 500–600 mg a day, and is much lower in the developing world (as low as 200 mg a day). From calcium tracer studies performed since the 1950s, intestinal calcium absorption ranges from 10% to 40% of ingested calcium, with a higher percentage absorption with lower calcium intakes. A large percentage (usually 70–80%) of dietary calcium is from milk and dairy products, which provides about 250 mg calcium per 8 oz (240 ml) glass of milk, and most studies show better absorption from dairy products than from vegetable sources. However, many people, especially non-Caucasians, develop relative lactose intolerance after childhood, and are reluctant to increase their dairy food intake.

Thus, attention has focused on whether supplementation or fortification with calcium, especially during adolescence, will ensure achievement of peak bone mass. Calcium supplementation in adolescent females has shown short-term increases in bone mineral density, but this may be because it increases mineralization in a limited amount of trabecular bone, and it remains to be seen whether this leads to long-term improvement or protection against future fractures. Also, most studies still assume that increased bone mineral density is synonymous with reduced fracture risk, although fractures may depend on many other factors such as optimal bone architecture and lack of falls. Although the

majority of scientific opinion probably favors increased dietary calcium intake in adolescence, the factors that control bone mineralization are not yet completely understood, and long-term protection against eventual bone loss and fractures remains to be demonstrated by randomized clinical trials.

Iron

Iron deficiency is one of the most common vitamin or mineral deficiencies in the world, affecting 20% or more of women and children especially in developing countries. Adolescent women who have started menses or who are pregnant are particularly at risk for developing iron deficiency, which may ldevelop long before iron stores are exhausted and anemia ensues. Anemia (low hemoglobin or red cell volume) may lead to reduced school and work performance and may affect cognitive function, as well as leading to cardiovascular and growth problems. Diagnosis is made most simply by hemoglobin level or packed red cell volume (hematocrit) and red cell morphology, or alternatively by transferrin saturation, serum ferritin, or serum iron level. Microscopic examination of a red cell smear typically shows red cells that are small (microcytic) and pale (hypochromic).

Folate

Folate is a vitamin that is responsible for one-carbon methyl transfer in a variety of cellular reactions, including formation of purines and pyrimidines, which make up DNA and RNA. Folate deficiency may result in megaloblastic anemia, as forming red cells fail to divide. As the best source of folate is in green leafy vegetables, folate nutrition may be marginal in many adolescents. Recent epidemiologic evidence suggests that folate supplementation, at levels that are higher than usual dietary intake ($200–400\,\mu g\,day^{-1}$), reduced the incidence of neural tube defects (anencephaly and spina bifida) in newborns. Supplementation needs to be started early in pregnancy, within the first 8 weeks and before most pregnancies are apparent, so should involve most women of child-bearing age. The recent decision to fortify grains and cereals with folic acid in the US will also reduce serum homocysteine levels, lowering the risk of cardiovascular disease.

Zinc and Other Minerals

Zinc is a component of many metalloenzymes including those needed for growth, pancreatic enzymes, and intestinal secretions. Although it is unusual to find a documented case of clinical zinc deficiency apart from occasional cases of acrodermatitis enteropathica, there has been recent concern over the possibility of relative zinc deficiency, especially among chronically ill patients with excessive intestinal secretions. Zinc deficiency could lead to impaired taste (hypogeusia) and appetite and immunodeficiency as well as affecting growth. A large group of adolescents in Shiraz, Iran was described to be of very short stature because of dietary zinc deficiency. Similarly, a group of people in Keshan, China was found to develop cardiomyopathy because of a selenium deficiency in the soil. Iodine deficiency is surprisingly common worldwide, perhaps involving up to half of the world population or 3 billion people, especially in areas of Southeast Asia where it is not supplemented in salt. It may cause hypothyroidism, goiter (neck masses), cretinism, or impaired intelligence if severe.

See also: **Adolescents**: Nutritional Requirements. **Anemia**: Iron-Deficiency Anemia. **Calcium**. **Eating Disorders**: Anorexia Nervosa; Bulimia Nervosa; Binge Eating. **Folic Acid**. **Iron**. **Obesity**: Definition, Etiology and Assessment. **Osteoporosis**. **Zinc**: Physiology.

Further Reading

(2002) Adolescent Nutrition: a springboard for health. *Journal of the American Dietetic Association* Supplement March.

Cheung LWY and Richmond JB (eds.) (1995) *Child Health, Nutrition, and Physical Activity,* Human Kinetics. Windsor, Ontario.

Ebbeling CB, Pawlak DB, and Ludwig DS (2002) Childhood obesity: public health crisis, common sense cure. *Lancet* **360**: 473–482.

Grand R, Sutphen J, and Dietz W (eds.) (1987) *Pediatric Nutrition*. London: Butterworth.

Heald F (1969) In *Adolescent Nutrition and Growth*. New York: Appleton Century Croft.

Hu FB, Manson JE, Stampfer MJ *et al.* (2001) Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *New England Journal of Medicine* **345**(11): 790–797.

Kleinman R (ed.) (2004) *Pediatric Nutrition Handbook*, 5th edn. Elk Grove Village, Illinois American Academy of Pediatrics.

Koletzko B, girardet JP, Klish W, and Tabacco O (2002) Obesity in children and adolescents worldwide. *Journal of Pediatric Gastroenterology and Nutrition* **202**: S205–S212.

McKigney J and Munro H (eds.) (1973) *Nutrient Requirements in Adolescents*. Cambridge: MIT Press.

Rickert VI (ed.) (1996) *Adolescent Nutrition: Assessment and Management*. Boston, MA: Jones and Bartlett.

Styne DM (2001) Childhood and adolescent obesity. *Pediatric Clinics of North America* **48**: 823–854.

Walker WA, Watkins J, and Duggan C (eds.) (2003) *Nutrition in Pediatrics*, 3rd edn. London: BC Decker.

AGING

P Hyland and Y Barnett, Nottingham Trent University, Nottingham, UK

Introduction

The aging processes, and interventions to ameliorate them, have fascinated humans since the dawn of civilization. Research into aging is now a vital area of human endeavor, as our species reaches the limits of its longevity and faces the prospect of an aging population.

This article aims to highlight the processes involved with aging and how they affect the entire hierarchical structure of living organisms, from molecules to cells, tissues, organs, and systems. Accordingly, many theories have evolved to explain the aging processes at each of these levels. A brief overview of these theories will highlight the framework for investigations into the aging processes with the ultimate aim of reducing their deleterious effects, such as age-related disease, perhaps with nutritional and molecular biological intervention strategies.

The term 'aging' can have a wide variety of different meanings in different circumstances. For example, the normal processes from birth, through growth and maturation, an extended period of adulthood, and on to senescence can be thought of as aging.

The term is used here to describe a progressive sequence of detrimental age-related changes that are observed to occur in every individual of a given species, although they may appear at different rates. These changes lead to a breakdown in the normal homeostatic mechanisms, with the result that the functional capacity of the body and its ability to respond to a wide variety of extrinsic and intrinsic agents is often decreased. This causes the degradation of structural elements within the cells, tissues, and organs of the body, leading eventually to the onset of age-related disorders and ultimately death.

Social and Demographic Considerations

An individual's life expectancy is contributed to by the interaction of intrinsic (genetic and epigenetic) factors with extrinsic (environmental and life style) factors (**Figure 1**). In the world's more developed countries (MDCs) the life expectancy at birth in the 1900s was around 47 years. By the end of the twentieth century this rose to a mean of 78 and 76 years in western Europe and north America, respectively, with many individuals living much longer. This dramatic increase in average life expectancy has been largely due to improvements in environmental conditions such as nutrition, housing, sanitation, and medical and social services, and has resulted in a large increase in the number of older people around the world. This change in the age structure of society is compounded by the decreasing fertility levels in the world's populations leading to large gains in worldwide median population ages. Our aging populations have a growing number and proportion of older people and, importantly, a growing number and proportion of very elderly people.

Based on the current rates and trends in population growth it has been predicted that by the year 2025 the elderly population (aged 65 and above) in the world's MDCs will increase by more than 50%, and will more than double worldwide. The elderly population itself is aging with the very elderly (aged 80 and above) being the fastest growing section of the elderly population. This

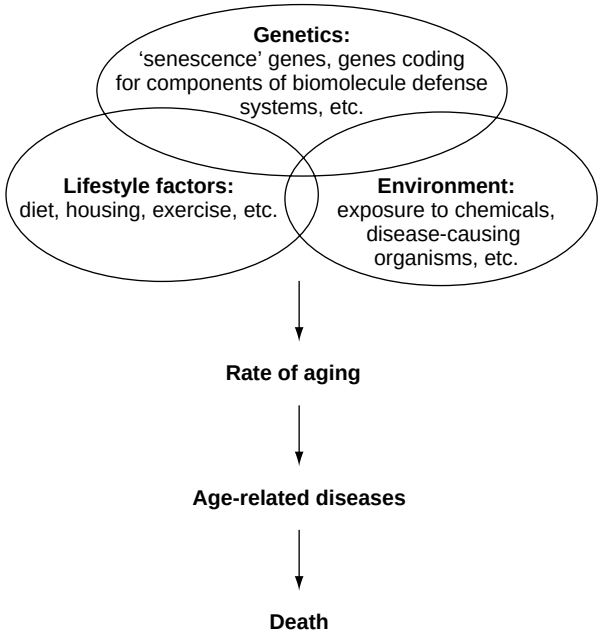


Figure 1 Interactive factors that contribute to the aging process. (Reproduced with permission from Barnett YA (1994) Nutrition and the ageing process. *British Journal of Biomedical Sciences* **51**: 278–287.)

changing demographic picture will result in a large increased prevalence worldwide of long-term illness, disability, and the degenerative diseases associated with aging. These alterations in the proportions of the population of working age and those beyond working age will have a significant impact on the funding and costs of healthcare for all nations, making research into aging of critical international importance.

Theories of Aging

The human body has a hierarchy of structure and function, ranging from cellular biomolecules, through to organelles and cells, and on to tissues, organs, and the body's various systems. The biological manifestations that occur with aging affect the entire hierarchical structure of living systems. Age-related effects are seen in the accumulation of damaged cellular biomolecules (e.g., advanced glycosylation end products, lipid peroxidation products, genetic damage, and mutation), damaged organelles (mitochondria), and loss of cellular function, which contributes to dysfunction of the body's tissues, organs, and systems. These hierarchical changes have paved the way for over 300 theories in an attempt to explain how and why aging occurs. These theories have previously been broadly categorized into: (1) programed or genetic theories; and (2) damage accumulation (stochastic) theories.

However, with ongoing research these categories have not proven to be entirely comprehensive or mutually exclusive and it is more likely that there is a shifting range throughout the life span that reflects a decreasing influence of genetic factors and an increasing influence of stochastic events.

Programed and Genetic Theories

Programed and genetic theories propose that the process of aging follows a biological timetable, perhaps a continuation of the one that regulates childhood growth and development. There are a number of lines of evidence supporting these theories.

Longevity genes It is clear that aging is controlled to some extent by genetic mechanisms. The distinct differences in life span among species are a direct indication of genetic control, at least at the species level. A number of genes have been identified in yeast, nematode worms (*Caenorhabditis elegans*), and fruit flies (*Drosophila melanogaster*) that significantly increase the organism's potential maximum life span. The products of these genes act in a diverse number of ways and are involved in stress response and resistance, development, signal transduction, transcriptional regulation, and metabolic activity.

However, the genetics of longevity have not been as revealing in mammalian studies. In mouse systems genes involved with immune response have been implicated in longevity, as has the 'longevity gene' $p66^{shc}$, which is involved in signal transduction pathways that regulate the cellular response to oxidative stress. In humans, a number of mitochondrial DNA polymorphisms are associated with longevity. Linkage analysis in humans systems has associated certain genes on chromosome 4 with exceptional longevity. Further support for human longevity genes may be provided by the observation that siblings and parents of centenarians live longer. The major histocompatibility complex (MHC), the master genetic control of the immune system, may also be one of the gene systems controlling aging, since a number of genetic defects that cause immunodeficiency shorten the life span of humans. Certain MHC phenotypes have also been associated with malignancy, autoimmune disease, Alzheimer's disease, and xeroderma pigmentosum in humans.

Accelerated aging syndromes No distinct phenocopy exists for normal aging, but there are several genetic diseases/syndromes that display some features of accelerated aging, including Hutchinson-Gilford syndrome (classic early onset Progeria), Werner's syndrome, and Down's syndrome. Patients with these syndromes suffer from many signs of premature aging including hair loss, early greying, and skin atrophy, and also suffer from premature age-related diseases such as atherosclerosis, osteoporosis, and glucose intolerance. The defined genetics involved in these syndromes provide strong evidence for the genetic basis of aging.

Neuroendocrine theories These theories propose that functional decrements in neurons and their associated hormones are pivotal to the aging process. An important version of this theory suggests that the hypothalamic-pituitary-adrenal (HPA) axis is the key regulator of mammalian aging. The neuroendocrine system regulates early development, growth, puberty, the reproductive system, metabolism, and many normal physiological functions. Functional changes to this system could exert effects of aging throughout an organism. However, the cells of the neuroendocrine system are subject to the normal cellular aging processes found in all cells, and the changes occurring in the

neuroendocrine system may be secondary expressions of the aging phenotype.

Immunologic theory and immunosenescence Deterioration of the immune system with aging ('immunosenescence') may contribute to morbidity and mortality due to decreased resistance to infection and, possibly, certain cancers in the aged. T-cell function decreases and autoimmune phenomena increase in elderly individuals.

Although the immune system obviously plays a central role in health status and survival, again the cells of the immune system are subject to the normal cellular aging processes found in all cells. Changes to the immune system may be secondary expressions of the aging phenotype.

Cellular senescence At the cellular level, most, if not all, somatic cell types have a limited replicative capacity *in vitro* before they senesce and die. The number of cell population doublings *in vitro* is inversely correlated with donor age. This is called the 'Hayflick phenomenon' after the scientist credited with its discovery. This limit in the capacity of a cell type or tissue to divide and replenish itself would have major repercussions *in vivo*. There is evidence that replicative senescence is related to *in vivo* aging, but definitive evidence that senescent cells accumulate *in vivo* is lacking to date. Many alterations to normal cellular physiology are exhibited with the senescent phenotype, indicating that senescent cells exist in a growth state that is quite distinct from that of young cells and are subject to a complex alteration to their cellular physiology.

A number of possible explanations for limiting the number of cell population doublings have been proposed, including a tumor suppressive mechanism. One proposal is that the shortening of telomeres, the sequences of noncoding DNA located at the end of chromosomes, is a measure of the number of cell divisions that a cell has experienced. These telomeres may act as specialized regions of the genome, a sacrificial 'sentinel' zone, for the detection of DNA damage being noncoding, more prone to damage, and less prone to repair than the genome as a whole. Damage to telomeres transposes to telomere shortening, and loss of telomere higher order structure may trigger senescence and/or apoptosis.

Studies involving fusion of normal cells (subject to senescence) with immortal cell lines *in vitro* have clearly demonstrated that the senescent phenotype is dominant, and that unlimited division potential results from changes in normal growth control mechanisms. These fusion studies have also revealed the existence of several dominant genes associated with the process of cellular senescence. These genes reside on a number of chromosomes, including 1,4, and X.

Disposable soma theory The disposable soma theory suggests that aging is due to stochastic background damage to the organism, i.e., damage that is not repaired efficiently because the energy resources of the somatic cells are limited. So, instead of wasting large amounts of energy in maintaining the whole body in good condition, it is far more economical to simply repair the heritable stem cell genetic material, in order to ensure the survival of the species. In this way the future of the species is secured at the expense of individual lives. When the somatic energy supply is exhausted, the body ages and dies, but the genetic material survives (in the next generation).

Damage Accumulation (Stochastic) Theories

The 'damage' or 'error' theories emphasize intrinsic and environmental insults to our cellular components that accumulate throughout life and gradually cause alterations in biological function and the physiological decline associated with aging.

Somatic mutation and DNA repair Damage to DNA occurs throughout the lifetime of a cell. If this damage is not repaired or removed then mutations may result. Mutations may result in the synthesis of aberrant proteins with altered or absent biological function; alterations to the transcriptional and translational machinery of a cell; and deregulation of gene control. The accumulation of mutations on their own, or in combination with other age-related changes, may lead to alterations in cellular function and ultimately the onset of age-related disease.

Error catastrophe This theory suggests that damage to mechanisms that synthesize proteins results in faulty proteins, which accumulate to a level that causes catastrophic damage to cells, tissues, and organs. Altered protein structure has been clearly demonstrated to occur with age; however, most of these changes are posttranslational in nature, and hence do not support this theory of aging. Such changes to protein structure may result in progressive loss of 'self-recognition' by the cells of the immune system and thus increase the likelihood that the immune system would identify self-cells as foreign and launch an immune attack. Indeed, the incidence of autoimmune episodes is known to increase with age.

Cross-linking The cross-linking theory states that an accumulation of cross-linked biomolecules caused by a covalent or hydrogen bond damages cellular and tissue function through molecular aggregation and decreased mobility. The modified malfunctional biomolecules accumulate and become increasingly resistant to degradation processes and may represent a physical impairment to the functioning of organs. There is evidence *in vitro* for such cross-linking over time in collagen and in other proteins, and in DNA. Many agents exist within the body that have the potential to act as cross-linking agents, e.g., aldehydes, antibodies, free radicals, quinones, citric acid, and polyvalent metals, to name but a few.

Free radicals The most popular, widely tested and influential of the damage accumulation theories of aging is the 'free radical' theory, first proposed by Harman in 1956. Free radicals from intrinsic and extrinsic sources (**Table 1**) can lead to activation of cytoplasmic and/or nuclear signal transduction pathways, modulation of gene and protein expression, and also alterations to the structure and ultimately the function of biomolecules. Free radicals may thus induce alterations to normal cell, tissue, and organ functions, which may result in a breakdown of homeostatic mechanisms and lead to the onset of age-related disorders and ultimately death. It can be predicted from this theory that the life span of an organism may be increased by slowing down the rate of initiation of random free radical reactions or by decreasing their chain length. Studies have demonstrated that it is possible to increase the life span of cells *in vitro* by culturing them with various antioxidants or free radical scavengers. Antioxidant supplementation with a spin-trapping agent has been demonstrated to increase the lifespan of the senescence accelerated mouse, although as yet there is little evidence for increasing the life span of a normal mammalian species by such strategies.

Table 1 Extrinsic and intrinsic sources of free radicals

Extrinsic sources	*Intrinsic sources*
Radiation: ionizing, ultraviolet	Plasma membrane: lipoxygenase, cycloxygenase, NADPH oxidase
Drug oxidation: paracetamol, carbon tetrachloride, cocaine	Mitochondria: electron transport, ubiquinone, NADH dehydrogenase
Oxidizing gases: oxygen, ozone, nitrogen dioxide	Microsomes: electron transport, cytochrome p450, cytochrome b_5
Xenobiotic elements: arsenic (As), lead (Pb), mercury (Hg), cadmium (Cd)	Peroxisomes: oxidases, flavoproteins
Redox cycling substances: paraquat, diquat, alloxan, doxorubicin	Phagocytic cells: neutrophils, macrophytes, eosinophils, endothelial cells
Heat shock	Auto-oxidation reactions: Metal catalyzed reactions
Cigarette smoke and combustion products	Other: hemoglobin, flavins, xanthine oxidase, monoamine oxidase, galactose oxidase, indolamine dioxygenase, tryptophan dioxygenase
	Ischemia – reperfusion

Mitochondrial DNA damage This hypothesis combines elements of several theories, covering both the stochastic and genetic classes of aging theories. It is proposed that free radical reactive oxygen species generated in the mitochondria contribute significantly to the somatic accumulation of mitochondrial DNA mutations. This leads to a downward spiral wherein mitochondrial DNA damage results in defective mitochondrial respiration that further enhances oxygen free radical production, mitochondrial DNA damage, and mutation. This leads to the loss of vital bioenergetic capacity eventually resulting in aging and cell death.

The absence of evidence that exclusively supports any one theory leaves no doubt that aging is due to many processes, interactive and interdependent, that determine life span and death.

Age-Related Diseases

Regardless of the molecular mechanisms that underlie the aging process, a number of well-characterized changes to the structure and therefore the function of the major cellular biomolecules (lipids, proteins, carbohydrates, and nucleic acids) are known to occur with age (**Table 2**). The age-related alterations to the structure and therefore the function of cellular biomolecules have physiological consequences and may directly cause or lead to an increased susceptibility to the development of a number of diseases (**Figure 2**).

Cellular biomolecules are constantly exposed to a variety of extrinsic and intrinsic agents that have the potential to cause damage. A number of defense systems exist, e.g., antioxidant enzymes and DNA repair systems, which aim to reduce, remove, or repair damaged biomolecules. These defense systems are not perfect, however, and biomolecular damage may still occur. Such damage can result in the degradation of structural elements within the cells, tissues, and organs of the body, leading to a decline in biological function and eventually to disease and death.

Table 2 Major age-related alterations in biomolecule structure and the resultant physiological consequences of such structural changes

Biomolecule	*Alteration*	*Physiological consequence*
Lipids	Lipid peroxidation	Oxidized membranes become rigid, lose selective permeability and integrity. Cell death may occur Peroxidation products can act as cross-linking agents and may play a role in protein aggregation, the generation of DNA damage and mutations, and the age-related pigment lipofuscin
Proteins	Racemization, deamination, oxidatation, and carbamylation	Alterations to long-lived proteins may contribute to aging and/or pathologies. For example, modified crystallins may aggregate in the lens of the eye thus leading to the formation of cataracts Cross-linking and formation of advance glycosylation end-products (AGEs), which can severely affect protein structure and function Effects on the maintenance of cellular homeostasis
Carbohydrates	Fragmentation, depolymerization Glucose auto-oxidation	Alters physical properties of connective tissue. Such alteration may be involved in the etiology and pathogenesis of osteoarthritis and other age-related joint disorders Glycosylation of proteins *in vivo* with subsequent alteration of biological function; for example, glycosylation of insulin in patients with diabetes may result in altered biological function of insulin and so contribute to the pathogenesis of the disease
Nucleic acids	Strand breaks Base adducts Loss of 5-methyl cytosine from DNA	Damage could be expected to interfere with the processes of transcription, translation, and DNA replication. Such interference may reduce a cell's capacity to synthesize vital polypeptides/proteins. In such circumstances cell death may occur. The accumulation of a number of hits in critical cellular genes associated with the control of cell growth and division has been shown to result in the process of carcinogenesis Dedifferentiation of cells (5-methylcytosine plays an important role in switching off genes as part of gene regulation) If viable, such dedifferentiated cells may have altered physiology and may contribute to altered tissue/organ function

The physiological alterations with age proceed at different rates in different individuals. Some of the common changes seen in humans are: the function of the immune system decreases by the age of 30 years of age, reducing defenses against infection or tumor establishment and increasing the likelihood of autoimmune disorders; metabolism starts to slow down at around 25 years of age; kidney and liver function decline; blood vessels lose their elasticity; bone mass peaks at age 30 years and drops about 1% per year thereafter; the senses fade; the epidermis becomes dry and the dermis thins; the quality of and need for sleep diminish; and the brain loses 20% of its weight, slowing recall and mental performance. A number of age-related diseases may develop as a consequence of the tissue, organ, and system deterioration (**Table 3**).

Modification of the Aging Process

Can the adverse consequences of aging be prevented? Down through the ages many have pursued the elixir of life. Attempts to increase the average life expectancy and quality of life in the elderly can only succeed by slowing the aging process itself. In humans, the rate of functional decline associated with aging may be reduced through good nutrition, exercise, timely health care, and avoidance of risk factors for age-related disease.

Nutritional Modification

It is clear that diet contributes in substantial ways to the development of age-related diseases and that modification of the diet can contribute to their prevention and thus help to improve the quality of life in old age. Macronutrient intake levels can play a significant part in the progression of age-related diseases and affect the quality of life. For example, the total and proportional intakes of polyunsaturated fatty acids and saturated fatty acids in the Western diet may have an effect on the incidence of atherosclerosis and cardiovascular diseases.

Our dietary requirements also change as we age and if such changes are not properly addressed this could lead to suboptimal nutritional status. This challenge is compounded by a decrease in the body's ability to monitor food and nutrient intakes. Dietary intake and requirements are complex issues, intertwined with many health and life style issues. However, most research points towards the need for

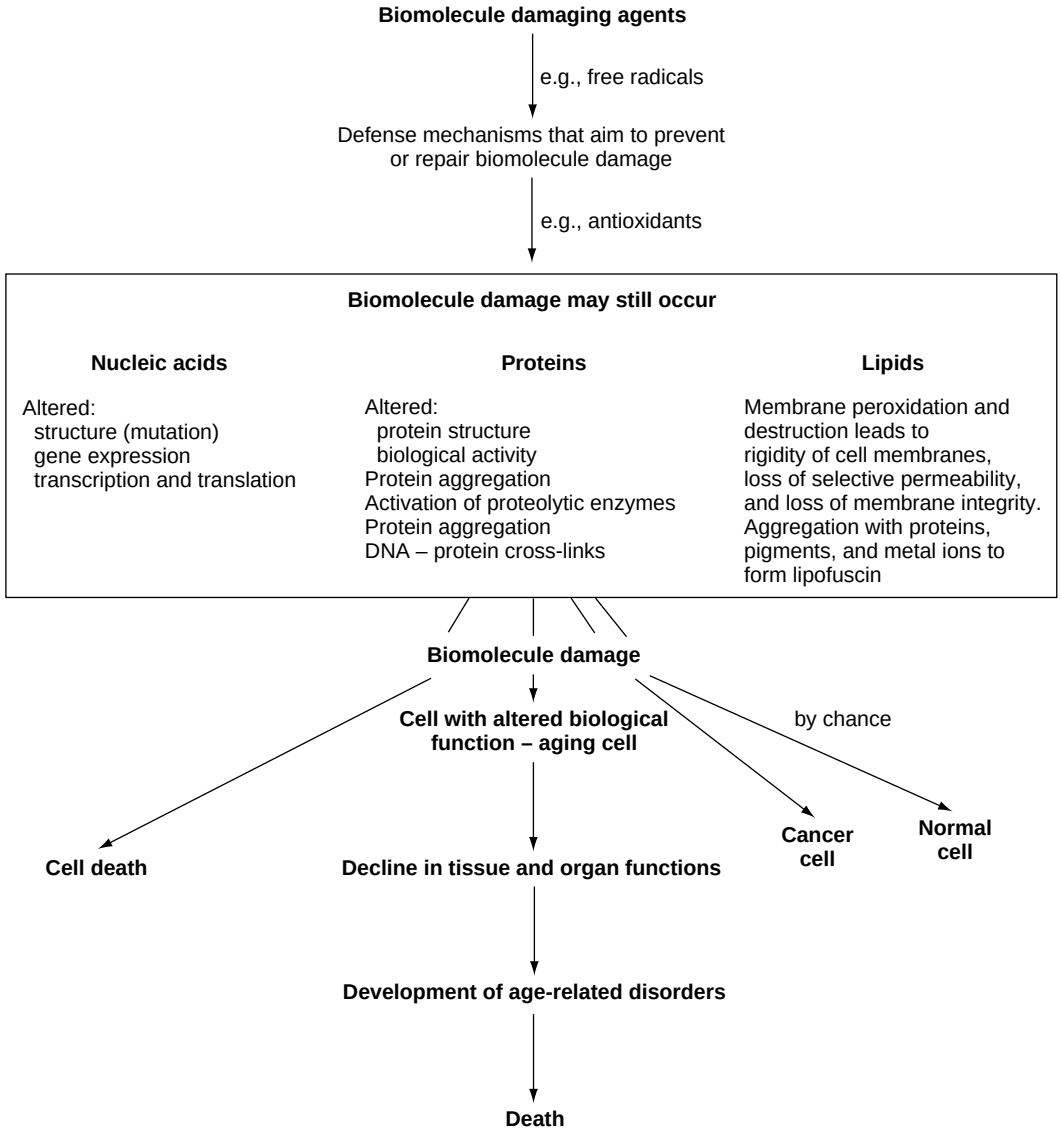


Figure 2 Biomolecule damage and the aging process. (Reproduced with permission from Barnett YA (1994) Nutrition and the aging process. *British Journal of Biomedical Sciences* **51**: 278–287.)

a varied diet as we age, with an increased emphasis on micronutrient intake levels.

An exemplary diet for healthy aging can be found in the traditional diet of Okinawa, Japan. Okinawans are the longest-living population in the world according to the World Health Organization, with low disability rates and the lowest frequencies of coronary heart disease, stroke, and cancer in the world. This has been attributed to healthy life style factors such as regular physical activity, minimal tobacco use, and developed social support networks as antistress mechanisms, all of which are underpinned by a varied diet low in salt and fat (with monosaturates as the principal fat) and high levels of micronutrient and antioxidant consumption.

Vitamins and micronutrients The mechanisms by which certain vitamins and micronutrients mediate their protective effect in relation to a number of age-related disorders is based in large part upon their abilities to prevent the formation of free radicals or scavenging them as they are formed, either directly (e.g., vitamins C, E, and β-carotene) or indirectly (e.g., copper/zinc superoxide dismutase, manganese-dependent superoxide dismutase, selenium-dependent glutathione peroxidase). **Table 4** summarizes the effects that a variety of vitamins and micronutrients can have on age-related disease. Only by exploring more fully the underlying molecular mechanisms of aging and the major classes of antioxidants will it be possible to establish the role

Table 3 Major age-related alterations *in vivo* and the resultant pathological conditions

Body system	*Pathological changes*
Cardiovascular	Atherosclerosis, coronary heart disease, hypertension
Central nervous system	Reduction of cognitive function, development of various dementias (e.g., Alzheimer's disease and Parkinson's disease)
Endocrine	Noninsulin-dependent diabetes, hypercortisolemia
Hemopoietic	Anemia, myelofibrosis
Immune	General decline in immune system function, particularly in T cells
Musculoskeletal	Osteoporosis, osteoarthritis, skeletal muscle atrophy
Renal	Glomerulosclerosis, interstitial fibrosis
Reproductive	Decreased spermatogenesis, hyalinization of semeniferous tubules
Respiratory	Interstitial fibrosis, decreased vital capacity, chronic obstructive pulmonary disease
Sense organs	Cataracts, senile macular degeneration, diabetic retinopathy
All systems	Cancer

of, and develop strategies for using various classes of antioxidants to reduce the effects of aging. Other dietary components may also have a beneficial effect in preventing or delaying the onset of age-related disease. For example, as a deterrent against the onset of osteoporosis, adults should ensure adequate calcium and vitamin D intakes.

Dietary energy restriction The effect of caloric restriction on life span has only been convincingly demonstrated in rodents to date. Feeding mice and rats diets that are severely deficient in energy (about 35% of that of animals fed ad libitum, after the initial period of growth) retards the aging of body tissues, inhibits the development of disease and tumors, and prolongs life span significantly. The exact mechanism of action of dietary energy restriction remains to be elucidated, but may involve modulation of free radical metabolism, or the reduced hormone excretion that occurs in dietary restricted animals may lower whole body metabolism resulting in less 'wear and tear' to body organs and tissues.

Current investigations into the effects of dietary energy restriction (of about 30%) on the life spans of primates, squirrels, and rhesus monkeys continue. Caloric restriction in rhesus monkeys leads to reductions in body temperature and energy expenditure consistent with the rodent studies. These investigations should have direct implications for a dietary energy restriction intervention aimed at slowing down the aging process in humans, should any humans wish to extend their life span at such a cost. Once the mechanisms of effects of caloric restriction on longevity are understood it may be possible to develop drugs that act through these mechanisms directly, mitigating the need for diets that interfere with the quality of life.

Table 4 Effects of vitamins and micronutrients on age-related disorders

Vitamin or micronutrient	*Possible effect on age-related disorder*
Vitamins B_6, E copper, zinc, and selenium	Impairment of immune function in older humans if inadequate amounts
Vitamins C, E, and carotenoids	Increased amount in the diet is associated with delayed development of various forms of cataract
	Protective effect against the development of lung cancer in smokers
Carotenoids and zinc	Dietary supplementation associated with a decreased risk of age-related macular degeneration
Selenium	Absolute or relative deficiency associated with development of a number of cancers (not breast cancer)
Vitamin C, β-carotene, α-tocopherol, and zinc	Dietary supplementation may decrease the rate of development of atherosclerosis
Selenium, copper, zinc, lithium, vanadium, chromium, and magnesium	Dietary deficits are associated with an increased risk of cardiovascular disease
Vitamins B_{12}, B_6, and folate	Adequate levels throughout a lifetime may prevent some of the age-related decrease in cognitive function
Chromium	Deficiency is associated with an increased risk of the development of type 2 diabetes mellitus

Molecular Biological Interventions and the Aging Process

Accelerated aging syndromes show degenerative characteristics similar to those appearing during normal aging. The mutations leading to these disorders are being identified and their roles in the aging process are being elucidated. Examining differences in the genetic material from normal elderly people and those with progeria should help to give a better understanding of the genetic mechanisms of aging. Identification of a control gene or genes that inhibit

the action of the genes producing the progeroid phenotype might make it possible to slow down aberrant protein production in normal people as well.

As an example, the genetic defect that predisposes individuals to the development of Werner's syndrome has now been elucidated. Individuals with this disease carry two copies of a mutant gene that codes for a helicase enzyme (helicases split apart or unwind the two strands of the DNA double helix). DNA helicases play a role in DNA replication and repair.

In light of the biological function of these enzymes it has been proposed that the reason for the premature aging in Werner's syndrome is that the defective helicase prevents DNA repair enzymes from removing background DNA damage, which thus becomes fixed as mutations, with consequent deleterious effects on cellular function. It remains to be determined whether increasing the fidelity or activity of helicases in cells will extend their life span.

Since it appears that the loss of telomeric DNA sequences can lead to replicative senescence in dividing cells, in theory by preventing such telomere loss the life span of the cell could be extended. A naturally occurring enzyme, telomerase, exists to restore telomeric DNA sequences lost by replication. Telomerase is normally only functional in germ cells. Manipulating certain cell types (e.g., cells of the immune system) to regulate the expression telomerase may extend their functional life span. Drugs that enhance telomerase activity in somatic cells are currently being developed. However, cellular senescence has been implicated as a tumor suppressor mechanism and it has been found that cancer cells express telomerase. An uncontrolled expression of this enzyme in somatic cells may lead to the onset of malignancy through uncontrolled cell proliferation. Thus, any intervention aiming to increase life span based on the cellular expression of telomerase must strike a balance between maintaining controlled cell division and uncontrolled proliferation.

A number of single gene mutations have been identified that affect metabolic function, hormonal signaling, and gene silencing pathways. In the future it may be possible to develop drugs to mimic the antiaging effects that these genes exert.

See also: **Antioxidants**: Diet and Antioxidant Defense; Observational Studies; Intervention Studies. **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Coronary Heart Disease**: Lipid Theory; Prevention. **Fats and Oils**. **Fatty Acids**: Monounsaturated; Saturated. **Growth and Development, Physiological Aspects**. **Lipids**: Chemistry and Classification; Composition and Role of Phospholipids. **Nucleic Acids**. **Nutrient Requirements, International Perspectives**. **Older People**: Nutritional Requirements; Nutrition-Related Problems; Nutritional Management of Geriatric Patients. **Protein**: Synthesis and Turnover; Requirements and Role in Diet; Digestion and Bioavailability. **Supplementation**: Role of Micronutrient Supplementation.

Further Reading

Barnett YA (1994) Nutrition and the ageing process. *British Journal of Biomedical Sciences* **51**: 278–287.

Bellamy D (ed.) (1995) *Ageing: A Biomedical Perspective*. Chichester: Wiley.

Esser K and Martin GM (1995) *Molecular Aspects of Ageing* Chichester: Wiley.

Finch CE (1991) *Longevity, Senescence and the Genome* Chicago: University of Chicago Press.

Hayflick L (1993) Aspects of cellular ageing. *Reviews in Clinical Gerontology* **3**: 207–222.

Kanungo MS (1994) In *Genes and Ageing*. Cambridge: Cambridge University Press.

Kirkland JL (2002) The biology of senescence: potential for the prevention of disease. *Clinics in Geriatric Medicine* **18**: 383–405.

Kirkwood TBL (1992) Comparative lifespans of species: why do species have the lifespans they do? *American Journal of Clinical Nutrition* **55**: 1191S–1195S.

Mera SL (1992) Senescence and pathology in ageing. *Medical Laboratory Sciences* **4**: 271–282.

(1995) Somatic mutations and ageing: cause or effect? *Mutation Research, DNAging* (special issue) **338**: 1–234.

Tominaga K, Olgun A, Smith JR, and Periera-Smith OM (2002) Genetics of cellular senescence. *Mechanisms of Ageing and Development* **123**: 927–936.

Troen BR (2003) The biology of ageing. *The Mount Sinai Journal of Medicine* **70**(1): 3–22.

US Bureau of the Census (1999) Report WP/98, World Population Profile. Washington, DC: US Government Printing Office.

von Zglinicki T, Bürkle A, and Kirkwood TBL (2001) Stress, DNA damage and ageing – an integrated approach. *Experimental Gerontology* **36**: 1049–1062.

ALCOHOL

Contents

Absorption, Metabolism and Physiological Effects

R Rajendram, R Hunter and V Preedy, King's College London, London, UK
T Peters, King's College Hospital, London, UK

After caffeine, ethanol is the most commonly used recreational drug worldwide. 'Alcohol' is synonymous with 'ethanol,' and 'drinking' often describes the consumption of beverages containing ethanol.

In the United Kingdom, a unit of alcohol (standard alcoholic drink; **Table 1**) contains 8 g of ethanol. The Department of Health (United Kingdom) and several of the medical Royal Colleges have recommended sensible limits for alcohol intake based on units of alcohol. However, because the amount of ethanol in one unit varies throughout the world (**Tables 2** and **3**), the unit system does not allow international comparisons.

Despite these guidelines, the quantity of alcohol consumed varies widely. Many enjoy the pleasant psychopharmacological effects of alcohol. However, some experience adverse reactions due to genetic variation of enzymes that metabolize alcohol. Misuse of alcohol undoubtedly induces pathological changes in most organs of the body. Some questionable data have suggested that alcohol may be beneficial in the reduction of ischaemic heart disease.

Many of the effects of alcohol correlate with the peak concentration of ethanol in the blood during a drinking session. It is therefore important to understand the factors that influence the blood ethanol concentration (BEC) achieved from a dose of ethanol.

Table 1 Unit system of ethanol content of alcoholic beverages[a]

Beverage containing ethanol	*Units of ethanol*
Half pint of low-strength beer (284 ml)	1
Pint of beer (568 ml)	2
500 ml of high-strength beer	6
Pint of cider	2
One glass of wine (125 ml)	1
Bottle of wine (750 ml)	6
One measure of spirits (e.g., whisky, gin, vodka)	1
Bottle of spirits (e.g., vodka; 750 ml)	36

[a]The unit system is a convenient way of quantifying consumption of ethanol and offers a suitable means to give practical guidance. However, there are several problems with the unit system. The ethanol content of various brands of alcoholic beverages varies considerably (for example, alcohol content of beers/ales is 0.5–9.0%—a pint may contain 2–5 units) and the amounts of alcohol consumed in homes bear little in common with standard measures.

Physical Properties of Ethanol

Ethanol is produced from the fermentation of glucose by yeast. Ethanol (**Figure 1**) is highly soluble in water due to its polar hydroxyl (OH) group. The nonpolar (C_2H_5) group enables ethanol to dissolve lipids and thereby disrupt biological membranes. As a relatively uncharged molecule, ethanol crosses cell membranes by passive diffusion.

Absorption and Distribution of Alcohol

The basic principles of alcohol absorption from the gastrointestinal (GI) tract and subsequent distribution are well understood. Beverages containing ethanol pass down the oesophagus into the stomach. The endogenous flora of the GI tract can also transform food into a mixture of alcohols including ethanol. This is particularly important if there are anatomical variations in the upper GI tract (e.g., diverticulae).

Alcohol continues down the GI tract until absorbed. The ethanol concentration therefore

Table 2 Geographical variation in the amount of ethanol in one unit[a]

Country	*Amount of alcohol (g)*
Japan	14
United States	12
Australia and New Zealand	10
United Kingdom	8

[a]The unit system does not permit international comparisons.

Table 3 Guidelines for the consumption of alcohol[a]

	Men (units)		Women (units)	
	Weekly[b]	Daily[c]	Weekly[b]	Daily[c]
Low risk	0–21	3–4	0–14	2–3
Hazardous	22–50	≥4	15–35	≥3
Harmful	>50		>35	≥1–2[d]

[a]Guidelines regarding the consumption of alcohol are designed to reduce harm. The Royal Colleges' (1995) guidelines are for weekly consumption rates, and the Department of Health's (1995) guidelines are for daily consumption.
[b]Recommendations of the Working Group of the Royal Colleges of Physicians, Psychiatrists and General Practitioners (UK).
[c]Recommendations of the Department of Health (UK).
[d]When pregnant or about to become pregnant, consumption of more than 1 or 2 units of alcohol, one or two times per week, is harmful.

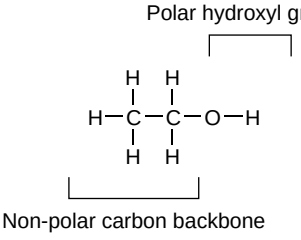


Figure 1 Chemical structure of ethanol.

decreases down the GI tract. There is also a concentration gradient of ethanol from the lumen to the blood. The concentration of ethanol is much higher in the lumen of the upper small intestine than in plasma (**Table 4**). Alcohol diffuses passively across the cell membranes of the mucosal surface into the submucosal space and then the submucosal capillaries.

Absorption occurs across all of the GI mucosa but is fastest in the duodenum and jejunum. The rate of gastric emptying is the main determinant of absorption because most ethanol is absorbed after leaving the stomach through the pylorus.

Alcohol diffuses from the blood into tissues across capillary walls. Ethanol concentration equilibrates between blood and the extracellular fluid within a single pass. However, equilibration between blood water and total tissue water may take several hours, depending on the cross-sectional area of the capillary bed and tissue blood flow.

Ethanol enters most tissues but its solubility in bone and fat is negligible. Therefore, in the postabsorption phase, the volume of distribution of ethanol reflects total body water. Thus, for a given dose, BEC will reflect lean body mass.

Table 4 Approximate ethanol concentrations in the gastrointestinal tract and in the blood after a dose of ethanol[a]

Site	Ethanol concentration	
	g/dl	mmol/l
Stomach	8	1740
Jejunum	4	870
Ileum	0.1–0.2	22–43
Blood (15–120 minutes after dosage)	0.1–0.2	22–43

[a]Ethanol appears in the blood as quickly as 5 minutes after ingestion and is rapidly distributed around the body. A dose of 0.8 g ethanol/kg body weight (56 g ethanol (7 units) consumed by a 70 kg male) should result in a blood ethanol concentration of 100–200 mg/dl (22–43 mmol/l) between 15 and 120 minutes after dosage. Highest concentrations occur after 30–90 minutes.

Metabolism of Alcohol

The rate at which alcohol is eliminated from the blood by oxidization varies from 6 to 10 g/h. This is reflected by the BEC, which falls by 9–20 mg/dl/h after consumption of ethanol. After a dose of 0.6–0.9 g/kg body weight without food, elimination of ethanol is approximately 15 mg/dl blood/h. However, many factors influence this rate and there is considerable individual variation.

Absorbed ethanol is initially oxidized to acetaldehyde (**Figure 2**) by one of three pathways (**Figure 3**):

1. Alcohol dehydrogenase (ADH)—cystosol
2. Microsomal ethanol oxidizing system (MEOS)—endoplasmic reticulum
3. Catalase—peroxisomes

Alcohol Dehydrogenase

ADH couples oxidation of ethanol to reduction of nicotinamide adenine dinucleotide (NAD^+) to NADH. ADH has a wide range of substrates and functions, including dehydrogenation of steroids and oxidation of fatty acids.

Alcohol Dehydrogenase Isoenzymes

ADH is a zinc metalloprotein with five classes of isoenzymes that arise from the association of eight different subunits into dimers (**Table 5**). A genetic model accounts for these five classes of ADH as

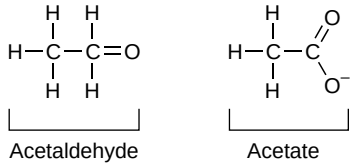


Figure 2 Chemical structures of acetaldehyde and acetate, the products of ethanol metabolism.

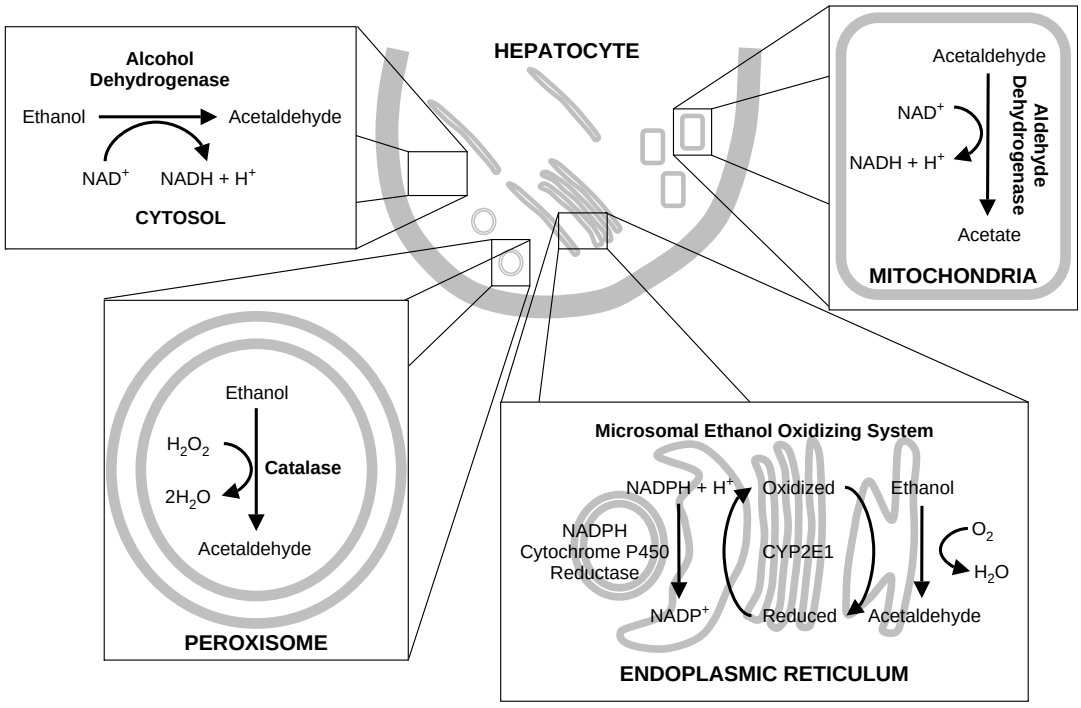


Figure 3 Pathways of ethanol metabolism.

products of five gene loci (ADH1–5). Class 1 isoenzymes generally require a low concentration of ethanol to achieve 'half-maximal activity' (low K_m), whereas class 2 isoenzymes have a relatively high K_m. Class 3 ADH has a low affinity for ethanol and does not participate in the oxidation of ethanol in the liver. Class 4 ADH is found in the human stomach and class 5 has been reported in liver and stomach. Whereas the majority of ethanol metabolism occurs in the liver, gastric ADH is responsible for a small portion of ethanol oxidation.

Table 5 Classes of alcohol dehydrogenase isoenzymes

Class	*Subunit*	*Location*	*K_m (mmol/l)*[a]	*V_{max}*
1				
ADH1	α	Liver	4	54
ADH2	β	Liver, lung	0.05–34	
ADH3	γ	Liver, stomach	0.6–1.0	
2				
ADH4	π	Liver, cornea	34	40
3				
ADH5	χ	Most tissues	1000	
4				
ADH7	σ, μ	Stomach, oesophagus, other mucosae	20	1510
5				
ADH6	—	Liver, stomach	30	

[a]K_m supplied is for ethanol; ADH also oxidizes other substrates.
Adapted with permission from Kwo PY and Crabb DW (2002) Genetics of ethanol metabolism and alcoholic liver disease. In: Sherman DIN, Preedy VR and Watson RR (eds.) *Ethanol and the Liver. Mechanisms and Management*, pp. 95–129. London: Taylor & Francis.

Catalase

Peroxisomal catalase, which requires the presence of hydrogen peroxide (H_2O_2), is of little significance in the metabolism of ethanol. Metabolism of ethanol by ADH inhibits catalase activity because H_2O_2 production is inhibited by the reducing equivalents produced by ADH.

Microsomal Ethanol Oxidizing System

Chronic administration of ethanol with nutritionally adequate diets increases clearance of ethanol from the blood. In 1968, the MEOS was identified. The MEOS has a higher K_m for ethanol (8–10 mmol/l) than ADH (0.2–2.0 mmol/l) so at low BEC, ADH is more important. However, unlike the other pathways, MEOS is highly inducible by chronic alcohol consumption. The key enzyme of the MEOS is cytochrome P4502E1 (CYP2E1). Chronic alcohol use is associated with a 4- to 10-fold increase of CYP2E1 due to increases in mRNA levels and rate of translation.

Acetaldehyde Metabolism

Acetaldehyde is highly toxic but is rapidly converted to acetate. This conversion is catalyzed by aldehyde

Table 6 Classes of aldehyde dehydrogenase isoenzymes

Class	*Structure*	*Location*	K_m (μmol/l)[a]
1			
ALDH1	α4	**Cytosolic**	
		Many tissues: highest in liver	30
2			
ALDH2	α4	**Mitochondrial**	
		Present in all tissues except red blood cells	1
		Liver > kidney > muscle > heart	

[a]K_m supplied is for acetaldehyde; ALDH also oxidizes other substrates.

Adapted with permission from Kwo PY and Crabb DW (2002) Genetics of ethanol metabolism and alcoholic liver disease. In: Sherman DIN, Preedy VR and Watson RR (eds.) *Ethanol and the Liver. Mechanisms and Management*, pp. 95–129. London: Taylor & Francis.

dehydrogenase (ALDH) and is accompanied by reduction of NAD^+ (**Figure 3**). There are several isoenzymes of ALDH (**Table 6**). The most important are ALDH1 (cytosolic) and ALDH2 (mitochondrial). The presence of ALDH in tissues may reduce the toxic effects of acetaldehyde.

In alcoholics, the oxidation of ethanol is increased by induction of MEOS. However, the capacity of mitochondria to oxidize acetaldehyde is reduced. Hepatic acetaldehyde therefore increases with chronic ethanol consumption. A significant increase of acetaldehyde in hepatic venous blood reflects the high tissue level.

Metabolism of Acetate

The final metabolism of acetate derived from ethanol remains unclear. However, some important principles have been elucidated:

1. The majority of absorbed ethanol is metabolized in the liver and released as acetate. Acetate release from the liver increases $2\frac{1}{2}$ times after ethanol consumption.
2. Acetyl-CoA synthetase catalyzes the conversion of acetate to acetyl-CoA via a reaction requiring adenosine triphosphate. The adenosine monophosphate produced is converted to adenosine in a reaction catalyzed by 5′-nucleosidase.
3. Acetyl-CoA may be converted to glycerol, glycogen, and lipid, particularly in the fed state. However, this only accounts for a small fraction of absorbed ethanol.
4. The acetyl-CoA generated from acetate may be used to generate adenosine triphosphate via the Kreb's cycle.
5. Acetate readily crosses the blood–brain barrier and is actively metabolized in the brain. The neurotransmitter acetylcholine is produced from acetyl-CoA in cholinergic neurons.
6. Both cardiac and skeletal muscle are very important in the metabolism of acetate.

Based on these observations, future studies on the effects of ethanol metabolism should focus on skeletal and cardiac muscle, adipose tissue, and the brain.

Blood Ethanol Concentration

The relationship between BEC and the effects of alcohol is complex and varies between individuals and with patterns of drinking. Many of the effects correlate with the peak concentration of ethanol in the blood and organs during a drinking session. Other effects are due to products of metabolism and the total dose of ethanol ingested over a period of time. These two considerations are not entirely separable because the ethanol concentration during a session may determine which pathways of ethanol metabolism predominate.

It is of considerable clinical interest to understand what factors increase the probability of higher maximum ethanol concentrations for any given level of consumption.

Factors Affecting Blood Ethanol Concentration

Gender Differences in Blood Ethanol Concentration

Women achieve higher peak BEC than men given the same dose of ethanol per kilogram of body weight. The volume of distribution of ethanol reflects total body water. Because the bodies of women contain a greater proportion of fat, it is not surprising that the BEC is higher in women. However, gender differences in the gastric metabolism of ethanol may also be relevant.

Period over which the Alcohol Is Consumed

Rapid intake of alcohol increases the concentration of ethanol in the stomach and small intestine. The greater the concentration gradient of alcohol, the faster the absorption of ethanol and therefore peak BEC. If alcohol is consumed and absorbed faster than the rate of oxidation, then BEC increases.

Effects of Food on Blood Ethanol Concentration

The peak BEC is reduced when alcohol is consumed with or after food. Food delays gastric emptying into

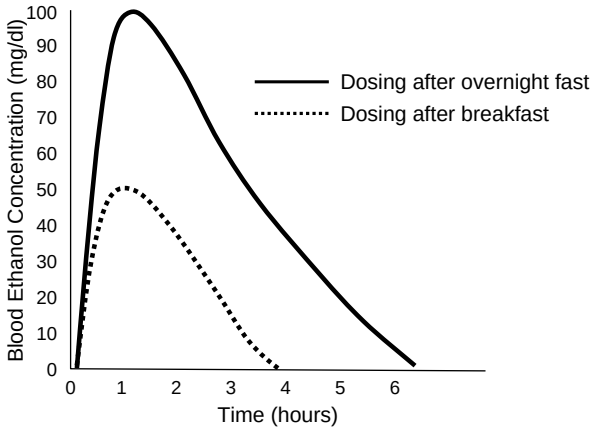


Figure 4 Blood ethanol concentration curve after oral dosing of ethanol. A subject injected 0.8 g/kg ethanol over 30 minutes either after an overnight fast or after breakfast. The peak blood ethanol concentration and the area under the curve are reduced if ethanol is consumed with food.

the duodenum and reduces the sharp early rise in BEC seen when alcohol is taken on an empty stomach. Food also increases elimination of ethanol from the blood. The area under the BEC/time curve (AUC) is reduced (**Figure 4**). The contributions of various nutrients to these effects have been studied, but small, often conflicting, differences have been found. It appears that the caloric value of the meal is more important than the precise balance of nutrients.

In animal studies ethanol is often administered with other nutrients in liquid diets. The AUC is less when alcohol is given in a liquid diet than with the same dose of ethanol in water. The different blood ethanol profile in these models may affect the expression of pathology.

However, food increases splanchnic blood flow, which maintains the ethanol diffusion gradient in the small intestine. Food-induced impairment of gastric emptying may be partially offset by faster absorption of ethanol in the duodenum.

Beverage Alcohol Content and Blood Ethanol Concentration

The ethanol concentration of the beverage consumed (**Table 7**) affects ethanol absorption and can affect BEC. Absorption is fastest when the concentration is 10–30%. Below 10%, the low ethanol concentration in the GI tract reduces diffusion and the greater volume of liquid slows gastric emptying. However, concentrations above 30% irritate the GI mucosa and the pyloric sphincter, increasing secretion of mucous and delaying gastric emptying.

Table 7 Alcohol content of selected beverages

Beverage	*Alcohol content*		
	g/dl (%)	*mmol/l*	*mol/l*
Low-strength beers	3–4	650–870	0.65–0.87
High-strength beers	8–9	1740–1960	1.74–1.96
Wine	7–14	1520–3040	1.52–3.04
Brandy	35–45	7610–9780	7.61–9.78
Vodka	35–50	7610–10870	7.61–10.87
Gin	35–50	7610–10870	7.61–10.87
Whisky	35–75	7610–16300	7.61–16.30

First-Pass Metabolism of Ethanol

The AUC is significantly lower after oral dosing of ethanol than after intravenous or intraperitoneal administration. The total dose of intravenously administered ethanol is available to the systemic circulation. The difference between AUC_{oral} and AUC_{iv} represents the fraction of the oral dose that was either not absorbed or metabolized before entering the systemic circulation (first-pass metabolism (FPM)). The ratio of AUC_{oral} to AUC_{iv} reflects the oral bioavailability of ethanol.

The investigation of ethanol metabolism has primarily focused on the liver and its relationship to liver pathology. However, gastric metabolism accounts for approximately 5% of ethanol oxidation and 2–10% is excreted in the breath, sweat, or urine. The rest is metabolized by the liver.

After absorption, ethanol is transported to the liver in the portal vein. Some is metabolized by the liver before reaching the systemic circulation. However, hepatic ADH is saturated at a BEC that may be achieved in an average-size adult after consumption of one or two units. If ADH is saturated by ethanol from the systemic blood via hepatic artery, ethanol in the portal blood must compete for binding to ADH. Although hepatic oxidation of ethanol cannot increase once ADH is saturated, gastric ADH can significantly metabolize ethanol at the high concentrations in the stomach after initial ingestion. If gastric emptying of ethanol is delayed, prolonged contact with gastric ADH increases FPM. Conversely, fasting, which greatly increases the speed of gastric emptying, virtually eliminates gastric FPM.

Physiological Effects of Alcohol

Ethanol or the products of its metabolism affect nearly all cellular structures and functions.

Effects of Alcohol on the Central Nervous System

Ethanol generally decreases the activity of the central nervous system. In relation to alcohol, the most

important neurotransmitters in the brain are glutamate, gamma-aminobutyric acid (GABA), dopamine, and serotonin.

Glutamate is the major excitatory neurotransmitter in the brain. Ethanol inhibits the *N*-methyl-D-aspartate (NMDA) subset of glutamate receptors. Ethanol thereby reduces the excitatory effects of glutamate. GABA is the major inhibitory neurotransmitter in the brain. Alcohol facilitates the action of the GABA-a receptor, increasing inhibition. Changes to these receptors seem to be important in the development of tolerance of and dependence on alcohol.

Dopamine is involved in the rewarding aspects of alcohol consumption. 'Enjoyable' activities such as eating or use of other recreational drugs also release dopamine in the nucleus accumbens of the brain. Serotonin is also involved in the in reward processes and may be important in encouraging alcohol use.

The most obvious effects of ethanol intoxication on the central nervous system begin with behavior modification (e.g., cheerfulness, impaired judgment, and loss of inhibitions). These 'excitatory' effects result from the disinhibition described previously (inhibition of cells in the brain that are usually inhibitory). As a result of these effects, it is well recognized that driving under the influence of ethanol is unsafe. However, the definition of what is safe or acceptable varies between countries (**Table 8**) and often changes.

The effects of ethanol are dose dependent (**Table 9**) and further intake causes agitation, slurred speech, memory loss, double vision, and loss of coordination. This may progress to depression of consciousness and loss of airway protective reflexes, with danger of aspiration, suffocation, and death.

Table 8 Legal limits of blood ethanol concentrations for driving[a]

Legal limit[b]	*Blood ethanol concentration*	
	mg/dl	*mmol/l*
Norway and Sweden	20	4.3
France, Germany, Italy, and Australia	50	11
United Kingdom, United States, and Canada	80	17
Russia	"Drunkenness"	

[a]Ethanol impairs judgment and coordination. It is well recognized that driving under the influence of ethanol is unsafe. However, the definition of what is safe or acceptable varies between countries and can change as a result of social, political, or scientific influences.
[b]Legislation regarding legal limits of blood ethanol for driving may change.

Table 9 Relationship between amount of ethanol consumed, blood ethanol concentration (BEC), and effect of ethanol on the central nervous system

Alcohol consumed (units)	*Possible BEC*	*Effect*
1–5	10–50 mg/dl 2–11 mmol/l	No obvious change in behavior
2–7	30–100 mg/dl 7–22 mmol/l	Increased self-confidence; loss of inhibitions Impaired judgment, attention, and control
	Euphoria	Mild sensorimotor impairment, delayed reaction times
	Sociability	Legal limits for driving generally fall within this range (see **Table 8**)
8–15	90–250 mg/dl 20–54 mmol/l	Loss of critical judgment Impairment of perception, memory, and comprehension Reduced visual acuity Reduced coordination, impaired balance Drowsiness
11–20	180–300 mg/dl 39–65 mmol/l	Disorientation Exaggerated emotional states Disturbances of vision and perception of color, form, motion, and depth
	Confusion	Increased pain threshold Further reduction of coordination, staggering gait, slurred speech
15–25	250–400 mg/dl 54–87 mmol/l	Loss of motor functions Markedly reduced response to stimuli Marked loss of coordination, inability to stand/walk
	Stupor	Incontinence Impaired consciousness
22–30	350–500 mg/dl 76–108 mmol/l	Unconsciousness Reduced or abolished reflexes Incontinence
	Coma	Cardiovascular and respiratory depression (death possible)
38	>600 mg/dl >130 mmol/l **Death**	Respiratory arrest

[a]Approximate amounts of alcohol required by a 70 kg male to produce the corresponding blood ethanol concentration and intoxicating effects of ethanol. One unit of alcohol contains 8g of ethanol.
Adapted with permission from Morgan MY and Ritson B (2003) *Alcohol and Health: A Handbook for Students and Medical Practitioners*, 4th edn. London: Medical Council on Alcohol.

This sequence of events is particularly relevant in the hospital setting, where patients may present intoxicated with a reduced level of consciousness. It is difficult to determine whether there is coexisting pathology such as an extradural hematoma or overdose of other

drugs in addition to ethanol. Although measurement of BEC is helpful (**Table 9**), it is safest to assume that alcohol is not responsible for any disturbance in consciousness and to search for another cause.

Neuroendocrine Effects of Alcohol

Alcohol activates the sympathetic nervous system, increasing circulating catecholamines from the adrenal medulla. Hypothalamic–pituitary stimulation results in increased circulating cortisol from the adrenal cortex and can, rarely, cause a pseudo-Cushing's syndrome with typical moon-shaped face, truncal obesity, and muscle weakness. Alcoholics with pseudo-Cushing's show many of the biochemical features of Cushing's syndrome, including failure to suppress cortisol with a 48-h low-dose dexamethasone suppression test. However, they may be distinguished by an insulin stress test. In pseudo-Cushing's, the cortisol rises in response to insulin-induced hypoglycemia, but in true Cushing's there is no response to hypoglycemia.

Ethanol affects hypothalamic osmoreceptors, reducing vasopressin release. This increases salt and water excretion from the kidney, causing polyuria. Significant dehydration may result particularly with consumption of spirits containing high concentrations of ethanol and little water. Loss of hypothalamic neurons (which secrete vasopressin) has also been described in chronic alcoholics, suggesting long-term consequences for fluid balance. Plasma atrial natriuretic peptide, increased by alcohol consumption, may also increase diuresis and resultant dehydration.

Alcoholism also affects the hypothalamic–pituitary–gonadal axis. These effects are further exacerbated by alcoholic liver disease. There are conflicting data regarding the changes observed. Testosterone is either normal or decreased in men, but it may increase in women. Estradiol is increased in men and women, and it increases as hepatic dysfunction deteriorates. Production of sex hormone-binding globulin is also perturbed by alcohol.

The development of female secondary sexual characteristics in men (e.g., gynaecomastia and testicular atrophy) generally only occurs after the development of cirrhosis. In women, the hormonal changes may reduce libido, disrupt menstruation, or even induce premature menopause. Sexual dysfunction is also common in men with reduced libido and impotence. Fertility may also be reduced, with decreased sperm counts and motility.

Effects of Alcohol on Muscle

Myopathy is common, affecting up to two-thirds of all alcoholics. It is characterized by wasting, weakness, and myalgia and improves with abstinence. Histology correlates with symptoms and shows selective atrophy of type II muscle fibers. Ethanol causes a reduction in muscle protein and ribonucleic acid content. The underlying mechanism is unclear, but rates of muscle protein synthesis are reduced, whereas protein degradation is either unaffected or inhibited. Attention has focused on the role of acetaldehyde adducts and free radicals in the pathogenesis of alcoholic myopathy.

Alcohol and Nutrition

The nutritional status of alcoholics is often impaired. Some of the pathophysiological changes seen in alcoholics are direct consequences of malnutrition. However, in the 1960s, Charles Lieber demonstrated that many alcohol-induced pathologies, including alcoholic hepatitis, cirrhosis, and myopathy, are reproducible in animals fed a nutritionally adequate diet. Consequently, the concept that all alcohol-induced pathologies are due to nutritional deficiencies is outdated and incorrect.

Myopathy is a direct consequence of alcohol or acetaldehyde on muscle and is not necessarily associated with malnutrition. Assessment of nutritional status in chronic alcoholics using anthropometric measures (e.g., limb circumference and muscle mass) may be misleading in the presence of myopathy.

Acute or chronic ethanol administration impairs the absorption of several nutrients, including glucose, amino acids, biotin, folate, and ascorbic acid. There is no strong evidence that alcohol impairs absorption of magnesium, riboflavin, or pyridoxine, so these deficiencies are due to poor intakes. Hepatogastrointestinal damage (e.g., villous injury, bacterial overgrowth of the intestine, pancreatic damage, or cholestasis) may impair the absorption of some nutrients such as the fat-soluble vitamins (A, D, E, and K). In contrast, iron stores may be adequate as absorption is increased.

Effects of Alcohol on the Cardiovascular System

Alcohol affects both the heart and the peripheral vasculature. Acutely, alcohol causes peripheral vasodilatation, giving a false sensation of warmth that can be dangerous. Heat loss is rapid in cold weather or when swimming, but reduced awareness leaves people vulnerable to hypothermia. The main adverse effect of acute alcohol on the cardiovascular system is the induction of arrhythmias. These are often harmless and experienced as palpitations but can rarely be fatal. Chronic ethanol consumption can cause systemic hypertension and

congestive cardiomyopathy. Alcoholic cardiomyopathy accounts for up to one-third of dilated cardiomyopathies but may improve with abstinence or progress to death.

The beneficial, cardioprotective effects of alcohol consumption have been broadcast widely. This observation is based on population studies of mortality due to ischemic heart disease, case–control studies, and animal experiments. However, there is no evidence from randomised controlled trials. The apparent protective effect of alcohol may therefore result from a confounding factor. Furthermore, on the population level, the burden of alcohol-induced morbidity and mortality far outweighs any possible cardiovascular benefit.

Effects of Alcohol on Liver Function

Central to the effects of ethanol is the liver, in which 60–90% of ethanol metabolism occurs. Ethanol displaces many of the substrates usually metabolized in the liver. Metabolism of ethanol by ADH in the liver generates reducing equivalents. ALDH also generates NADH with conversion of acetaldehyde to acetate. The NADH/NAD+ ratio is increased, with a corresponding increase in the lactate/pyruvate ratio. If lactic acidosis combines with a β-hydroxybutyrate predominant ketoacidosis, the blood pH can fall to 7.1 and hypoglycemia may occur. Severe ketoacidosis and hypoglycemia can cause permanent brain damage. However, in general the prognosis of alcohol-induced acidosis is good. Lactic acid also reduces the renal capacity for urate excretion. Hyperuricemia is exacerbated by alcohol-induced ketosis and acetate-mediated purine generation. Hyperuricemia explains, at least in part, the clinical observation that alcohol misuse can precipitate gout.

The excess NADH promotes fatty acid synthesis and inhibits lipid oxidation in the mitochondria, resulting in fat accumulation. Fatty changes are usually asymptomatic but can be seen on ultrasound or computed tomography scanning, and they are associated with abnormal liver toxicity tests (e.g., raised activities of serum γ-glutamyl transferase, aspartate aminotransferase, and alanine transaminases).

Progression to alcoholic hepatitis involves invasion of the liver by neutrophils with hepatocyte necrosis. Giant mitochondria are visible and dense cytoplasmic lesions (Mallory bodies) are seen. Alcoholic hepatitis can be asymptomatic but usually presents with abdominal pain, fever, and jaundice, or, depending on the severity of disease, patients may have encephalopathy, ascites, and ankle oedema.

Continued alcohol consumption may lead to cirrhosis. However, not all alcoholics progress to cirrhosis. The reason for this is unclear. It has been suggested that genetic factors and differences in immune response may play a role.

In alcoholic cirrhosis there is fibrocollagenous deposition, with scarring and disruption of surrounding hepatic architecture. There is ongoing necrosis with concurrent regeneration. Alcoholic cirrhosis is classically said to be micronodular, but often a mixed pattern is present. The underlying pathological mechanisms are complex and are the subject of debate. Induction of the MEOS and oxidation of ethanol by catalase result in free radical production. Glutathione (a free radical scavenger) is reduced in alcoholics, impairing the ability to dispose of free radicals. Mitochondrial damage occurs, limiting their capacity to oxidize fatty acids. Peroxisomal oxidation of fatty acids further increases free radical production. These changes eventually result in hepatocyte necrosis, and inflammation and fibrosis ensue. Acetaldehyde also contributes by promoting collagen synthesis and fibrosis.

Alcohol and Facial Flushing

Genetic variations in ADH and ALDH may explain why particular individuals develop some of the pathologies of alcoholism and others do not. For example, up to 50% of Orientals have a genetically determined reduction in ALDH2 activity ('flushing' phenotype). As a result, acetaldehyde accumulates after ethanol administration, with plasma levels up to 20 times higher in people with ALDH2 deficiency. Even small amounts of alcohol produce a rapid facial flush, tachycardia, headache, and nausea. Acetaldehyde partly acts through catecholamines, although other mediators have been implicated, including histamine, bradykinin, prostaglandin, and endogenous opioids.

This is similar to the disulfiram reaction due to the rise of acetaldehyde after inhibition of ALDH. Disulfiram is used therapeutically to encourage abstinence in alcohol rehabilitation programs. The aversive effects of acetaldehyde may reduce the development of alcoholism and the incidence of cirrhosis in 'flushers.' However, some alcoholics with ALDH2 deficiency and, presumably, higher hepatic acetaldehyde levels develop alcoholic liver disease at a lower intake of ethanol than controls.

Effects of Acetaldehyde

Acetaldehyde is highly toxic and can bind cellular constituents (e.g., proteins including CYP2E1, lipids, and nucleic acids) to produce harmful acetaldehyde adducts (**Figure 5**). Adduct formation changes

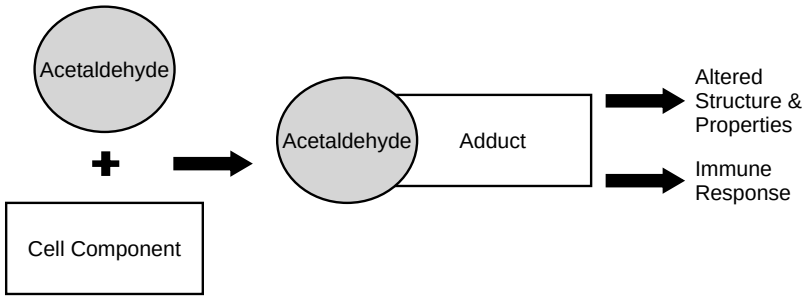


Figure 5 Formation of acetaldehyde adducts.

the structure and the biochemical properties of the affected molecules. The new structures may be recognized as foreign antigens by the immune system and initiate a damaging response.

Adduct formation leads to retention of protein within hepatocytes, contributing to the hepatomegaly, and several toxic manifestations, including impairment of antioxidant mechanisms (e.g., decreased glutathione (GSH)). Acetaldehyde thereby promotes free radical-mediated toxicity and lipid peroxidation. Binding of acetaldehyde with cysteine (one of the three amino acids that comprise GSH) and/or GSH also reduces liver GSH content. Chronic ethanol administration significantly increases rates of GSH turnover in rats. Acute ethanol administration inhibits GSH synthesis and increases losses from the liver. Furthermore, mitochondrial GSH is selectively depleted and this may contribute to the marked disruption of mitochondria in alcoholic cirrhosis.

Effects of Acetate

The role of acetate in alcohol-induced pathology is not well understood. The uptake and utilization of acetate by tissues depend on the activity of acetyl-CoA synthetase. Acetyl-CoA and adenosine are produced from the metabolism of acetate. Acetate crosses the blood–brain barrier easily and is actively metabolized in the brain. Many of the central nervous system depressant effects of ethanol may be blocked by adenosine receptor blockers. Thus, acetate and adenosine may be important in the intoxicating effects of ethanol.

Ethanol increases portal blood flow, mainly by increasing GI tract blood flow. This effect is reproduced by acetate. Acetate also increases coronary blood flow, myocardial contractility, and cardiac output. Acetate inhibits lipolysis in adipose tissue and promotes steatosis in the liver. The reduced circulating free fatty acids (a source of energy for many tissues) may have significant metabolic consequences. Thus, many of the effects of alcohol may be due to acetate.

Summary

Ethanol is probably the most commonly used recreational drug worldwide. Taken orally, alcohol is absorbed from the GI tract by diffusion and is rapidly distributed throughout the body in the blood before entering tissues by diffusion. Ethanol is metabolized to acetaldehyde mainly in the stomach and liver. Acetaldehyde is highly toxic and binds cellular constituents, generating harmful acetaldehyde adducts. Acetaldehyde is further oxidized to acetate, but the fate of acetate and its role in the effects of ethanol are much less clear. Ethanol and the products of its metabolism affect nearly every cellular structure or function and are a significant cause of morbidity and mortality.

See also: **Alcohol**: Disease Risk and Beneficial Effects; Effects of Consumption on Diet and Nutritional Status. **Liver Disorders**.

Further Reading

Department of Health (1995) *Sensible Drinking: The Report of an Inter-Departmental Working Group*. London: Department of Health.

Gluud C (2002) Endocrine system. In: Sherman DIN, Preedy VR, and Watson RR (eds.) *Ethanol and the Liver. Mechanisms and Management*, pp. 472–494. London: Taylor & Francis.

Haber PS (2000) Metabolism of alcohol by the human stomach. *Alcoholism: Clinical & Experimental Research* **24**: 407–408.

Henderson L, Gregory J, Irving K and Swan G (2003) The National Diet and Nutrition Survey: adults aged 19–64 years. Volume 2: Energy, protein, carbohydrate, fat and alcohol intake. London: TSO.

Israel Y, Orrego H, and Carmichael FJ (1994) Acetate-mediated effects of ethanol. *Alcoholism: Clinical & Experimental Research* **18**(1): 144–148.

Jones AW (2000) Aspects of in-vivo pharmacokinetics of ethanol. *Alcoholism: Clinical & Experimental Research* **24**: 400–402.

Kwo PY and Crabb DW (2002) Genetics of ethanol metabolism and alcoholic liver disease. In: Sherman DIN, Preedy VR, and Watson RR (eds.) *Ethanol and the Liver. Mechanisms and Management*, pp. 95–129. London: Taylor & Francis.

Lader D and Meltzer H (2002) *Drinking: Adults' Behaviour and Knowledge in 2002*. London: Office for National Statistics.

Lieber CS (1996) The metabolism of alcohol and its implications for the pathogenesis of disease. In: Preedy VR and Watson RR (eds.) *Alcohol and the Gastrointestinal Tract*, pp. 19–39. New York: CRC Press.
Lieber CS (2000) Alcohol: Its metabolism and interaction with nutrients. *Annual Review of Nutrition* **20**: 395–430.
Mezey E (1985) Effect of ethanol on intestinal morphology, metabolism and function. In: Seitz HK and Kommerell B (eds.) *Alcohol Related Diseases in Gastroenterology*, pp. 342–360. Berlin: Springer-Verlag.
Morgan MY and Ritson B (2003) *Alcohol and Health: A Handbook for Students and Medical Practitioners*, 4th edn. London: Medical Council on Alcohol.
Peters TJ and Preedy VR (1999) Chronic alcohol abuse: Effects on the body. *Medicine* **27**: 11–15.
Preedy VR, Adachi J, Ueno Y *et al.* (2001) Alcoholic skeletal muscle myopathy: Definitions, features, contribution of neuropathy, impact and diagnosis. *European Journal of Neurology* **8**: 677–687.
Preedy VR, Patel VB, Reilly ME *et al.* (1999) Oxidants, antioxidants and alcohol: Implications for skeletal and cardiac muscle. *Frontiers in Bioscience* **4**: 58–66.
Royal Colleges (1995) Alcohol and the heart in perspective. Sensible limits reaffirmed. A Working Group of the Royal Colleges of Physicians, Psychiatrists and General Practitioners. *Journal of the Royal College of Physicians of London* **29**: 266–271.

Disease Risk and Beneficial Effects

M Grønbæk, National Institute of Public Health, Copenhagen, Denmark

Alcohol has for hundreds of years been part of the diet for many people. When enjoyed in small amounts and together with meals, alcohol may have positive effects on health, especially on the prevention of coronary heart disease. In larger amounts, and especially drunk in binges, alcohol is a toxic and dependence-inducing substance, with many short- and long-term detrimental effects. The latter, combined with the high alcohol intake in subsets of the population, implies that alcohol has a major impact on public health in most Western countries. A higher alcohol intake results in higher rates of certain cancer, cirrhosis, suicide, traffic accidents, abuse, and a number of socioeconomic conditions.

Alcohol and Mortality

Amount of Alcohol

Several large prospective population studies from many countries have described the impact of alcohol intake on mortality as J-shaped, indicating both the beneficial effect of a light to moderate alcohol intake and a detrimental effect of a high alcohol intake (**Figure 1**).

Some have explained the J shape as an artefact due to misclassification or confounding. Prevailing beliefs among these researchers is that abstainers comprise a mix of former heavy drinkers, underreporting drinkers, ill people who have stopped drinking, and people with an especially unhealthy lifestyle apart from abstaining. However, most researchers attribute the 'J' to a combination of beneficial and harmful effects of ethanol. This is based on findings from population studies of alcohol-related morbidity and cause-specific mortality that show a decreased relative risk of coronary heart disease, and an increased risk of certain cancers and cirrhosis, with increased alcohol intake. Further evidence derives from studies in which people who were ill at baseline were excluded, and these confirmed the previously mentioned findings.

Benefits—Coronary Heart Disease

A large number of investigators have studied the relation between alcohol intake and coronary heart disease. Studies indicate that the descending leg of the curve is mainly attributable to death from coronary heart disease, as mentioned previously. The lowest risk seems to be among subjects reporting an

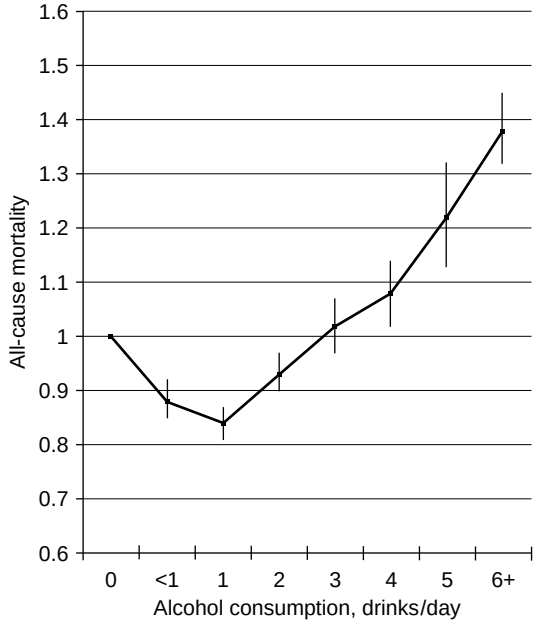


Figure 1 Relative risk of death from all causes according to total alcohol intake. Relative risk is set at 1.00 among nondrinkers (0 drinks/week). (Reproduced with permission from Boffetta P and Garfinkel L (1990) Alcohol drinking and mortality among men enrolled in an American Cancer Society prospective study. *Epidemiology* **1**: 342–348.)

average intake of one to four drinks daily. Several studies have found plausible mechanisms for the apparent cardioprotective effect of a light to moderate intake of alcohol. Subjects with a high alcohol intake have a higher level of high-density lipoprotein, which has been found to be a mediator of the effect of alcohol on coronary heart disease. Thus, 40–60% of the effect of alcohol on coronary heart disease is likely to be attributable to the effect on high-density lipoprotein. Furthermore, drinkers have a lower low-density lipoprotein. Also, alcohol has a beneficial effect on platelet aggregation, and thrombin level in blood is higher among drinkers than among nondrinkers. Ultimately, a few small-scale intervention studies have indicated that alcohol has a beneficial effect on fibronolytic factors.

Risks—Large Number of Somatic Diseases

At the other end of the range of intake, the ascending leg has been explained by the increased risk of cirrhosis and development of certain types of cancers with a high alcohol intake. The mechanisms by which alcohol induces cirrhosis have been intensively studied but sparsely enlightened. It is well documented that women, most likely due to smaller size and different distribution of body fat and water, are at higher risk of developing cirrhosis than men, but other risk factors for alcoholic cirrhosis are not well established (**Figure 2**).

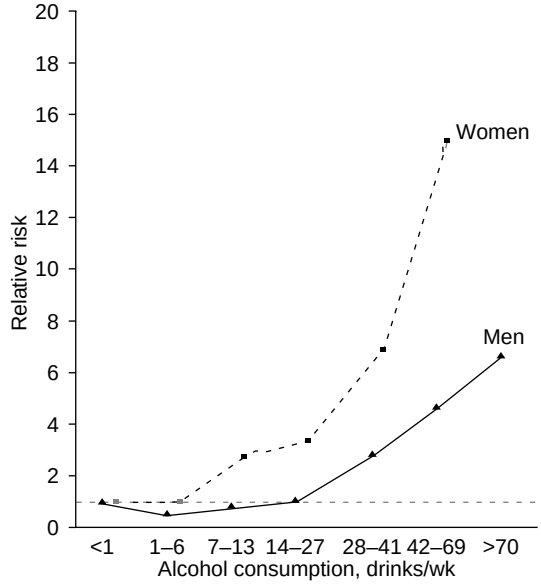


Figure 2 Relative risk of alcohol-induced cirrhosis according to sex and alcohol intake. Relative risk is set at 1.00 among nondrinkers (<1 drink/week). (Reproduced with permission from Becker U *et al.* (1996) Prediction of risk of liver disease in relation to alcohol intake, sex and age: A prospective population study. *Hepatology* **23**: 1025–1029.)

The types of cancer related to a high alcohol intake are those in direct contact with the alcohol; those of the oropharynx and oesophagus and those related to cirrhosis (liver cancer). There is a strong dose-dependent increase in risk of upper digestive tract cancer with increasing alcohol intake. Heavy drinkers of alcohol (5–10 drinks per day) have a 10–15 times higher risk of these relatively rare cancers. Of larger public health relevance are the more frequent cancers—breast and colorectal cancer, which have both been suggested to be related to alcohol. Hence, the risk of breast cancer is doubled for heavy drinking women compared to that for nondrinking women. It is controversial whether a small, frequent daily intake implies an increased risk, although meta-analyses have suggested a 7–9% increased risk per drink per day. Also, the risk of colon cancer is increased among heavy drinkers. The relative risk is twice as high for heavy drinkers compared to nondrinkers, but it is very likely that only colorectal cancer risk is increased, and newer studies have suggested that the risk is mainly increased among beer drinkers. Although not directly related to somatic diseases, other more frequent causes of death among heavy alcohol drinkers, such as traffic accidents, violence, and suicides, substantially add to the ascending leg of the J-shaped curve.

Modifiers of the J Shape

During the past decade, a number of factors that may influence the shape of the curve describing the relation between alcohol and morbidity and mortality have been identified.

Age and Risk Factor Profile

A few studies have indicated that subjects already at high risk of coronary disease experience a greater beneficial effect of drinking alcohol moderately; conversely, only in those with a high risk level is coronary heart disease prevented. Hence, the large Nurses Health Study found that the J-shaped relation was significant only in women older than 50 years of age, whereas younger women who had a light alcohol intake did not differ from abstainers with regard to mortality. Fuchs *et al.* found that women at high risk for coronary heart disease (due to risk factors such as older age, diabetes, family history of coronary heart disease, high cholesterol, and hypertension) who had a light alcohol intake were at a lower risk of death than women who were at the same risk level but did not drink alcohol. In a study by the American Cancer Society, the finding by Fuchs *et al.* was confirmed among men,

and the different mortality risk functions for different age groups were emphasized.

Drinking Pattern

It seems quite obvious that a small, frequent intake (steady) of alcohol has different health implications than a high, irregular (binge) one, and that many of the results from studies measuring only average weekly intake, for example, are imprecise. A few studies have been able to distinguish between frequency and amount of intake, and these studies have supported the previous statement both with regard to all-cause mortality and with regard to the apparent beneficial effect of alcohol on coronary heart disease.

A very large study on all-cause mortality confirmed the J-shaped relation but also clearly showed that those who had an infrequent high alcohol intake had a higher risk of death than those with a similar average intake who had a frequent pattern (**Figure 3**).

One of the mechanisms by which alcohol is assumed to exert its beneficial effect on coronary heart disease is by lowering high-density lipoprotein. Studies in rats have shown that a steady small intake of alcohol implies an increase in high-density lipoprotein level, whereas a peak intake of the same average amount of alcohol does not. Australian and US studies have shown that drinking pattern—steady versus binge drinking—plays a role in the apparent cardioprotective effect of alcohol.

Drinking with Meals

Drinking with meals has been shown to positively affect fribronolysis and lipids. The issue has been sparsely studied in free-living populations, and the results are not consistent. An Italian study showed that drinkers of wine outside meals exhibited higher death rates from all causes, noncardiovascular diseases, and cancer compared to drinkers of wine with meals. However, a larger US study could not confirm these results.

Type of Alcohol

Correlational studies suggest that there may be different effects of the different types of alcoholic beverages. They have shown that mortality from coronary heart disease is lower in countries where wine is the predominant type of alcohol than in countries where beer or spirits are the beverages mainly ingested. These results have been supported from population studies from many countries, suggesting that wine drinkers are at lower risk of death from all causes, including coronary heart disease and cancer, than beer and spirits drinkers (**Figures 4–6**).

One way in which the different types of beverages may exert their different effects on the development of coronary heart disease is via abdominal obesity. It has been suggested that beer drinkers are at a higher risk of developing abdominal obesity than wine drinkers (**Figure 7**). These beverage-specific differences may be explained by either the traits of the drinker or the different substances in the different beverages. Wine consumption in many populations is related to higher socioeconomic status, higher education, and more optimal health behaviour in general compared with beer and spirits consumption. Because these factors are negatively associated

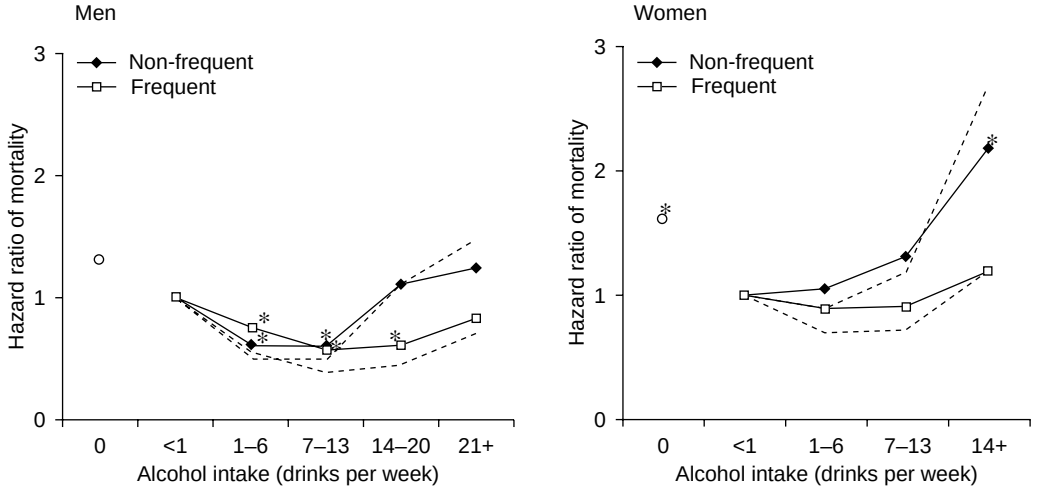


Figure 3 Hazard ratios for all-cause mortality according to quantity and frequency of alcohol intake in men and women (* = $P<0.05$ compared to reference, frequent = at least 2 drinking days per week; nonfrequent = less than 2 drinking days per week). Adjusted for education, smoking, body mass index, physical activity, diet, and diseases before baseline. Reference category is drinkers of less than one but more than zero drinks per week. (Reproduced with permission from Tolstrup J *et al.* (2004) Drinking pattern and mortality in middle-aged men and women. *Addiction* **99**: 323–330.)

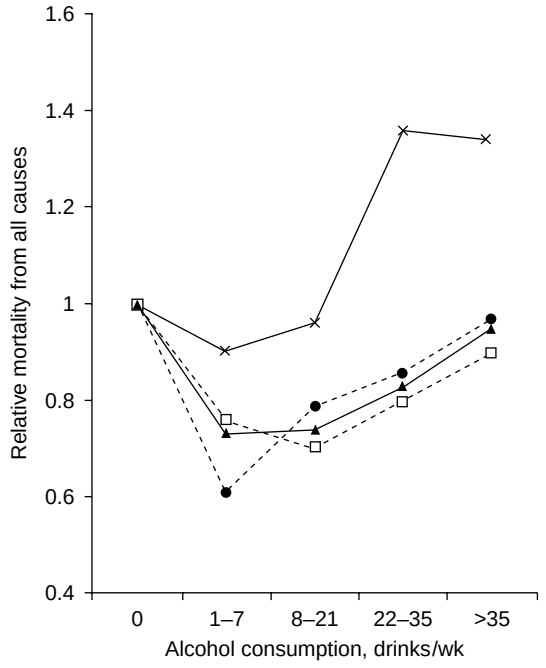


Figure 4 Relative risk of death from all causes according to type of alcohol intake. Data pertain to non-wine drinkers (crosses), wine drinkers (triangles), drinkers for whom wine made up 1–30% of their total alcohol consumption (circles), and drinkers for whom wine made up more than 30% of their total alcohol intake (squares). Relative risk is set at 1.00 among nondrinkers (<1 drink/week). Estimates were adjusted for age, sex, educational level, smoking status, body mass index, and physical activity. (Reproduced with permission from Grønbæk M *et al.* (2000) Type of alcohol consumed and mortality from all causes, coronary heart disease, and cancer. *Annals of Internal Medicine* **133**: 411–419.)

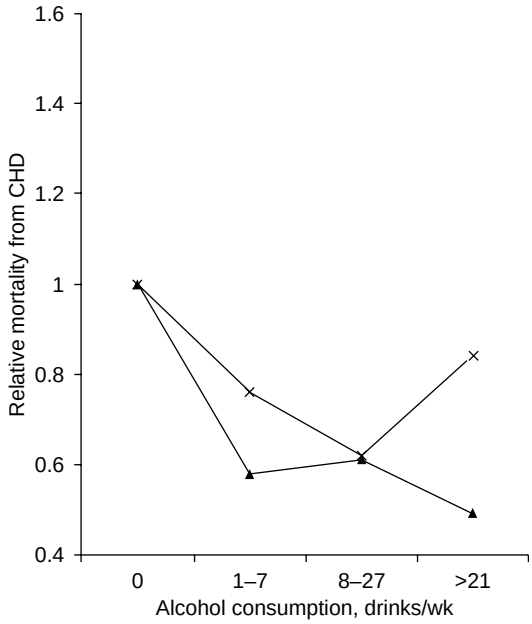


Figure 5 Relative risk of death from coronary heart disease (CHD) according to type of alcohol intake. Data pertain to non-wine drinkers (crosses) and wine drinkers (triangles). Relative risk is set at 1.00 among nondrinkers (<1 drink/week). Estimates were adjusted for age, sex, educational level, smoking status, body mass index, and physical activity. (Reproduced with permission from Grønbæk M *et al.* (2000) Type of alcohol consumed and mortality from all causes, coronary heart disease, and cancer. *Annals of Internal Medicine* **133**: 411–419.)

with mortality, it has been proposed that an unequal distribution according to beverage type may explain the beverage-specific differences in mortality observed in some studies. Several of the components in wine may have antioxidant properties. Hence, flavonoids such as quercetin, rutin, catechin, and epicatechin are present in red wine, responsible for the color of the wine. These compounds have been found to inhibit eicosanoid synthesis and platelet aggregation *in vitro*. Frankel *et al.* found flavonoids to be 10–20 times more potent than vitamin E, and they found an inhibition of low-density lipoprotein oxidation in humans by these phenolic substances. Hertog *et al.* found a preventive effect of dietary flavonoids on risk of developing ischemic heart disease. A high intake of fruits, vegetables, and fish and a low intake of saturated fat have been suggested to reduce the risk of cardiovascular disease. The Mediterranean diet, which includes fruits and vegetables, has been found to have a weak protective effect on cardiovascular disease in 6 of 10 cohort studies. Therefore, diet may play a role in the complex relation between alcoholic beverage type and coronary heart disease mortality. In the Danish Diet Cancer and Health Study, preference of wine was associated with a higher intake of fruit, fish, vegetables, and salad and a higher frequency of use of olive oil for cooking compared with preference of beer or spirits in both men and women. However, sensitivity analysis of the effect of a potential confounder shows that such a confounder, or conglomerate of confounders, should be very strong to explain the previous findings.

Abstainers

Abstainers may have stopped drinking due to ill health. Empiric evidence for the argument is sparse, but it does seem reasonable that some subjects may stop drinking when they are seriously ill. Another reason to quit drinking is alcoholic dependence; some alcoholics can only keep away from drinking by total abstinence. These people will be more ill and thus more likely to die than others. Both situations will confound the relation between alcohol intake and mortality; ill health is the confounder, unequally distributed among intake groups and associated with mortality. A large number of

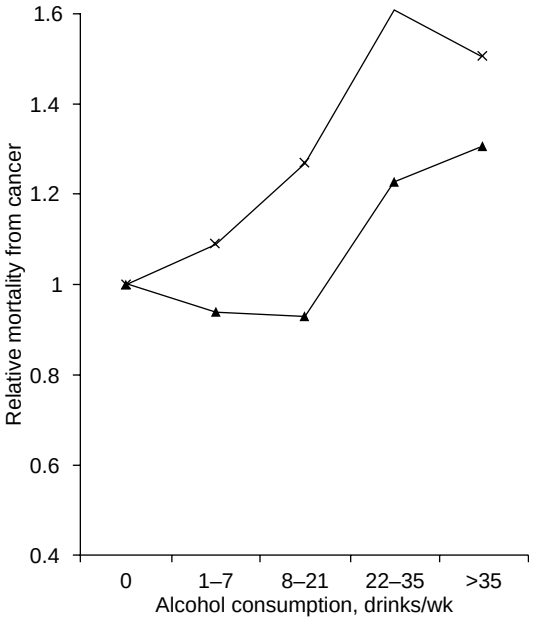


Figure 6 Relative risk of death from cancer according to type of alcohol intake. Data pertain to non-wine drinkers (crosses) and wine drinkers (triangles). Relative risk is set at 1.00 among nondrinkers (<1 drink/week). Estimates were adjusted for age, sex, educational level, smoking status, body mass index, and physical activity. (Reproduced with permission from Grønbæk M *et al.* (2000) Type of alcohol consumed and mortality from all causes, coronary heart disease, and cancer. *Annals of Internal Medicine* **133**: 411–419.)

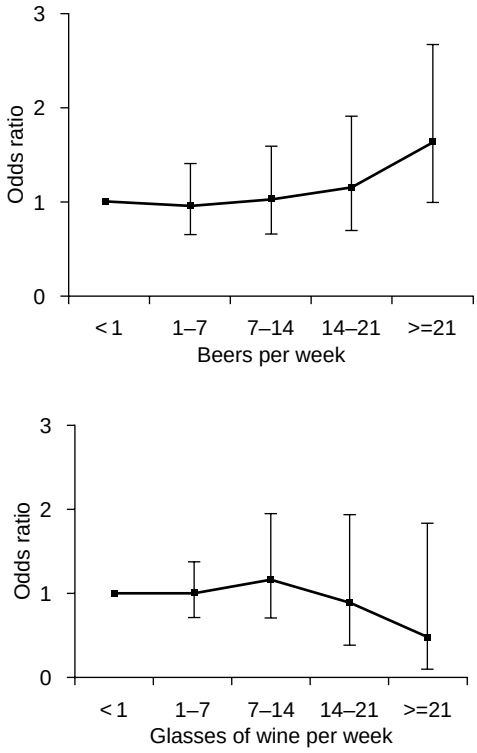


Figure 7 Odds ratio for developing abdominal obesity (waist measure >102 cm) among men. (Reproduced with permission from Vadstrup E *et al.* (2003) Waist circumference in relation to history of amount and type of alcohol: Results from the Copenhagen City Heart Study. *International Journal of Obesity* **27**: 238–246.)

studies, however, have tried to exclude such subjects from analyses, without notable differences; that is, abstainers, or nondrinkers, still seem to be at a higher risk of death than light to moderate drinkers.

Validity of Alcohol Intake

Reporting bias by high-intake or low-intake consumers could, to some extent, explain the apparent lower mortality among light to moderate drinkers. In the type of studies included in this review,– with an emphasis on prospective population studies, one obvious source of bias is misclassification of subjects according to their self-reported alcohol intake. Studies of the validity of self-reported total alcohol intake have mainly concentrated on validating total alcohol intake in suspected alcoholics, whereas intake validity among low-intake consumers in the general population is poorly studied. No reference of alcohol intake (sales reports, collateral information, biological markers, etc.) has been identified. Some biochemical markers of alcohol intake have been suggested, such as γ-glutamyl transferase, high-density lipoprotein, and carbohydrate-deficient transferrin, the latter being one of the most promising. However, in a study from Copenhagen, it was shown that carbohydrate-deficient transferrin was an invalid marker of self-reported alcohol intake in a general population. With regard to information on alcohol intake from the general population, the need for any such marker can further be questioned. First, participants in prospective cohort studies sampled from the general population have less reason to underreport, or deny, their alcohol intake than alcoholics or insurance populations. Second, in a study on alcoholic cirrhosis, it was found that self-reported alcohol intake in the questionnaire used in most of the studies included in the overview was a reliable measure of 'true alcohol intake' since self-reported alcohol intake is a valid predictor of this outcome (**Figure 2**).

Beverage-Specific Reporting Bias

Differential beverage-specific reporting bias by high-intake or low-intake consumers using the frequency questionnaire may, to some extent, explain the apparent lower mortality among wine drinkers than among beer and spirits drinkers. A few validation studies have shown the correlation between total alcohol consumption reported by questionnaire and interview to be 0.8. With regard to type of

beverage, there was an overall agreement between frequency questionnaire and dietary interview. Thus, most subjects in one consumption category of any type of beverage according to the frequency questionnaire also responded to this category in the interview. Mean differences between intake of all three types of beverages were very small or zero, and there were no systematic differences at different levels of average intakes of any of the types of beverages. These are not true validation studies because neither of the two methods can be considered as reference or 'gold standard.' Thus, alcohol intake may have been underreported, and subjects may have reported intake of the three types of beverages differentially. Nevertheless, the close agreement for most individuals suggests that in the range of a small to moderate intake of different types of beverages, the more simple questionnaire approach is not disadvantageous to the expensive and time-consuming personal interview.

Conclusions

The risks and benefits of alcohol on health describe a J-shaped relation. This relation between alcohol intake and all-cause mortality is influenced by several factors, including age, since the J shape seems to persist among the elderly but not among young subjects; sex, since the ascending leg of the curve seems to be steeper for women than for men; drinking pattern, since a small daily intake seems to imply a decreased mortality from cardiovascular disease, whereas binge drinking does not; and type of alcohol, since wine drinkers in some studies seem to be at a lower risk than beer and spirits drinkers.

See also: **Alcohol**: Absorption, Metabolism and Physiological Effects; Effects of Consumption on Diet and Nutritional Status. **Cancer**: Epidemiology and Associations Between Diet and Cancer; Effects on Nutritional Status. **Cholesterol**: Sources, Absorption, Function and Metabolism. **Coronary Heart Disease**: Prevention. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology. **Hypertension**: Dietary Factors. **Obesity**: Definition, Etiology and Assessment; Complications. **Older People**: Nutrition-Related Problems.

Further Reading

Becker U *et al.* (1996) Prediction of risk of liver disease in relation to alcohol intake, sex and age: A prospective population study. *Hepatology* **23**: 1025–1029.

Criqui MH and Rigel BL (1994) Does diet or alcohol explain the French paradox? *Lancet* **344**: 1719–1723.

Fagrell B *et al.* (1999) The effects of light to moderate drinking on cardiovascular diseases. *Journal of International Medicine* **246**: 331–340.

Fuchs CS *et al.* (1995) Alcohol consumption and mortality among women. *New England Journal of Medicine* **332**: 1245–1250.

Grønbæk M *et al.* (1994) Influence of sex, age, body mass index, and smoking on alcohol and mortality. *British Medical Journal* **308**: 302–306.

Grønbæk M *et al.* (1998) Population based cohort study of the association between alcohol intake and cancer of the upper digestive tract. *British Medical Journal* **317**: 844–848.

Grønbæk M *et al.* (2000) Type of alcohol consumed and mortality from all causes, coronary heart disease, and cancer. *Annals of Internal Medicine* **133**: 411–419.

Hendriks HF *et al.* (1994) Effect of moderate dose of alcohol with evening meal on fibrinolytic factors. *British Medical Journal* **308**: 1003–1006.

McElduff P and Dobson AJ (1997) How much alcohol and how often? Population based case–control study of alcohol consumption and risk of a major coronary event. *British Medical Journal* **314**: 1159–1164.

Mukamal KJ *et al.* (2003) Roles of drinking pattern and type of alcohol consumed in coronary heart disease in men. *New England Journal of Medicine* **348**: 109–118.

Thun MJ *et al.* (1997) Alcohol consumption and mortality among middle-aged and elderly U.S. adults. *New England Journal of Medicine* **337**: 1705–1714.

Tolstrup J *et al.* (2004) Drinking pattern and mortality in middle-aged men and women. *Addiction* **99**: 323–330.

Vadstrup E *et al.* (2003) Waist circumference in relation to history of amount and type of alcohol: Results from the Copenhagen City Heart Study. *International Journal of Obesity* **27**: 238–246.

Wannamethee SG and Shaper AG (1999) Type of alcoholic drink and risk of major coronary heart disease events and all-cause mortality. *American Journal of Public Health* **89**: 685–690.

Effects of Consumption on Diet and Nutritional Status

C H Halsted, University of California Davis, Davis, CA, USA

Introduction

Alcohol is a component of the diet that provides 7.1 kcal per gram and on average 5.6% of total dietary energy in the US. When consumed in moderation, alcoholic beverages protect against cardiovascular disease, but when alcohol is consumed in excess it can become an addictive drug with potential for displacement of beneficial components of the diet, damage to several organ systems including the liver, brain, and heart, and increased risk of several cancers. The consumption of excessive amounts of alcohol contributes to generalized

malnutrition, with particular effects on the availability and metabolism of both water- and fat-soluble vitamins including folate, thiamine, pyridoxine, and vitamins A and D. All of the effects of alcoholism on nutritional status are magnified in the presence of alcoholic liver disease. This entry will address the benefits and risks of alcohol consumption and the effects of drinking alcohol on human nutritional status.

Effects of Alcohol Consumption on the Diet

Alcohol is consumed by about two-thirds of adult Americans, and the estimated per capita annual consumption of alcohol exceeds 2 gallons for each US citizen over age 14 years. In the US, young adults between 18 and 25 years of age consume more alcohol than any other age group, and the preferred beverages are wine, beer, and spirits in that order. Men and teenage boys consume about 3 times more alcohol than teenage girls and adult women. Among alcohol consumers, most are moderate drinkers, while about 10% are heavy drinkers at risk of addiction and organ damage. Moderate drinking can be defined as no more than 2 drinks per day for men or 1 drink per day for women, where 1 drink is equivalent to 12–15 g of alcohol. Heavy drinking is defined as consuming more than 5 drinks on any given day per week in men or 4 drinks on any given day per week for women. Chronic alcoholics are addicts who typically consume excessive amounts of alcohol on a daily basis. Binge drinkers are chronic alcoholics who escalate their alcohol intake over weeks or months, typically to the exclusion of the essential components of their regular diets. Alcoholic beverages differ in their alcohol content, such that spirits contain about 40 g/100 ml, wine about 12 g/100 ml, and beer about 4.5 g/100 ml. Thus, the amount of alcohol in 12 oz of beer (16 g) is roughly equivalent to the amount found in 5 oz wine or 1.5 oz spirits.

The effects of alcohol on the diet depend upon the amount consumed each day and changes in overall eating behavior. Although alcohol contains 7.1 kcal per gram, it is rapidly metabolized to acetaldehyde in the liver at rates up to $50\,g\,h^{-1}$, and none is stored as energy equivalents in the body. Furthermore, the metabolism of alcohol influences the metabolism of dietary fat and carbohydrate. There are three metabolic routes for the disposal of alcohol by the body: two in the liver and one in the stomach. Alcohol dehydrogenase (ADH) is present in the cytosol of hepatocytes and metabolizes the relatively low levels of alcohol that would be expected after moderate drinking. The metabolism of alcohol by ADH causes a redox change that promotes lipid synthesis in the liver as well as reduced gluconeogenesis and increased lactate production. Thus, even moderate drinking can cause fatty liver with elevated serum triglyceride levels and, in the absence of dietary carbohydrate, may result in low blood glucose levels that impair concentration and even consciousness. The second liver enzyme, CYP2E1, is part of the cytochrome P450 family, and metabolizes alcohol at levels to be expected after heavy drinking. During metabolism of high levels of alcohol, CYP2E1 utilizes adenosine triphosphate (ATP) energy units and thus 'wastes' stored calories, with resultant potential for weight loss. Another form of this enzyme, gastric CYP2E1, exists in the stomach and, as the first of the three alcohol-metabolizing enzymes to encounter alcohol, accounts for about 30% of all alcohol metabolism in men, but only 10% in women. This gender difference may explain why women's tolerance to alcohol is much less than men's, hence the recognized lower 'safe' level for moderate drinking in women.

The Potential Benefits of Moderate Alcohol Consumption

In 1992, French scientists published a report that indicated that cardiovascular mortality was much less among predominantly wine-drinking residents of the Mediterranean southern provinces of France than in northern provinces where wine is less frequently preferred, in spite of similar overall dietary components and rates of consumption of alcoholic beverages (**Table 1**). This report on the 'French paradox' was assumed to confer specific cardioprotective benefit to wine, but was soon tempered by *in vitro* studies, which showed that the protective effect of wine on the oxidation of low-density lipoprotein could be mimicked by constitutive antioxidant flavonoids present not only in grapes but in many other fruits and vegetables. Another epidemiological study concluded that the lower mortality risk among wine drinkers compared to non-wine drinkers could be attributed in large part to a better life style, including less smoking, more exercise, and better diet. Subsequent population studies defined J-shaped curves for alcohol-related mortality, where mortality is increased in abstainers and progressively increased in those who consume more than one (women) or two (men) drinks per day. It can now be concluded that the benefits of moderate drinking are confined to reductions in incidences of coronary vessel occlusions and ischemic strokes, but not to hemorrhagic

Table 1 Benefits and risks of alcohol consumption

	Minimal amount or duration (drink units per day)	*Mechanism*
Benefits		
Coronary disease protection	1–2 (women),	Flavonoid antioxidants
Cerebrovascular disease	2–4 (men)	Elevated HDL lipoprotein
(nonhemorrhagic) protection		Reduced platelet adhesiveness
Risks		
Cancer		
Oropharynx and esophagus	>2 (women), >4 (men)	Unknown; higher risk in smoking alcoholics
Breast (women)	>2	Increases estrogen production
Colon	>2 (women), >4 (men)	Risk increases with low folate
Alcoholic liver disease		
Fatty liver	>2	Increased liver fat synthesis
Alcoholic hepatitis	>3 (women) × 10 years >6 (men) × 15 years	Toxicity of alcohol metabolism
Alcoholic cirrhosis	>3 (women) × 15 years >6 (men) × 20 years	Increased collagen synthesis
Pancreas		
Pancreatitis	~10 years	Acute inflammation of pancreas
Pancreatic insufficiency	~10–15 years	Loss of exocrine and endocrine pancreatic cells
Cardiomyopathy	Binge drinking	Mitochondrial damage of muscle cells or thiamine deficiency
Neurological		
Acute trauma, e.g., motor vehicle accidents	1–2 in social setting	Legal intoxication
Coma and death	10–20 in rapid succession	Severe toxicity
Withdrawal syndrome	Follows binge	Neuronal hyperexcitability
Wernicke-Korsokoff syndrome	10–15 years	Thiamine deficiency
Anemia	5–10 years	Combinations of iron, folate and pyridoxal deficiencies

strokes. Whereas red and white wine both contain protective antioxidant flavonoids, moderate amounts of alcohol also improve the circulating lipid profile by increasing levels of high-density lipoprotein and tissue plasminogen activator while reducing platelet adhesiveness.

The risks of Excessive Alcohol Consumption

Unlike other abused drugs, chronic alcohol in excess affects many different organ systems, which include the liver, pancreas, heart, and brain (**Table 1**). Excessive chronic alcohol use also increases the risk of certain cancers. While these risks are apparent among the 7% of US citizens over aged 14 who abuse alcohol, their prevalence is generally no less in countries such as France, Italy, and Spain where drinking wine with meals is considered part of the culture. The organ damage from chronic alcoholism may impact on processes of nutrient assimilation and metabolism, as is the case with chronic liver and pancreatic disease, or may be modulated in large part by nutrient deficiencies, as with thiamine and brain function. This section will consider specific effects of alcohol abuse on certain organs as a background for consideration of specific effects on nutritional status.

Alcoholic Liver Disease

Alcoholic liver disease is among the top ten causes of mortality in the US with somewhat higher mortality rates in western European countries where wine is considered a dietary staple, and is a leading cause of death in Russia. Among the three stages of alcoholic liver disease, fatty liver is related to the acute effects of alcohol on hepatic lipid metabolism and is completely reversible. By contrast, alcoholic hepatitis usually occurs after a decade or more of chronic drinking, is associated with inflammation of the liver and necrosis of liver cells, and carries about a 40% mortality risk for each hospitalization. Alcoholic cirrhosis represents irreversible scarring of the liver with loss of liver cells, and may be associated with alcoholic hepatitis. The scarring process greatly alters the circulation of blood through the liver and is associated with increased blood pressure in the portal (visceral) circulation and shunting of blood flow away from the liver and through other organs such as the esophagus. The potentially lethal complications of portal hypertension include rupture of esophageal varices, ascites or accumulation of fluid

in the abdominal cavity, and the syndrome of hepatic encephalopathy, which is due to inadequate hepatic detoxification of substances in the visceral blood that is shunted around the liver. The risk of developing alcoholic cirrhosis is dependent upon the amount of alcohol exposure independent of the presence or absence of malnutrition. For example, a study of well-nourished German male executives found that the incidence of alcoholic cirrhosis was directly related to the daily amount and duration of alcohol consumption, such that daily ingestion of 160 g alcohol, equivalent to that found in a pint of whisky, over a 15-year period predicted a 50% risk of cirrhosis on liver biopsy. Other worldwide demographic data indicate that mortality rates from cirrhosis of the liver can be related to national per capita alcohol intake. These studies have defined the threshold risk for eventual development of alcoholic cirrhosis as 6 drinks per day for men, and about half that for women.

Pancreatitis and Pancreatic Insufficiency

Pancreatitis occurs less frequently than liver disease in chronic alcoholics, and is characterized by severe attacks of abdominal pain due to pancreatic inflammation, while pancreatic insufficiency is due to the eventual destruction of pancreatic cells that secrete digestive enzymes and insulin. This destructive process is associated with progressive scarring of the pancreas together with distortion and partial blockage of the pancreatic ducts, which promote recurrent episodes of acute inflammatory pancreatitis. Since the pancreas is the site of production of proteases and lipases for protein and lipid digestion, destruction of more than 90% of the pancreas results in significant malabsorption of these major dietary constituents, as well as diabetes secondary to reduced insulin secretion. Consequently, patients with pancreatic insufficiency exhibit severe loss of body fat and muscle protein. Since the absorption of fat-soluble vitamins is dependent upon pancreatic lipase for solubilization of dietary fat, these patients are also at risk for deficiencies of vitamins A, D, and E.

Cancers

Chronic alcoholics are at increased risk for cancer of the oro-pharynx and esophagus, colon, and breast. The risk of oro-pharyngeal cancer is greatest when heavy smoking is combined with excessive daily alcohol. Increased risk of squamous cell cancer of the esophagus is also compounded by smoking and may be associated with deficiencies of vitamin A and zinc. Breast cancer in women may be mediated through increased estrogen production during heavy alcohol intake. Colon cancer risk is greatest among alcoholics with marginal folate deficiency.

Heart

Although coronary disease risk is decreased by alcohol consumption, excessive alcohol use also impairs cardiac muscle function. Episodic heavy drinking bouts can lead to arrhythmias in the 'holiday heart' syndrome. Chronic alcoholics are prone to left-sided heart failure secondary to decreased mitochondrial function of cardiac muscle cells, possibly mediated by abnormal fatty acid metabolism. A specific form of high output heart failure, or 'wet beriberi,' occurs in association with thiamine deficiency.

Neurological Effects

The many neurological effects of acute and chronic alcohol abuse can be categorized as those related directly to alcohol, those secondary to chronic liver disease, and those mediated by thiamine deficiency. The stages of acute alcohol toxicity progress upward from legal intoxication with reduced reaction time and judgment, as occurs with blood levels greater than $0.08\,g\,dl^{-1}$ that usually define legal intoxication, to coma and death with levels greater than $0.4\,g\,dl^{-1}$. While mild intoxication is common with social drinking, coma and death have been described among college age males who consume excessive amounts of alcohol in a very short period of time. Automobile accidents, which account for a large portion of alcohol-related deaths, are more common in drunken pedestrians than drivers. Intoxication also leads to frequent falls and head trauma, and subdural hematoma can present with delayed but progressive loss of cognition, headaches, and eventual death. Chronic alcoholics are prone to episodes of alcohol withdrawal, which can be characterized according to stages of tremulousness, seizures, and delirium tremens with hyper-excitability and hallucinations at any time up to 5 days after the last drink. This state of altered consciousness is distinct from hepatic encephalopathy, which results from diversion of toxic nitrogenous substances around the scarred cirrhotic liver and is associated with progressive slowing of cerebral functions with stages of confusion, loss of cognition, and eventual coma and death. Progressive altered cognition and judgment can also result from cerebral atrophy following years of heavy drinking, and may also be mediated by thiamine deficiency as described in greater detail below.

Anemia

Chronic alcoholics who substitute large amounts of alcohol for other dietary constituents are at risk for developing anemia. The causes of anemia in chronic alcoholics are multifactorial, including iron deficiency secondary to bleeding from episodic gastritis or other gastrointestinal sites, folate deficiency from inadequate diet or malabsorption, and deficiency of pyridoxine (vitamin B_6) due to abnormal effects on its metabolism. Consequently, the bone marrow may demonstrate absent iron and mixtures of megaloblastosis from folate deficiency and sideroblastosis from pyridoxine deficiency.

The Effects of Alcohol Consumption on Nutritional Status

Body Weight and Energy Balance

The effects of alcohol on body weight are dependent upon the timing and amount of alcohol consumption in relation to meals and on the presence or absence of organ damage, in particular alcoholic liver disease (**Table 2**). Whereas body weight is usually unaffected by moderate alcohol consumption, chronic alcoholics who drink daily while substituting alcohol for other dietary constituents lose weight due to the energy neutral effect of alcohol in the diet. Moderate drinkers on weight loss regimens are less likely to lose weight while consuming alcohol with their meals since one effect of alcohol is to decrease restraint over food intake. At the same time, those who consume alcohol with high-fat meals are more likely to gain weight due to an acute effect of alcohol on reducing the oxidation of fat at the same time as it promotes its storage.

The presence of alcoholic liver disease results in significant changes in body composition and energy balance. Although fatty liver is fully reversible, progression to alcoholic hepatitis can have profound effects on nutritional status. According to large multicenter studies, alcoholic hepatitis patients demonstrate universal evidence for protein calorie malnutrition, according to the physical findings of muscle wasting and edema, low levels of serum albumin and other visceral proteins, and decreased cell-mediated immunity, whereas their 6-month mortality is related in part to the severity of malnutrition. Anorexia is a major cause of weight loss in alcoholic liver disease, and may be caused by increased circulating levels of leptin. Furthermore, active alcoholic hepatitis contributes to increased resting energy expenditure as another cause of weight loss. On the other hand, resting energy expenditure is normal in stable alcoholic cirrhotics who are also typically underweight or malnourished in part due to preferential metabolism of endogenous fat stores. At the same time, the digestion of dietary fat is decreased in cirrhotic patients due to diminished secretion of bile salts and pancreatic enzymes.

Table 2 Effects of alcohol on body weight

Drinking behavior	*Explanation*
Moderate drinking	
Reduce weight	Substitution of carbohydrate by alcohol; more likely in women
Increase weight	Decreased dietary restraint
Heavy drinking	
Reduce weight	Substitution of nonalcohol calories by alcohol calories, which are 'wasted' during metabolism
Increase weight	Alcohol metabolism decreases lipid metabolism, promotes fat storage

Micronutrient Deficiencies

The chronic exposure to excessive amounts of ethanol is associated with deficiencies of multiple nutrients, in particular thiamine, folate, pyridoxine, vitamin A, vitamin D, and zinc (**Table 3**). The frequency of these deficiencies is increased in the presence of alcoholic liver disease, which results in decreased numbers of hepatocytes for vitamin storage and metabolism. Many of the clinical signs of alcoholic liver disease are related to vitamin deficiencies.

Thiamine

Low circulating levels of thiamine have been described in 80% of patients with alcoholic cirrhosis. Thiamine pyrophosphate is a coenzyme in the intermediary metabolism of carbohydrates, in particular for transketolases, which play a role in cardiac and neurological functions. While alcoholic beverages are essentially devoid of thiamine, acute exposure to alcohol decreases the activity of intestinal transporters required for thiamine absorption. The major neurological signs and symptoms of thiamine deficiency in alcoholics include peripheral neuropathy, partial paresis of ocular muscles, wide-based gait secondary to cerebellar lesions, cognitive defects, and severe memory loss. The presence of peripheral neuropathy is sometimes referred to as 'dry beriberi,' while the other symptoms constitute the Wernicke-Korsokoff syndrome. Whereas abnormal eye movements can be treated acutely by thiamine injections, the other signs are often permanent and contribute to the dementia that often afflicts

Table 3 Micronutrient deficiencies in chronic alcoholic patients

Deficiency	*Cause*	*Effect*
Thiamine	• Poor diet • Intestinal malabsorption	• Peripheral neuropathy • Wernicke-Korsokoff syndrome • High output heart failure
Folate	• Poor diet • Intestinal malabsorption • Decreased liver storage • Increase urine excretion	• Megaloblastic anemia • Hyperhomocysteinemia • Neural tube defect • Altered cognition
Pyridoxine (vitamin B_6)	• Poor diet • Displacement from circulating albumin promotes urine excretion	• Peripheral neuropathy • Sideroblastic anemia
Vitamin A	• Malabsorption • Increased biliary secretion	• Night blindness • May promote development of alcoholic liver disease
Vitamin D	• Malabsorption • Decteased sun exposure	• Calcium deficiency • Metabolic bone disease
Zinc	• Poor diet • Increaded urine excretion	• Night blindness • Decreased taste • Decreased immune funtion
Iron	• Gastrointestinal bleeding	• Anemia

alcoholics after years of drinking. 'Wet beriberi' refers to the high-output cardiac failure that can also occur in thiamine-deficient alcoholics, and is responsive to thiamine therapy in addition to conventional treatment. Since endogenous thiamine is used during carbohydrate metabolism, acute cardiac failure can be precipitated by the administration of intravenous glucose to malnourished and marginally thiamine-deficient patients by depletion of remaining thiamine stores. This process can be prevented by the addition of soluble vitamins including thiamine to malnourished chronic alcoholic patients who are undergoing treatment for medical emergencies.

Folate

Folates are polyglutamylated in their dietary forms and circulate in the methylated and reduced monoglutatate form. Folates function in DNA synthesis and cell turnover, and play a central role in methionine metabolism as substrate for the enzyme methionine synthase in the conversion of homocysteine to methionine. While originally recognized as a cause of megaloblastic anemia, the expanding consequences of folate deficiency are related to elevated circulating homocysteine and include increased risk for neural tube defects and other congenital abnormalities in newborns and altered cognition in the elderly. Prior to folate fortification in the US, the incidence of low serum folate levels in chronic alcoholics was at about 80%. Megaloblastic anemia, due to the negative effects of folate deficiency on DNA synthesis, has been described in about one-third of patients with alcoholic liver disease. Excessive alcohol use is associated with reversible hyperhomocysteinemia in chronic alcoholics because of the inhibitory effect of alcohol or its metabolite acetaldehyde on methionine synthase. Furthermore, folate deficiency may play a role in the pathogenesis of alcoholic liver disease by exacerbating abnormalities in the metabolism of *S*-adenosylmethionine.

The causes of folate deficiency in chronic alcoholism are multiple. With the exception of beer, all alcoholic beverages are devoid of folate, and the typical diet of the chronic alcoholic does not include its fresh vegetable sources. Chronic alcoholism is associated with intestinal folate malabsorption, decreased liver folate uptake, and accelerated folate excretion in the urine. In addition, alcoholic liver disease results in decreased liver stores of folate, so the duration of time for development of folate deficiency with marginal diet is shortened.

Pyridoxine Deficiency

Pyridoxine (vitamin B_6) is required for transamination reactions, including the elimination of homocysteine. Pyridoxine deficiency in chronic alcoholism is caused by poor diet, whereas displacement of pyridoxal phosphate from circulating albumin by the alcohol metabolite acetaldehyde increases its urinary excretion. Low serum levels of pyridoxal phosphate are common in chronic alcoholics, and pyridoxine deficiency is manifest by peripheral neuropathy and sideroblastic anemia. In alcoholic hepatitis, the serum level of alanine transaminase (ALT) is disproportionately low compared to aspartate

transaminase (AST), due to the requirement of pyridoxine for ALT activity.

Vitamin B_{12}

The incidence of vitamin B_{12} deficiency in chronic alcoholism is undefined, since serum levels are often normal or increased due to the presence of B_{12} analogs in alcoholic liver disease. Nevertheless, the intestinal absorption of vitamin B_{12} is decreased in chronic alcoholics due to defective uptake at the ileum. Presumed low levels of vitamin B_{12} in the liver may contribute to abnormal hepatic methionine metabolism with elevated serum homocysteine, since this vitamin is a cofactor for methionine synthase.

Vitamin A

Although serum levels of vitamin A are usually normal in chronic alcoholics, liver retinoids are progressively lowered through the stages of alcoholic liver disease.

Retinoids may play a central role in hepatic function, where vitamin A is stored as retinyl esters in fat-storing transitional Ito cells. The process of transformation of Ito cells to collagen-producing, hepatic stellate cells is associated with depletion of retinyl esters, which may be implicated in the development of alcoholic liver disease. The causes of vitamin A deficiency in alcoholic liver disease include malabsorption, which is due to decreased secretion of bile and pancreatic enzymes necessary for the digestion of dietary retinyl esters and their incorporation into water-soluble micelles prior to intestinal transport. In addition, the transport of retinol is impaired due to decreased hepatic production of retinol-binding protein. Thirdly, the metabolism of alcohol induces microsomal enzymes that promote the production of polar retinol metabolites, which are more easily excreted in the bile. The signs of vitamin A deficiency include night blindness with increased risk of automobile accidents and increased risk of esophageal cancer due to abnormal squamous cell cycling. Conversely, patients with alcoholic liver disease are more susceptible to vitamin A hepatotoxicity so that supplemental doses should be used with caution.

Vitamin D and Calcium

Chronic alcoholic patients are at increased risk for metabolic bone disease due to low vitamin D and hence decreased absorption of calcium. Alcoholic liver disease increases the likelihood of low circulating levels of 25-hydroxy vitamin D because of decreased excretion of bile required for absorption of this fat-soluble vitamin, poor diet, and often decreased sun exposure. Calcium deficiency results from low levels of vitamin D that are required to regulate its absorption, and also because the fat malabsorption that often accompanies alcoholic liver disease results in increased binding of calcium to unabsorbed intestinal fatty acids.

Zinc

Zinc is a cofactor for many enzymatic reactions including retinol dehydrogenase, is stored in the pancreas, and circulates in the blood bound mainly to albumin. Chronic alcoholic patients are frequently zinc deficient because of poor diet, deficiency of pancreatic enzymes, and increased urine excretion due to low zinc-binding albumin in the circulation. The consequences of zinc deficiency include night blindness from decreased production of retinal, decreased taste, and hypogonadism, which may result in lowered testosterone levels and increased risk of osteoporosis in men. Since zinc is required for cellular immunity, its deficiency may contribute to increased infection risk in alcoholic patients.

Iron

Chronic alcoholic patients are often iron deficient because of increased frequency of gastrointestinal bleeding, typically due to alcoholic gastritis or esophageal tears from frequent retching and vomiting, or from rupture of esophageal varices in patients with cirrhosis and portal hypertension. The major consequence of iron deficiency is anemia, which may be compounded by the concurrent effects of folate and pyridoxine deficiencies. Conversely, increased exposure to iron, e.g., from cooking in iron pots, increases the likelihood and severity of alcoholic liver disease, since the presence of iron in the liver promotes oxidative liver damage during the metabolism of alcohol.

See also: **Ascorbic Acid**: Deficiency States. **Calcium**. **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Folic Acid**. **Iron**. **Liver Disorders**. **Thiamin**: Physiology. **Vitamin A**: Biochemistry and Physiological Role. **Vitamin B_6**. **Vitamin E**: Metabolism and Requirements. **Zinc**: Physiology.

Further Reading

Halsted CH (2004) Nutrition and alcoholic liver disease. *Seminars in Liver Diseases* **24**: 289–304.

Halsted CH (1995) Alcohol and folate interactions: clinical implications. In: Bailey LB (ed.) *Folate in Health and Disease*, pp. 313–327. New York: M. Decker, Inc.

Klatsky AL (2002) Alcohol and cardiovascular diseases: a historical overview. *Annals of the New York Academy of Science* **957**: 7–15.

Lieber CS (1992) In *Medical and Nutritional Complications of Alcoholism: Mechanisms and Management*. New York and London: Plenum Medical Book Company.

Lieber CS (2000) ALCOHOL: its metabolism and interaction with nutrients. *Annual Review of Nutrition* **20**: 395–430.

Lieber CS (2004) New concepts of the pathogenesis of alcoholic liver disease lead to novel treatments. *Current Gastroenterology Reports* **6**: 60–65.

McClain CJ, Hill DB, Song Z, Chawla R, Watson WH, Chen T, and Barve S (2002) S-Adenosylmethionine, cytokines, and alcoholic liver disease. *Alcohol* **27**: 185–192.

Mendenhall C, Roselle GA, Gartside P, and Moritz T (1995) Relationship of protein calorie malnutrition to alcoholic liver disease: a reexamination of data from two Veterans Administration Cooperative Studies. *Alcoholism: Clinical and Experimental Research* **19**: 635–641.

Mezey E (1991) Interaction between alcohol and nutrition in the pathogenesis of alcoholic liver disease. *Seminars in Liver Disease* **11**: 340–348.

Nanji A (1993) Role of eicosanoids in experimental alcoholic liver disease. *Alcohol* **10**: 443–446.

Secretary of Health and Human Services (2000) *Tenth Special Report to the U.S.Congress on Alcohol and Health*. US Department of Health and Human Services, National Institute of Alcohol Abuse and Alcoholism.

ALUMINUM

N D Priest, Middlesex University, London, UK

Occurrence in Food and the Environment

Properties and Natural Occurrence

Aluminum was discovered in 1825 by the Danish chemist Oersted. It is a soft, ductile, malleable, silvery metal. Its atomic number is 13, and it has one stable isotope, ^{27}Al. Aluminum belongs to group 3a of the periodic table, along with boron, indium, gallium, and thallium. It most commonly forms trivalent ionic (Al^{3+}) compounds, but it has some covalent characteristics. Aluminum is the most common metal in the earth's crust and is the third most common element. It is too reactive to occur in nature as the free metal.

Aluminum occurs in natural systems as the trivalent ion and in these it has no oxidation-reduction chemistry. In aqueous solution, the chemistry is complicated by the formation of several pH-dependent complex ions. These ions—$Al(OH)^{2+}$, $Al(OH)_2^+$, and $Al(OH)_4^-$—compete with Al^{3+} and $Al(OH)_3$ within aquatic systems. Aluminum is minimally soluble in water at approximately pH 6, when the $Al(OH)_2^+$ ion dominates, but solubility increases at lower and higher pH values. At pH 7 and higher, the most important ion is $Al(OH)_4^-$, whereas at low pH values Al^{3+} dominates.

In contrast to its abundance in the earth's crust, most natural waters contain very little dissolved aluminum (often $<10\,\mu g\,l^{-1}$), reflecting the low solubility of minerals and the deposition of Al^{3+} in sediments as the hydroxide. Seawater contains only $1\,\mu g\,l^{-1}$ of aluminum, and much of this is thought to be bound within the skeletons of diatoms. Where natural waters have either been acidified by acid rain or treated with aluminum sulfate to produce drinking water, the levels of the metal are higher. Concentrations in acidified lakes and rivers (up to $700\,\mu g\,l^{-1}$) commonly exceed levels toxic to fish. In acidic well water, concentrations $>1\,mg\,l^{-1}$ may occur. Aluminum concentrations in tap water should not exceed $200\,\mu g\,l^{-1}$a guideline specified by the World Health Organization (WHO) on esthetic grounds.

Air concentrations of aluminum range from less than $1\,\mu g\,m^{-3}$ in rural environments to as high as $10\,\mu g\,m^{-3}$ in urban, industrialized areas. The higher levels in the latter result from the dust-creating activities of urban man.

Nonfood Uses

Aluminum compounds are widely utilized by industry. They are used in the paper industry, for water purification, in the dye industry, in missile fuels, in paints and pigments, in the textile industry, as a catalyst in oil refining, in the glass industry, and as components of cosmetic and pharmaceutical preparations. Of these, the uses within the cosmetic/pharmaceutical industry are of particular significance since they provide the most likely sources of aluminum uptake by the body.

The following are major cosmetic/pharmaceutical uses of aluminum compounds:

- Aluminum hydroxide as an antacid, particularly for patients suffering from peptic and duodenal ulcers
- Aluminum hydroxide as an effective, nonabsorbed phosfate binder for patients with longstanding kidney failure

- As a component of buffered aspirin
- Aluminum hydroxide and monostearates as components of some vaccines/injection solutions
- Aluminum chloride, aluminum zirconium glycine complex, and aluminum chlorohydrate as the active ingredients of antiperspirants

Many of these applications are under review and their use is discouraged where alternatives of equal efficacy are available and where the potential for high aluminum uptakes exists. For example, both calcium carbonate and lanthanum sulfate are possible alternatives to the long-term use of aluminum hydroxide as a phosfate binder.

Food Uses of Aluminum Compounds

Aluminum compounds that may be employed as food additives are listed in **Table 1**. Although most are present in foods as trace components, others may be present in significant quantities. For example, aluminum-based baking powders, employing sodium aluminum phosfate (SALP), may contain more than $10\,mg\,g^{-1}$ of aluminum, and bread or cake made with these may contain 5–15 mg of the element per slice. American processed cheese may contain as much as 50 mg of aluminum per slice due to the addition of Kasel, an emulsifying agent. Pickled cucumbers may contain 10 mg of aluminum per fruit when alum has been employed as a firming agent. Aluminum anticaking agents may also be present in significant quantities in common table salt.

Table 1 Permitted aluminum-containing food additives and uses

Compound	*Use*
Aluminum	Metallic color for surface treatment
Aluminum ammonium sulfate (ammonium alum)	Acidic compound used as a neutralizing agent and as a buffer
Aluminum potassium sulfate (potassium alum)	Acidic compound used as a neutralizing agent, a buffer, and a firming agent
Aluminum sodium sulfate (soda/sodium alum)	Buffer, neutralizing agent, and firming agent
Aluminum sulfate (alum)	Firming agent in pickling
Aluminum calcium silicate	Anticaking agent for powders
Aluminum sodium silicate	Anticaking agent for powders
Sodium calcium aluminosilicate	Anticaking agent for powders
Kaolin (contains aluminum oxide)	Anticaking agent for powders
Sodium aluminum phosfate (acidic), SALP	Acid, raising (leavening) agent for flour
Sodium aluminum phosfate (basic), Kasel	Emulsifying salt

Natural Aluminum in Food

Even though concentrations of aluminum in soil are high (3–10%), most food plants contain little aluminum. Reports describe diverse levels in different foods and reported values vary for similar foods. Much of this variation results from either the inadequate removal of soil and/or contamination of foods with soil prior to analysis or the use of poor analytical techniques. A selection of results for plant foods is given in **Table 2**. This shows that most uncooked plant foods contain $<5\,\mu g\,g^{-1}$. However, reported concentrations in herbs and spices are higher due to their dehydrated state and their content of aluminum-containing grinding materials.

The concentration of natural aluminum in animal-derived foods is influenced by the low concentration of aluminum in animal feeds, the poor biological uptake of 'food aluminum' by food animals, and the limited ability of body aluminum to transfer to products such as eggs and milk. Consequently, most animal-based foods contain $<1\,\mu g\,g^{-1}$ of aluminum (**Table 3**).

Of the remaining miscellaneous foods in common usage, few contain significant aluminum. Beer stored in aluminum cans for up to 1 year contains $<29\,\mu g\,cm^{-3}$ and the maximum level recorded in beverages within aluminum cans is 1.5 mg per 375-cm^3 can. Tea, which when dry contains large quantities of aluminum ($1.28\,mg\,g^{-1}$), contains relatively little when steeped ($2.8\,\mu g\,cm^{-3}$). For comparison, brewed coffee contains less than $0.4\,\mu g\,cm^{-3}$.

The aluminum concentration of acidic foods may be increased by cooking in aluminum vessels. In one study the increment to the aluminum content of an average-sized serving of rhubarb was 25 mg. Onions boiled in aluminum saucepans similarly accumulate the metal. In contrast, the cooking of nonacidic food products in aluminum utensils has little effect on their aluminum content. Similarly, the use of aluminum foil for wrapping foods adds little to their metal content.

Total Dietary Intake of Aluminum

Regarding measurements of the aluminum content of individual foods, the measurement of total daily diets is complicated by problems of analysis and sample contamination. However, relatively reliable data are available for some countries. In a Finnish study, daily intake of aluminum from food was calculated to be 6.7 mg. Studies performed by the US Food and Drug Administration suggest average daily intake values of between 9 and 14 mg

Table 2 Concentrations of natural aluminum in plant foods

Food	*Concentration ($\mu g\,g^{-1}$)*
Cereals	
Barley	5.0–7.0
Maize (corn)	0.4–3.1
Oats	5.1
Wheat	4.0–16
Rye	4.8
Whole wheat bread	5.4
Cheerios	4.7
Corn Flakes	<2
Macaroni	<2
Rolled oats	<2–5.0
Rice Krispies	<2
Spaghetti	<2
Vegetables	
Asparagus	1.7–9.0
Green beans	8.0
Cabbage, inner leaf	5.7
Carrots	3.8
Cauliflower head	4.0
Leek	15
Onions	5.0–10
Potatoes	<1–20
Spinach	6.9
Tomatoes	0.2–1.1
Fruit	
Apples	0.2–0.9
Bananas	<0.4
Grapes	<0.5
Honeydew melon	0.2
Oranges	<0.4
Plums	<0.3–0.5
Rhubarb	0.8–4.8
Pineapples	<0.3
Herbs and spices	
Allspice	51–101
Basil	167–450
Cinnamon	48–115
Marjoram	>500–1000
Mustard	5–10
Nutmeg	5–11
Paprika	49–700
Black pepper	48–237
Sesame seed	5–<10
Thyme	>500–<1000
Nuts	
Peanuts	<2
Walnuts	<2
Oils	
Olive	0.1–0.4
Sugars	
White sugar	<2
Brown sugar	<2
Cuban raw	5.3
Molasses	110

(**Table 4**). In general, it is likely that daily intakes of aluminum in North America are higher than in Europe due to a higher utilization of SALP and Kasel in the preparation of processed foods.

Table 3 Concentration of natural aluminum in animal foods

Food	*Concentration ($\mu g\,g^{-1}$)*
Milk	0.1–0.7
Cottage cheese	<2
Swiss cheese	19
Beef	<1
Steak	2.3–8.4
Bovine kidney	0.4–1.0
Bovine liver	<2
Lamb	<1
Pork	<1
Bacon	<2
Veal	<1
Chicken	<1
Eggs	0.2–1.4
Turkey	<1
Carp	0.7–1.0
Haddock	<1
Salmon (canned)	8.2
Sole	<1

Table 4 Sources of dietary aluminum intake

Food component	*Al content (mg/standard serving)*
US male and female mean Al intake	9–14/day
Cornbread (homemade)	18
Processed cheese, yellow cake (iced)	10–11
Fish sticks, muffins, hamburger	4–6
Pancakes, spinach	2–3
Lasagne	1

Bioavailability and Metabolism

Given the ubiquitous nature of aluminum in the environment, it is surprising that the human body contains little and that it has no function as an essential trace mineral. To a large extent, this level reflects the integrity of the body's barriers to metal ion intake.

Bioavailability of Ingested Aluminum

Metal ions enter the body by one of three main routes: via the gut wall, by inhalation, and through wounds. For most individuals, the important route of aluminum uptake is through the gut wall. This is true even though the bioavailability of aluminum may be higher by other routes.

Like most polyvalent metal ions, most ingested aluminum passes through the intestinal tract without being absorbed. An estimate of the daily aluminum uptake by the body from all sources, based on an estimate of average daily intake and the level of

excreted aluminum, indicated an absorbed fraction of 0.001.

More precise measurements of absorption have been made using the isotope ^{26}Al. The first reported study utilizing ^{26}Al indicated that in the presence of excess citrate, 1% of the metal was absorbed. This result was considered consistent with the ability of citrate to complex metal ions, holding them in solution at physiological pH values, but unrepresentative because of the large amount of citrate employed. Moreover, the study estimated aluminum uptake from the results of single blood analyses, which may not provide a true measure of uptake. Later studies employing this method have indicated lower uptake values, typically 0.0005. They have also shown that some subpopulations, including those suffering from Alzheimer's disease and some 'normal' individuals, absorb more aluminum than average members of the population and that the coingestion of silicic acid inhibits aluminum absorption by a factor of approximately 3.

Complete balance studies using the ^{26}Al tracer have been undertaken to determine bioavailability. These showed that the fractional uptake of aluminum following administration as a citrate solution was 0.005 and following its intake as hydroxide was 0.0001. The coadministration of citrate with aluminum hydroxide enhanced aluminum uptake by a factor of approximately 10. A later study measured the bioavailability of aluminum in drinking water, and a fractional uptake of 0.002 was determined. It follows that at a maximum concentration of aluminum in drinking water of $200\,\mu g\,l^{-1}$, this source will normally account for approximately 6% of total (nonmedical) aluminum uptake.

In addition to the concentration of citrate in ingested food, other factors have been shown to affect the bioavailability of aluminum: age—metal uptake in milk-fed infants is higher than average and uptake may also be greater in the elderly; gut contents reduce metal bioavailability; silicic acid binds strongly to aluminum, reducing its bioavailability; and local gut conditions affect the ability of the gut wall to sequester aluminum, changing the time available for its uptake. Overall, results suggest that aluminum bioavailability varies between approximately 0.01 and 0.0001.

Biokinetics of Aluminum in Blood

The biochemistry of aluminum is, to a large extent, determined by its valency, ion size, and redox chemistry. The effective ionic radius of Al^{3+} is 54 pm. This is sufficiently similar to that of Fe^{3+} (65 pm) for it to follow some of its metabolic pathways. However, its progress through these is halted at stages where iron is transformed to its divalent, Fe^{2+} state. Aluminum is also sufficiently similar to calcium to codeposit in the skeleton. In addition, aluminum binds particularly strongly to phosfates, giving it the potential to bind with DNA, ATP, and many other biomolecules.

Within the blood, transferrin, the iron-transport protein, binds aluminum most avidly. However, compared to Fe^{3+} ions, the strength of this affinity is low. Consequently, aluminum will not displace iron from transferrin. Aluminum also binds with low-molecular-weight proteins and citrate. Aluminum complexes with low-molecular-weight species may leak from blood vessels into surrounding tissue fluids. The extent of binding to low-molecular-weight molecules is uncertain, but studies indicate that 50% or less of blood aluminum may be bound to them. It has also been suggested that silicic acid in blood may also bind aluminum to form aluminosilicate colloidal particles, which would then deposit within reticuloendothelial organs.

Aluminum is initially rapidly lost from the blood to other body fluids and to excretion, reflecting the weakness/kinetics of the binding of the Al^{3+} ion to proteins. Subsequently, the rate of loss slows. Volunteer studies showed that more than half of the ion had left the blood by 15 minutes postadministration, and that by 1 h an average of 68% had been lost. At 1 day, approximately 2% remained in the blood and by 5 days only 0.4% remained. These variations make the interpretation of isolated blood and serum aluminum levels particularly difficult. At 1 h after uptake, little or no aluminum in the blood is associated with red blood cells. However, at 880 days after intake 14% is associated with these cells. Unlike those in plasma, aluminum deposits in red blood cells are cleared with a long half-life and they may provide a basis for an aluminum assay.

Aluminum Excretion and Body Retention

^{26}Al studies have shown that most aluminum entering blood is excreted in the urine, with only approximately 1% lost in feces. In these studies, intersubject variability was conspicuous, such that 1 and 5 days after intake the range of fractional aluminum excretions was 0.5–0.8 and 0.6–0.9, respectively. In another study employing a single volunteer, the long-term retention of aluminum was determined. At early times this volunteer showed a retention pattern consistent with the mean of that found in the short-term study. However, approximately 4% of injected aluminum was retained for years. This finding indicates that under conditions of

continuous intake, aluminum is accumulated by the body, even in subjects with normal kidney function.

From a retention equation, and assuming daily systemic uptakes of 15 μg of aluminum, terminal body burdens of the metal were predicted. The calculations suggested that 50 years of continuous uptake of aluminum should give rise to body burdens of 2–7 mg. This estimate is lower than others based on the results of chemical analyses (35–60 mg). It is suggested that the most likely reasons for this discrepancy are errors in the extrapolation of body burden from the results of the chemical analysis of small pieces of tissue and errors due to the measurement of samples that have become contaminated with environmental aluminum (i.e., dust).

Aluminum Deposition in Tissues

Most metals are deposited to a much greater extent than average in a few organs: liver, kidneys, and skeleton. However, the proportion of the total body burden deposited in these is variable and depends on many factors, including the chemical properties of the ion and the age, sex, and metabolic status of the individual. The major site of deposition of aluminum is the skeleton. Skeletal deposits of aluminum have been demonstrated in normal bone using chemical analysis and are easily detected in bone from renal failure patients using histochemical staining techniques.

The levels of aluminum deposited in the liver are uncertain. Published measurements of total liver aluminum suggest values of 6–9 mg for normal adults, indicating that a significant fraction of the aluminum body burden is present in this organ. However, external counting of ^{26}Al did not indicate large liver deposits. Moreover, a comparative failure of aluminum to deposit in the normal liver would be consistent with the low levels of fecal excretion seen and with the low concentrations of aluminum found in the livers of some dialyed renal patients. Also, at least in rats, the relative depositions of trivalent metals in the skeleton and liver have been shown to be a function of ion size, with appreciable liver deposition occurring only where the ion size is large. The ion size of aluminum is very small. However, the observation that with time aluminum levels build up in red blood cells may indicate the presence of a delayed pathway of aluminum accumulation by the liver since this organ is involved in the breakdown of hemoglobin.

Within the skeleton, aluminum, in common with most other polyvalent metal ions, initially deposits as a very thin layer on bone surfaces. The mechanisms of deposition have not been investigated, but one report suggests that three may be involved: entrapment of aluminum ions within the hydration shell of existing bone mineral; incorporation of aluminum into new bone mineral at sites of bone apposition; and binding of the metal to acidic organic components of the bone matrix, such as phosphoproteins.

Subsequently, aluminum remains on bone surfaces until it back-exchanges into tissue fluids, the bone surface is removed by osteoclasts, or the bone surface is buried by the apposition of new bone. These processes will result in the gradual loss of bone aluminum and in a transfer of aluminum from bone surfaces to the volume of the bone matrix. Such volume deposits are clearly seen in stained biopsy sections from dialysis patients.

The back-exchange of deposited aluminum, from bone surfaces to blood, may occur relatively quickly and its rate will be an important determinant of the rate of early loss of aluminum from the body. At longer times after deposition, more aluminum will be removed by bone turnover. Given that in adult man the rate of bone turnover is very low (3–20% per year), most firmly deposited metal may be expected to be retained in the body for tens of years. This accounts for the reported body retention of 2% of the injected ^{26}Al at 5 years postinjection and its low rate of loss at this time (equivalent retention half-time greater than 5 years). In children, who exhibit high rates of bone growth and turnover, aluminum is lost more rapidly.

Toxicity of Systemic Aluminum

The toxicity of aluminum has been extensively reviewed both by WHO and by the US Department of Health and Human Services. Exposure to aluminum at environmental levels produces no known adverse effects in man. There is little evidence to suggest that aluminum may produce adverse effects under conditions of chronic, excess, occupational exposure. Under conditions of high medical exposure, resulting in large aluminum body burdens, the metal is toxic. Aluminum intoxication is characterized by aluminum-induced bone disease (AIBD), microcytic anemia, and encephalopathy. Most information concerning these has been obtained by the study of dialyzed renal failure patients. These patients had lost their ability to excrete aluminum and accumulated large body burdens of aluminum by transfer of the metal from contaminated dialyzates (most commonly tap water) during hemodialysis. The amount of transfer, and resultant body burdens, depended on the duration of treatment and the concentration of aluminum in the dialyzate.

In addition, toxic effects of aluminum have been demonstrated in four groups of patients with normal

kidney function: patients supported by total parenteral feeding, patients with hepatic insufficiency receiving aluminum antacids, premature infants receiving prolonged intravenous therapy, and other patients receiving parenteral therapy. Aluminum-induced toxicity has also been claimed in some occupationally exposed groups, but evidence supporting these claims is not conclusive.

Studies on developing mice and rats that had been exposed to aluminum either during gestation or during lactation indicate that this metal may have an adverse effect on the development of some regions of the brain. In such animals, adverse effects on reflexes and simple motor behaviors, but not consistently on learning and memory, have been demonstrated.

Aluminum-Induced Bone Disease

AIBD is characterized either by a low turnover osteomalacia or by an aplastic disease. Chemical analyses have shown these conditions to be present when bone aluminum levels are between 12 and $500\,\mu g\,g^{-1}$. High levels of the metal in diseased bones have also been demonstrated using aluminum-specific histochemical bone stains.

Aluminum-induced osteodystrophic osteomalacia develops in the absence of hypophosphatemia. The condition does not respond to vitamin D therapy, but it may be prevented by hyperparathyroidism. The disease is progressive and produces a variety of symptoms, including severe bone pain, muscle pains, and multiple nontraumatic fractures. It is normal for AIBD patients to remain asymptomatic for many years before physical signs of the disease, including funnel chest deformity, sternal bowing, and loss of height, become evident.

At the histological level, AIBD is characterized by a low rate of bone formation. The disease is variable, but bone removed from most patients show an increased amount of unmineralized osteoid; an increase in bone volume; a very low rate of bone apposition; a patchy, irregular pattern of calcification; a reduction in the number of active osteoblasts and osteoclasts; and irregular, misshaped bone trabeculae.

Although the causation of AIBD is not firmly established, it seems likely that it is produced by impaired bone matrix mineralization and by decreased osteoclastic activity. The progress of AIBD may be halted, and even reversed, by repeated administration of the chelating agent desferrioximine.

Microcytic Anemia

Microcytic anemia, a disease characterized by the presence of small red blood cells in blood, has been described in renal patients. This disease occurs in the absence of iron deficiency and is reversed following the use of purified dialyzate. The mechanism by which aluminum induces microcytic anemia is uncertain, but it has been suggested that a disturbance in the hem biosynthetic pathway may be involved. In this context, the red blood cells of intoxicated dialysis patients may contain a considerable fraction of the total blood aluminum content.

Encephalopathy

Several neurological effects have been attributed to aluminum intoxication. In weanling rats, dietary aluminum fed at high levels to their dams has been demonstrated to delay brain maturation. This effect has not been described in man.

Aluminum-induced impairment of cognitive function following the occupational exposure of gold miners to inhaled aluminum, the exposure of members of the general public to ingested aluminum sulfate, and, recently, exposure of workers and ex-workers in the aluminum industry has been claimed but not proven. No convincing evidence has been produced to either support or refute the existence of neurological effects at low levels of aluminum uptake. However, there is sufficient evidence that such effects may occur in man at some levels of uptake. For example, dialysis patients exposed to lower than average levels of aluminum sometimes demonstrate disturbed cerebral function compared to controls.

In one study of dialysis patients, correlations were sought between cognitive function and exposure to aluminum in both dialysis source water and administered oral gut phosfate binders. The results of the study were confusing since it found negative correlations between the cognitive measures and source water aluminum but positive correlations with the level of orally administered phosfate binder. Again, in a gold miner study, some results indicated cognitive impairment, whereas others were less conclusive. It has been noted that most studies performed to date are flawed, because they have failed to include normal aging controls. Attempts to improve cognitive scores by chelation therapy have met with very limited success.

The most important neurological effects produced by aluminum occur at large body burdens. These include ataxia, dysarthria, dysphagia, myoclonia, convulsions, and dementia. Epidemiological studies have shown that aluminum-induced encephalopathy (dialysis dementia) was absent at dialysis centers using water with aluminum concentrations less than $50\,\mu g\,l^{-1}$. In contrast, encephalopathy was common in those that employed water with aluminum

concentrations greater than 200 $\mu g\,l^{-1}$. At these centers, the prevalence of the disease increased significantly with increasing cumulative exposure to aluminum and was often a direct cause of death. In terminal cases, facial grimacing, myotonic spasms, and dysphagia interfere with eating and lead to inhalation pneumonia and death. The recorded concentrations of aluminum in the brains of such patients are highly variable, but values of 15–100 $mg\,kg^{-1}$ are typical. Experience has shown that chelation therapy with desferrioximine is effective in reversing neurological effects in renal patients.

The mechanisms of encephalopathy are not clear. Most evidence suggests that aluminum likely crosses the blood–brain barrier by a transferrin-mediated mechanism. Imaging secondary ion mass spectrometry has shown aluminum to be deposited within the brain cortex as focal deposits at sites known to be rich in transferrin receptors. These sites, corresponding to the distribution of pyramidal neurones, have a high demand for iron in the synthesis of respiratory chain enzymes. It is suggested that damage at these sites results in the neuropathy.

Evidence for a Role in Alzheimer's Disease

Alzheimer's disease (AD) is a progressive, often insidious, dementing disease occurring in mid- to late life. Its incidence increases with age, such that at age 85+ approximately 20% of people suffer from the condition. AD causes neurone death and a reduction in brain volume. The progression of the disease (which in most cases means approximately 7 years of intellectual and personal decline until death) cannot be arrested and eventually patients become bedridden. At this stage, concomitant bedsores, feeding difficulties, and pneumonia result in death.

The diagnosis of AD in made on the basis of the histological examination of brain tissues. These show the presence of widespread accumulations of β-amyloid senile plaques and neurofibrillary tangles throughout the limbic system and in parts of the cerebral isocortex and brain stem. β-Amyloid plaque formation occurs in the majority of nondemented elderly individuals, but neurofibrillary tangles are rare. Consequently, it has been suggested that in AD plaque formation precedes and may predispose to neurofibrillary tangle formation.

The etiology of the disease is complex and incompletely understood. Two risk factors for the disease have been identified and confirmed: old age and family history. 'Familial' AD is genetically variable but mutations in chromosome 21 are often involved. These cases show a marked tendency toward early onset and can be positively identified to represent only a small proportion of the total AD population.

The etiology of nonfamilial, sporadic AD is unknown. However, cases have been attributed to head injury and environmental factors, including aluminum. Involvement of aluminum in AD has been suggested because (1) of the similar symptomologies of AD and dialysis dementia; (2) the administration of aluminum to animals produces histological changes within the brain that are, in some respects, similar to those seen in the brains of AD patients; (3) of some reports indicating the presence of aluminum within the cores of senile plaques; (4) of the results of some epidemiological studies that have linked AD incidence either with aluminum levels in drinking water or with its consumption as medicines; and (5) a disease similar to AD is prevalent in some Pacific islands (Guam), where the levels of aluminum in soils and water are high.

However, (1) the pathologies of AD and dialysis dementia are different; (2) the histomorphological changes seen in experimental animals differ, in important respects, from those seen in the brains of AD patients; (3) not all studies have indicated the presence of aluminum within the cores of senile plaques, and attempts to demonstrate enhanced levels of aluminum in the brain of AD patients have mostly failed; (4) the results of the epidemiological studies are conflicting and have been criticised on methodological and logical grounds; and (5) Guam disease and AD are clinically different.

It follows that it is now generally accepted that in the absence of a clear association between exposure to aluminum and the disease and/or an identified mechanism for disease induction by the metal, there is insufficient evidence to suggest that aluminum is causative with respect to AD.

See also: **Aging**. **Bone**. **Food Safety**: Heavy Metals.

Further Reading

Ackrill P and Day JP (1993) The use of desferrioximine in dialysis-associated aluminium disease. *Contributions to Nephrology* **102**: 125–134.

Day JP, Drumm PV, Edwardson JA *et al.* (1994) Biological chemistry of aluminium studied using ^{26}Al and accelerator mass spectrometry. *Nuclear Instruments and Methods in Physics Research B* **92**: 463–468.

Doll R (1993) Review: Alzheimer's disease and environmental aluminium. *Age and Ageing* **22**: 138–153.

Edwardson JA, Moore PB, Ferrier IN *et al.* (1993) Effect of silicon on gastrointestinal absorption of aluminium. *Lancet* **342**: 211–212.

Gardner MJ and Gunn AM (1995) Speciation and bioavailability of aluminium in drinking water. *Chemical Speciation and Bioavailability* **7**: 9–16.

National Research Council (2003) *Food and Nutrition Board: Food Chemicals Codex, 5th edn* Washington, DC: National Academy Press.

Nieboer E and Gibson BL (1993) In *Health Effects of Aluminum: A Critical Review with Emphasis on Aluminum in Drinking Water*. Toronto: Ontario Ministry of Health.

Pennington JAT and Schoen SA (1995) Estimates of dietary exposures to aluminium. *Food Additives and Contaminants* **5**: 119–128.

Powell JJ and Thompson RPH (1993) The chemistry of aluminium in the intestinal lumen and its uptake and absorption. *Proceedings of the Nutrition Society* **52**: 241–253.

Priest ND (1993) The bioavailability and metabolism of aluminium compounds in man. *Proceedings of the Nutrition Society* **52**: 231–240.

Priest ND (2004) The biological behaviour and bioavailability of aluminium in man, with special reference to studies employing aluminium-26 as a tracer: Review and study update. *Journal of Environmental Monitoring* **6**: 1–30.

Priest ND, Newton D, Day JP *et al.* (1995) Human metabolism of aluminium-26 and gallium-67 injected as citrates. *Human & Environmental Toxicology* **14**: 287–293.

Priest ND, Talbot RJ, Austin JG *et al.* (1996) The bioavailability of aluminium-26 labelled aluminium citrate and aluminium hydroxide in volunteers. *Biometals* **9**: 221–228.

Rowan MJ (1993) Recent research on the causes of Alzheimer's disease. *Proceedings of the Nutrition Society* **52**: 255–262.

Talbot RJ, Newton D, Priest ND *et al.* (1995) Intersubject variability in the metabolism of aluminium following intravenous injection as citrate. *Human & Experimental Toxicology* **14**: 595–599.

US Department of Health and Human Services (1999) *Toxicological Profile for Aluminum*. Atlanta: US Department of Health and Human Services, Agency for Toxic Substances and Disease Registry, Public Health Service.

World Health Organization (1997) *Environmental Health Criteria 194: Aluminum*. Geneva: World Health Organization, International Programme on Chemical Safety.

AMINO ACIDS

Contents

Chemistry and Classification

P W Emery, King's College London, London, UK

Amino acids are a series of small organic molecules whose prime importance lies in the fact that they are the monomers from which proteins are made. The form and functions of proteins depend on the sequence in which the amino acids are joined together since each amino acid has specific chemical and physical properties. In this article, the structures and chemical properties of each amino acid are outlined, with an indication of how this affects the metabolic role of the free amino acid and how it affects the behavior of the amino acid residue within a protein. These chemical properties also form the basis for methods of analysis of amino acids. Some amino acids can be synthesized within the body from other molecules, whereas others cannot, so the final section explains the basis of the classification into essential and nonessential amino acids.

Chemical Structures and Nomenclature

Amino acids are small organic molecules with the general formula shown in **Figure 1**.

The central carbon atom in this structure is called the α-carbon, and the amino and carboxyl groups attached to it are known as the α-amino group and the α-carboxyl group, respectively. The R groups of the 20 amino acids that can be incorporated into proteins are shown in **Table 1**; these R groups give the different amino acids their specific chemical and physical properties.

The α-amino group acts as a weak base and is always protonated at physiological pH; similarly, the α-carboxyl group acts as weak acid and at physiological pH is always ionized. Thus, free amino acids in biological material exist as zwitterions, as shown in **Figure 2**.

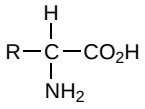
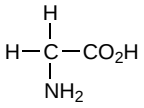
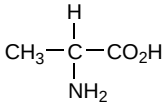
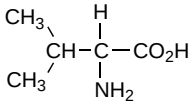
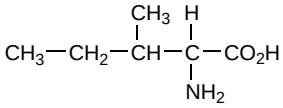
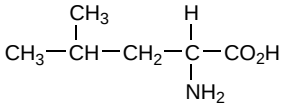
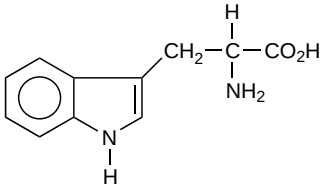
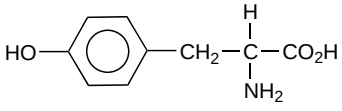
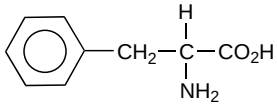
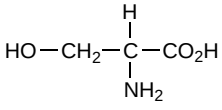
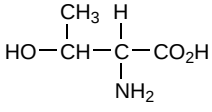
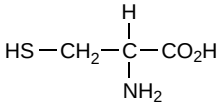
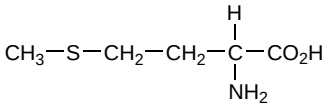


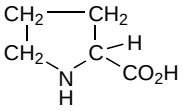
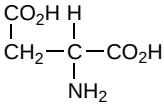
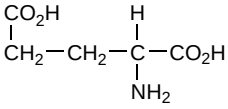
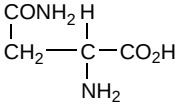
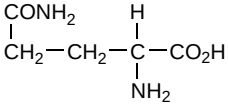
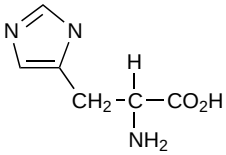
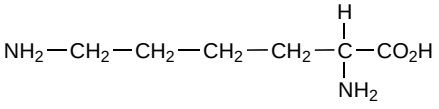
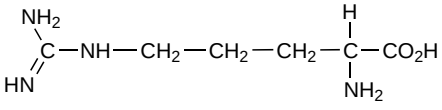
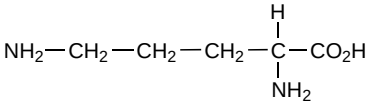
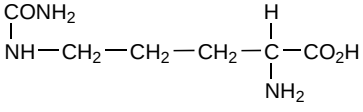
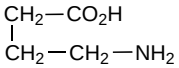
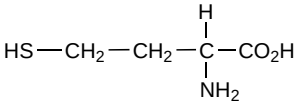
Figure 1 Amino acid structure.

Table 1 Amino acid characteristics

Name (3 letter code; 1 letter code)	*Structure*	*Molecular weight*	*pK_a*
Small neutral amino acids Glycine (Gly; G)		75	2.35 9.78
Alanine (Ala; A)		89	2.35 9.87
Branched-chain amino acids Valine (Val; V)		117	2.29 9.74
Isoleucine (Ile; I)		131	2.32 9.76
Leucine (Leu; L)		131	2.33 9.74
Aromatic amino acids Tryptophan (Trp; W)		204	2.43 9.44
Tyrosine (Tyr; Y)		181	2.20 9.11 10.13
Phenylalanine (Phe; F)		165	2.16 9.18
Hydroxyl-containing amino acids Serine (Ser; S)		105	2.19 9.21
Threonine (Thr; T)		119	2.09 9.10
Sulfur-containing amino acids Cysteine (Cys; C)		121	1.92 8.35 10.46
Methionine (Met; M)		149	2.13 9.28

Continued

Table 1 Continued

Name (3 letter code; 1 letter code)	*Structure*	*Molecular weight*	*pK_a*
Imino acid Proline (Pro; P)		115	1.95 10.64
Acidic side chains Aspartic acid (Asp; D)		133	1.99 3.90 9.90
Glutamic acid (Glu; E)		147	2.10 4.07 9.47
Amides Asparagine (Asn; N)		132	2.10 8.84
Glutamine (Gln; Q)		146	2.17 9.13
Basic side chains Histidine (His; H)		155	1.80 6.04 9.76
Lysine (Lys; K)		146	2.16 9.18 10.79
Arginine (Arg; R)		174	1.83 8.99 12.48
Nonprotein amino acids Ornithine		132	1.71 8.69 10.76
Citrulline		175	Not determined
γ-Aminobutyric acid (GABA)		103	4.03 10.56
Homocysteine		117	2.22 8.87 10.86

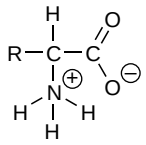


Figure 2 Zwitterionic structure of an amino acid.

The α-carbon atom is asymmetric so that amino acids show stereoisomerism; the exception to this is glycine, in which the R group is a second hydrogen atom. Most of the amino acids found in nature are in the L form, and only L-amino acids can be used for protein synthesis in higher organisms. However, D-amino acids may be ingested from bacterial sources, and if high concentrations accumulate they may be toxic. The human body has a D-amino acid oxidase enzyme, found in the liver and the kidney, that disposes of these molecules by oxidative deamination.

The most important common chemical property of the amino acids is their ability to form peptide bonds with one another. The α-amino group of one amino acid reacts with the α-carboxyl group of another to form a peptide bond with the elimination of water (**Figure 3**). The results of this process are conventionally known as peptides or oligopeptides if they contain 2–20 amino acid residues or as polypeptides, which may contain 21 to several thousand amino acid residues. The polypeptides may undergo further processing, including chemical modification, before taking up their final conformation as proteins.

Each amino acid also has specific chemical properties that depend on the nature of the R group. This affects the behavior of the free amino acids and the corresponding amino acid residues in peptides and polypeptides. For convenience, the amino acids may be considered in groups according to some common properties.

Small Neutral Amino Acids: Glycine and Alanine

The small side chains, a hydrogen atom and a methyl group, respectively, have little effect on the shape of a peptide chain. The free amino acids tend to be heavily involved as metabolic intermediates. Glycine is a precursor of purines, porphyrins, bile acids, and creatine; it acts as a neurotransmitter and as a conjugating substance that aids the excretion of xenobiotics by making them more water-soluble. Alanine is the transamination product of pyruvic acid and is thus closely associated with the metabolism of carbohydrates, acting as a major precursor for gluconeogenesis.

Branched-Chain Amino Acids: Valine, Leucine, and Isoleucine

These have bulky, nonpolar side chains, so they are often found within the hydrophobic core of proteins. Isoleucine has an extra chiral center, so four optical isomers are theoretically possible, but only L-isoleucine (and not L-allo-isoleucine) is found in proteins. Branched-chain amino acids are metabolized initially in muscle and adipose tissue rather than liver, where most of the other amino acids are metabolized.

Aromatic Amino Acids: Tryptophan, Tyrosine, and Phenylalanine

These are also bulky and nonpolar, and they may interact with other hydrophobic molecules. The phenolic hydrogen of tyrosine is weakly acidic and can form hydrogen bonds to create cross-links or can be donated during catalysis. Tyrosine residues on certain membrane-bound receptors become phosphorylated by tyrosine kinase domains, thereby initiating a signal transduction cascade. Tryptophan is important as a precursor of the neurotransmitter 5-hydroxytryptamine (serotonin) and of the nicotinamide-containing coenzymes NAD and NADP. Phenylalanine can be converted to tyrosine in the body, but not vice versa. Tyrosine is a precursor of the catecholamines and the thyroid hormones and also the pigment melanin.

Hydroxyl-Containing Amino Acids: Serine and Threonine

These are polar, very weakly acidic molecules but uncharged at neutral pH. They can form hydrogen bonds and are thus quite soluble. Threonine has an additional chiral centre, but again only L-threonine is found in proteins. Serine is found at the active centre of some enzymes. It is also the usual site of

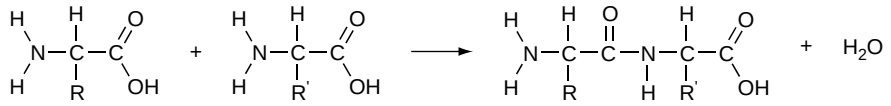


Figure 3 Peptide bond formation.

attachment for the carbohydrate residues in glycoproteins and for the phosphoryl groups in phosphoproteins.

Sulfur-Containing Amino Acids: Cysteine and Methionine

Methionine is nonpolar, but cysteine is polar. Cysteine can form weak hydrogen bonds with oxygen and nitrogen; it is also weakly acidic and is sometimes found at the active site of enzymes. Cysteine also acts as a reducing agent within the cell, both as the free amino acid and in the form of the antioxidant tripeptide glutathione. The sulfydril groups of two cysteine residues can be oxidized to form the double amino acid cystine, and this is the predominant form of the amino acid in extracellular fluid. When the same reaction occurs between cysteine residues in adjacent polypeptide chains, a strong, covalent disulfide bond is formed that gives the protein a rigid structure. This appears to be particularly important in stabilizing extracellular or secreted proteins. Methionine can be converted to *S*-adenosyl-methionine, the donor of methyl groups in transmethylation reactions. Methionine can be converted to cysteine in the body, but not vice versa. Selenium can replace sulfur in some cysteine and methionine residues, particularly when selenium intake is high. The antioxidant protein glutathione peroxidase requires a selenocysteine residue at its active site.

Imino Acid: Proline

Since its structure contains a secondary amine rather than a primary amine, proline is actually an imino acid rather than an amino acid, but it forms peptide bonds and is incorporated into proteins just like an amino acid. It causes an abrupt and rigid change of direction in the polypeptide chain, and this has a major effect on the final conformation of the protein. The carbon at the 4 position can be hydroxylated to form hydroxyproline. Every third residue of the structural protein collagen is a hydroxyproline residue.

Acidic Side Chains: Aspartic Acid and Glutamic Acid

These are dicarboxylic acids, although at physiological pH they exist almost entirely in the anionic form and so should be referred to as aspartate and glutamate. They are mainly found on the surfaces of proteins. The free amino acids play a central role in transamination reactions, equilibrating rapidly with their corresponding keto acids oxaloacatate and 2-oxoglutarate. Glutamate is a precursor for the inhibitory neurotransmitter γ-aminobutyric acid (GABA). The monosodium salt of glutamate is used in the food industry as a flavor enhancer.

Amides: Asparagine and Glutamine

Although they are uncharged, these molecules are strongly polar. They are often found on the surface of proteins, where they can form hydrogen bonds with water or with other polar molecules. The conversion of glutamate to glutamine is central to the disposal of ammonia and to the maintenance of acid–base balance. Glutamine is a precursor for the synthesis of purines and pyrimidines. It is also a precursor for gluconeogenesis, and it is the main source of energy for enterocytes and leucocytes. There is evidence that glutamine may play a role in the control of protein metabolism and that it may be beneficial in augmenting the immune response in critically ill patients.

Basic Side Chains: Histidine, Lysine, and Arginine

These are hydrophilic amino acids that are positively charged at neutral pH. The imidazole group of histidine has a pK_a just below 7, so it is weakly ionized at physiological pH, giving it some buffering capacity and making it useful at the active site of many enzymes. Histidine is also a precursor for the physiologically active amine histamine. Arginine is an intermediate in the urea cycle and a precursor for polyamine synthesis. It is also the precursor for nitric oxide, which appears to have many physiologically important properties, including that of an endothelial-derived relaxing factor and a cell signaling molecule in the coordination of the inflammatory response. Although arginine, like glutamine, is a nonessential amino acid, there is evidence that increasing the dietary supply of arginine can improve clinical outcome in critically ill patients. Lysine is the limiting amino acid in cereals and cereal-based diets.

Posttranslational Modification

Some amino acid residues may become chemically modified after they have been incorporated into polypeptide chains. They will thus be present when the protein is degraded but cannot be reutilized for protein synthesis.

Hydroxylation of proline to hydroxyproline is mainly associated with collagen. Hydroxylysine is also found in collagen.

The side chain nitrogen atoms of the dibasic amino acids (histidine, arginine, and lysine) can all

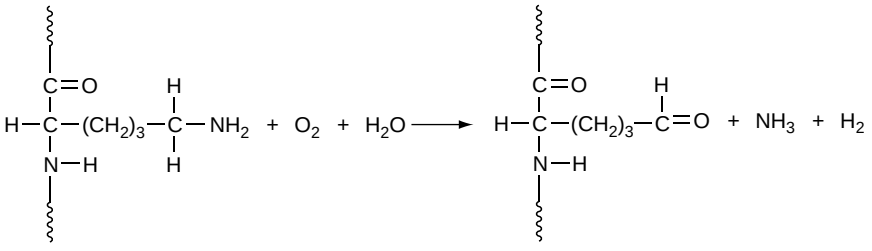
be methylated. For example, N^{τ}-methylhistidine (3-methylhistidine) is found mainly in the contractile proteins actin and myosin so that detection of N^{τ}-methylhistidine in a food sample usually indicates the presence of meat. It has also been suggested that measurement of the urinary excretion of N^{τ}-methylhistidine could provide an index of the rate of breakdown of myofibrillar proteins in skeletal muscle, although interpretation is complicated by the presence of N^{τ}-methylhistidine derived from other tissues.

The hydroxyl groups of serine, threonine, and tyrosine can all be phosphorylated. Phosphoserine residues bind calcium and are found in proteins such as casein. Another calcium-binding residue is γ-carboxyglutamic acid, which is found in prothrombin.

The hydroxyl groups of serine can also be glycosylated to form glycoproteins and proteoglycans. The amide group of asparagine can also be glycosylated.

The ε-N of certain lysine residues can be oxidized by the copper-containing enzyme lysyl oxidase to form allysine. Four allysine residues in adjacent polypeptide chains may then condense to form desmosine (**Figure 4**). This covalent link gives considerable strength and elasticity to the connective tissue protein elastin.

The ε-N of lysine residues is also susceptible to chemical reactions within food systems. It undergoes the Maillard reaction with carbonyl groups of carbohydrates to form a series of brown and slightly bitter products. This is an integral part of the baking process when producing bread, cakes, and biscuits, although there is evidence that large quantities of some Maillard products may be toxic or carcinogenic. On the other hand, since the lysine in Maillard products is not biologically available when the food is ingested, this can seriously reduce the protein quality of heat-treated animal feedstuffs.



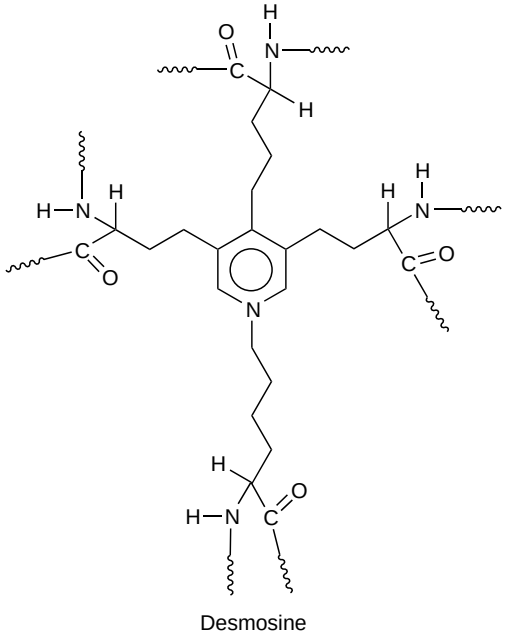


Figure 4 Formation of allysine and structure of desmosine.

Proteins within living systems can also be damaged by covalent binding to other molecules (usually reactive biochemicals) to form adducts, thereby rendering the protein inoperative or immunogenic. Adducts can be formed by the reaction of an aldehyde function with a receptive nucleophilic centre in the protein, particularly the ε-amino groups on lysine residues but also the α-amino terminus, the thiol groups on cysteine residues, the imidazole groups on histidine residues, and the phenolic groups on tyrosine residues. The aldehydes that may be involved in adduct formation include malondialdehyde and 4-hydroxy-2-nonenal, which are produced by free radical damage to polyunsaturated fatty acids in cell membranes, and acetaldehyde, which is produced when alcohol is metabolized. Adduct formation may play a role in the pathological processes leading to diseases such as alcoholic cirrhosis and coronary heart disease.

Nonprotein Amino Acids

There are several amino acids found in biological systems that are not incorporated into proteins. Ornithine and citrulline, for example, are intermediates of the urea cycle; GABA is a neurotransmitter.

Homocysteine is an intermediate in the transulphuration pathway for the conversion of methionine to cysteine. Homocystinuria is an inborn error of metabolism that is characterized by the accumulation of high concentrations of homocysteine, and this leads to severe cardiovascular disease at an early age. However, there is a much more common mutation in the enzyme 5,10-methylenetetrahydrofolate reductase that causes a moderate increase in plasma homocysteine concentration in more than 10% of the population. A high plasma homocysteine concentration appears to be an independent risk factor for cardiovascular disease in the population as a whole, although the mechanism is not known. Supplementing the diet with folic acid is often effective in reducing plasma homocysteine concentration because methyltetrahydrofolate is a substrate for the remethylation of homocysteine by the vitamin B_{12}-dependent enzyme methionine synthase. An inverse relationship has been observed between plasma homocysteine concentration and folate status in many studies, and this has led to the proposal that plasma homocysteine concentration may be used as a biomarker of folate intake.

Peptides

In addition to free amino acids and proteins, significant amounts of amino acids are present in physiological systems as small peptides. One of the most important is the tripeptide glutathione (γ-glutamylcysteinylglycine), which acts as an intracellular antioxidant.

Dipeptides found within the cell include carnosine (β-alanylhistidine) and its methylated derivatives anserine and balenine. These may act as buffers; no other physiological role has been identified.

Peptides are also used in food systems. For example, cysteine-containing peptides, or cysteine itself, are used as improvers in bread making to speed up the cross-linking that is required to give the bread its texture.

Another peptide used in the food industry is aspartame, which is composed of aspartic acid and phenylalanine. It is a very powerful sweetener that does not have the bitter aftertaste of some other intense sweeteners.

Analysis

The analysis of amino acids is based on chromatographic techniques. Traditional amino acid analyzers involved separation of the amino acid mixture on a column of ion-exchange resin using a series of sodium or lithium citrate buffers of increasing pH. The column effluent was then reacted with ninhydrin and passed through a spectrophotometer that would detect and quantify a series of peaks. This method is still used, although high-performance liquid chromatography (HPLC) hardware is usually employed. Other postcolumn detection systems can be used, replacing the ninhydrin reagent with orthophthalaldehyde (OPA) or fluorescamine and detecting the product fluorimetrically, thereby increasing the sensitivity.

Amino acids can also be separated by HPLC on a reversed-phase column. The mobile phase is usually based on an aqueous buffer with a gradient of increasing concentration of acetonitrile. In this case, the amino acids are usually converted to a fluorimetrically detectable (or ultraviolet-absorbing) form before being injected onto the column. A wide variety of derivatizing agents can be used for this, including OPA, 1-fluoro-2,4-dinitrobenzene, dansyl chloride, phenylisothiocyanate, and 9-fluorenylmethyl chloroformate.

It is also possible to measure amino acids using gas–liquid chromatography, but this has never been popular, perhaps because the sample cleanup and derivatization steps are more laborious. The amino acids have to be converted to volatile derivatives before analysis, commonly either *N*-trifluoroacetyl-*n*-butyl or *N*-heptafluorobutyl-isobutyl esters. Gas–liquid chromatography is potentially a very sensitive method. It can also be coupled with mass spectrometry for identification of unknown compounds or

for measurement of tracer enrichment when carrying out metabolic studies with stable isotopes.

These analytical methods can be applied equally to the measurement of amino acids in proteins, after hydrolysis, or free amino acids in physiological fluids such as plasma, urine, or tissue extracts. For physiological fluids, the protein must first be removed, and this is usually accomplished by precipitating with an acid such as sulfosalicylic acid or an organic molecule such as acetonitrile. The chromatographic requirements for physiological fluids are more demanding than for protein hydrolysates because there are many more contaminating substances producing extra peaks from which the amino acid peaks must be resolved, so the run time is generally longer.

Proteins have to be hydrolyzed before their amino acid composition can be measured. This is done by heating to 110 °C with an excess of 6 M HCl, either under nitrogen or in a vacuum. Proteins are usually hydrolyzed for 24 h, but this actually represents a compromise since some amino acids, including valine and isoleucine, may take longer than 24 h to liberate completely, whereas others, including tyrosine, threonine, and serine, are progressively destroyed. Thus, for complete accuracy a protein should be hydrolysed for different lengths of time (usually between 16 and 72 h) and appropriate extrapolations made to the analytical values for each amino acid.

Acid hydrolysis destroys tryptophan, so a separate alkaline hydrolysis is needed to measure this amino acid. The sulfur-containing amino acids are also partially oxidized during acid hydrolysis, so the protein may be oxidized with performic acid before hydrolysis and the oxidation products of cysteine and methionine measured. Finally, acid hydrolysis converts the amides glutamine and asparagine to their parent dicarboxylic acids, so values are often reported as total [glutamic acid plus glutamine] and [aspartic acid plus asparagine]. If separate values are required for the amides, the protein must be subjected to enzymic hydrolysis.

Classification

From a nutritional standpoint, the most important classification of amino acids is the division between those that are essential (or indispensable) and those that are nonessential (or dispensable). Essential amino acids may be defined as those that the body cannot synthesize in sufficient quantities.

This classification is based on work carried out by W. C. Rose in the 1930s. Young, rapidly growing rats were fed purified diets from which one amino acid was removed. For some of the amino acids, this made no difference to the rats' growth rate—these are the nonessential amino acids shown in **Table 2**. For the essential amino acids removal from the diet resulted in immediate cessation of growth, followed by loss of weight, decline in food intake, and eventual death of the rats. The response to the removal of arginine was less dramatic because the rats continued to grow, but at a reduced rate. Thus, it appeared that the rat can synthesize arginine, but not at a high enough rate to support maximal growth.

Table 2 Essential amino acids for the rat

Valine
Isoleucine
Leucine
Tryptophan
Phenylalanine
Threonine
Methionine
Histidine
Lysine
(Arginine)

It has subsequently been shown that the reason why certain amino acids are essential is that their carbon skeletons cannot be synthesized in mammalian cells. As long as the carbon skeletons are present, all amino acids except threonine and lysine can be formed by transamination. It should be noted, however, that tyrosine can only be synthesized from phenylalanine, and cysteine can only be synthesized from methionine.

Rose also determined which amino acids are essential for man by carrying out nitrogen balance experiments on healthy young adult volunteers. He showed that nitrogen balance could be maintained on a diet in which the only source of nitrogen was a mixture of the 10 amino acids that are essential for the rat. He then found that histidine and arginine could also be removed without affecting nitrogen balance. Thus, the 8 amino acids that are essential for adult man are shown in **Table 3**.

More recent work has identified certain circumstances, usually associated with disease or recovery from malnutrition, in which the addition of particular nonessential amino acids to an otherwise

Table 3 Essential amino acids for man

Valine
Isoleucine
Leucine
Phenylalanine
Tryptophan
Threonine
Methionine
Lysine

adequate diet appears to cause an unexpected improvement in either nitrogen balance or growth rate. It is hypothesized that the rate at which the body can synthesize these particular amino acids is limited, and that in extreme circumstances the requirement for them becomes greater than the rate at which they can be synthesized. These amino acids are thus sometimes called conditionally essential amino acids, and these include glycine, arginine, histidine, and glutamine.

See also: **Amino Acids**: Metabolism; Specific Functions. **Protein**: Synthesis and Turnover; Requirements and Role in Diet; Digestion and Bioavailability; Quality and Sources; Deficiency.

Further Reading

Bender DA (1985) *Amino Acid Metabolism*, 2nd edn. Chichester, UK: John Wiley & Sons.

Bigwood EJ (ed.) (1972) Protein and amino acid functions. In *International Encyclopaedia of Food and Nutrition*. Oxford: Pergamon Press.

Gehrke CW and Zumwalt RW (1987) Symposium on chromatography of amino acids. *Journal of the Association of Official Analytical Chemists* **70**: 146–147.

Laidlaw SA and Kopple JD (1987) Newer concepts of the indispensable amino acids. *American Journal of Clinical Nutrition* **46**: 593–605.

Metzler DE (1977) *Biochemistry* New York: Academic Press.

Reeds PJ (2000) Dispensable and indispensable amino acids for humans. *Journal of Nutrition* **130**: 1835S–1840S.

Rose WC (1957) The amino acid requirements of adult man. *Nutrition Abstracts and Reviews* **27**: 631–647.

Williams AP (1988) Determination of amino acids. In: Macrae R (ed.) *HPLC in Food Analysis*, 2nd edn., pp. 441–470. London: Academic Press.

Metabolism

P W Emery, King's College London, London, UK

Amino acids are generated within the body from three different sources. They enter the body from protein in the diet, and nonessential (dispensable) amino acids are synthesized from other metabolic intermediates, but by far the largest quantities of free amino acids arise from the breakdown of tissue proteins. Similarly, there are three metabolic fates for amino acids. Amino acid disposal is dominated by protein synthesis, but amino acids are also oxidized to carbon dioxide, water, and urea, or they may be metabolized to other small molecules. The pathways involved in each of these processes are considered, followed by a discussion of the movement of amino acids between different compartments within the body (**Figure 1**).

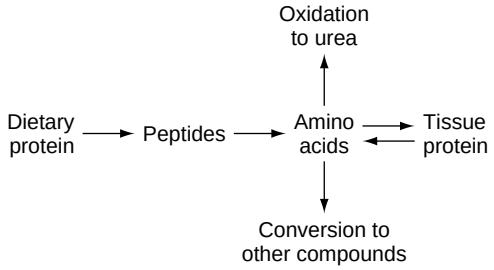


Figure 1 Overview of amino acid metabolism.

Amino Acid Supply

Dietary Intake

Protein is digested in the stomach by pepsins and in the small intestine by proteolytic enzymes from the pancreas. The products of digestion are mainly small peptides, which are then taken up by the intestinal epithelium and hydrolyzed to free amino acids. The portal circulation transports these amino acids to the liver, where approximately 75% of the amino acids are metabolized. The remaining 25% then enter the systemic circulation for transport to other tissues.

Amino Acid Biosynthesis

The essential (indispensable) amino acids must be supplied in the diet because their carbon skeletons cannot be synthesized in the human body, whereas the nonessential amino acids can be synthesized from common intermediates of the central metabolic pathways within the cell (i.e., glycolysis, the pentose phosphate pathway and the TCA cycle). As long as the keto-analogs are present, almost all amino acids can be generated by the process of transamination. The exceptions are threonine and lysine. Threonine is a poor substrate for mammalian transaminase enzymes, whereas the keto-analog of lysine, α-oxo-ε-aminocaproate, is unstable and cyclizes spontaneously to pipecolic acid.

Glutamic acid, glutamine, proline, and arginine Glutamic acid is synthesized by transamination of 2-oxoglutarate, a TCA cycle intermediate. This reaction represents the first stage in the catabolism of many other amino acids, particularly the branched-chain amino acids. Vitamin B_6 is a cofactor for all transamination (aminotransferase) reactions. Glutamine is made from glutamic acid and ammonium in an energy-requiring reaction catalyzed by glutamine synthetase. The synthesis of glutamine plays an important role in the removal of the ammonium

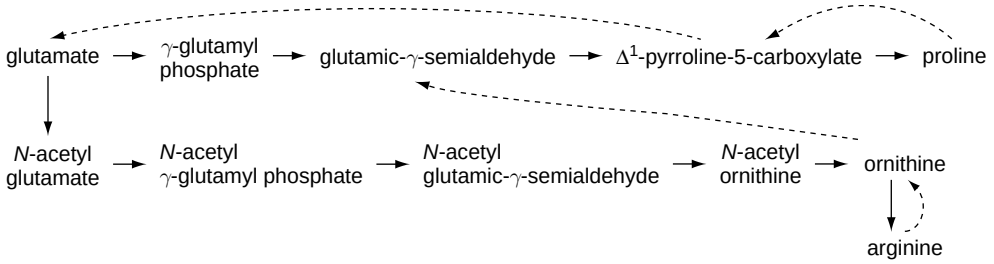


Figure 2 Synthesis and catabolism of proline and arginine. Solid lines indicate biosynthetic pathways; broken lines indicate catabolic pathways.

formed in peripheral tissues by deamination of amino acids as it is transported to the liver and used for urea synthesis.

Glutamic acid can be phosphorylated to γ-glutamyl phosphate by ATP, and this can then be dephosphorylated to glutamic-γ-semialdehyde. This undergoes nonenzymic cyclization to Δ^1-pyroline-5-carboxylate, which can then be reduced to proline (**Figure 2**).

Arginine is made from ornithine via the reactions of the urea cycle. Ornithine can theoretically be made by transamination of glutamic-γ-semialdehyde, but as mentioned previously this cyclizes spontaneously to pyroline-5-carboxylate. Thus, in practice glutamate is first acetylated by acetyl CoA to *N*-acetyl glutamate so that when this is converted to *N*-acetyl glutamic-γ-semialdehyde the amino group is blocked and cannot cyclize. The *N*-acetyl glutamic-γ-semialdehyde is then transaminated to *N*-acetyl ornithine, and this is deacetylated to ornithine (**Figure 2**).

Aspartic acid and asparagine Aspartic acid is derived from transamination of oxaloacetic acid, a TCA cycle intermediate. As with glutamic acid synthesis, this represents a common mechanism for removing amino groups from many other amino acids. Asparagine is made from aspartic acid by transfer of the amide group from glutamine.

Alanine Alanine is made by transamination of pyruvic acid, which is generated by glycolysis.

Serine and glycine Serine and glycine are readily interconvertible via methylene tetrahydrofolate, which either condenses with a glycine molecule to yield serine or is cleaved to yield glycine and tetrahydrofolate (**Figure 3**). However, there are also separate biosynthetic pathways for both molecules. Glycine can be synthesized by transamination of glyoxylate, which arises from the pentose phosphate pathway. Serine can be made by dephosphorylation of 3-phosphoserine, which is made by sequential dehydrogenation and transamination of 3-phosphoglycerate, a glycolytic intermediate (**Figure 3**).

Histidine Histidine is synthesized by a relatively long pathway that has no branch points and does not lead to the formation of any other important intermediates. The main precursors are phosphoribosyl pyrophosphate and ATP, with the α-amino group arising by transamination from glutamate (**Figure 4**).

Cysteine In man and other animals, cysteine can only be synthesized from the essential amino acid methionine. Methionine reacts with ATP to form *S*-adenosylmethionine, an important methylating agent within the cell. Transfer of the methyl group results in the formation of *S*-adenosylhomocysteine, which is then converted to homocysteine. Homocysteine can condense with serine to form cystathionine, which is then cleaved by cystathionase to yield cysteine (**Figure 5**).

An alternative fate for homocysteine is remethylation to methionine. The methyl donor for this reaction can be either methyltetrahydrofolate, in a

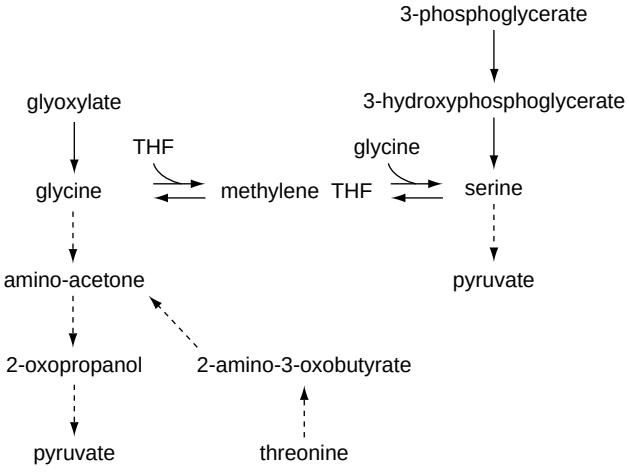


Figure 3 Synthesis and catabolism of glycine, serine, and threonine. Solid lines indicate biosynthetic pathways; broken lines indicate catabolic pathways. THF, tetrahydrofolate.

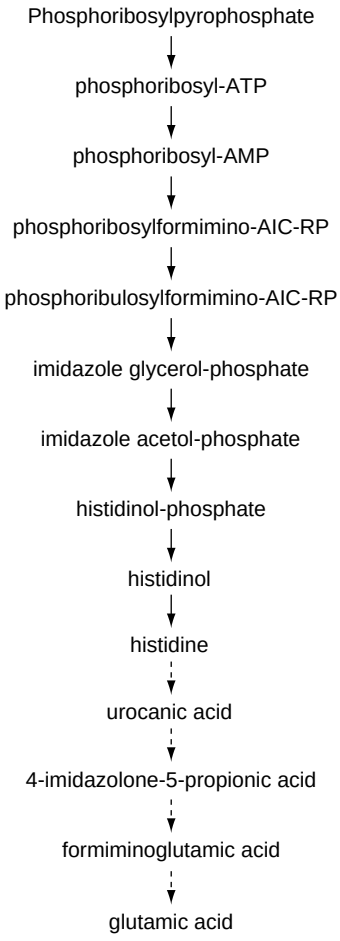


Figure 4 Synthesis and catabolism of histidine. Solid lines indicate biosynthetic pathways; broken lines indicate catabolic pathways.

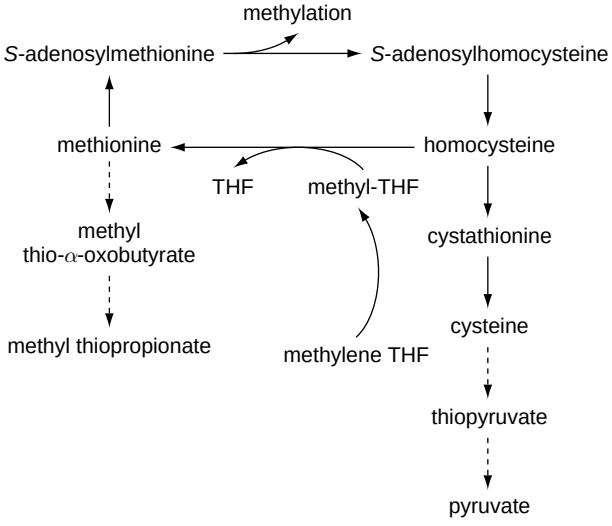


Figure 5 Synthesis and catabolism of methionine and cysteine. Solid lines indicate biosynthetic pathways; broken lines indicate catabolic pathways. THF, tetrahydrofolate.

reaction for which vitamin B_{12} is a cofactor, or betaine. Remethylation seems to be quite sensitive to folate status, and plasma homocysteine is becoming accepted as a biomarker of nutritional status with respect to folate.

Homocystinuria is an important inborn error of metabolism that is caused by impaired activity of cystathionine synthetase, the enzyme that catalyzes the condensation of homocysteine with serine. One of the consequences of homocystinuria is premature cardiovascular disease. There is considerable evidence that milder elevations of plasma homocysteine, caused by poorly active variants of the methylenetetrahydrofolate reductase enzyme (which is required to make the methyl donor methyltetrahydrofolate) or by low folic acid status, may be an important risk factor for cadiovascular disease throughout the population.

Tyrosine In mammals, including man, tyrosine can only be formed by hydroxylation of the essential amino acid phenylalanine. The inborn error of metabolism phenylketonuria is caused by a failure of the enzyme phenylalanine hydroxylase.

Protein Breakdown

Amino acids are continously released by the hydrolysis of proteins. This occurs by several different mechanisms. Much intracellular proteolysis occurs within lysosomes, which provide the acidic environment within which enzymes such as cathepsins operate. However, there are also cytosolic proteolytic enzymes that operate at neutral or alkaline pH. These include the enzymes that hydrolyze proteins bound to ubiquitin. There are also extracellular proteinases that degrade extracellular proteins such as collagen.

Disposal of Amino Acids

Protein Synthesis

Protein synthesis represents the major route of disposal of amino acids. Amino acids are activated by binding to specific molecules of transfer RNA and assembled by ribosomes into a sequence that has been specified by messenger RNA, which in turn has been transcribed from the DNA template. Peptide bonds are then formed between adjacent amino acids. Once the polypeptide chain has been completed, the subsequent folding, posttranslational amino acid modifications, and protein packaging are all determined by the primary sequence of amino acids. The rate of protein synthesis is controlled by the rate of transcription of specific genes,

the number and state of aggregation of ribosomes, and modulation of the rate of initiation of peptide synthesis.

Amino Acid Catabolism

Many amino acids can be converted to other useful molecules within the cell, and the same pathways may also lead to oxidation of the amino acid. It is therefore convenient to consider these metabolic fates together.

Glycine, serine, and threonine The interconversion of glycine and serine has already been mentioned (**Figure 3**), and this can act as a mechanism for disposal of either amino acid. In quantitative terms, however, the main tendency is for both to be converted to the common intermediate methylene tetrahydrofolate, which acts as a methyl donor in many important biosynthetic reactions, including the conversion of dUMP to dTMP for DNA synthesis.

An alternative pathway for serine catabolism is deamination to pyruvate. However, the K_m of this enzyme is relatively high, so the pathway would only operate at high serine concentrations (**Figure 3**).

Another pathway of glycine catabolism is by condensation with acetyl CoA to form amino-acetone. This is then transaminated and dehydrogenated to yield carbon dioxide and pyruvate. Amino-acetone is also formed by the NAD-linked dehydrogenation of threonine, followed by the spontaneous decarboxylation of the unstable intermediate 2-amino-3-oxo-butyrate, and this appears to be the main pathway of catabolism of threonine in mammals (**Figure 3**).

Glycine is also an important precursor for several larger molecules. Purines are synthesized by a pathway that begins with the condensation of glycine and phosphoribosylamine. Porphyrins, including hem, are synthesized from glycine and succinyl CoA via δ-aminolaevulinic acid. Creatine synthesis involves the addition of the guanidino nitrogen from arginine to glycine. Glycine is also used to conjugate many foreign compounds, allowing them to be excreted in the urine. Glycine also conjugates with cholic acid to form the major bile acid glycocholic acid.

Glutamic acid, glutamine, proline, and arginine Glutamic acid can be transaminated to 2-oxoglutarate, which can enter the TCA cycle. The amino group would be tranferred to aspartate, which would then enter the urea cycle. Alternatively, glutamate can be deaminated by glutamate dehydrogenase, with the resulting ammonium entering the urea cycle as carbamoyl phosphate. Decarboxylation of glutamate yields γ-aminobutyric acid, an important inhibitory neurotransmitter.

Glutamine is deaminated to glutamic acid in the kidney; this process is central to the maintenance of acid–base balance and the control of urine pH. Glutamine also acts as a nitrogen donor in the synthesis of purines and pyrimidines.

Proline is metabolized by oxidation to glutamic acid, although the enzymes involved are not the same as those that are responsible for the synthesis of proline from glutamic acid (**Figure 2**).

Arginine is an intermediate of the urea cycle and is metabolized by hydrolysis to ornithine. Ornithine can transfer its δ-amino group to 2-oxoglutarate, forming glutamic-γ-semialdehyde, which can then be metabolized to glutamate (**Figure 2**). Ornithine can also be decarboxylated to putrescine, which in turn can be converted to other polyamines such as spermidine and spermine.

Arginine can also be oxidized to nitric oxide and citrulline. Nitric oxide appears to be an important cellular signaling molecule that has been implicated in numerous functions, including relaxation of the vascular endothelium and cell killing by macrophages. In the vascular endothelium, nitric oxide is made by two different nitric oxide synthase isozymes, one of which is inducible and the other acts constitutitively.

Aspartic acid and asparagine Aspartic acid can be transaminated to oxaloacetic acid, a TCA cycle intermediate. Alternatively, when aspartic acid feeds its amino group directly into the urea cycle, the resulting keto acid is fumarate, another TCA cycle intermediate. Aspartic acid is also the starting point for pyrimidine synthesis. Asparagine is metabolized by deamidation to aspartic acid.

Lysine In mammals, lysine is catabolized by condensing with 2-oxoglutarate to form saccharopine, which is then converted to α-aminoadipic acid and glutamate. The α-aminoadipic acid is ultimately converted to acetyl CoA. In the brain, some lysine is metabolized via a different pathway to pipecolic acid (**Figure 6**). Lysine is also the precursor for the synthesis of carnitine, which carries long-chain fatty acids into the mitochondrion for oxidation. In mammals this process starts with three successive methylations of a lysine residue in a protein. The trimethyl lysine is then released by proteolysis before undergoing further reactions to form carnitine.

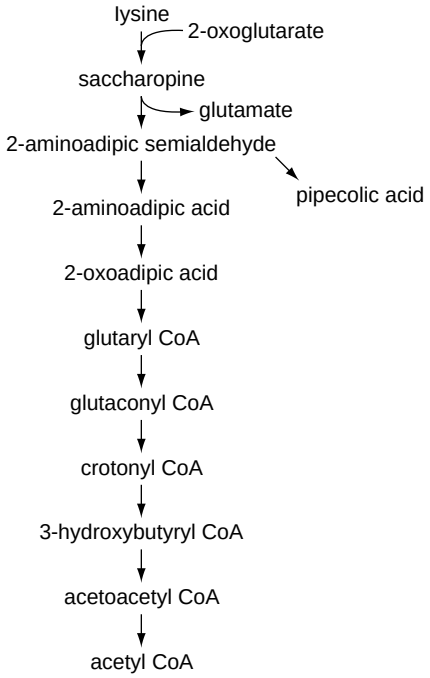


Figure 6 Metabolism of lysine.

Methionine and cysteine The conversion of methionine to cysteine via the so-called transsulfuration pathway has already been mentioned (**Figure 5**). This pathway appears to act mainly as a biosynthetic pathway for the synthesis of cysteine. There is an alternative pathway for methionine catabolism that involves transamination to methyl thio-α-oxobutyrate and then to methyl thiopropionate.

Cysteine can be transaminated to thiopyruvate, which then undergoes desulfuration to pyruvate and hydrogen sulfide (**Figure 5**). Cysteine can also be oxidized to cysteine sulfinic acid, which can then be decarboxylated to hypotaurine, and this is then oxidized to taurine. High concentrations of taurine are found within most cells of the body, although its role is far from clear. In the liver the main fate of taurine is the production of taurocholic acid, which acts as an emulsifier in the bile. Another key role for cysteine is in the synthesis of the tripeptide glutathione, which is an important intracellular antioxidant.

Leucine, isoleucine, and valine The branched-chain amino acids are unusual in that the first step in their metabolism occurs in muscle rather than liver. This step is transamination, producing α-oxoisocaproic acid, α-oxo-β-methyl valeric acid, and α-oxoisovaleric acid. These ketoacids are then transported to the liver for decarboxylation and dehydrogenation. Subsequent catabolism yields acetyl CoA and acetoacetate in the case of leucine, acetyl CoA and propionyl CoA from isoleucine, and succinyl CoA from valine (**Figure 7**).

Histidine The first step in histidine metabolism is deamination to urocanic acid. Subsequent metabolism of this compound can follow several different pathways, but the major pathway is the one that involves formiminoglutamic acid (FIGLU), which is demethylated by a terahydrofolic acid-dependent

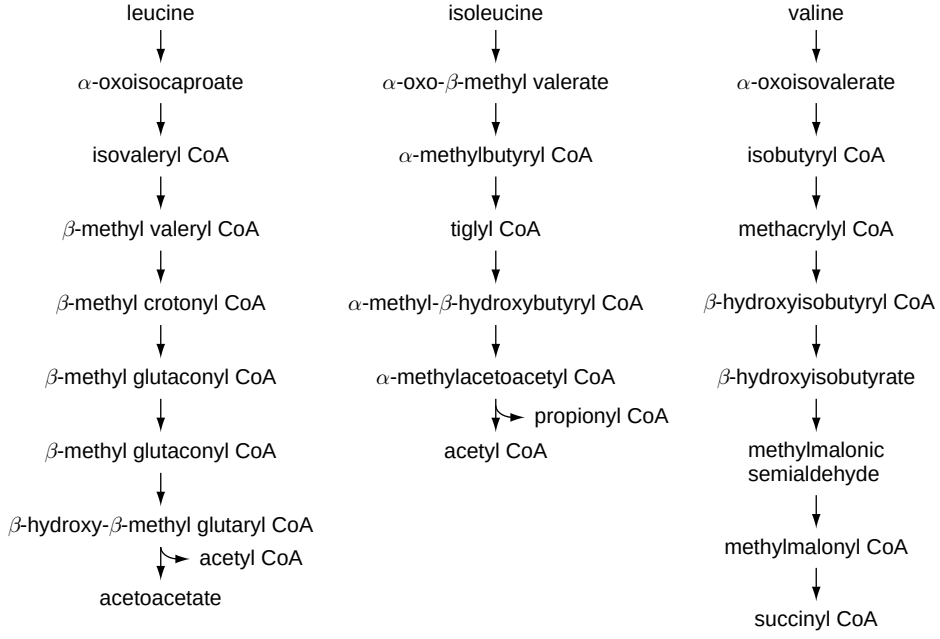


Figure 7 Metabolism of the branched-chain amino acids.

reaction to glutamic acid (**Figure 4**). This forms the basis of the FIGLU test for folate status. Another physiologically important pathway of histidine metabolism is decarboxylation to histamine, for which vitamin B_6 is a cofactor.

Phenylalanine and tyrosine Since mammalian enzymes cannot break open the benzene ring of phenylalanine, the only important pathway for catabolism of this amino acid is through hydroxylation to tyrosine. If the phenylalanine hydroxylase enzyme is lacking, as in phenylketonuria, a high concentration of phenylalanine accumulates and it is converted to phenylpyruvate, phenyllactate, and phenylacetate, which are toxic.

Tyrosine is transaminated to *p*-hydroxyphenylpyruvate, which is then decarboxylated to homogentisic acid. This is subsequently metabolized to acetoacetic acid and fumaric acid (**Figure 8**). Small amounts of tyrosine are hydroxylated to 3,4-dihydroxyphenylalanine (DOPA), which is then decarboxylated to the catecholamines dopamine, noradrenaline, and adrenaline. DOPA can also be converted to the pigment melanin. In the thyroid gland, protein-bound tyrosine is iodinated to the thyroid hormones tri-iodothyronine and thyroxine.

Tryptophan Tryptophan is oxidized by the hormone-sensitive enzyme tryptophan oxygenase to *N*-formyl kynurenine, which then follows a series of steps to yield amino-carboxymuconic semialdehyde. Most of this undergoes enzymic decarboxylation, leading ultimately to acetyl CoA. However, a small proportion undergoes nonenzymic cyclization to quinolic acid, which leads to the formation of NAD. This is why excess dietary tryptophan can meet the requirement for the vitamin niacin (**Figure 9**).

One of the steps in the catabolism of tryptophan is catalyzed by the vitamin B_6-dependent enzyme kynureninase. If vitamin B_6 status is inadequate

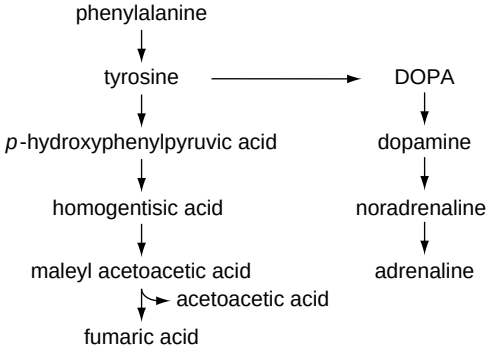


Figure 8 Metabolism of phenylalanine and tyrosine.

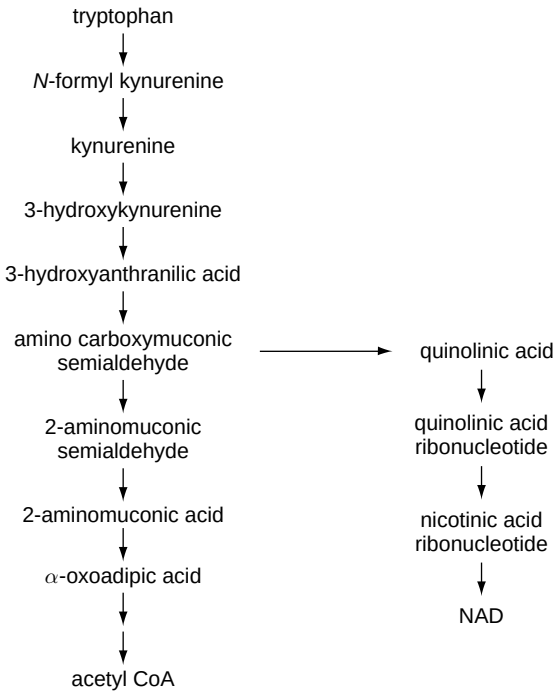


Figure 9 Metabolism of tryptophan.

and a large dose of tryptophan is administered, much of the tryptophan will be metabolized by an alternative pathway to kynurenic and xanthurenic acids, which will be excreted in the urine. This is the basis of the tryptophan load test for vitamin B_6 status.

A small amount of tryptophan undergoes hydroxylation to 5-hydroxytryptophan, which is then decarboxylated to the physiologically active amine 5-hydroxytryptamine (serotonin).

Alanine Alanine is metabolized by transamination to pyruvate.

Urea Cycle

From the previous discussion, it can be seen that the metabolism of most amino acids involves removal of the amino groups by transamination. 2-oxoglutarate is the main acceptor of these amino groups, being converted to glutamate, which can then be deaminated to release ammonium. However, ammonium is highly toxic and cannot be allowed to accumulate, so it is converted to urea, which is the form in which most of the nitrogen derived from protein is excreted from the body. Urea is formed in the liver by the cyclic series of reactions shown in **Figure 10**. It can be seen that only one of the nitrogen atoms in the urea molecule is actually derived from ammonium, via carbamyl phosphate. The other nitrogen atom comes from

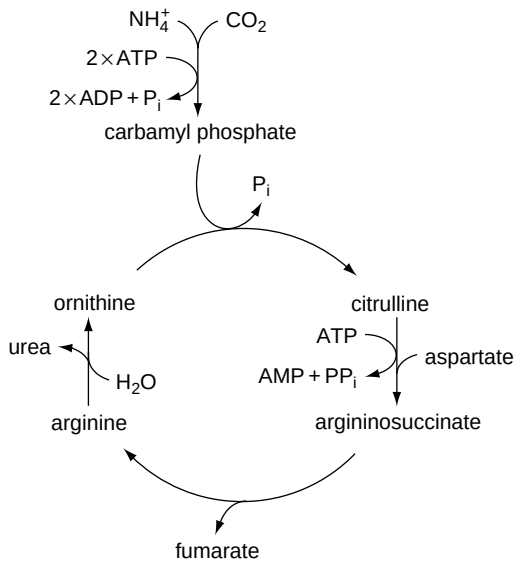


Figure 10 The urea cycle.

aspartic acid, which is formed by transamination of oxaloacetic acid.

The rate of production of urea by the liver is normally greater than the rate of urea excretion in the urine. This is because some of the urea diffuses into the colon, where it is hydrolyzed to ammonia by bacteria. The ammonia can be absorbed and taken up by the liver, where it can be reincorporated into amino acids, thereby augmenting the net supply of nonessential amino acids. The colonic bacteria can also use ammonia to synthesize essential amino acids, and there is evidence that some of these essential amino acids can also be absorbed and utilized by the human body. However, the rate at which this happens is clearly not sufficient to meet the body's requirements for essential amino acids.

Glucogenic and Ketogenic Amino Acids

The carbon skeletons that are left after amino acids have been transaminated are converted to common intermediates of the central metabolic pathways of the cell and so are ultimately used to provide energy. Clearly, for an adult in energy and nitrogen balance, energy will be derived from amino acids in the same proportion as protein is present in the diet, and for most human diets this is 10–15% of energy.

In certain circumstances, such as starvation, diabetes, or a high-fat diet, the body may need to synthesize glucose from amino acids rather than oxidize them directly. Experiments with diabetic dogs fed on single amino acids have shown that most of the amino acids can be converted to glucose and are therefore classified as glucogenic. However, leucine and lysine cannot be converted to glucose, and in these circumstances they give rise to acetoacetic acid, so they are classified as ketogenic. This classification can be related to the catabolic pathways outlined previously. The ketogenic amino acids are those that are metabolized only to acetyl CoA, whereas those that are metabolized to pyruvate or TCA cycle intermediates are glucogenic. Tryptophan, phenylalanine, tyrosine, isoleucine, methionine, and cysteine are both glucogenic and ketogenic.

Interorgan Exchange of Amino Acids

Amino Acid Pools

Free amino acids make up only approximately 2% of the total amino acid content of the body, with the rest being present as protein. The concentrations of free amino acids are regulated largely by modulation of their catabolic pathways, although in the case of nonessential amino acids there is also some regulation of the rate at which they are synthesized. There is evidence that the rates of protein synthesis and degradation are regulated by amino acid supply, and that this is another homeostatic mechanism acting to maintain free amino acid concentrations within safe limits. Protein degradation is suppressed following a meal containing protein, and the rate of protein synthesis may be increased so that there is net storage of amino acids as protein. Subsequently, in the postabsorptive state the changes in the rates of protein synthesis and breakdown are reversed so that there is net release of amino acids from protein. In nongrowing adults these changes balance out over a 24-h period so that there is no net change in body protein content. The amplitude of these diurnal changes in the rates of protein synthesis and degradation appears to vary in direct proportion to the amount of protein that an individual habitually consumes.

Free amino acids are found in all cells of the body and in extracellular fluid. They are transported between tissues in the plasma and into cells by a variety of transport mechanisms that are relatively specific for particular groups of amino acids. Amino acids are also present in red blood cells, but their role in interorgan transport appears to differ from that of plasma. For example, the plasma amino acid concentration increases as blood traverses the gastrointestinal tract after a meal, whereas the amino acid content of blood cells actually decreases.

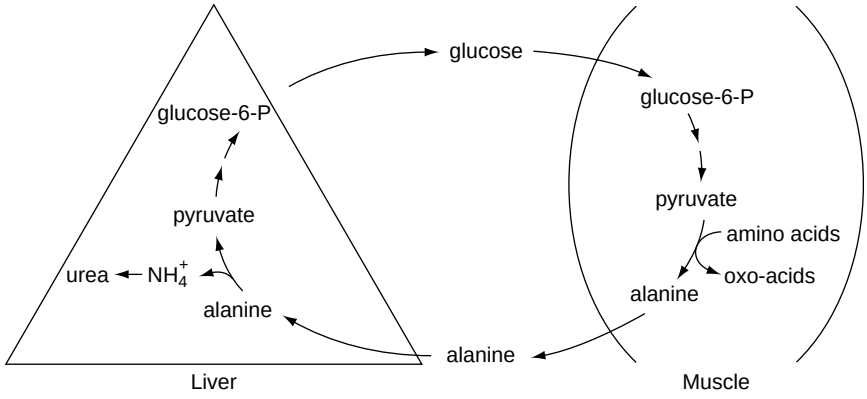


Figure 11 The glucose–alanine cycle.

Metabolism in Different Organs

The liver is responsible for most of the deamination of amino acids, except for the branched-chain amino acids, which are transaminated in muscle. Oxidation of amino acids is one of the main sources of energy for the liver. The liver is also the main site of gluconeogenesis, extracting large amounts of glutamine and alanine from the plasma for this purpose. The liver is the only site of urea synthesis.

Skeletal and cardiac muscle and adipose tissue are the main sites for transamination of the branched-chain amino acids, and the resulting ketoacids are transported to the liver for oxidation. However, in fasting and diabetes the capacity of muscle to oxidize branched-chain ketoacids increases markedly. In the postabsorptive state there is a net loss of amino acids from muscle, whereas in the fed state there is net uptake, reflecting the changes in net protein deposition and loss. However, at all times there is net output of alanine and glutamine from muscle, representing the disposal of the amino groups from the branched-chain amino acids. Muscle also takes up glucose, which is metabolized to supply the carbon skeletons for alanine and glutamine. Thus, there is a well-recognized glucose–alanine cycle between muscle and liver (**Figure 11**).

The kidney is a prime site of glutamine deamidation, producing ammonium to maintain acid–base balance and regulate the pH of the urine. Glutamine also serves as a substrate for gluconeogenesis in the kidney.

Glutamine is the major energy source for the small intestine, and at least part of the glutamine is derived from the lumen of the gut. Much of the glutamine is metabolized to pyruvate, which is then transaminated and exported to the liver as alanine. Some glutamine is also converted by the gut to citrulline, which then circulates to the kidney to be converted to arginine. Glutamine is also a major energy source for lymphocytes and monocytes when the immune system is activated.

See also: **Amino Acids**: Chemistry and Classification; Specific Functions. **Folic Acid**. **Inborn Errors of Metabolism**: Classification and Biochemical Aspects; Nutritional Management of Phenylketonuria. **Niacin**. **Protein**: Synthesis and Turnover; Digestion and Bioavailability. **Vitamin B_6**.

Further Reading

Bender DA (1985) *Amino Acid Metabolism*, 2nd edn. Chichester, UK: John Wiley and Sons.

Finkelstein JD (1990) Methionine metabolism in mammals. *Journal of Nutritional Biochemistry* **1**: 228–237.

McCully KS (1996) Homocysteine and vascular disease. *Nature Medicine* **2**: 386–389.

Millward DJ (2003) An adaptive metabolic demand model for protein and amino acid requirements. *British Journal of Nutrition* **90**: 249–260.

Munro HN and Allison JB (1964 and 1970) *Mammalian Protein Metabolism*, vols. 1–4. London: Academic Press.

Newsholme EA and Leech AR (1983) *Biochemistry for the Medical Sciences* Chichester, UK: John Wiley and Sons.

Waterlow JC and Stephen JML (1981) *Nitrogen Metabolism in Man* London: Applied Science.

Specific Functions

M C G van de Poll, Y C Luiking, C H C Dejong and P B Soeters, University Hospital Maastricht, Maastricht, The Netherlands

Introduction

Apart from being the building blocks of proteins, many amino acids are indispensable for certain vital functions or have specific functions of their own. They can function as neurotransmitters, as precursors for neurotransmitters and other important metabolites, including crucial oligo- and polypeptides, as a stimulus for hormonal release, and in inter-organ nitrogen transport and nitrogen excretion. Consequently, manipulation of free amino acid levels by dietary or topical supplementation may support and modulate these specific functions.

Amino Acid Flux, Concentration, and Function

Many amino acids have specific functions or support specific functions by serving as precursors or substrates for reactions in which vital end products are produced. The availability of amino acids to serve these purposes is determined by the rate at which they are released into the plasma and other pools in which these reactions take place, as well as by the rate of disappearance through excretion, protein synthesis, or conversion to other amino acids. The rate of this release, referred to as amino acid flux, is determined by the breakdown of (dietary) proteins or the conversion from other amino acids. Increased demand for one or more amino acids generally leads to an increased flux of the required amino acids across specific organs. Since it is the flux of an amino acid that determines its availability for metabolic processes, the flux is far more important for maintenance of specific functions than the plasma concentration. In fact it is striking that fluxes of some amino acids can double without significantly affecting plasma levels despite the fact that the plasma pool may be quantitatively negligible compared to the flux per hour. Plasma amino acid concentrations must therefore be subject to strong regulatory mechanisms. Increased demand and utilization of a specific amino acid may lead to decreased plasma and tissue concentrations, which may act as a signal to increase flux. Thus, a low plasma concentration in itself does not necessarily imply that the supply of the amino acid in question is inadequate, but it may indicate that there is increased turnover of the amino acid and that deficiencies may result when dietary or endogenous supply is inadequate. Other factors determining amino acid concentration are induction of enzymes and stimulation or blocking of specific amino acid transporters affecting the exchange and distribution of amino acids between different compartments. The regulation of plasma and tissue concentrations of specific amino acids may also be executed by the fact that release of the amino acid by an organ (e.g., muscle) and the uptake of that amino acid by another organ (e.g., liver) are subject to a highly integrated network including the action of cytokines and other hormones.

By repeated conversion of one amino acid to another, metabolic pathways arise by which (part of) the carbon backbone of a single amino acid can pass through a succession of different amino acids. Because of this interconvertibility, groups of amino acids rather than one specific amino acid contribute to specific functions. Apart from the rate at which these amino acids interconvert, the rate at which they gain access to the tissue where the specific end products exert their functions is also an important determinant of deficiencies of amino acids.

Amino acid Deficiencies and Supplementation

In many diseases and during undernutrition diminished turnover of amino acids can occur. These deficiencies may concern specific amino acids in certain diseases or a more generalized amino acid deficiency. The resulting functional deficits can contribute to the symptoms, severity, and progress of the disease. In some instances these deficits can be counteracted by simple supplementation of the deficient amino acids. Amino acid supplementation is also applied to enhance turnover and improve amino acid function in nondeficient patients. However, amino acid supplementation in nondeficient states does not necessarily lead to an increased function since the organism utilizes what is programed by regulating hormones and cytokines. An additional factor to consider is that metabolic processes can be subject to counter-regulatory feedback mechanisms. Some important metabolic processes served by a specific amino acid require only a marginal part of the total flux of that amino acid. The question may be raised whether true shortages may arise in such pathways, and supplemented amino acids may be

Table 1 Specific functions of amino acids and their intermediate products

Amino acid	*Intermediate products*	*Function*	*Supplementation efficacy*
Alanine	Pyruvate	Gluconeogenesis Nitrogen transport	Data too limited
Arginine	Nitric oxide	Vasodilation Immunomodulation Neurotransmission	Positive effects of arginine-containing immunonutrition on morbidity in surgical and trauma patients suggested; further research required
	Urea	Ammonia detoxification	
	Creatine	Muscle constituent/fuel	
	Agmatine	Cell signaling Ornithine precursor	
Citrulline	Arginine production		
Ornithine	Polyamines	Cell differentiation Proline precursor	Improves healing of burns (ornithine α-ketoglutarate)
Proline	Hydroxyproline	Hepatocyte DNA, protein synthesis Collagen synthesis	
Asparagine		Aspartic acid precursor	(Asparaginase-induced asparagine depletion is therapeutic in leukemia)
Aspartic acid	Oxaloacetate, fumarate	Gluconeogenesis	
Methionine		Cysteine precursor	
	Creatine	(*see* arginine)	
Cysteine (Cystine)	Glutathione	Antioxidant	Improves antioxidant status in undernutrition, inflammatory diseases
	Taurine	Bile acid conjugation, neuronal cell development, regulation of membrane potential, calcium transport, antioxidant	Reduces contrast-induced nephropathy in renal failure Mucolysis, symptom reduction in COPD Hepatoprotective in acetaminophen intoxication
Glutamic acid	Glutamine	Ammonia disposal	
	α-ketoglutarate	Gluconeogenesis	
	Glutathione	Antioxidant	
	γ-aminobutyric acid	Inhibition CNS Excitation CNS (NMDA receptor)	
Glutamine	Ammonia	Inter-organ nitrogen transport Renal HCO_3^- production	Reduces infectious morbidity in trauma patients, burn patients, and surgical patients
	Purines, pyrimidines	RNA synthesis, DNA synthesis Glutamic acid precursor	
Glycine		Inhibition CNS (glycine receptor) Excitation CNS (NMDA receptor)	Adjuvant to antipsychotics, probably reduces negative symptoms of schizophrenia
	Glutathione	Antioxidant	
	Creatine	(*see* arginine)	
		Serine precursor	
Serine	D-serine	Excitation CNS (NMDA receptor) Glycine precursor Cysteine precursor	Adjuvant to antipsychotics, probably reduces negative symptoms of schizophrenia
Threonine	Glycine Serine	Brain development	
Histidine	Histamine	Immunomodulation Gastric acid secretion	
Lysine	Carnitine	Mitochondrial oxidation of long-chain fatty acids	Reduces chronic stress-induced anxiety
	Glutamate		
Branched chain amino acids			
Isoleucine	α-keto-β-methylvaleric acid		Upper gastrointestinal hemorrhage
Leucine	α-ketoisocaproic acid	Important in regulation of energy and protein metabolism Substrate for glutamine synthesis	Improve protein malnutrition and restore amino acid and neurotransmitter balance in hepatic failure and hepatic encephalopathy (supplemented BCAA)
Valine	α-ketoisovaleric acid		

Continued

Table 1 Continued

Amino acid	*Intermediate products*	*Function*	*Supplementation efficacy*
Aromatic amino acids			
Phenylalanine		Tyrosine precursor	
Tyrosine	L-dopa	Dopamine synthesis	Possible slight improvement of cognitive functions after physical or mental exhaustion. Metabolites are powerful pharmacotherapeutic drugs
	Dopamine	Movement, affect on pleasure, motivation	
	Noradrenaline, adrenaline	Activation of sympathetic nervous system (fight-or-flight response)	
	Tri-iodothyronine, thyroxine	Regulation of basal metabolic rate	
Tryptophan	Kynureninic acid	CNS inhibition	No scientific evidence for beneficial effects of supplementation
	Quinolinic acid	CNS excitation	
	Serotonin	Mood regulation Sleep regulation Intestinal motility	
	Melatonin	Regulation of circadian rhythms	

Different fonts indicate: nonessential amino acids, **essential amino acids**, and ***conditionally essential amino acids***.

disposed of in pathways other than those serving to improve a specific function.

Assessment of Amino Acid Function

The effectiveness of amino acid supplementation, particularly with respect to clinical effectiveness, can be assessed at four levels. First, the intervention should lead to an increased local or systemic concentration of the amino acid in question. The conversion of amino acids in (interorgan) metabolic pathways can lead to an increase in the levels of amino acids other than the one supplemented, increasing or mediating its functionality. Alternatively, supplementation of one amino acid may decrease the uptake of other amino acids because they compete for a common transporter. Second, the metabolic process for which the supplemented amino acid forms the substrate should be stimulated or upregulated by this increased amino acid availability. Third, this enhanced metabolic activity must lead to physiological changes. Fourth, these changes must be clinically effective in a desirable fashion.

Alanine

Alanine and glutamine are the principal amino acid substrates for hepatic gluconeogenesis and ureagenesis. Alanine is produced in peripheral tissues in transamination processes with glutamate, branched chain amino acids, and other amino acids; following its release in the systemic circulation, alanine is predominantly taken up by the liver and to a lesser extent by the kidney. Here, alanine can be deaminated to yield pyruvate and an amino group, which can be used for transamination processes, ureagenesis, or can be excreted in urine. Thus, the alanine released from peripheral tissues may be converted to glucose in the liver or kidney and eventually become a substrate for peripheral (mainly muscular) glycolysis. This so-called glucose-alanine cycle may be especially relevant during metabolic stress and critical illness when the endogenous alanine release from peripheral tissues is increased. Simultaneously, alanine serves as a nitrogen carrier in this manner. Alanine is often used as the second amino acid in glutamine dipeptides that are applied to increase solubility and stability of glutamine in nutritional solutions.

Supplementation

No clinical benefits have been ascribed to supplementation with alanine, although it has never been considered whether the beneficial effects of the dipeptide alanine-glutamine, which are generally ascribed to glutamine, may also be due to alanine. In this context, it should be realized, however, that alanine itself constitutes the strongest drive for hepatic ureagenesis (leading to breakdown of alanine).

Arginine, Citrulline, Ornithine, and Proline (Figure 1)

Arginine is a nitrogen-rich amino acid because it contains three nitrogen atoms and is the precursor for nitric oxide (NO). The conversion to NO is catalyzed by the enzyme nitric oxide synthase (NOS), and results in coproduction of the amino acid citrulline. Depending on its site of release, NO exerts several functions including stimulation of the pituitary gland, vasodilation, neurotransmission, and immune modulation. Arginine is also a precursor for urea synthesis in the urea cycle, which has an important function in the detoxification of ammonia and excretion of waste nitrogen from the body. A full urea cycle is only present in the liver, but the arginase enzyme that converts arginine to urea and ornithine is to a limited extent also found in other tissues and cells, such as brain, kidney, small intestine, and red blood cells. Ornithine is utilized for the formation of proline, polyamines (putrescine, spermine, and spermidine), glutamic acid, and glutamine. Arginine is involved in collagen formation, tissue repair, and wound healing via proline, which is hydroxylated to form hydroxyproline. This role in wound healing may additionally be mediated by stimulation of collagen synthesis by NO, although this claim is still under investigation. It is currently thought that arginine availability is regulated by the balance between NOS and arginase enzyme activity, which subsequently determines substrate availability for NO and ornithine production. Proline also stimulates hepatocyte DNA and protein synthesis. Polyamines are potent inducers of cell differentiation.

In addition to synthesis of NO, urea, and ornithine, arginine is used for synthesis of creatine, which is an important constituent of skeletal muscle and neurons and acts as an energy source for these tissues. Furthermore, arginine may be catabolyzed to agmatine, which acts as a cell-signaling molecule. Arginine not only acts as an intermediate in the synthesis of functional products, but also is a potent stimulus for the release of several hormones, such as insulin, glucagon, somatostatin, and growth hormone, illustrating its pharmacological characteristics.

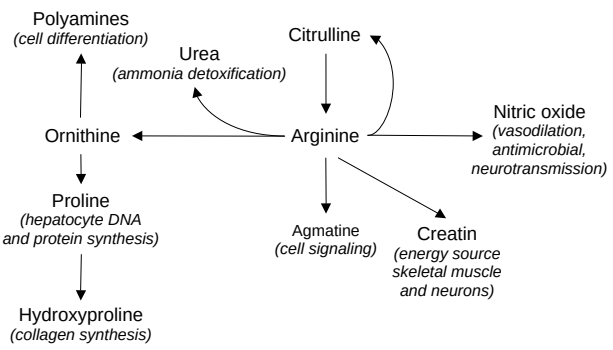


Figure 1 Specific functions of arginine metabolism.

Arginine can be synthesized by the body from citrulline. However, since virtually all arginine produced in the liver is trapped within the urea cycle, the kidney is the only arginine-synthesizing organ that significantly contributes to the total body pool of free arginine. Diminished renal arginine synthesis has been found in patients with renal failure and in highly catabolic conditions, like sepsis, burn injury, or trauma (which may be related to concomitant renal failure). In these situations arginine may be considered a conditionally essential amino acid and it has been suggested that arginine supplementation can become useful in these situations.

Citrulline is formed from glutamine, glutamic acid, and proline in the intestine. Plasma citrulline concentration reflects intestinal metabolic function and has recently been introduced as a potential marker for (reduced) enterocyte mass.

Supplementation

Based on its pluripotent functions, arginine has been widely used in supplemental nutrition for surgical patients, patients with burns, and patients with sepsis and cancer in order to modify the inflammatory response, to enhance organ perfusion, and to stimulate wound healing. However, the benefits of arginine supplementation in these conditions are not uniformly proven and accepted. Moreover, arginine is never given alone but is always provided in a mixture of amino acids and other nutrients. The use of NO donors that have vasodilatory actions is an established therapeutic modality in coronary artery disease and for erectile dysfunction. Given this fact it remains worthwhile to clarify the need for arginine supplementation as the natural substrate for NO synthesis in other conditions.

Using citrulline as an arginine-delivering substrate has been suggested, but has not been applied clinically. Ornithine is supplied as part of the ornithine-α-ketoglutarate molecule (see glutamine). Creatine is widely used by professional and recreational athletes as a nutritional supplement, although the ascribed performance-enhancing effects have not been proven.

Asparagine and Aspartic Acid

Asparagine can be converted by asparaginase to ammonia and aspartic acid, which is the precursor

of the citrate cycle intermediates oxaloacetate and fumarate; this reaction is reversible. In fasting humans asparagine and aspartic acid are utilized as precursors for *de novo* synthesis of glutamine and alanine in muscle.

Supplementation

The claim that asparagine or aspartic acid supplementation improves endurance has not been confirmed in human studies. Asparaginase, which degrades asparagine, is widely used in the treatment of pediatric leukemia since the resulting asparagine depletion leads to apoptosis of leukemic cells.

Cysteine, Cystine, Methionine, and Taurine (Figure 2)

Methionine is converted to cysteine and its dipeptide cystine. In addition methionine is a precursor for creatine (see arginine). The potential for formation of disulfide bonds between its thiol (-SH) groups makes protein-bound cysteine important in the folding and structural assembly of proteins. Reduced cysteine thiol groups are found in protein (albumin), free cysteine, and in the principal intracellular antioxidant tripeptide glutathione (see glycine, glutamic acid) for which free cysteine is the synthesis rate-limiting constituent. Through the formation of disulfides (e.g., cystine, cysteinyl-glutathione, glutathione disulfide, mercaptalbumin) thiol-containing molecules can scavenge oxygen-derived free radicals. The ratio between oxidized and reduced thiol groups reflects the cellular redox state. Owing to its small pool size cysteine deficiencies rapidly occur during malnutrition.

Cysteine is also the precursor for taurine, which is abundant in all mammalian cells, particularly in neuronal cells and lymphocytes, but is not a true amino acid and is not incorporated in proteins. Taurine is involved in the conjugation of bile acids and may act as an antioxidant. Moreover, taurine is an osmolyte by virtue of the fact that through its transporter its intracellular concentrations are between 50 and100-fold higher than in the extracellular compartment. This gradient contributes to the maintenance of the cellular hydration state. Similarly, it has been proposed that taurine is involved in stabilization of cell membrane potential and regulation of Ca^{2+} transport through several calcium-ion channels. Based upon these characteristics it has been suggested that taurine is involved in the control of cardiac muscle cell contraction, which has led to the addition of taurine to commercially available energy drinks. Its high level in lymphocytes suggests an important role in immunological resistance to infections. Taurine plays an important part in the development and maintenance of neuronal and especially retinal cells.

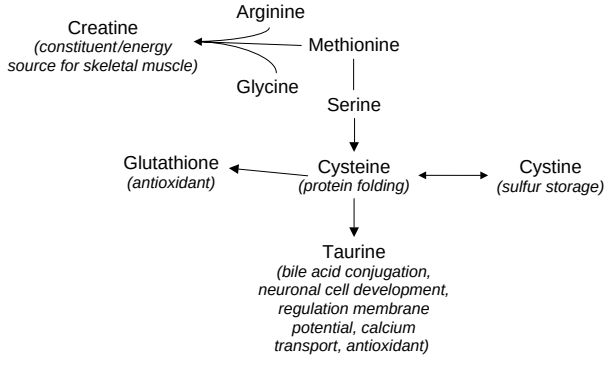


Figure 2 Specific functions of sulfur-containing amino acids.

Supplementation

Although methionine is the only sulfur-containing essential amino acid, it has not been considered as part of supplementation regimes. Since cysteine easily oxidizes to cystine, which has a poor solubility, it is generally supplemented in the form of *n*-acetylcysteine (NAC). Both directly and indirectly, as a precursor for glutathione, NAC has attracted attention as a potentially protective agent against oxidative injury in numerous conditions including endurance exercise, ischemia reperfusion injury, adult respiratory distress syndrome (ARDS), and cystic fibrosis. In addition, NAC has mucolytic properties in chronic obstructive pulmonary disease (COPD) patients by reducing disulfide bonds of polymers in mucus, blocking their reactivity. Currently, only robust evidence exists for the usefulness of NAC supplementation in the protection against nephropathy, induced by administration of iodine-containing contrast agents for radiological imaging in patients with chronic renal failure, in the reduction of the number of exacerbations and disability in COPD patients, and in the treatment of liver injury induced by acetaminophen intoxication. On the other hand it has been suggested that glutathione depletion by buthionine sulfoximine administration potentiates the effect of radiotherapy by increasing the susceptibility of tumor cells to radiation-induced oxidative injury.

In a few studies it has been demonstrated that taurine supplementation improves retinal development in premature babies receiving parenteral nutrition. Human data on the efficacy of taurine supplementation in so-called energy drinks are very limited. In the absence of taurine supplementation in

children taurine concentrations drop, suggesting its conditional indispensability also in the postneonatal period. This has led to the addition of taurine to standard feeding formulas for infants and growing children.

Glutamine, Glutamic acid, and Ornithine α-Ketoglutarate (Figure 3)

Glutamine is the most abundant amino acid in plasma and in tissue. In glutamine-consuming cells it is readily converted by the enzyme glutaminase to form ammonia and glutamic acid, which is the primary intermediate in almost all routes of glutamine degradation. In the presence of ammonia this process can occur in reverse, catalyzed by the enzyme glutamine synthetase. In contrast to glutamic acid, glutamine can easily pass through the cellular membrane, thus exporting waste nitrogen out of the cell and serving as an inter-organ nitrogen carrier. In the kidney glutamine donates NH_3, which is the acceptor for protons released from carbonic acid, to form NH_4^+ and thus facilitates the formation of HCO_3^-, which is essential in plasma pH regulation.

Following conversion to glutamic acid and subsequently α-ketoglutarate, glutamine may supplement intermediates of the citrate cycle. In this manner glutamine serves as the preferred fuel for rapidly dividing cells of, for example, the immune system cells and intestinal mucosa. In the brain glutamic acid is the most abundant excitatory neurotransmitter and the precursor for gamma-aminobutyric acid, which is an important inhibitory neurotransmitter. Glutamine is a direct precursor for purine and pyrimidine and therefore is involved in RNA and DNA synthesis and cell proliferation. In addition it is a constituent of the tripeptide glutathione, which is the principal intracellular antioxidant in eukaryotes (see also sections on cysteine and glycine).

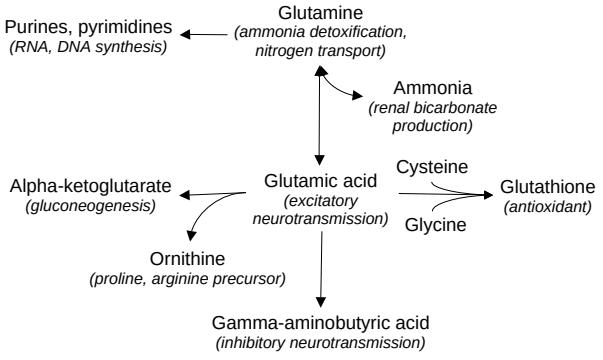


Figure 3 Specific functions of glutamine and glutamine degradation products.

Supplementation

Of all the compounds discussed above glutamine is the most extensively applied in clinical and experimental amino acid supplementation, often in the form of the more soluble and stable dipeptides alanyl- and glycyl-glutamine. Glutamic acid and α-ketoglutarate are less ideally suited for use in feeding formulas because of poor inward transport of glutamic acid and poor solubility and stability of α-ketoglutarate. Moreover, glutamic acid has been related to the 'Chinese restaurant syndrome,' characterized by light-headiness and nausea after consumption of Chinese food containing glutamic acid for flavor improvement. However, scientific evidence is weak. Numerous experimental and clinical studies have suggested that glutamine supplementation has positive effects on immune function, intestinal mucosal integrity, nitrogen balance, and glutathione concentration in a wide variety of conditions. Nevertheless, the true benefit of glutamine supplementation is difficult to quantify in clinical practice. Its benefit has especially been claimed in the critically ill and surgical patients in whom clinical outcome is multifactorial. Recent meta-analyses support the view that glutamine supplementation is safe and may reduce infectious morbidity and hospital stay in surgical patients. A positive effect of glutamine supplementation on morbidity and mortality in critical illness, trauma patients, and burn patients has been demonstrated in a few well-designed clinical trials. However, due to the paucity of such trials reliable meta-analyses are not possible in these latter patient categories. It has been demonstrated in some small clinical series that supplementation with ornithine α-ketoglutarate may improve wound healing in burn patients, benefiting from the combined actions of both α-ketoglutarate and ornithine (see sections on arginine and ornithine).

Glycine, Serine, and Threonine

Threonine is an essential amino acid, which can be converted to glycine in the liver and subsequently to serine. Glycine is a constituent of glutathione (see also sections on cysteine and glutamic acid) and is a versatile neurotransmitter in the central nervous system. Through the glycine receptor it has a direct inhibitory neurotransmitter function but it is also a ligand for the glycine site at the *N*-methyl-D-aspartate (NMDA) glutamic acid receptor. Activation of this glycine site is needed for NMDA activation, which makes glycine a mediator in the excitatory neurotransmitter effects of glutamic acid. Besides a role in the central nervous system, glycine is also

thought to possess anti-inflammatory properties, but to date these properties have only been demonstrated in the test tube. Furthermore, glycine can react with arginine and methionine to form creatine (see section on arginine). Finally, glycine, like taurine, is a conjugate for bile acids.

Glycine is convertible to serine in a reversible reaction, which can be converted to its stereoisomeric form D-serine; this is also a ligand for the glycine site at the NMDA receptor. Furthermore, serine is an intermediate in the pathway from methionine to cysteine and a precursor for pyrimidines and purines and as such is involved in cell proliferation. It is also a precursor for gluconeogenesis, albeit of lesser importance than glutamine and alanine.

Supplementation

Based upon their excitatory effects on the central nervous system both glycine and D-serine have been implicated in the treatment of schizophrenia. As adjuvant therapy to standard psychopharmacological treatment they may reduce the negative symptoms of the disease.

High doses of threonine in adults have been used as tentative therapy for spastic syndromes, a therapy that probably acts through increased glycine formation. A negative effect of excessive threonine, which is abundant in bovine infant formula nutrition, has been considered in experimental studies on brain development, and it has been suggested that this happens through its conversion to glycine and serine, or through competition of amino acid transport across the blood–brain barrier.

Histidine

Histidine is the precursor for histamine, which is important for the immune system by mediating growth and functionality of immune cells. Excessive release of histamine from mast cells induces the clinical signs of allergy (dilation of capillaries and larger blood vessels, increased capillary permeability and swelling, itching, and anaphylactic shock). These phenomena are effected via the H_1 receptor, which is found in smooth muscle cells of the vascular wall and bronchi, among others. Furthermore, histamine acts as a neurotransmitter and mediates gastric acid production. The latter occurs via the H_2 receptor found in gastric mucosa. There is no literature available on the potential relationship between histidine availability and histamine production and action.

Supplementation

H_1 receptor antagonists are applied in the treatment of allergy and H_2 receptor antagonists have been shown to be very effective in the inhibition of gastric acid secretion and have greatly improved the treatment of individuals with peptic ulcer disease and acid reflux esophagitis. Histamine is present in abundance in many dietary sources; no beneficial effects of supplementation of either histidine or histamine are known.

Branched Chain Amino Acids (Isoleucine, Leucine, Valine)

Branched chain amino acids (BCAAs) are essential amino acids, which together compose approximately a third of the daily amino acid requirement in humans. BCAAs, and especially leucine, play an important role in the regulation of energy and protein metabolism. BCAAs are primarily oxidized in skeletal muscle and not in the liver. BCAAs donate their amino groups to furnish glutamic acid in muscle in transamination reactions yielding the α-ketoacids α-ketoisocaproic acid, α-keto-β-methylvaleric acid, and α-ketoisovaleric acid. These transamination products of BCAAs can enter the citrate cycle and contribute to ATP production by aerobic substrate oxidation, which is important during the change from rest to exercise. After consumption of protein-containing meals, a large part of the BCAA passes through the liver and is taken up by muscle where it primarily contributes to protein synthesis and the synthesis of glutamine, which accounts for about 70% of the amino acid release from muscle. The importance of the essential branched chain amino acids for protein synthesis is strikingly exemplified by the negative nitrogen balance and catabolism that follows upper gastrointestinal bleeding caused by ingestion of large amounts of hemoglobin (which lacks isoleucine). Leucine has been suggested to regulate the turnover of protein in muscle cells by inhibiting protein degradation and enhancing protein synthesis. This has led to a worldwide interest in the possible use of BCAAs in general, and leucine in particular, for metabolic support.

In liver failure the plasma concentrations of the aromatic amino acids (AAAs) tyrosine, phenylalanine, and tryptophan increase, probably because they are predominantly broken down in the liver, whereas the plasma levels of BCAAs decrease while they are degraded in excess in muscle as a consequence of hepatic failure-induced catabolism. As AAAs and BCAAs are all neutral amino acids and share a common transporter across the blood–brain barrier (system L carrier), changes in their plasma ratio are reflected in the brain, subsequently disrupting the neurotransmitter profile of

the catecholamines and indoleamines (see sections on tyrosine and tryptophan). It has been hypothesized that this disturbance contributes to the multifactorial pathogenesis of hepatic encephalopathy. In line with this hypothesis it has been suggested that normalization of the amino acid pattern by supplementing extra BCAAs counteracts hepatic encephalopathy.

Supplementation

Specialized formulas that are widely used for hepatic failure and hepatic encephalopathy are based on a high content of BCAAs to improve protein malnutrition and restore the amino acid and neurotransmitter balance. Although BCAA-enriched formulas have been proven to improve neurological status in comatose liver patients it is not certain that this is achieved by the addition of BCAAs specifically, because of a lack of adequate control groups.

Since BCAAs compete with tryptophan for uptake by the brain, they have (in line with the ascribed benefits in hepatic encephalopathy) been applied as competitive antagonists for tryptophan transport, reducing tryptophan-induced cognitive impairment (see also section on tryptophan).

Isoleucine, which is absent in the hemoglobin molecule, can be supplemented to patients with upper gastrointestinal bleeding to restore the balance of amino acids that are taken up by the splanchnic organs. This has been demonstrated to improve mainly protein synthesis in liver and muscle in small observational studies. Prospective randomized clinical trials are, however, still lacking.

Lysine

Lysine is an essential amino acid that is mainly provided by meat products and is therefore limited in diets where wheat is the primary protein source. Lysine is also the first rate-limiting amino acid in milk-fed newborns for growth and protein synthesis. Lysine is catabolized to glutamate and acetyl-CoA and is also the precursor for the synthesis of carnitine, which is needed for mitochondrial oxidation of long-chain fatty acids.

Supplementation

Lysine supplementation in patients with renal failure is contraindicated, as the amino acid shows some degree of nephrotoxicity.

Phenylalanine and Tyrosine

Phenylalanine is hydroxylated to tyrosine by the enzyme phenylalanine hydroxylase. The inborn disease phenylketonuria is characterized by a deficiency of this enzyme.

Tyrosine is the precursor for dihydroxyphenylalanine (dopa), which can successively be converted to the catecholamines dopamine, noradrenaline (norepinephrine) and adrenaline (epinephrine). Although only a small proportion of tyrosine is used in this pathway, this metabolic route is extremely relevant. Dopamine is an important neurotransmitter in different parts of the brain and is involved in movement and affects pleasure and motivation. Disruption of dopamine neurons in the basal ganglia is the cause of Parkinson's disease. Noradrenaline and ardrenaline are the most important neurotransmitters in the sympathetic nervous system. The sympathetic nervous system becomes activated during different forms of emotional and physical arousal, and results in the induction of phenomena such as increased blood pressure and heart rate, increased alertness, and decreased intestinal motility (fight-or-flight response). Besides acting as a precursor for catecholamines, tyrosine can be iodinated and as such is the precursor for the thyroid hormones triiodothyronine and thyroxine. These hormones are important regulators of general whole body rate of metabolic activity.

Supplementation

The processes described in the paragraph above quantitatively contribute only marginally to total tyrosine turnover and the limited data on tyrosine supplementation in phenylketonuria suggest that tyrosine deficiency is not causal in the development of cognitive dysfunction in the disease. In two studies tyrosine supplementation has been found to modestly increase mental status and cognitive performance following exhausting efforts such as prolonged wakefulness and intensive military training. In contrast, tyrosine derivatives (L-dopa, noradrenaline, adrenaline) have strong pharmacological properties. L-dopa is the direct precursor of dopamine synthesis and has been found to have strong beneficial effects in Parkinson's disease. The fact that administration of tyrosine as the physiological precursor of catecholamines has no or minor effects on catecholamine-induced sympathic activity, whereas the effects of the catecholamines or more direct precursors is very strong, suggests that tyrosine hydroxylation to L-dopa is not limited by substrate availability.

Tryptophan

Functional end products of the essential amino acid tryptophan arise mainly through two distinctive pathways. The major pathway is degradation of tryptophan by oxidation, which fuels the kynurenine pathway (See 00011). The second and quantitatively minor pathway is hydroxylation of tryptophan and its subsequent decarboxylation to the indoleamine 5-hydroxytryptamine (serotonin) and subsequently melatonin. The metabolites of the kynurenine pathway, indicated as kynurenines, include quinolic acid and kynurenic acid. Quinolinic acid is an agonist of the NMDA receptor (see also section on glutamic acid), while kynurenic acid is a nonselective NMDA-receptor antagonist with a high affinity for the glycine site of the NMDA receptor (see also section on glycine), and as such is a blocker of amino acid-modulated excitation of the central nervous system. Imbalance between kynurenic acid and quinolinic acid can lead to excitotoxic neuronal cell death and is believed to play a role in the development of several neurological diseases such as Huntington's chorea and epilepsy. In addition, an immunomodulatory role is suggested for several metabolites of the kynurenine pathway.

Serotonin is synthesized in the central nervous system and is involved in the regulation of mood and sleep. In addition it is found in high quantities in neurons in the gastrointestinal tract where it is involved in regulation of gut motility. Tryptophan competes with BCAAs for transport across the blood–brain barrier and the ratio between tryptophan and BCAAs therefore determines the uptake of both (groups of) amino acids by the brain (see section on BCAAs). Since albumin has a strong tryptophan-binding capacity, the plasma albumin concentration is inversely related to the plasma concentration of free tryptophan and as such influences the BCAA to tryptophan ratio and hence the brain uptake of both BCAAs and tryptophan. It has been suggested that increased plasma AAAs (tyrosine, phenylalanine, and tryptophan) levels in patients with liver failure are caused by the inability of the liver to degrade these amino acids. The resulting change in the ratio between AAA and BCAA plasma levels has been implied in the pathogenesis of hepatic encephalopathy since this may cause marked disturbances in transport of both AAAs and BCAAs across the blood–brain barrier, leading to disturbed release of indoleamines and catecholamines in the brain (see also section on BCAAs). High tryptophan concentrations have been associated with chronic fatigue disorders and hepatic encephalopathy while low tryptophan plasma concentrations have been implicated in the etiology of mood disorders, cognitive impairment, and functional bowel disorders. Melatonin, which is produced in the degradation pathway of serotonin during the dark period of the light-dark cycle, is an important mediator of circadian rhythms.

Supplementation

Inhibition of serotonin reuptake from the neuronal synapse and the subsequent increase in its functionality is one of the mainstays of the pharmacological treatment of depression. Like many amino acids, tryptophan is commercially available as a nutritional supplement or as a so-called smart drug, claiming to reduce symptoms of depression, anxiety, obsessive-compulsive disorders, insomnia, fibromyalgia, alcohol withdrawal, and migraine. However, no convincing clinical data are available to support these claims. In contrast tryptophan depletion induced by ingestion of a tryptophan-deficient amino acid mixture, is widely used in experimental psychiatry to study the biological background of various psychiatric disorders.

See also: **Amino Acids**: Chemistry and Classification; Metabolism. **Brain and Nervous System**. **Carbohydrates**: Regulation of Metabolism. **Cytokines**. **Electrolytes**: Acid-Base Balance; Water–Electrolyte Balance. **Glucose**: Metabolism and Maintenance of Blood Glucose Level. **Inborn Errors of Metabolism**: Classification and Biochemical Aspects; Nutritional Management of Phenylketonuria. **Protein**: Synthesis and Turnover; Requirements and Role in Diet; Quality and Sources. **Stomach**: Structure and Function.

Further Reading

Cynober LA (ed.) (2004) *Metabolic and Therapeutic Aspects of Amino Acids in Clinical Nutrition*, 2nd edn. Boca Raton: CRC Press.

Fürst P and Young V (2000) *Proteins, Peptides and Amino Acids in Enteral Nutrition.* Nestlé Nutrition Workshop Series Clinical & Performance Program, vol. 3. Vevey: Nestec Ltd and Basel: Karger.

Guyton AC and Hall JE (1996) *Textbook of Medical Physiology*, 9th edn. Philadelphia: W.B. Saunders.

Labadarios D and Pichard C (2002) *Clinical Nutrition: early Intention.* Nestlé Nutrition Workshop Series Clinical & Performance Program, vol. 7. Vevey: Nestec Ltd and Basel: Karger.

Newsholme P, Procopio J, Lima MMR, Ptihon-Curi TC, and Curi R (2003) Glutamine and glutamate – their central role in cell metabolism and function. *Cell Biochemistry and Function* **21**: 1–9.

Wu G and Morris SM (1998) Arginine metabolism: nitric oxide and beyond. *Biochemical Journal* **336**: 1–17.

Young V, Bier DM, Cynober L, Hayashi Y, and Kadowaki M (eds.) (2003) The third Workshop on the Assessment of adequate Intake of Dietary Amino Acids. *J Nutr Supplement* **134**: 1553S–1672S.

ANEMIA

Contents
Iron-Deficiency Anemia
Megaloblastic Anemia

Iron-Deficiency Anemia

K J Schulze and M L Dreyfuss, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Anemia is defined by abnormally low circulating hemoglobin concentrations. A variety of etiologies exist for anemia, including dietary deficiencies of folate or vitamin B_{12} (pernicious or macrocytic anemia), infections and inflammatory states (anemia of chronic disease), and conditions that result in insufficient production of red blood cells (aplastic anemia) or excessive destruction of red blood cells (hemolytic anemia). However, worldwide, the most prevalent form of anemia is that of iron deficiency, which causes anemia characterized by hypochromic and normo- or microcytic red blood cells. Iron deficiency anemia remains a health problem in both the developed and the developing world. This article discusses the metabolism of iron; the assessment of iron deficiency; iron requirements across the life span; and the consequences, prevention, and treatment of iron deficiency and iron deficiency anemia.

Iron Metabolism

The adult body contains 2.5–5 g of iron, approximately two-thirds of which is present in hemoglobin. Other essential iron-containing systems include muscle myoglobin (3%) and a variety of iron-containing enzymes (5–15%), including cytochromes. In addition to the role of iron in oxygen transfer via hemoglobin and myoglobin, iron is involved in energy metabolism and also affects neural myelination and neurotransmitter metabolism. Iron stores vary considerably but may represent up to 30% of body iron, and iron that circulates with transferrin represents less than 1% of body iron. Men have a higher concentration of iron per kilogram body weight than women because they have larger erythrocyte mass and iron stores.

More than 90% of body iron is conserved through the recycling of iron through the reticuloendothelial system (**Figure 1**). Iron is transported through the body by the protein transferrin, which carries up to two iron atoms. The distribution of iron to body tissues is mediated by transferrin receptors (TfRs), which are upregulated in the face of increased tissue demand for iron. The transferrin/TfR complex is internalized via cell invagination, iron is released into the cell cytosol, and transferrin is recycled back to the cell surface.

In hematopoietic cells, iron is used to produce hemoglobin through its combination with zinc protoporphyrin to form heme. Therefore, protoporphyrin accumulates relative to hemoglobin in red blood cells during iron deficiency. Mature red blood cells circulate in the body for approximately 120 days before being destroyed. Macrophage cells of the liver and spleen phagocytize senescent red blood cells and the iron released in this process is recycled back to the circulation or, when iron is readily available, incorporated with ferritin or hemosiderin for storage. A typical ferritin molecule may contain 2000 iron atoms. Hemosiderin is a less soluble variant of ferritin that may contain even greater amounts of iron.

The production of transferrin receptors and ferritin is regulated by iron response proteins (IRPs) that 'sense' intracellular iron concentrations and interact with iron response elements (IREs) of protein mRNA. When cellular iron concentrations are low, the IRP–IRE interaction works to prevent translation of mRNA to ferritin or to stabilize mRNA to enhance the translation of transferrin receptors. Identifying other proteins regulated through the IRP–IRE interaction is an area of particular interest. Although body iron is highly conserved, daily basal losses of iron of ~1 mg/day do occur even in healthy individuals. These basal losses occur primarily through the gastrointestinal tract (in bile, sloughing of ferritin-containing enterocytes, and via blood loss), and sweat and urine are additional minor sources of iron loss (**Figure 1**). Iron losses are not strictly regulated; rather, iron balance is achieved through the regulation of dietary iron absorption.

Dietary Iron Absorption

The efficiency of iron absorption depends on both the bioavailability of dietary iron and iron status.

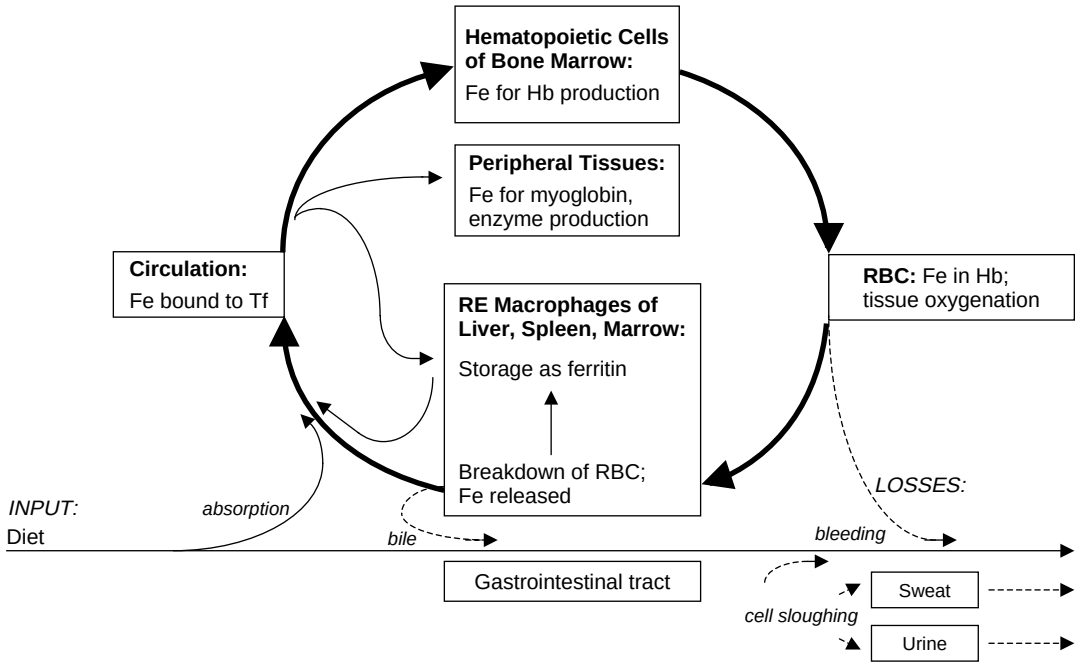


Figure 1 Iron metabolism and balance: inputs, losses, and recycling of iron through the reticuloendothelial system. Fe, iron; Tf, transferrin; Hb, hemoglobin; RBC, red blood cell; RE, reticuloendothelial.

Typically, 5–20% of the iron present in a mixed diet is absorbed. Dietary iron exists in two forms, heme and non-heme. Heme iron is derived from animal source food and is more bioavailable than non-heme iron, with approximately 20–30% of heme iron absorbed via endocytosis of the entire heme molecule. Iron is then released into the enterocyte by a heme oxidase.

Non-heme iron exists in plant products and its bioavailability is compromised by the concurrent ingestion of tannins, phytates, soy, and other plant constituents, that decrease its solubility in the intestinal lumen. Bioavailability of non-heme iron is increased by concurrent ingestion of ascorbic acid and meat products. Non-heme iron is reduced from the ferric to the ferrous form in the intestinal lumen and transported into enterocytes via the divalent metal transporter (DMT-1). Once inside the enterocyte, iron from heme and non-heme sources is similarly transported through the cell and across the basolateral membrane by the ferroportin transporter in conjunction with the ferroxidase hephaestin after which it can be taken up by transferrin into the circulation. The regulation of iron across the basolateral membrane of the enterocyte is considered the most important aspect of iron absorption.

The absorption efficiency of non-heme iron in particular is also inversely related to iron status. The factor responsible for communicating body iron status to the enterocyte to allow for the up- or downregulation of iron absorption remained elusive until recently, when the hormone hepcidin was identified. Hepcidin declines during iron deficiency, and its decline is associated with an increased production of the DMT-1 and ferroportin transporters in a rat model, although its exact mode of action is unknown. Hepcidin may also regulate iron absorption and retention or release of iron from body stores during conditions of enhanced erythropoiesis and inflammation.

Iron Requirements

Iron requirements depend on iron losses and growth demands for iron across life stages. To maintain iron balance or achieve positive iron balance, therefore, the amount of iron absorbed from the diet must equal or exceed the basal losses plus any additional demands for iron attributable to physiologic state (e.g., growth, menstruation, and pregnancy) and/or pathological iron losses (e.g., excess bleeding). When iron balance is negative, iron deficiency will occur following the depletion of the body's iron reserves. Thus, ensuring an adequate supply of dietary iron is of paramount importance. The risk of iron deficiency and iron deficiency anemia varies across the life cycle as iron demand and/or the likelihood of consuming adequate dietary iron changes.

Basal Iron Loss

Because basal iron losses are due to cell exfoliation, these losses are relative to interior body surfaces,

totaling an estimated 14 μg/kg body weight/day, and are approximately 0.8 mg/day for nonmenstruating women and 1.0 mg/day for men. Basal losses in infants and children have not been directly determined and are estimated from data available on adult men. Basal losses are reduced in people with iron deficiency and increased in people with iron overload. The absorbed iron requirement for adult men and nonmenstruating women is based on these obligate iron losses.

Infancy and Childhood

The iron content of a newborn infant is approximately 75 mg/kg body weight, and much of this iron is found in hemoglobin. The body iron of the newborn is derived from maternal–fetal iron transfer, 80% of which occurs during the third trimester of pregnancy. Preterm infants, with less opportunity to establish iron stores, have a substantially reduced endowment of body iron at birth than term infants.

During the first 2 months of life, there is a physiologic shift of body iron from hemoglobin to iron stores. For the first 6 months of life, the iron requirement of a term infant is satisfied by storage iron and breast milk iron, which is present in low concentrations but is highly bioavailable (50–100%) to the infant. However, by 6 months of age in term infants, and even earlier in preterm infants, iron intake and body stores become insufficient to meet the demands for growth (expanding erythrocyte mass and growth of body tissues), such that negative iron balance will ensue at this time without the introduction of iron supplements or iron-rich weaning foods.

A full-term infant almost doubles its body iron content and triples its body weight in the first year of life. Although growth continues through childhood, the rate of growth declines following the first year of life. Similarly, the requirement for iron expressed per kilogram body weight declines through childhood from a high of 0.10 mg/kg in the first 6 months to 0.03 mg/kg/d by 7–10 years of age until increasing again during the adolescent growth spurt. Throughout the period of growth, the iron concentration of the diet of infants and children must be greater than that of an adult man in order to achieve iron balance.

Adolescence

Adolescents have very high iron requirements, and the iron demand of individual children during periods of rapid growth is highly variable and may exceed mean estimated requirements. Boys going through puberty experience a large increase in erythrocyte mass and hemoglobin concentration. The growth spurt in adolescent girls usually occurs in early adolescence before menarche, but growth continues postmenarche at a slower rate. The addition of menstrual iron loss to the iron demand for growth leads to particularly high iron requirements for postmenarchal adolescent girls.

Menstruation

Although the quantity of menstrual blood loss is fairly constant across time for an individual, it varies considerably from woman to woman. The mean menstrual iron loss is 0.56 mg/day when averaged over a monthly cycle. However, menstrual blood losses are highly skewed so that a small proportion of women have heavy losses. In 10% of women, menstrual iron loss exceeds 1.47 mg/day and in 5% it exceeds 2.04 mg/day. Therefore, the daily iron requirement for menstruating women is set quite high to cover the iron needs of most of the population. Menstrual blood loss is decreased by oral contraceptives but increased by intrauterine devices. However, recent progesterone-releasing versions of the device lead to decreased menstrual blood loss or amenorrhea.

Pregnancy and Lactation

The body's iron needs during pregnancy are very high despite the cessation of menstruation during this period. Demand for iron comes primarily from the expansion of the red blood cell mass (450 mg), the fetus (270 mg), the placenta and cord (90 mg), and blood loss at parturition (150 mg). However, the requirement for iron is not spread evenly over the course of pregnancy, as depicted in **Figure 2**, with iron requirements actually reduced in the first trimester because menstrual blood loss is

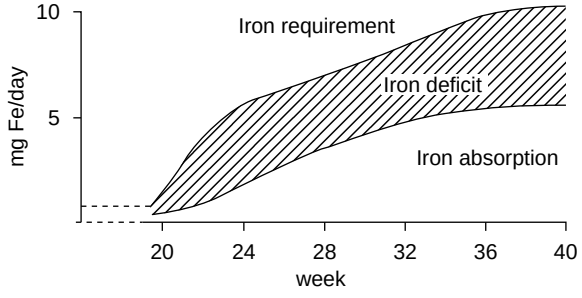


Figure 2 The discrepancy between iron requirements and availability of iron from dietary absorption in pregnant women beyond 20 weeks of gestation. The resulting iron deficit is maintained as pregnancy progresses into the second and third trimesters. (Reproduced with permission from the Food and Agriculture Organization of the United Nations (2001) Iron. In *Human Vitamin and Mineral Requirements: Report of a Joint FAO/WHO Expert Consultation, Bangkok, Thailand*, pp. 195–221. Rome: FAO.)

absent and fetal demand for iron is negligible. Iron requirements increase dramatically through the second and third trimesters to support expansion of maternal red blood cell mass and fetal growth. The maternal red cell mass expands approximately 35% in the second and third trimesters to meet increased maternal oxygen needs. When iron deficiency is present, the expansion of the red cell mass is compromised, resulting in anemia. Furthermore, an expansion of the plasma fluid that is proportionately greater than that of the red cell mass results in a physiologic anemia attributable to hemodilution.

To attempt to meet iron requirements during pregnancy, iron absorption becomes more efficient in the second and third trimesters. Iron absorption nearly doubles in the second trimester and can increase up to four times in the third trimester. Despite this dramatic increase in iron absorption, it is virtually impossible for pregnant women to acquire sufficient iron through diet alone because of the concurrent increase in iron requirements during the latter half of pregnancy (**Figure 2**).

There is also an iron cost of lactation to women of approximately 0.3 mg/day as iron is lost in breastmilk. However, this is compensated by the absence of menstrual iron losses and the gain in iron stores achieved when much of the iron previously invested in expansion of the red cell mass is recovered postpartum.

Pathological Losses

Conditions that cause excessive bleeding additionally compromise iron status. Approximately 1 mg of iron is lost in each 1 ml of packed red blood cells. Excessive losses of blood may occur from the gastrointestinal tract, urinary tract, and lung in a variety of clinical pathologies, including ulcers, malignancies, inflammatory bowel disease, hemorrhoids, hemoglobinuria, and idiopathic pulmonary hemosiderosis. In developing countries, parasitic infestation with hookworm and schistosomiasis can contribute substantially to gastrointestinal blood loss and iron deficiency.

Recommended Nutrient Intakes for Iron

Recommended intakes of dietary iron are based on the requirement for absorbed iron and assumptions about the bioavailability of iron in the diet. They are meant to cover the iron needs of nearly the entire population group. Thus, the amount of dietary iron necessary to meet an iron requirement depends in large part on the bioavailability of iron in the diet (**Figure 3**). Americans consume approximately 15 mg of iron daily from a diet that is considered moderately to highly bioavailable (10–15%) due to the meat and ascorbic acid content. Studies in European countries suggest that iron intake

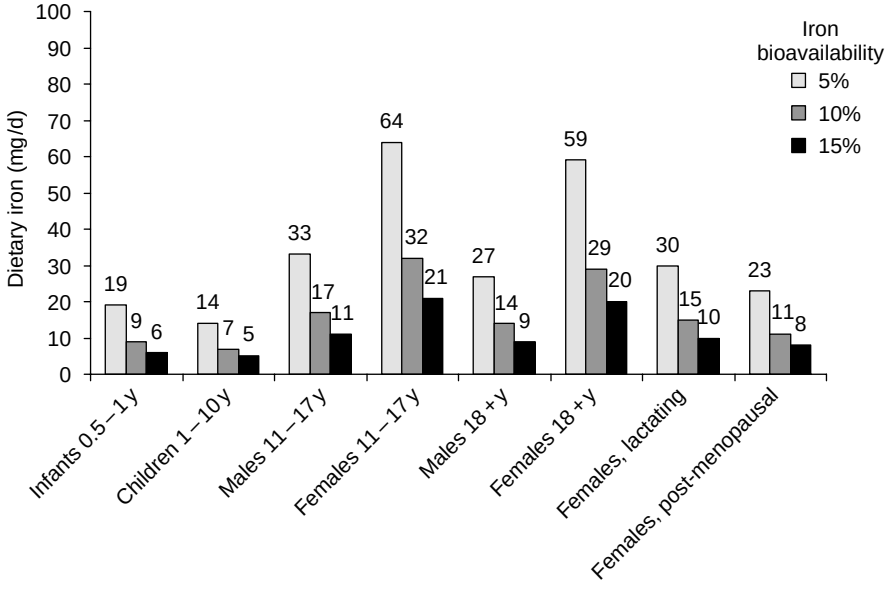


Figure 3 The Recommended Nutrient Intake (RNI) for iron given different levels of bioavailability of iron in the diet: 5%, low; 10%, moderate; and 15%, high. The RNI is based on the amount of iron necessary to meet requirements of 95% of the population for each age/sex group. Because typical iron intakes range from 10 to 15 mg/day, iron requirements are nearly impossible to meet on low-bioavailability diets. (Data from the Food and Agriculture Organization of the United Nations (2001) Iron. In *Human Vitamin and Mineral Requirements: Report of a Joint FAO/WHO Expert Consultation, Bangkok, Thailand*, pp. 195–221. Rome: FAO.)

averages 10 mg/day, representing a decline in dietary iron. Although estimates of total iron intake in developing countries are not substantially lower than that, iron is often consumed in plant-based diets that inhibit its absorption and contain few animal products to counterbalance that effect, such that the bioavailability of iron is closer to 5%. Thus, **Figure 3** demonstrates the total amount of dietary iron that would be necessary to meet the iron requirements of various population groups based on its bioavailability. Where intakes are sufficient and bioavailability adequate, dietary iron can meet the iron needs of adolescent boys and adult men and also lactating and postmenopausal women. However, regardless of bioavailability, iron requirements are not met by many adolescent girls and adult menstruating women who have above average menstrual blood loss. Few if any population groups can achieve iron intakes sufficient to meet iron requirements when bioavailability of iron is poor.

Dietary recommendations for infants are based on the iron content and bioavailability of human milk. The iron in infant formula is much less bioavailable (10%) than that of human milk and is thus present in greater concentrations than that of human milk. Infants who are not breast-fed should consume iron-fortified formula. Complementary foods offered after 6 months of age can potentially meet iron needs if they have a high content of meat and ascorbic acid. This is rarely the case in developing or developed countries, and fortified infant cereals and iron drops are often introduced at this time in developed countries. In developing countries where diets are poor in bioavailable iron, iron-fortified weaning foods are not commonly consumed, and iron supplements are rarely given to infants and children.

Pregnant women rarely have sufficient iron stores and consume diets adequate to maintain positive iron balance, particularly in the latter half of pregnancy, as previously discussed. They cannot meet their iron requirements through diet alone even in developed countries, where high iron content diets with high bioavailability are common. Supplementation is universally recommended for pregnant women, as discussed later.

Indicators of Iron Deficiency and Anemia

Indicators of iron deficiency can be used to distinguish the degree of iron deficiency that exists across the spectrum from the depletion of body iron stores to frank anemia (**Table 1**). Indicator cutoffs vary by age, sex, race, and physiologic state (e.g., pregnancy), so using a proper reference is important when interpreting indicators of iron deficiency.

Table 1 Indicators for assessing the progression of iron deficiency from depletion of iron stores to iron deficiency anemia

Stage of iron deficiency	*Consequence*	*Indicator*
Depletion	Decline in storage iron	↓ Serum ferritin
Deficiency	Decreased circulating iron	↓ Serum iron ↑ Total iron binding capacity ↓ Transferrin saturation
	Insufficient tissue iron	↑ Transferrin receptor
	Impaired heme synthesis	↑ Protoporphyrin/heme
Depletion	Impaired red blood cell production	↓ Hemoglobin ↓ Hematocrit ↓ Red blood cell indices

Serum ferritin is directly related to liver iron stores—a gold standard for iron deficiency that is infrequently used due to the invasive nature of the test. Different sources place the cutoff for serum ferritin concentrations indicative of depleted stores at 12 or 15 μg/l. Once iron stores are exhausted, serum ferritin is not useful for determining the extent of iron deficiency. Serum ferritin is also useful for diagnosing iron excess. A major limitation of serum ferritin is the fact that it acts as an acute phase reactant and therefore is mildly to substantially elevated in the presence of inflammation or infection, complicating its interpretation when such conditions exist.

Transferrin saturation is measured as the ratio between total serum iron (which declines during iron deficiency) and total iron binding capacity (which increases during iron deficiency). Typically, transferrin is approximately 30% saturated, and low transferrin saturation (<16%) is indicative of iron deficiency. Transferrin saturation concentrations higher than 60% are indicative of iron overload associated with hereditary hemochromatosis. The use of transferrin saturation to distinguish iron deficiency is limited because of marked diurnal variation and its lack of sensitivity as an indicator.

Elevated circulating TfRs are a sensitive indicator of the tissue demand for iron. Circulating TfR is not affected by inflammation, a limitation of other indicators of iron status. Furthermore, expressing TfR as a ratio with ferritin appears to distinguish with a great deal of sensitivity iron deficiency anemia from

anemia of chronic disease, making this combined measure potentially very useful in settings in which these conditions coexist.

Elevated erythrocyte zinc protoporphyrin indicates iron-deficient erythropoiesis. Protoporphyrin concentrations may also be elevated by inflammation and lead exposure.

Finally, although hemoglobin concentrations or percentage hematocrit are not specific for iron deficiency, these measures are used most frequently as a proxy for iron deficiency in field settings because of their technical ease. Anemia is defined as a hemoglobin concentration of less than 110 g/l for those 6 months to 5 years old and for pregnant women, 115 g/l for those 5–11 years old, 120 g/l for nonpregnant females older than 11 years and for males 12–15 years old, or 130 g/l for males older than 15 years of age. Other measures of red blood cell characteristics include total red blood cell counts, mean corpuscular volume, and mean hemoglobin volume.

The choice of indicators and the strategy for assessment will depend on technical feasibility and whether a screening or survey approach is warranted. When more than 5% of a population is anemic, iron deficiency is considered a public health problem, and population-based surveys may be useful for assessing and monitoring the prevalence of iron deficiency. When anemia is less prevalent, screening for iron deficiency in high-risk groups or symptomatic individuals is a more efficient approach. Hemoglobin alone would be insufficient to diagnose iron deficiency in an individual, but hemoglobin distributions can offer clues as to the extent to which anemia is attributable to iron deficiency in a population. Preferred indicators, such as transferrin and/or ferritin, may not be feasible due to blood collection requirements, cost, or technical difficulty in a population survey, but they may be indispensable for characterizing iron status of a population subgroup or individual.

Prevalence of Iron Deficiency and Iron Deficiency Anemia

Although iron deficiency anemia is considered the most prevalent nutritional deficiency globally, accurate prevalence estimates are difficult to obtain. Worldwide, prevalence estimates for iron deficiency anemia have ranged from 500 million to approximately 2 billion people affected. However, most global prevalence estimates are based on anemia surveys, which will overestimate the amount of anemia attributable to iron deficiency but underestimate the prevalence of less severe iron deficiency. There is clearly a disparity in anemia prevalence between the developing and developed world, with ~50% of children and nonpregnant women in the developing world considered anemic compared with ~10% in the developed world. The prevalence of anemia increases during pregnancy, with ~20% of US women anemic during pregnancy and estimates of anemia prevalence in some developing countries exceeding 60%.

Data from the US NHANES III (1988–1994) survey, which used a variety of indicators of iron status, showed that 9% of US toddlers were iron deficient and 3% had iron deficiency anemia. Eleven percent of adolescent females and women of reproductive age were iron deficient, and 3–5% of these women had iron deficiency anemia. Iron deficiency in the developed world is more common among low-income minorities.

Consequences of Iron Deficiency and Iron Deficiency Anemia

Iron deficiency anemia has been implicated in adverse pregnancy outcomes, maternal and infant mortality, cognitive dysfunction and developmental delays in infants and children, and compromised physical capacity in children and adults. However, data to support causal relationships with some of these outcomes is limited, and the extent to which outcomes are associated with iron deficiency specifically or more generally with anemia regardless of the etiology is the subject of debate.

A variety of observational studies demonstrate an association of maternal hemoglobin concentrations during pregnancy with birth weight, the likelihood of low birth weight and preterm birth, and perinatal mortality, such that adverse pregnancy outcomes are associated in a 'U-shaped' manner with the lowest and highest maternal hemoglobin concentrations. Anemia during pregnancy may not be specific to iron deficiency, and randomized trials utilizing a strict placebo to firmly establish a causal link between iron status per se and adverse pregnancy outcomes are rare because of ethical concerns about denying women iron supplements during pregnancy. However, two randomized trials, one conducted in a developed country and the other in a developing country, have shown a positive impact of iron supplementation during pregnancy on birth weight. In the developed country study, control women with evidence of compromised iron stores were offered iron supplements at 28 weeks of gestation after 8 weeks of randomized iron

supplementation. In the developing country study, the control group received supplemental vitamin A and the intervention group received folic acid in addition to iron.

Mortality among pregnant women and infants and children also increases with severe anemia. However, most data showing this relationship are observational and clinic based. Anemia in such circumstances is unlikely to be attributable to iron deficiency alone; furthermore, the degree to which mild to moderate anemia influences mortality outcomes is not well established.

Iron deficiency and iron deficiency anemia have been associated with impaired cognitive development and functioning. These effects of iron deficiency may be mediated in part by the deprivation of functional iron in brain tissue, and the impact of iron deprivation may vary depending on its timing in relation to critical stages of brain development. Iron interventions in anemic school-age children generally result in improved school performance. The results of iron interventions in infants and preschool-age children are less clear, perhaps in part because cognition is more difficult to measure in this age group.

Studies have shown a negative impact of iron deficiency anemia on work productivity among adult male and female workers in settings requiring both strenuous labor (rubber plantation) and less intensive efforts (factory work). The impact of iron deficiency anemia on performance may be mediated by a reduction in the oxygen-carrying capacity of blood associated with low hemoglobin concentration and by a reduction in muscle tissue oxidative capacity related to reductions in myoglobin and effects on iron-containing proteins involved in cellular respiration.

Interventions: Prevention and Treatment of Iron Deficiency Anemia

Supplementation

Iron supplementation is the most common intervention used to prevent and treat iron deficiency anemia. Global guidelines established by the International Nutritional Anemia Consultative Group, the World Health Organization, and UNICEF identify pregnant women and children 6–24 months of age as the priority target groups for iron supplementation because these populations are at the highest risk of iron deficiency and most likely to benefit from its control. However, recommendations are given for other target groups, such as children, adolescents, and women of reproductive age, who may also benefit from iron supplementation for the prevention of iron deficiency. The recommendations are given in **Table 2**. Recommended dose and/or duration of supplementation are increased for populations where the prevalence of anemia is 40% or higher. The recommended treatment for severe anemia (Hb <70 g/l) is to double prophylactic doses for 3 months and then to continue the preventive supplementation regimen.

Ferrous sulfate is the most common form of iron used in iron tablets, but fumarate and gluconate are also sometimes used. A liquid formulation is available for infants, but it is not used often in anemia control programs in developing countries because of the expense compared to tablets. Crushed tablets can be given to infants and young children as an alternative, but this has not been very successful programmatically.

Efforts to improve the iron status of populations worldwide through supplementation have met with mixed success. Given the frequency with which iron

Table 2 Guidelines for iron supplementation to prevent iron deficiency anemia

Target group	*Dose*	*Duration*
Pregnant women	60 mg iron + 400 μg folic acid daily	6 months in pregnancy[a,b]
Children 6–24 months (normal birth weight)	12.5 mg iron[c] + 50 μg folic acid daily	6–12 months of age[d]
Children 6–24 months (low birth weight)	12.5 mg iron + 50 μg folic acid daily	2–24 months of age
Children 2–5 years	20–30 mg iron[c] daily	
Children 6–11 years	30–60 mg iron daily	
Adolescents and adults	60 mg iron daily[e]	

[a]If 6-months' duration cannot be achieved during pregnancy, continue to supplement during the postpartum period for 6 months or increase the dose to 120 mg iron daily during pregnancy.
[b]Continue for 3 months postpartum where the prevalence of pregnancy anemia is $\geq$40%.
[c]Iron dosage based on 2 mg iron/kg body weight/day.
[d]Continue until 24 months of age where the prevalence of anemia is $\geq$40%.
[e]For adolescent girls and women of reproductive age, 400 μg folic acid should be included with iron supplementation.
Adapted with permission from Stoltzfus RJ and Dreyfuss ML (1998) *Guidelines for the Use of Iron Supplements to Prevent and Treat Iron Deficiency Anemia*. Washington, DC: International Nutritional Anemia Consultative Group.

tablets must be taken to be effective, a lack of efficacy of iron supplementation in research studies and programs has often been attributed to poor compliance and the presence of side effects such as nausea and constipation. Ensuring compliance in some settings also requires extensive logistical support. Although in developing countries the maximum coverage of iron supplementation programs for pregnant women is higher than 50%, other high-risk groups are less frequently targeted for iron supplementation.

Comparative trials have demonstrated that both weekly and daily iron supplementation regimens significantly increase indicators of iron status and anemia, but that daily supplements are more efficacious at reducing the prevalence of iron deficiency and anemia, particularly among pregnant women and young children who have high iron demands. Therefore, daily iron supplements continue to be the recommended choice for pregnant women and young children because there is often a high prevalence of iron deficiency anemia in these populations. Weekly supplementation in school-age children and adolescents holds promise for anemia prevention programs because it reduces side effects, improves compliance, and lowers costs. Further assessment of the relative effectiveness of the two approaches is needed to determine which is more effective in the context of programs.

Fortification

Iron fortification of food is the addition of supplemental iron to a mass-produced food vehicle consumed by target populations at risk of iron deficiency anemia. Among anemia control strategies, iron fortification has the greatest potential to improve the iron status of populations. However, its success has been limited by technical challenges of the fortification process: (i) the identification of a suitable iron compound that does not alter the taste or appearance of the food vehicle but is adequately absorbed and (ii) the inhibitory effect of phytic acid and other dietary components that limit iron absorption. Water-soluble iron compounds, such as ferrous sulfate, are readily absorbed but cause rancidity of fats and color changes in some potential food vehicles (e.g., cereal flours). In contrast, elemental iron compounds do not cause these sensory changes but are poorly absorbed and are unlikely to benefit iron status. Research on iron compounds and iron absorption enhancers that addresses these problems has yielded some promising alternatives. Encapsulated iron compounds prevent some of the sensory changes that occur in fortified food vehicles. The addition of ascorbic acid enhances iron absorption from fortified foods, and NaFe–EDTA provides highly absorbable iron in the presence of phytic acid.

Many iron-fortified products have been tested for the compatibility of the fortificant with the food vehicle and for the bioavailability of the fortified iron, but few efficacy or effectiveness trials have been done. Iron-fortified fish sauce, sugar, infant formula, and infant cereal have been shown to improve iron status. In contrast, attempts to fortify cereal flours with iron have met with little success because they contain high levels of phytic acid and the characteristics of these foods require the use of poorly bioavailable iron compounds.

In the developed world, iron fortification has resulted in decreased rates of iron deficiency and anemia during the past few decades. Some debate remains, however, about the potential for the acquisition of excess iron, which has been associated with increased chronic disease risk in some studies. In Europe, Finland and Denmark have recently discontinued food fortification programs because of concerns of iron overload. Individuals with hereditary hemochromatosis, ~5/1000 individuals in populations of European descent, are at particular risk of iron overload.

Control of Parasitic Infections

Because geohelminths such as hookworm also contribute to iron deficiency, programs that increase iron intakes but do not address this major source of iron loss are unlikely to be effective at improving iron status. Other infections and inflammation also cause anemia, as does malaria, and the safety of iron supplementation during infection or malaria has been debated. Where malaria and iron deficiency coexist, the current view is that iron supplementation is sufficiently beneficial to support its use. Ideally, however, where multiple etiologies of anemia coexist, these etiologies need to be recognized and simultaneously addressed.

Other Micronutrients

Other nutrients and their deficiencies that can impact iron status, utilization, or anemia include vitamin A, folate, vitamin B_{12}, riboflavin, and ascorbic acid (vitamin C). Improving iron status can also increase the utilization of iodine and vitamin A from supplements. On the other hand, it is increasingly recognized that simultaneous provision of iron and zinc in supplements may decrease the benefit of one or both of these nutrients. These complex micronutrient interactions and their implications for nutritional interventions are incompletely understood but

have significant implications for population-based supplementation strategies.

Summary

Iron deficiency anemia exists throughout the world, and pregnant women and infants 6–24 months old are at highest risk because of their high iron requirements. Women of reproductive age, school-age children, and adolescents are also high-risk groups that may require attention in anemia control programs. Although numerous indicators exist to characterize the progression of iron deficiency to anemia, difficulties remain with their use and interpretation, particularly in the face of other causes of anemia. Despite the proven efficacy of iron supplementation and fortification to improve iron status, there are few examples of effective anemia prevention programs. More innovative programmatic approaches that aim to improve iron status, such as geohelminth control or prevention of other micronutrient deficiencies, deserve more attention. Challenges remain in preventing and controlling iron deficiency anemia worldwide.

See also: **Adolescents**: Nutritional Requirements. **Breast Feeding**. **Children**: Nutritional Requirements. **Folic Acid**. **Infants**: Nutritional Requirements. **Iron**. **Lactation**: Dietary Requirements. **Pregnancy**: Nutrient Requirements.

Further Reading

ACC/SCN (2001) Preventing and treating anaemia. In: Allen LH and Gillespie SR (eds.) *What Works? A Review of the Efficacy and Effectiveness of Nutrition Interventions*. Geneva: ACC/SCN in collaboration with the Asian Development Bank, Manila.

Anonymous (2001) Supplement II: Iron deficiency anemia: Reexamining the nature and magnitude of the public health problem. *Journal of Nutrition* **131**: 563S–703S.

Anonymous (2002) Supplement: Forging effective strategies to combat iron deficiency. *Journal of Nutrition* **132**: 789S–882S.

Bothwell TH, Charlton RW, Cook JD, and Finch CA (1979) *Iron Metabolism in Man*. Oxford: Blackwell Scientific.

Cook JD, Skikne BS, and Baynes RD (1994) Iron deficiency: The global perspective. *Advances in Experimental Medicine and Biology* **356**: 219–228.

Eisenstein RS and Ross KL (2003) Novel roles for iron regulatory proteins in the adaptive response to iron deficiency. *Journal of Nutrition* **133**: 1510S–1516S.

Fairbanks VP (1999) Iron in medicine and nutrition. In: Shils M, Olson JA, Shike M, and Ross AC (eds.) *Modern Nutrition in Health and Disease*, 9th edn, pp. 193–221. Philadelphia: Lippincott Williams & Wilkins.

Food and Agriculture Organization of the United Nations (2001) Iron. In: *Human Vitamin and Mineral Requirements: Report of a Joint FAO/WHO Expert Consultation, Bangkok, Thailand*, pp. 195–221. Rome: FAO.

Frazer DM and Anderson GJ (2003) The orchestration of body iron intake: How and where do enterocytes receive their cues? *Blood Cells, Molecules, and Diseases* **30**(3): 288–297.

Koury MJ and Ponka P (2004) New insights into erythropoiesis: The roles of folate, vitamin B_{12}, and iron. *Annual Review of Nutrition* **24**: 105–131.

Leong W-I and Lonnerdal B (2004) Hepcidin, the recently identified peptide that appears to regulate iron absorption. *Journal of Nutrition* **134**: 1–4.

Roy CN and Enns CA (2000) Iron homeostasis: New tales from the crypt. *Blood* **96**(13): 4020–4027.

Stoltzfus RJ and Dreyfuss ML (1998) *Guidelines for the Use of Iron Supplements to Prevent and Treat Iron Deficiency Anemia*. Washington, DC: International Nutritional Anemia Consultative Group.

World Health Organization (2001) *Iron Deficiency Anaemia Assessment, Prevention and Control: A Guide for Programme Managers*. Geneva: WHO.

Megaloblastic Anemia

J M Scott, Trinity College Dublin, Dublin, Ireland
P Browne, St James's Hospital, Dublin, Ireland

Introduction

A major distinction in diagnosis and classification of anemias is whether the eventual red cells that appear in the circulation are smaller (microcytic) or larger (macrocytic) than the usual normal cell size (normocytic). The most important example of the former is iron deficiency anemia where it appears that the red cell precursors, during their replication in the bone marrow from an original pluripotent stem cell undergo a higher than normal number of divisions. Since each such division results in two daughter cells that are slightly smaller, an increase in the number of divisions in the marrow compartment will result in smaller red cells in the circulation. In iron deficiency this is thought to happen because the usual progressive inactivation of the nucleus after each division occurs at a slower than normal rate.

The most characteristic example of a macrocytic anemia occurs because there is an abnormally slow rate of DNA biosynthesis in the developing red cell. Such reduced synthesis delays the rate of development of the nucleus and with it the rate of cell division during replication in the bone marrow compartment. Thus, by the time such cells have differentiated to the point at which they receive a signal to leave the bone marrow, they have undergone fewer than usual cell divisions, resulting in cells that are larger than normal

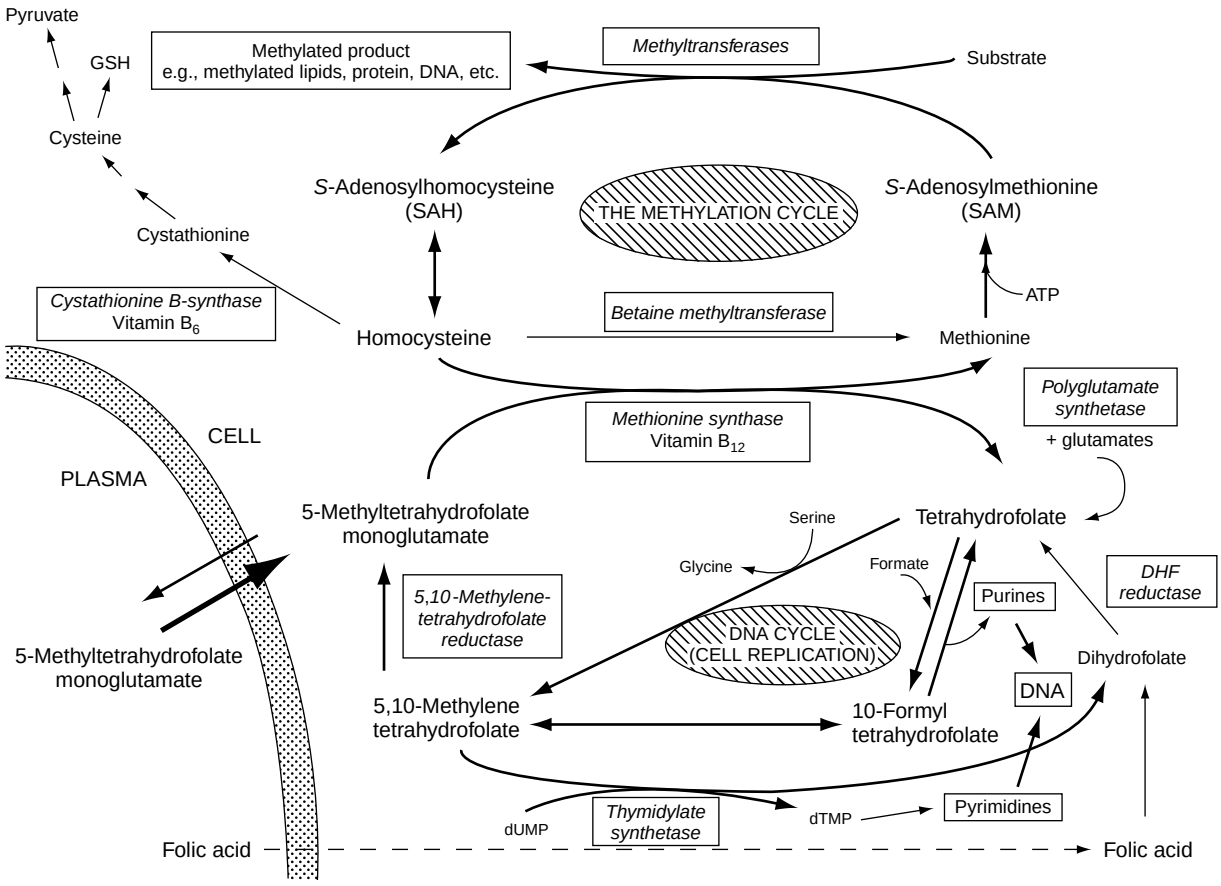


Figure 1 The role of folate cofactors in the DNA and methylation cycles.

or macrocytic. The other unique characteristic of such arrest in DNA biosynthesis is evidenced by the cells that are present in the bone marrow itself. The red cell precursors have a very different appearance from that of the normally developing red cell series (normoblasts). The nuclei are much larger than usual and are far less differentiated. These characteristic cells are called megaloblasts and only occur where there has been a slow down or arrest of DNA biosynthesis. This occurs in only three circumstances: folate deficiency, vitamin B_{12} deficiency, or during therapy with drugs that interfere directly or indirectly with DNA biosynthesis (**Figure 1**).

Definition

As the name suggests and as discussed above the unique feature that defines megaloblastic anemia is the presence of abnormal red cell precursors called megaloblasts in the bone marrow. Therefore, bone marrow examination by a competent hematologist remains the gold standard for diagnosis of megaloblastic anemia. As discussed below, such morphological examinations are no longer routinely part of the diagnosis. However, despite the availability of other tests, the presence of megaloblasts in bone marrow aspirate remains the only way to achieve a definitive diagnosis and is still required if the patient fails to respond to treatment.

Biochemical Aspects of the Megaloblastic Anemias

The biochemical interrelationships between vitamin B_{12} and folate are described in **Figure 1** and discussed elsewhere in the chapters on cobalamins and folic acid. The folate cofactors are essential for the provision of so-called carbon one units for the biosynthesis of purines and pyrimidines and thus for DNA. Folate in the form of 5-methyltetrahydrofolate (5-methyl THF) is also needed to supply the methylation cycle with methyl groups (**Figure 1**). These are needed to regenerate methionine and *S*-adenosylmethionine (SAM) in cells, the latter being used to donate methyl groups to the three dozen or so methyltransferases present in all cells. In hepatocytes, part of the methylation cycle is used to degrade the 60% or so excess of methionine

present in the diet over and above daily requirements. When folate status is reduced there will be a reduced capacity in cells to make DNA and thus to replicate. This will be most easily seen in rapidly dividing cells such as those of the bone marrow, hence the emergence of the very characteristic megaloblastic anemia with megaloblasts being seen in bone marrow aspirates. Clearly one would also expect to see a reduction in the methylation cycle, which could in turn reduce the activity of the numerous methyltransferases. The effects of such a reduction are less obvious and contrast sharply with what happens when the methylation cycle is interrupted by deficiency of vitamin B_{12} (see below). Vitamin B_{12} is involved in two enzymatic reactions in man, methylmalonyl CoA mutase and methionine synthase. As discussed later deficiency of the former leads to a raised level of methylmalonyl CoA in cells, which is seen in the circulation and the urine as methylmalonic acid (MMA). What is of very great interest is the clinical sequence of events during vitamin B_{12} deficiency and how they arise. There are two such sequences: the development of a megaloblastic anemia identical to that seen in folate deficiency and a neuropathy not usually associated with folate deficiency.

At a biochemical level, the explanation for the anemia is encapsulated in the methyl trap hypothesis, first put forward by Victor Herbert as early as 1961. The biosynthesis of 5-methyl THF by the enzyme 5,10-methylenetetrahydrofolate reductase (MTHFR) (**Figure 1**) is irreversible *in vivo*. Thus, once formed this folate cofactor can only be used by the vitamin B_{12}-dependent enzyme methionine synthase. The activity of the enzyme is reduced or absent in the bone marrow of patients with vitamin B_{12} deficiency. Progressively more and more of the folate cofactors become metabolically trapped as 5-methyl THF reducing the intracellular levels of 10-formyl THF and 5,10-methylene THF needed for the biosynthesis of purines and pyrimidines and thus for DNA and cell division. Although such cells contain folate, they are unable to use it and suffer from a kind of pseudo folate deficiency, thus producing an identical megaloblastic anemia to that seen in folate deficiency. One might question why cells do such an apparently destructive thing. The answer is that cells perceive vitamin B_{12} deficiency through an ever-reducing level of SAM. This essential methyl donor normally downregulates the amount of 5-methyl THF synthesized in cells by reducing the level of the enzyme MTHFR. Falling levels of SAM in vitamin B_{12} deficiency by contrast are met with an ever-increasing activity of MTHFR and diversion of the folate cofactors into the trapped form, namely 5-methyl THF, which, of course, cannot be regenerated into THF because of the absence of methionine synthase.

A very characteristic neuropathy is also associated with vitamin B_{12} deficiency. This neuropathy is due to a reduction or interruption of the methylation cycle. This is clear from two independent lines of evidence. Firstly, inactivation of methionine synthase in experimental animals (monkeys and pigs) leads to the classical so-called subacute combined degeneration of the spinal cord (SCD) seen in patients with severe vitamin B_{12} deficiency. Secondly, patients with genetically very rare dramatic reductions in the enzyme MTHFR have the classical signs and symptoms of SCD. The most plausible explanation is that as a result of reduced MTHFR levels they are unable to supply the methyl groups needed for the methylation cycle. It is of interest that such patients do not get megaloblastic anemia, presumably because while their folate metabolism is interfered with, there is no trapping of the folate cofactors metabolically as 5-methyl THF. It seems probable that a reduction in activity of one or more of the methyltransferases, the activity of which is compromised by an interruption of the methylation cycle, causes the characteristic neuropathy. It is unclear which specific methyltransferase is involved.

Diagnosis of Megaloblastic Anemia

As mentioned above the definitive diagnosis requires identification of the presence of megaloblasts in bone marrow aspirate. The taking of such an aspirate (usually from the hip bone) involves some discomfort for the patient and must be performed by an appropriately trained practitioner. Frequently, the routine diagnosis of unexplained macrocytic anemia falls to general physicians or general practitioners. In this situation, where there is clear evidence that the macrocytic anemia is due to deficiency of vitamin B_{12} or folate, it is not necessary to obtain a bone marrow aspirate to confirm megaloblastic changes. However, when bone marrow is not examined initially and a patient is treated for deficiency of vitamin B_{12} or folic acid, it is essential to verify that their response to treatment includes correction of anemia and macrocytosis. If there is any doubt, a bone marrow aspirate must be performed to exclude other possible underlying hematological disorders.

The first stage of diagnosis is based on the result of a full blood count (FBC) (also called complete blood count (CBC) in some countries) using an automatic instrument such as a Coulter counter. An FBC is done on virtually every patient admitted

to hospital. Frequently, an FBC would also form part of an outpatient work-up or might be ordered by a GP through an associated hospital or laboratory. Where the hemoglobin level is below the reference value with respect to sex and age indicating anemia, the mean corpuscular volume (MCV) is assessed. This parameter essentially gives a mean of the size of red blood cells in the circulation. Megaloblastic anemia usually results in larger than normal red cells in the circulation and thus a raised MCV; however, sometimes quite advanced stages of megaloblastic anemia can be accompanied by a normal and, infrequently, even below normal MCV. This can arise because of the concomitant presence of iron deficiency. A raised MCV accompanying the anemia seen in the FBC (macrocytic anemia) moves the diagnosis to being one of megaloblastic anemia, although other causes of macrocytosis such as hypothyroidism or excess alcohol consumption may need to be considered also. Conventionally, the next step is to carry out a bone marrow aspirate to verify if megaloblasts are present, but, as mentioned earlier, this step can be omitted if the diagnosis of vitamin B_{12} or folate deficiency can be made rapidly and accurately. After a positive bone marrow aspirate, or in its absence if this step is omitted, the next analysis would be the determination of circulatory levels of folate and vitamin $B_{12.}$ If only one of the vitamin levels is in the deficient range, most clinicians would embark upon the regimen of therapy discussed below. As mentioned above, antifolate or anti-DNA drugs, such as methotrexate, 5-fluorouracil, or cyclophosphamide, will also arrest DNA biosynthesis and cause megaloblastic anemia; however, it is usually known when patients are on such anticancer chemotherapy.

The circulating levels of folate and vitamin B_{12} can be measured in serum or plasma samples by a number of methods. Most regard microbiological assays using *Lactobacillus casei* for folate and *Lactobacillus leichmannii* for vitamin B_{12} as the 'gold standard.' However, these assays are difficult to perform and most laboratories use methods based on enzyme linked immunosorbent assays (ELISA) or competitive binding assays using a natural binder such as intrinsic factor for vitamin B_{12} or β-lactoglobulin for folate. While very low plasma or serum levels of $<2.0\,\mu g\,l^{-1}$ (4.5 nM) for folate and $<120\,ng\,l^{-1}$ (88 pM) for vitamin $B_{12,}$ are considered as being diagnostic of deficiency, there is a gray area for both assays 2.0–2.7 $\mu g\,l^{-1}$ (4.5–6.1 nM) for serum folate and 120–200 $ng\,l^{-1}$ (88–148 pM) for vitamin B_{12} indicating possible deficiency. Values above $2.7\,\mu g\,l^{-1}$ (6.1 nM) for folate or $200\,ng\,l^{-1}$ (148 pM) for vitamin B_{12} usually indicate the absence of deficiency. Some laboratories also offer red cell folate levels. The red cell during its maturation in the bone marrow incorporates a level of folate commensurate with what is present in the circulation during that period. When the red cell passes from the bone marrow into the circulation it can neither take up nor lose folate until the end of its life, usually 120 days later. Thus, the circulatory red cells give an average of the folate level over the previous 4 months. Unlike the plasma or serum level the red cell folate level is not influenced by recent fluctuation in dietary intake. Thus, low red cell folate levels of $<100\,\mu g\,l^{-1}$ (226 nM) are a very good indication of folate deficiency with a range of 100–150 $\mu g\,l^{-1}$ (226–340 nM) where there is possible deficiency and values above $150\,\mu g\,l^{-1}$ (340 nM) generally indicating the absence of folate deficiency. While red cell folate levels have significant advantages over serum folate levels they have one very significant drawback. Red cell folate levels are also significantly reduced in vitamin B_{12} deficiency. This is because the bone marrow cells take up the predominant circulating form of folate, namely 5-methyl THF. However, this form, which has just a single glutamate, is not retained by the cells unless it is converted into a predominant cellular form of folate with on average five glutamate residues. The enzyme that adds these glutamates does not use 5-methyl THF as a substrate; therefore, 5-methyl THF must be converted to THF before it can be converted to a polyglutamate. The only enzyme in the cell that converts 5-methyl THF to THF is the vitamin B_{12}-dependent methionine synthase. As mentioned above, its activity is reduced or absent in vitamin B_{12}-deficient bone marrow. Thus, such cells have an inability to conjugate and retain the circulating form of folate and as a result have reduced red cell folate levels. Thus, a low red cell folate level may lead to the misdiagnosis of vitamin B_{12} deficiency as folate deficiency, a circumstance which for the reasons discussed later must be avoided at all costs. Conseqently, it is always necessary to measure the level of plasma or serum folate . If it is also low or deficient and accompanied by a low red cell folate this is indicative of folate rather than vitamin B_{12} deficiency. This is because the circulating folate levels tends to back up in the serum resulting in higher rather than lower serum folate levels in vitamin B_{12} deficiency.

Before therapy, further investigations could be undertaken. These largely depend upon the availability of such tests in any particular clinical context. Elevated plasma homocysteine levels occur in both vitamin B_{12} and folate deficiency and raised homocysteine does not establish which vitamin is deficient. This is because such elevation is due to a reduction in

the flux of homocysteine back to methionine as part of the methylation cycle (**Figure 1**). The enzyme that is compromised is methionine synthase, which uses vitamin B_{12} as a cofactor (**Figure 2**) and 5-methyltetrahydrofolate (**Figure 3**) and homocysteine as its substrates. This enzyme, and consequently the methylation cycle, thus requires both a normal folate and a normal vitamin B_{12} status for optimum activity. Thus reduction in the status of either vitamin is always accompanied by an elevation of plasma homocysteine. Homocysteine is also elevated in other circumstances, most notably in impaired renal function. This can, to some extent, be corrected for the creatinine level. Homocysteine is also elevated in vitamin B_6 deficiency and common C→T677 MTHFR polymorphism. Thus, while elevated plasma homocysteine confirms the presence of megaloblastic anemia, establishing which vitamin is deficient still relies on measurement of the circulating levels of the vitamins involved.

The measurement of plasma, serum, or urine MMA is very helpful in confirming a diagnosis of vitamin B_{12} deficiency. This analyte is elevated due to a reduction in the activity of methylmalonyl CoA mutase, the other vitamin B_{12}-dependent enzyme in man (**Figure 4**). It appears that it is not possible to be functionally deficient in vitamin B_{12} without a concomitant elevation in MMA, and so a false negative result is not really an issue. However, MMA like plasma homocysteine is also elevated during renal impairment, and while this can to some extent be corrected for by a raised creatinine, it cannot be assumed that elevation of MMA is due to vitamin B_{12} deficiency. While the estimation of plasma homocysteine is widely available the estimation of MMA requires gas chromatography mass spectroscopy (GC-MS) and has very limited availability in practice. Newer methods to measure vitamin B_{12} on its transport protein TC II are under development.

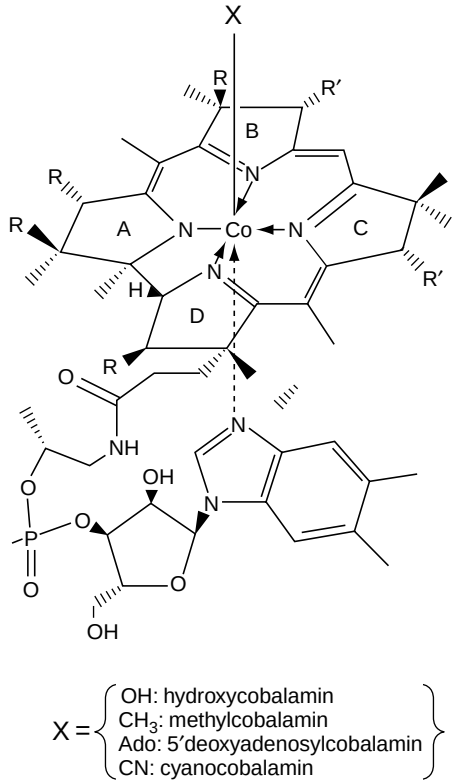


Figure 2 The structure of naturally occurring vitamin B_{12} (hydroxycobalamin), its synthetic form cyanocobalamin, and its two cofactor forms methylcobalamin and 5′deoxyadenosylcobalamin. Hydroxycobalamin, X = Co-hydroxide; cyanocobalamin, X = Co-cyanide; methylcobalamin, X = Co-CH_3; deoxyadenosylcobalamin, X = Co5′deoxyadenosyl.

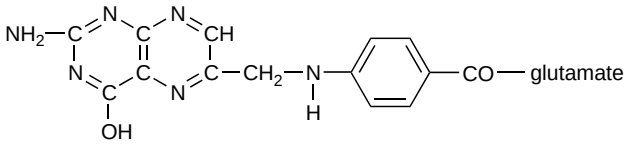
For the reasons given above, it is essential that vitamin B_{12} deficiency is not confused with folate deficiency. As mentioned previously, both conditions present with a morphologically indistinguishable megaloblastic anemia. The inappropriate treatment of vitamin B_{12} deficiency with folic acid is to be avoided at all costs (see below). Apart from using biochemical assays to measure circulatory levels of the two vitamins and looking for an elevation of the biomarkers plasma homocysteine and MMA, further tests can also implicate vitamin B_{12} malabsorption, the most common type of severe vitamin B_{12} deficiency. These include the Schilling test and the detection of antibodies against either intrinsic factor or the parietal cells that manufacture it.

In practice, if vitamin B_{12} deficiency cannot be ruled out, many clinicians will treat patients with vitamin B_{12} if uncertain about the diagnosis. If this is followed by a reticulocyte response and complete disappearance of the anemia, it confirms a diagnosis of vitamin B_{12} deficiency. The appropriate treatment regimen can then be implemented (see later). If treatment with vitamin B_{12} does not result in improvement of the anemia then the patient is treated for folic acid deficiency, but only after vitamin B_{12} deficiency has been excluded by all means at the clinician's disposal.

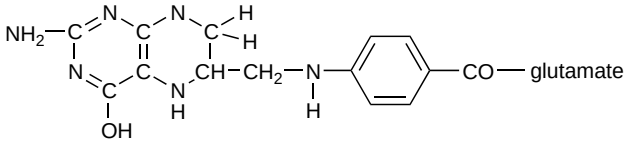
Causes of Folate Deficiency

Dietary

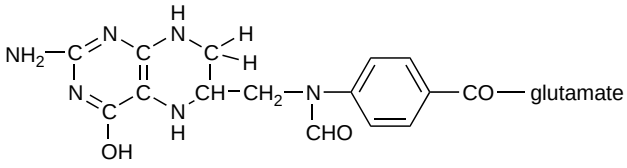
The most common cause of folate deficiency is undoubtedly due to inadequate dietary intake. The naturally occurring folates, unlike the synthetic form of the vitamin folic acid, are chemically unstable (**Figure 3**). The folate in food after harvesting or during processing is subject to deterioration.



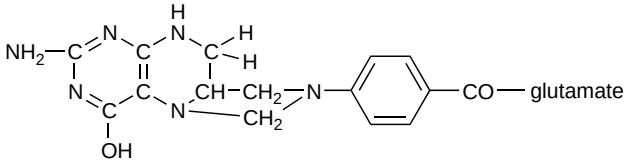
Folic acid



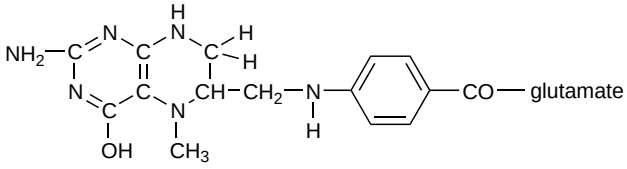
Tetrahydrofolate



10-Formyltetrahydrofolate



5,10-Methylenetetrahydrofolate



5-Methyltetrahydrofolate

Figure 3 The structure of synthetic folic acid and the naturally occurring forms of the vitamin.

Furthermore, folate can be lost from food during cooking. While the prevalence of overt folate deficiency in those on adequate mixed diets is relatively uncommon, it is now clear that many people have elevations of the biochemical marker homocysteine, which can be decreased by increased status of folic acid or folate, certainly where such diets are not fortified with folic acid. This has led many to conclude that more people than previously suspected are at increased risk of impaired folate function, such as might put them at increased risk of cardiovascular disease and other chronic diseases.

Malabsorption

Normally, the folate cofactors seem to be relatively bioavailable but in some circumstances malabsorption causing deficiency can occur, such as in celiac disease or tropical sprue.

Alcohol Abuse

Chronic alcoholics often have evidence of less than optimal folate status. It is unclear if this is due to inadequate dietary intake, some direct toxic effect of alcohol on folate metabolism in the bone marrow, or increased renal loss.

Drugs

Antifolate drugs such as methotrexate inhibit the enzyme dihydrofolate reductase, which is necessary for maintaining pyrimidine biosynthesis. A known side effect of methotrexate is megaloblastic anemia if given inappropriately.

Pregnancy

It is well established that many women are at risk of reduced folate status or even deficiency in their third trimester of pregnancy. This is probably due to an

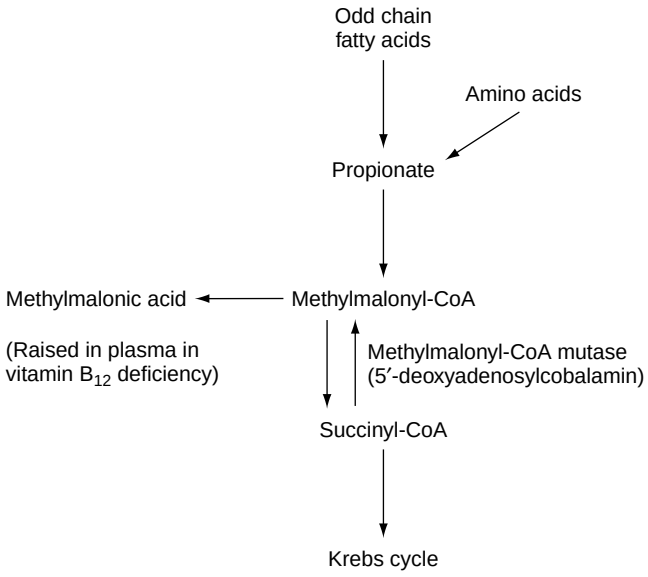


Figure 4 The role of vitamin B_{12} in the metabolism of propionates, odd chain fatty acids, and certain amino acids.

increased breakdown or catabolism of the vitamins associated with the rapid growth of the fetus/placenta rather than the transfer of maternal folate to the fetus, which is quantitatively small. Rapid cell division would result in an increased flux through tetrahydro and dihydrofolate forms of the vitamin, which are known to be the two most chemically unstable forms of the vitamin. In many countries folic acid is given in the latter stages of pregnancy to protect against this risk of megaloblastic anemia, the emergence of which very much depends upon the mother entering the pregnancy with poor stores. These events in the third trimester should not to be confused with the more recent incontrovertible evidence that the maternal periconceptional ingestion of folic acid prevents the majority of cases of spina bifida and other neural tube defects, which take place within the first 4 weeks postconception.

Causes of Vitamin B_{12} Deficiency

Dietary

No plant material can synthesize vitamin B_{12}. Apart from reports that some algae can synthesize vitamin B_{12} its origin in the food chain seems to be exclusively due to its biosynthesis by microorganisims. Thus, most vitamin B_{12} enters the human food chain from biosynthesis by microorganisms in herbivorous animals. Meat and products such as milk, cheese, or eggs introduce vitamin B_{12} into the human food chain. Chickens ingest food contaminated with microbes and introduce the vitamin via their meat and eggs. Vegetarians who have milk or eggs (lacto ovo vegetarians) as part of their diet and thus a source of some, albeit reduced, dietary vitamin B_{12} still have reduced vitamin B_{12} status. Yet other communities who for religious or other reasons are strict vegetarians (vegans) have no source of vitamin B_{12} and are at high risk of deficiency. This risk can be reduced in some of these communities where fermented food is eaten, in which bacteria have introduced vitamin B_{12}; also, it has been suggested that in some circumstances the food is contaminated by bacteria. However, vegans and in particular babies born to and weaned by strict vegan women are established to be at risk of vitamin B_{12} deficiency and such babies have been reported on several occasions to show the signs and symptoms of the neuropathy associated with such deficiency.

Malabsorption

The majority of cases of vitamin B_{12} deficiency, particularly severe deficiency, are due to malabsorption. While vitamin B_{12} is a water-soluble vitamin it is extremely large and only between 1 and 3% of any specific dose will cross the intestinal wall by diffusion. Thus, the normal physiological absorption of vitamin B_{12} is dependent upon it forming a complex with a glycoprotein that is secreted by the parietal cells of the stomach called intrinsic factor (IF). The most classical case where IF is deficient or absent is in the autoimmune pernicious anemia (PA). The most usual presentation of this condition is where antibodies are produced against the parietal cells rendering them incapable of secretion not only of IF but also hydrochloric acid (HCl) leading to

hypochlorhydria. Yet another form sees autoantibodies produced against IF itself rendering it incapable of binding vitamin B_{12} with consequent malabsorption of the IF-B_{12} complex in the ileum, where specific receptors are responsible for the active absorption of vitamin B_{12}. There is some evidence that many elderly people, perhaps even the majority, suffer to varying degrees from gastric atrophy. In such circumstances while they may still have an adequate supply of IF they lack a competent secretion of HCl. It is suggested that this acid and the accompanying action of pepsin is necessary to release vitamin B_{12} from the form in which it is present in food. The consequences would be varying degrees of malabsorption of food-bound vitamin B_{12} but an ability to absorb the free form of the vitamin present in foods fortified with vitamin B_{12} or from supplements. Other now infrequent causes of vitamin B_{12} malabsorption are resections of the stomach or removal of the ileum, the site of absorption.

Treatment of Folate Deficiency

If the deficiency is nutritional it is usually treated in the first instance with dietary supplements. In the past, daily supplements of $5.0\,mg\,day^{-1}$ have been used but more recent evidence suggests that such high levels would only be appropriate for the immediate treatment of an overt deficiency. More long-term treatment would recommend dietary changes to improve folate intake. In practice, to achieve effective changes is very difficult so the recommendation might be to improve intake through foods fortified with the synthetic form of the vitamin, namely folic acid, or the use of supplements of folic acid. In both of these instances the aim is to achieve a maximum increased intake via folic acid of $400\,\mu g\,day^{-1}$. Long-term ingestion of larger amounts are not recommended because of their ability to mask the diagnosis of vitamin B_{12} deficiencies (discussed above). Other causes of folate deficiency are treated by removing the cause, e.g., alcohol abuse.

Treatment of Vitamin B_{12} Deficiency

If the cause of the deficiency is nutritional, dietary supplements containing vitamin B_{12} (usually $2.0\,\mu g\,day^{-1)}$ should be taken.

If the deficiency is due to malabsorption, apart from where the vitamin B_{12} deficiency may be due to gastric atrophy, a dietary remedy is not effective because it will be malabsorbed. Such malabsorption conditions must be treated by regular monthly or bimonthly injections of $1000\,\mu g$ of vitamin B_{12} for life. If these injections are discontinued a return to the vitamin B_{12} deficiency state is inevitable. This is not only due to malabsorption of dietary vitamin B_{12} but also the 1 or $2\,\mu g$ of vitamin B_{12} secreted daily in the bile will be malabsorbed leading to negative balance. If the deficiency is due to gastric atrophy, supplements providing $500–1000\,\mu g\,day^{-1}$ can replete stores and maintain B_{12} status.

Inappropriate Treatment of Vitamin B_{12} Deficiency with Folic Acid

As mentioned above, megaloblastic anemia caused by folate deficiency should not be confused with that caused by vitamin B_{12} deficiency. The subsequent inappropriate treatment with folic acid could have serious and frequently irreversible consequences.

Historically, before there was a clearer understanding of folate metabolism, in vitamin B_{12} deficiency, synthetic folic acid (**Figure 3**) was used in many instances to treat vitamin B_{12} deficiency. This at first appeared to be successful in that continued treatment with folic acid largely reversed the anemia. However, it became clear that at best this masked the underlying concomitant development of the neuropathy, and some data suggest that folic acid exacerbated the neuropathy. In any event, the inappropriate treatment of vitamin B_{12} with folic acid masks the emergence of the anemia. Historically, it appears that about one-third of patients with vitamin B_{12} deficiency present with anemia, one-third with the neuropathy, and one-third with both. In addition, the signs and symptoms associated with the anemia are easily recognizable while those of the neuropathy are less so. The earlier features of the neuropathy such as loss of balance, tingling of the fingers, and mild ataxia can easily be confused with advancing years, which coincides with the usual development of vitamin B_{12} deficiency. This masking of the presence of anemia in vitamin B_{12}-deficient patients by folic acid therapy, potentially causes an early diagnosis to be missed in up to two-thirds of patients. When the vitamin B_{12} deficiency is eventually recognized through the onward and progressive development of the neuropathy some of the pathological features may have reached the stage at which they are irreversible.

See also: **Alcohol**: Disease Risk and Beneficial Effects. **Anemia**: Iron-Deficiency Anemia. **Cobalamins**. **Celiac Disease**. **Folic Acid**.

Further Reading

Bailey LB (1995) *Folate in Health and Disease*. New York: Marcel Dekker.
Chanarin I (1979) *Megaloblastic Anaemias*, 2nd edn. Oxford: Blackwell Scientific.
Chanarin I (1990) *The Megaloblastic Anaemias*, 3rd edn. Oxford: Blackwell Scientific.
Cuskelly CJ, McNulty H, and Scott JM (1996) Effect of increasing dietary folate on red-cell folate; implications for prevention of neural tube defects. *Lancet* **347**: 657–659.
Scott JM (1992) Folate-vitamin B_{12} interrelationships in the central nervous system. *Proceedings of the Nutrition Society* **51**: 219–224.
Scott JM (1997) Bioavailability of vitamin B_{12}. *European Journal of Clinical Nutrition* **51**(supplement 1): S49–S53.
Scott JM and Weir DG (1994) Folate vitamin B_{12} interrelationships. *Essays in Biochemistry* **28**: 63–72.
Scott JM and Weir DG (1996) Homocysteine and cardiovascular disease. *Quarterly Journal of Medicine* **89**: 561–563.
Wickramasinghe SM (1995) Megaloblastic anaemia. In *Baillière's Clinical Haematology*, vol. 8. London: Baillière Tindall.

ANTIOXIDANTS

Contents

Diet and Antioxidant Defense

I F F Benzie, The Hong Kong Polytechnic University, Hong Kong SAR, China
J J Strain, University of Ulster, Coleraine, UK

Introduction

Oxygen is an essential 'nutrient' for most organisms. Paradoxically, however, oxygen damages key biological sites. This has led to oxygen being referred to as a double-edged sword. The beneficial side of oxygen is that it permits energy-efficient catabolism of fuel by acting as the ultimate electron acceptor within mitochondria. During aerobic respiration, an oxygen atom accepts two electrons, forming (with hydrogen) harmless water. The less friendly side of oxygen is the unavoidable and continuous production of partially reduced oxygen intermediates within the body. These 'free radicals' (reactive oxygen species; ROS) are more reactive than ground-state oxygen and cause oxidative changes to carbohydrate, DNA, lipid, and protein. Such changes can affect the structures and functions of macromolecules, organelles, cells, and biological systems. This induces oxidant stress if allowed to proceed unopposed.

The human body is generally well equipped with an array of 'antioxidative' strategies to protect against the damaging effects of ROS. Our endogenous antioxidants are inadequate, however, as we are unable to synthesize at least two important antioxidant compounds, vitamin C and vitamin E. Ingestion of these, and perhaps other, antioxidants is needed to augment our defenses and prevent or minimize oxidative damage. In this article, the causes and consequences of oxidant stress and the types and action of antioxidants will be described, the source and role of dietary antioxidants will be discussed, and current evidence relating to dietary antioxidants and human health will be briefly reviewed.

Oxidant Stress

Oxidant, or oxidative, stress is a pro-oxidant shift in the oxidant–antioxidant balance caused by a relative or absolute deficiency of antioxidants (**Figure 1**). A pro-oxidant shift promotes damaging oxidative changes to important cellular constituents, and this may, in turn, lead to cellular dysfunction and, ultimately, to aging, disability, and disease.

Molecular oxygen is relatively unreactive in its ground state. However, molecular oxygen can be reduced in several ways within the body to produce more reactive species (**Table 1**). These species include radical and nonradical forms of oxygen, some of which contain nitrogen or chlorine. A 'free radical' is capable of independent existence and has a single (unpaired) electron in an orbital. Electrons

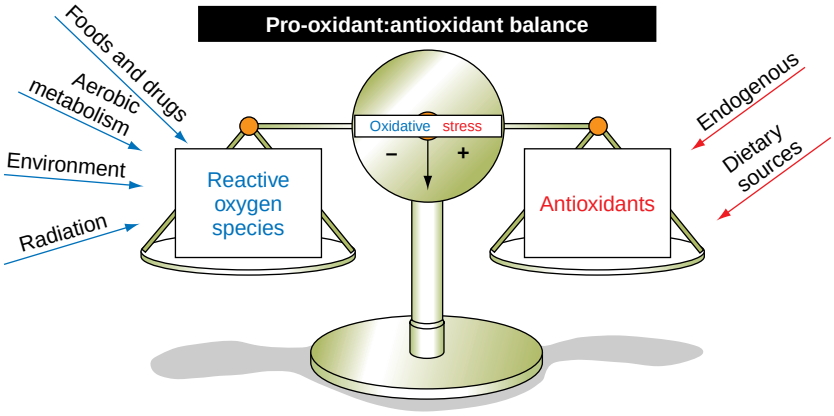


Figure 1 Antioxidant defenses balance reactive oxygen species load and oppose oxidant stress.

stabilize as pairs with opposing spins within an orbital. An unpaired electron seeks a partner for stability, and this increases the reactivity of the radical. A partner electron can be obtained by removing ('abstracting') an electron from another species or co-reactant. The result of this interaction may be either quenching by reduction (electron addition) of the radical with the production of a new radical by oxidation (electron loss) of the reductive ('antioxidant') co-reactant or quenching of two radicals if the co-reactant is also a radical (one quenched by reduction (electron addition) and one by oxidation (electron removal)).

Free radicals produced *in vivo* include superoxide, the hydroxyl radical, nitric oxide, oxygen-centered organic radicals such as peroxyl and alkoxyl radicals, and sulfur-centered thiyl radicals. Other oxygen-containing reactive species that are not radicals are also formed. These include hydrogen peroxide, peroxynitrite, and hypochlorous acid. While these are not radical species, they are actually or potentially damaging oxidants. The collective term ROS is often used to describe both radical and nonradical species.

What Causes Oxidant Stress?

Oxidant stress is caused by the damaging action of ROS. There are two main routes of production of ROS in the body: one is deliberate and useful; the other is accidental but unavoidable (**Figure 2**). Deliberate production of ROS is seen, for example,

Table 1 Reactive oxygen species found *in vivo* in general order of reactivity (from lowest to highest)

Name of species	*Sign/formula*	*Radical (R) or nonradical (NR)*	*Comment*
Molecular oxygen	O_2	R	Biradical, with two unpaired electrons; these are in parallel spins, and this limits reactivity
Nitric oxide	$NO^{\cdot}$	R	Important to maintain normal vasomotor tone
Superoxide	$O_2^{\cdot -}$	R	Single electron reduction product of O_2; large amounts produced *in vivo*
Peroxyl	$ROO^{\cdot}$	R	R is often a carbon of an unsaturated fatty acid
Singlet oxygen	$^1\Delta_g O_2$ $^1\Sigma_g^+ O_2$	NR	'Energized' nonradical forms of molecular oxygen; one unpaired electron is transferred to the same orbital as the other unpaired electron
Hydrogen peroxide	H_2O_2	NR	Small, uncharged, freely diffusible ROS formed by dismutation of superoxide
Hydroperoxyl	$HOO^{\cdot}$	R	Protonated, more reactive form of superoxide; formed at sites of low pH
Alkoxyl	$RO^{\cdot}$	R	R is carbon in a carbon-centered radical formed by peroxidation of unsaturated fatty acid
Hypochlorous	HOCl	NR	Formed in activated phagocytes to aid in microbial killing
Peroxynitrite	$HNOO^-$	NR	Highly reactive product of nitric oxide and superoxide
Hydroxyl	$^{\cdot}OH$	R	Fiercely, indiscriminately reactive radical

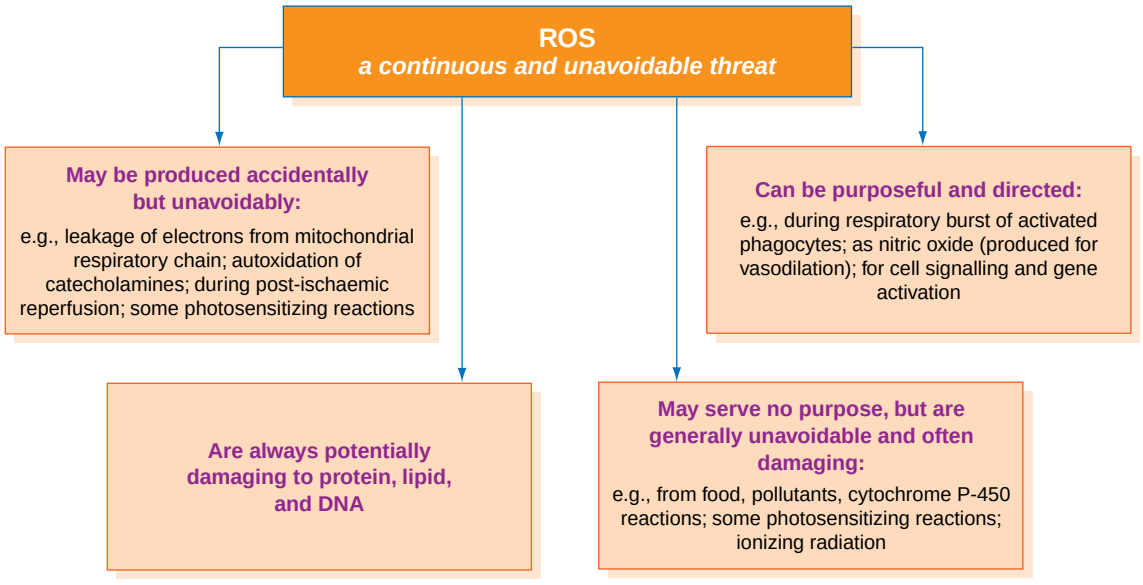


Figure 2 Sources of reactive oxygen species found *in vivo*.

during the respiratory burst of activated phagocytic white cells (macrophages, neutrophils, and monocytes). Activated phagocytes produce large amounts of superoxide and hypochlorous acid for microbial killing. The ROS nitric oxide is produced constitutively and inducibly, is a powerful vasodilator, and is vital for the maintenance of normal blood pressure. Nitric oxide also decreases platelet aggregability, decreasing the likelihood of the blood clotting within the circulation. Hydrogen peroxide is produced enzymatically from superoxide by the action of the superoxide dismutases (SODs) and is recognized increasingly as playing a central role in cell signalling and gene activation. Nonetheless, while some ROS are physiologically useful, they are damaging if they accumulate in excess as a result of, for example, acute or chronic inflammation or ischaemia.

Accidental, but unavoidable, production of ROS occurs during the passage of electrons along the mitochondrial electron transport chain. Leakage of electrons from the chain leads to the single-electron reduction of oxygen, with the consequent formation of superoxide. This can be regarded as a normal, but undesirable, by-product of aerobic metabolism. Around 1–3% of electrons entering the respiratory chain are estimated to end up in superoxide, and this results in a large daily ROS load *in vivo*. If anything increases oxygen use, such as exercise, then more ROS will be formed, and oxidant stress may increase owing to a pro-oxidant shift. Significant amounts of ROS are also produced during the metabolism of drugs and pollutants by the mixed-function cytochrome P-450 oxidase (phase I) detoxifying system and as a consequence of the transformation of xanthine dehydrogenase to its truncated oxidase form, which occurs as a result of ischemia. This causes a flood of superoxide to be formed when the oxygen supply is restored. In addition, if free iron is present (as may happen in iron overload, acute intravascular hemolysis, or cell injury), there is a risk of a cycle of ROS production via iron-catalyzed 'autoxidation' of various constituents in biological fluids, including ascorbic acid, catecholamines, dopamine, hemoglobin, flavins, and thiol compounds such as cysteine or homocysteine. Preformed reactive species in food further contribute to the oxidant load of the body, and ROS are also produced by pathological processes and agents such as chronic inflammation, infection, ionizing radiation, and cigarette smoke. Breathing oxygen-enriched air results in enhanced production of ROS within the lungs, and various toxins and drugs, such as aflatoxin, acetominophen, carbon tetrachloride, chloroform, and ethanol, produce reactive radical species during their metabolism or detoxification and excretion by the liver or kidneys. Clearly, all body tissues are exposed to ROS on a regular or even constant basis. However, sites of particularly high ROS loads within the human body include the mitochondria, the eyes, the skin, areas of cell damage, inflammation, and post-ischemic reperfusion, the liver, the lungs (especially if oxygen-enriched air is breathed), and the brain.

What does Oxidant Stress Cause?

A sudden and large increase in ROS load can overwhelm local antioxidant defenses and induce severe

oxidant stress, with cell damage, cell death, and subsequent organ failure. However, less dramatic chronic oxidant stress may lead to depletion of defenses and accumulation of damage and ultimately cause physiological dysfunction and pathological change resulting in disability and disease. This is because oxidant stress causes oxidative changes to DNA, lipid, and protein. These changes lead in turn to DNA breaks, mutagenesis, changed phenotypic expression, membrane disruption, mitochondrial dysfunction, adenosine triphosphate depletion, intracellular accumulation of non-degradable oxidized proteins, increased atherogenicity of low-density lipoproteins, and crosslinking of proteins with subsequent loss of function of specialized protein structures, for example, enzymes, receptors, and the crystallins of the ocular lens. In addition, the aldehydic degradation products of oxidized polyunsaturated fatty acids (PUFAs) are carcinogenic and cytotoxic. Increased oxidant stress can also trigger apoptosis, or programed cell death, through a changed redox balance, damage to membrane ion-transport channels, and increased intracellular calcium levels (**Figure 3**).

Oxidant stress, through its effects on key biological sites and structures, is implicated in chronic noncommunicable diseases such as coronary heart disease, cancer, cataract, dementia, and stroke (**Figure 4**). Oxidant stress is also thought to be a key player in the aging process itself. A cause-and-effect relationship between oxidant stress and aging and disease has not been confirmed, however, and it is very unlikely that oxidant stress is the sole cause of aging and chronic degenerative disease. Nonetheless, there is evidence that oxidant stress contributes substantially to age-related physiological decline and pathological changes. Consequently, if it is accepted that oxidant stress is associated with aging and degenerative disease, then opposing oxidant stress by increasing antioxidant defense offers a potentially effective means of delaying the deleterious effects of aging, decreasing the risk of chronic disease, and achieving functional longevity. For this reason, there has been great interest in recent years in the source, action, and potential health benefits of dietary antioxidants.

Antioxidant Defense

An antioxidant can be described in simple terms as anything that can delay or prevent oxidation of a susceptible substrate. Our antioxidant system is complex, however, and consists of various intracellular and extracellular, endogenous and exogenous, and aqueous and lipid-soluble components that act in concert to prevent ROS formation (preventative antioxidants), destroy or inactivate ROS that are formed (scavenging and enzymatic antioxidants), and terminate chains of ROS-initiated peroxidation of biological substrates (chain-breaking antioxidants). In addition, metals and minerals (such as selenium, copper, and zinc) that are key components of antioxidant enzymes are often referred to as antioxidants.

There are many biological and dietary constituents that show 'antioxidant' properties *in vitro*. For an antioxidant to have a physiological role, however, certain criteria must be met.

1. The antioxidant must be able to react with ROS found at the site(s) in the body where the putative antioxidant is found.
2. Upon interacting with a ROS, the putative antioxidant must not be transformed into a more reactive species than the original ROS.
3. The antioxidant must be found in sufficient quantity at the site of its presumed action *in vivo* for it to make an appreciable contribution to defense at that site: if its concentration is very low, there must be some way of continuously recycling or resupplying the putative antioxidant.

Antioxidants Found Within the Human Body

The structures of the human body are exposed continuously to a variety of ROS. Humans have evolved an effective antioxidant system to defend against these damaging agents. Different sites of the body contain different antioxidants or contain the same antioxidants but in different amounts. Differences are likely to reflect the different requirements and characteristics of these sites.

Human plasma and other biological fluids are generally rich in scavenging and chain-breaking antioxidants, including vitamin C (ascorbic acid) and 'vitamin E.' Vitamin E is the name given to a group of eight lipid-soluble tocopherols and tocotrienols. In the human diet, γ-tocopherol is the main form of vitamin E, but the predominant form in human plasma is α-tocopherol. Bilirubin, uric acid, glutathione, flavonoids, and carotenoids also have antioxidant activity and are found in cells and/or plasma. Scavenging and chain-breaking antioxidants found *in vivo* are derived overall from both endogenous and exogenous sources. Cells contain, in addition, antioxidant enzymes, the SODs, glutathione peroxidase, and catalase. The transition

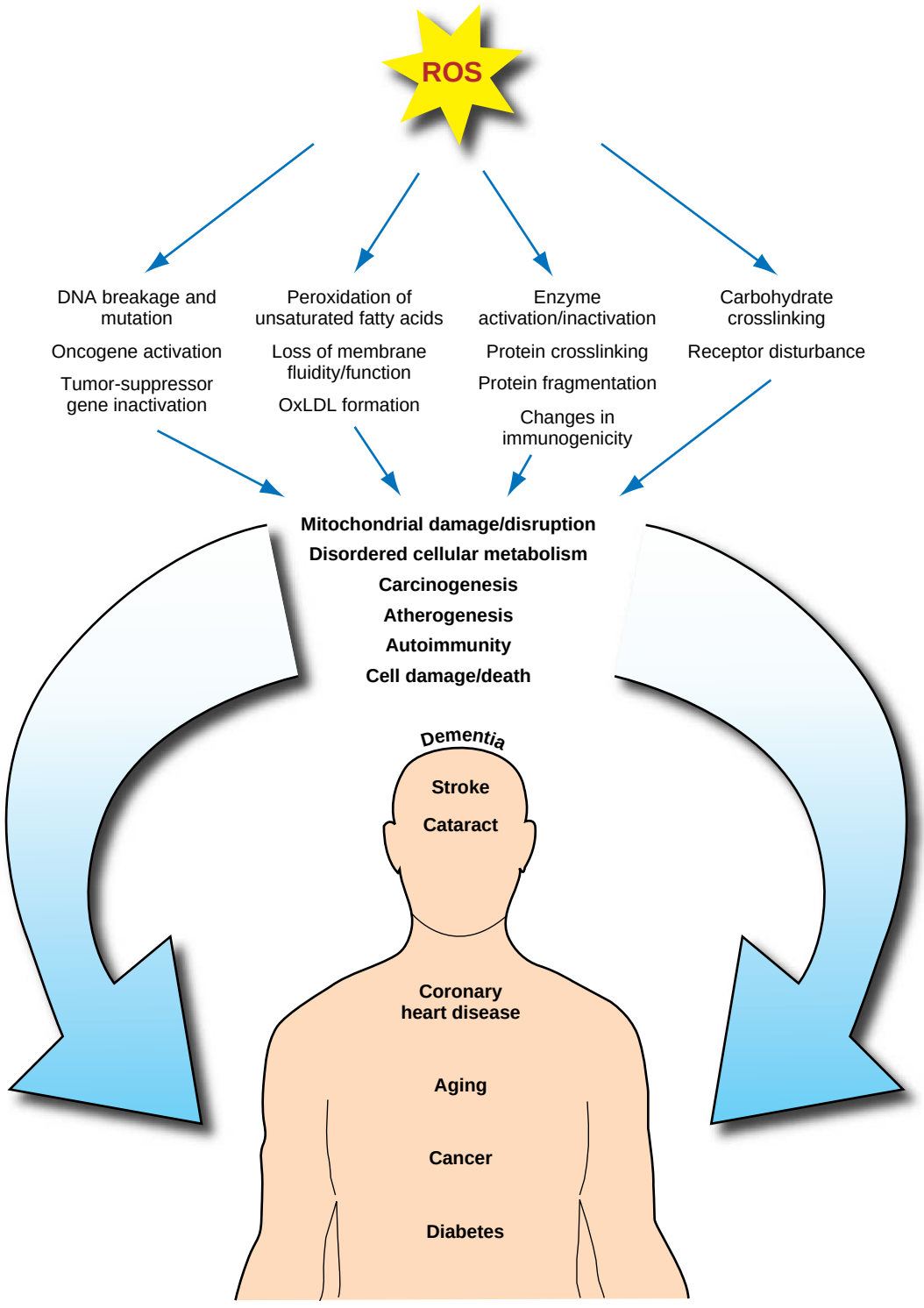


Figure 3 Possible involvement of reactive oxygen species (ROS) in ageing and chronic degenerative disease. OxLDL, oxidized low density lipoprotein.

metals iron and copper, which can degrade pre-existing peroxides and form highly reactive ROS, are kept out of the peroxidation equation by being tightly bound to, or incorporated within, specific proteins such as transferrin and ferritin (for iron) and caeruloplasmin (for copper). These proteins are regarded as preventive antioxidants. Caeruloplasmin ferroxidase activity is also important for

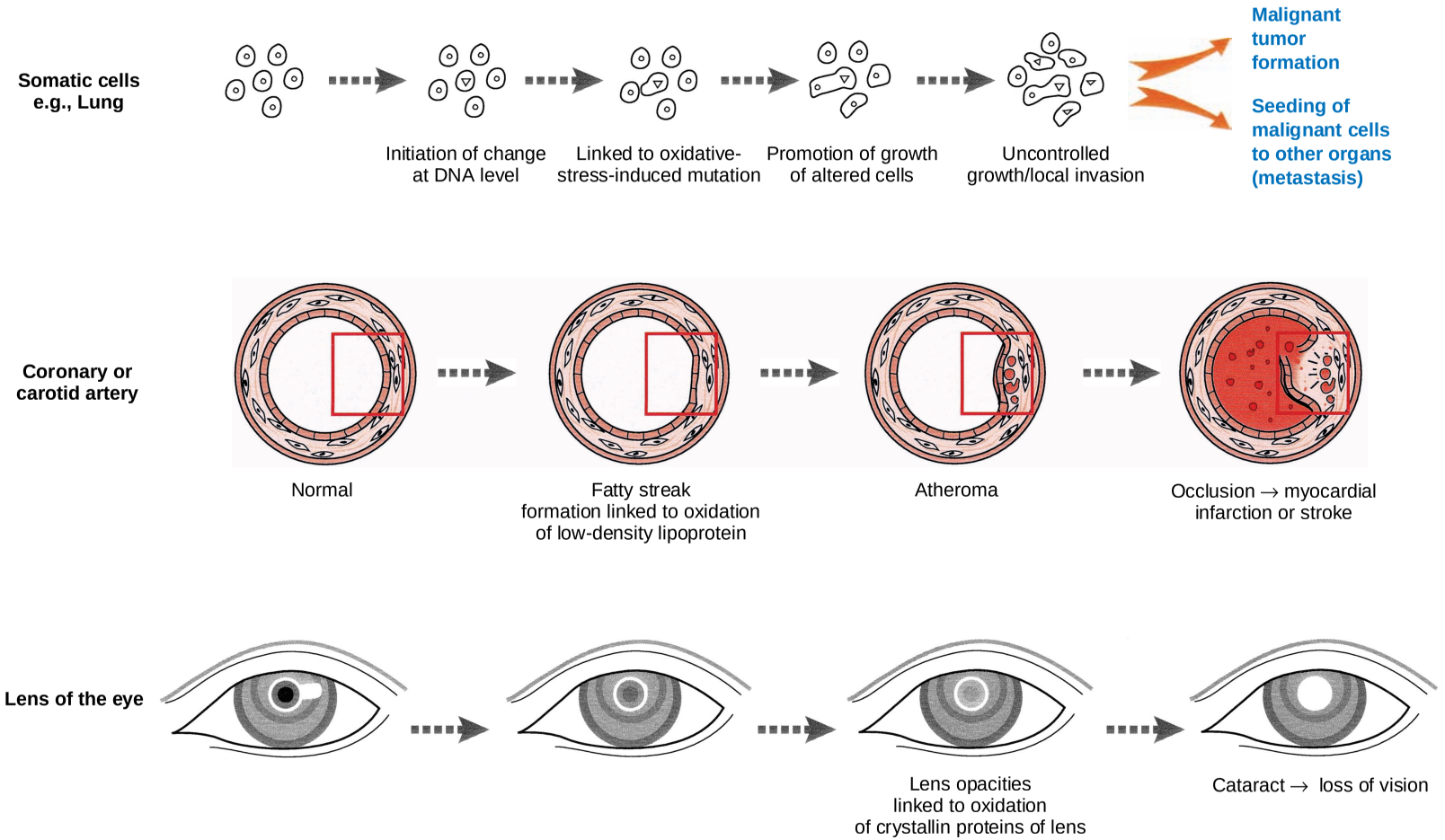


Figure 4 Antioxidants may help prevent the long-term oxidative changes to DNA, lipid, and protein that lead to age-related disease.

Table 2 Types of antioxidants

Physical barriers *prevent* ROS generation or ROS access to important biological sites; e.g., UV filters, cell membranes
Chemical traps or sinks 'absorb' energy and electrons and *quench* ROS; e.g., carotenoids, anthocyanidins
Catalytic systems *neutralize or divert* ROS, e.g., the antioxidant enzymes superoxide dismutase, catalase, and glutathione peroxidase
Binding and redox inactivation of metal ions *prevent generation* of ROS by inhibiting the Haber–Weiss reaction; e.g., ferritin, caeruloplasmin, catechins
Sacrificial and chain-breaking antioxidants *scavenge and destroy* ROS; e.g., ascorbic acid (vitamin C), tocopherols (vitamin E), uric acid, glutathione, flavonoids

ROS, reactive oxygen species.

the non-ROS-producing route of ferrous (Fe(II)) to ferric (Fe(III)) oxidation and for incorporating released iron into ferritin for 'safe' iron storage. Haptoglobin (which binds released hemoglobin), hemopexin (which binds released hem), and albumin (which binds transition-metal ions and localizes or absorbs their oxidative effects) can also be regarded as antioxidants in that they protect against metal-ion-catalyzed redox reactions that may produce ROS. An overview of the major types of antioxidants within the body and their interactions is given in **Table 2** and **Figure 5**.

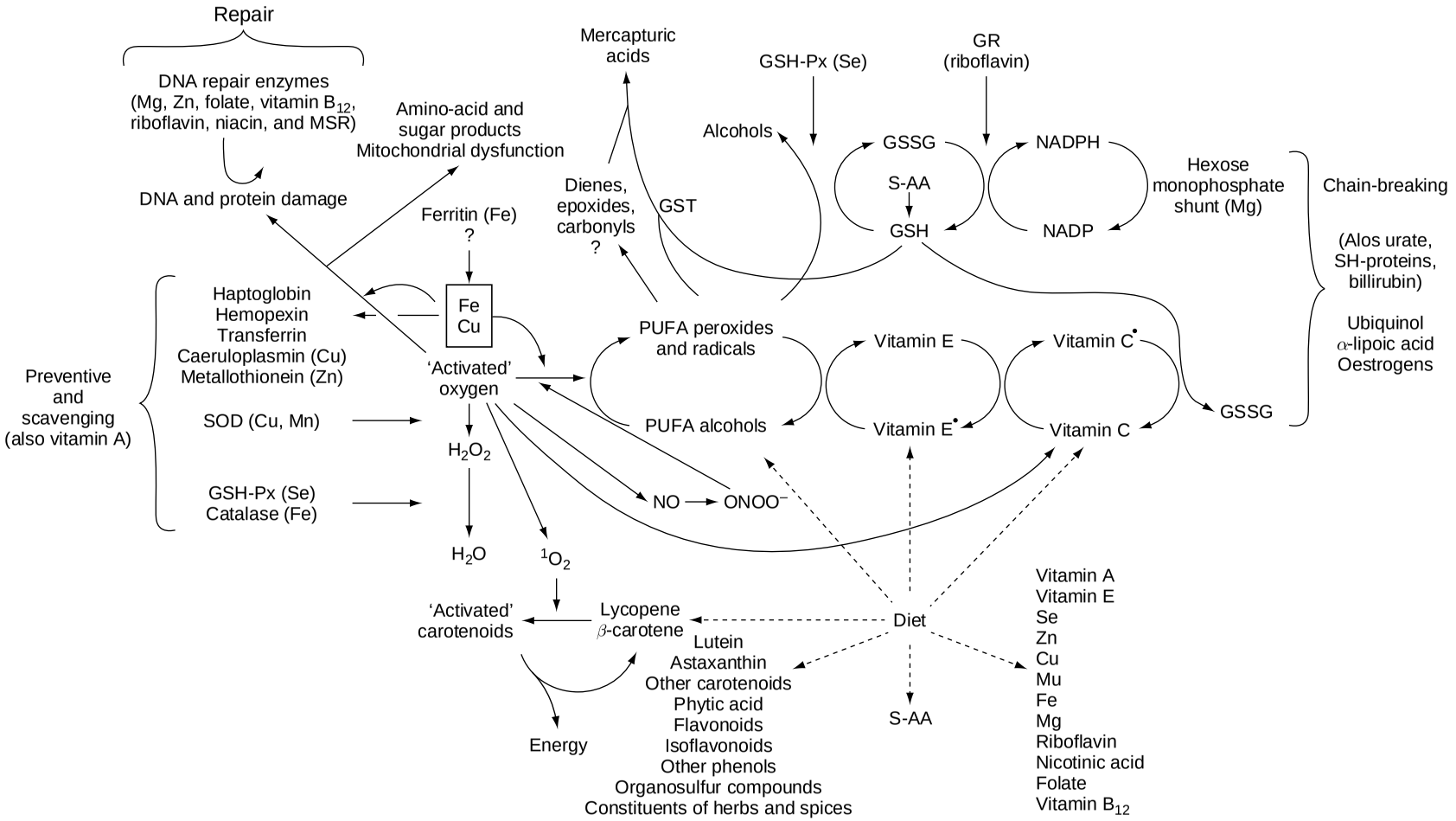
Dietary Antioxidants

The human endogenous antioxidant system is impressive but incomplete. Regular and adequate dietary intakes of (largely) plant-based antioxidants, most notably vitamin C, vitamin E, and folic acid, are needed. Fresh fruits and vegetables are rich in antioxidants (**Figure 6**), and epidemiological evidence of protection by diets rich in fruits and vegetables is strong. To decrease the risk of cancer of various sites, five or more servings per day of fruits and vegetables are recommended. However, it is not known whether it is one, some, or all antioxidant(s) that are the key protective agents in these foods. Furthermore, it may be that antioxidants are simple co-travellers with other, as yet unidentified, components of antioxidant-rich foods. Perhaps antioxidants are not 'magic bullets' but rather 'magic markers' of protective elements. Nonetheless, the US recommended daily intakes (RDIs) for vitamin C and vitamin E were increased in 2000 in recognition of the strong evidence that regular high intakes of these antioxidant vitamins are associated with a decreased risk of chronic disease and with lower all-cause mortality.

To date, research on dietary antioxidant micronutrients has concentrated mainly on vitamin C and vitamin E. This is likely to be because humans have an undoubted requirement for these antioxidants, which we cannot synthesize and must obtain in regular adequate amounts from food. However, there are a plethora of other dietary antioxidants. Some or all of the thousands of carotenoids, flavonoids, and phenolics found in plant-based foods, herbs, and beverages, such as teas and wines, may also be important for human health, although there are currently no RDIs for these. Furthermore, while there are recommended intakes for vitamin C, vitamin E, and folic acid, these vary among countries, and there is currently no agreement as regards the 'optimal' intake for health. In addition, there is growing evidence that other dietary constituents with antioxidant properties, such as quercetin and catechins (found in teas, wines, apples, and onions), lycopene, lutein, and zeaxanthin (found in tomatoes, spinach, and herbs) contribute to human health. Zinc (found especially in lamb, leafy and root vegetables, and shellfish) and selenium (found especially in beef, cereals, nuts, and fish) are incorporated into the antioxidant enzymes SOD and glutathione peroxidase, and the elements are themselves sometimes referred to as antioxidants.

The levels of ascorbic acid, α-tocopherol, folic acid, carotenoids, and flavonoids within the body are maintained by dietary intake. While the role and importance of dietary antioxidants are currently unclear, antioxidant defense can be modulated by increasing or decreasing the intake of foods containing these antioxidants. There are a number of reasons for recommending dietary changes in preference to supplementation for achieving increased antioxidant status, as follows.

1. It is not clear which antioxidants confer protection.
2. The hierarchy of protection may vary depending on body conditions.
3. A cooperative mix of antioxidants is likely to be more effective than an increased intake of one antioxidant.
4. Antioxidants, including vitamin A, β-carotene, vitamin C, selenium, and copper, can be harmful in large doses or under certain circumstances.
5. Antioxidant status is likely to be affected by the overall composition of the diet, e.g., the fatty-acid and phytochemical mix.
6. The iron status of the body, environmental conditions, and lifestyle undoubtedly affect antioxidant demand.



Fe Cu, transition-metal-catalyzed oxidant damage to biomolecules; ?, biological relevance.

Figure 5 The integrated antioxidant defense system comprises both endogenous and dietary-derived antioxidants. GR, glutathione reductase (EC1.6.4.2); GSH, reduced glutathione; GSH-Px, glutathione peroxidase (EC1.11.1.9); GSSG, oxidized glutathione; GST, glutathione-S-transferase (EC 2.5.1.18); MSR, methionine sulfoxide reductase (EC1.8.4.5); NADPH and, NADP, are, respectively, the reduced and oxidized forms of the co-factor nicotinamide adenine dinucleotide phosphate; PUFA, polyunsaturated fatty acid; S-AA, sulfur amino-acids; SH, sulfydryl; SOD, superoxide dismutase (EC1.15.1.1).

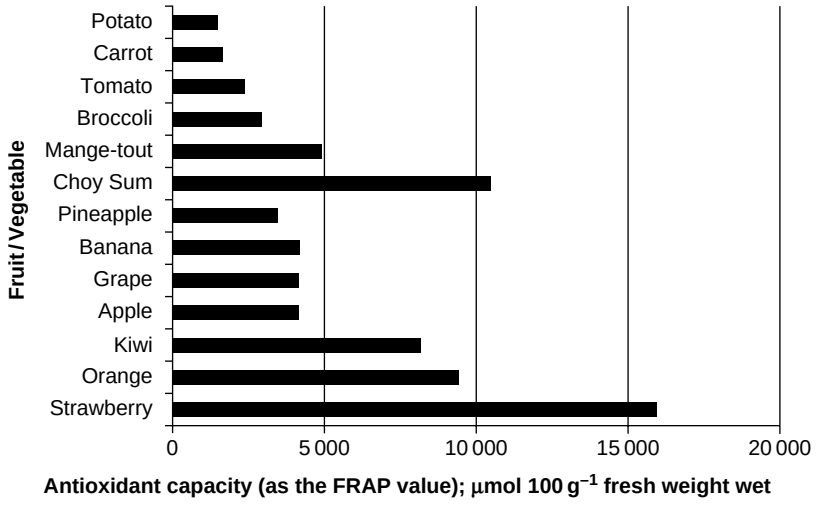


Figure 6 Antioxidant capacity varies among different fruits and vegetables. FRAP, Ferric Reducing/Anti-oxidant Power.

Antioxidant defense, therefore, is likely to be optimized through a balanced intake of a variety of antioxidants from natural sources rather than by pharmacological doses of one or a few antioxidants.

Dietary Recommendations for Increased Antioxidant Defense

Dietary recommendations that would result in increased antioxidant defense are not inconsistent with accepted recommendations for healthy eating. The recommendation to increase the consumption of plant-based foods and beverages is one that is widely perceived as health promoting, and the consistent and strong epidemiological links between high fruit and vegetable intake and the greater life expectancy seen in various groups worldwide whose diet is high in plant-based foods indicate that more emphasis should be given to this particular dietary recommendation. Vitamin C, vitamin E, various carotenoids, flavonoids, isoflavonoids, phenolic acids, organosulfur compounds, folic acid, copper, zinc, and selenium are all important for antioxidant defense, and these are found in plant-based foods and beverages such as fruits, vegetables, nuts, seeds, teas, herbs, and wines. Dietary strategies for health promotion should be directed towards optimizing the consumption of these items.

It is recommended generally that at least five servings of fruits and vegetables are eaten each day. This recommendation is based on a wealth of epidemiological evidence that, overall, indicates that 30–40% of all cancers can be prevented by diet. However, it is estimated that most individuals in developed countries eat less than half this amount of fruits and vegetables, and intake by people in developing nations is often very low. Furthermore, the antioxidant contents (both of individual antioxidants and in total) of foods vary widely among different food items and even within the same food item, depending on storage, processing, and cooking method. In addition, the issues of bioavailability and distribution must be considered, and it is of interest to see where dietary antioxidants accumulate (**Table 3**). Vitamin C is absorbed well at low doses and is concentrated in nucleated cells and in the eyes, but relative absorption within the gastrointestinal tract decreases as dose ingested increases. Of the eight isomers of 'vitamin E,' α-tocopherol and γ-tocopherol are distributed around the body and are found in various sites, including skin and adipose tissue. Vitamin E protects lipid systems, such as membranes and lipoproteins. While α-tocopherol is by far the predominant form in human lipophilic structures, there is limited information on the bioavailabilities and roles of the other isomers. Gastrointestinal absorption of catechins (a type of flavonoid found in high quantity in tea) is very low, and, although it has been shown that plasma antioxidant capacity increases after ingesting catechin-rich green tea, catechins appear to be excreted via the urine fairly rapidly. Some are likely to be taken up by membranes and cells, although this is not clear, but most of the flavonoids ingested are likely to remain within the gastrointestinal tract. However, this does not necessarily mean that they have no role to play in antioxidant defense, as the unabsorbed antioxidants may provide local defense to the gut lining (**Figure 7**).

With regard to plasma and intracellular distributions of dietary antioxidants, if it is confirmed that

Table 3 Dietary antioxidants: source, bioavailability, and concentrations in human plasma

	Dietary source	*Bioavailability*	*Concentration*	*Comment*
Ascorbic acid (vitamin C)	Fruits and vegetables, particularly strawberries, citrus, kiwi, Brussels sprouts, and cauliflower	100% at low doses (<100 mg) decreasing to <15% at >10 g	25–80 μmol l^{-1}	Unstable at neutral pH, concentrated in cells and the eye
'Vitamin E' (in humans mainly α-tocopherol)	Green leafy vegetables, e.g., spinach, nuts, seeds, especially wheatgerm, vegetable oils, especially sunflower	10–95%, but limited hepatic uptake of absorbed tocopherol	15–40 μmol l^{-1} (depending on vitamin supply and lipid levels)	Major tocopherol in diet is γ form, but α form is preferentially taken up by human liver
Carotenoids (hundreds)	Orange/red fruits and vegetables (carrot, tomato, apricot, melon, yam), green leafy vegetables	Unclear, dose and form dependent, probably <15%	Very low (<1 μmol l^{-1})	Lutein and zeaxanthin are concentrated in macula region of the eye
Flavonoids (enormous range of different types)	Berries, apples, onions, tea, red wine, some herbs (parsley, thyme), citrus fruits, grapes, cherries	Most poorly absorbed, quercetin absorption 20–50%, catechins <2%, dependent on form and dose	No data for most, likely <3 μmol l^{-1} in total	Quercetin and catechins may be most relevant to humans health as intake is relatively high, there is some absorption, possible gastrointestinal-tract protection by unabsorbed flavonoids

increasing defense by dietary means is desirable, frequent small doses of antioxidant-rich food may be the most effective way to achieve this. Furthermore, ingestion of those foods with the highest antioxidant contents may be the most cost-effective strategy. For example, it has been estimated that around 100 mg of ascorbic acid (meeting the recently revised US RDI for vitamin C) is supplied by one orange, a few strawberries, one kiwi fruit, two slices of pineapple, or a handful of raw cauliflower or uncooked spinach leaves. Interestingly, apples, bananas, pears, and plums, which are probably the most commonly consumed fruits in Western countries, are very low in vitamin C. However, these, and other, fruits contain a significant amount of antioxidant power, which is conferred by a variety of other scavenging and chain-breaking antioxidants (**Figure 6**).

Dietary Antioxidants and Human Health

Plants produce a very impressive array of antioxidant compounds, including carotenoids, flavonoids, cinnamic acids, benzoic acids, folic acid, ascorbic acid, tocopherols, and tocotrienols, and plant-based foods are our major source of dietary antioxidants. Antioxidant compounds are concentrated in the oxidation-prone sites of the plant, such as the oxygen-producing chloroplast and the PUFA-rich seeds and oils. Plants make antioxidants to protect their own structures from oxidant stress, and plants increase antioxidant synthesis at times of additional need and when environmental conditions are particularly harsh.

Humans also can upregulate the synthesis of endogenous antioxidants, but this facility is very limited. For example, production of the antioxidant enzyme SOD is increased with regular exercise, presumably as an adaptation to the increased ROS load resulting from higher oxygen use. However, an increase in other endogenous antioxidants, such as bilirubin and uric acid, is associated with disease, not with improved health. Increasing the antioxidant status of the body by purposefully increasing the production of these antioxidants, therefore, is not a realistic strategy. However, the concept that increased antioxidant intake leads to increased antioxidant defense, conferring increased protection against oxidant stress and, thereby, decreasing the risk of disease, is a simple and attractive one. Antioxidant defense can be modulated by varying the dietary intake of foods rich in natural antioxidants. It has been shown that following ingestion of an antioxidant-rich food, drink, or herb the antioxidant

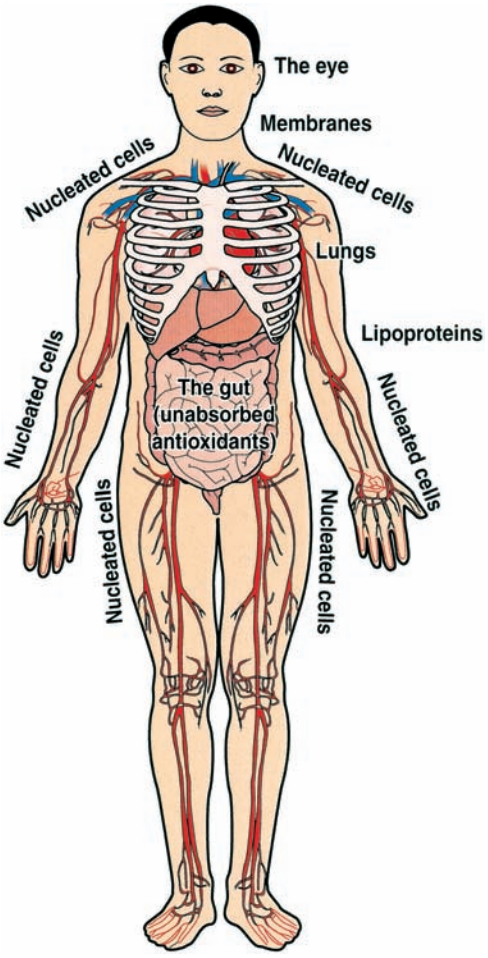


Figure 7 Dietary antioxidants are absorbed and distributed to various sites within the human body.

status of the plasma does indeed increase. The question remains, however, as to whether increasing the antioxidant defense of the body by dietary means, while achievable, is a desirable strategy to promote human health and well-being.

There are many age-related disorders that, in theory at least, may be prevented or delayed by increased antioxidant defense. These disorders include arthritis, cancer, coronary heart disease, cataract, dementia, hypertension, macular degeneration, the metabolic complications of diabetes mellitus, and stroke. The rationale for prevention of disease by antioxidants is based on the following facts.

1. Epidemiological evidence shows that a high intake of antioxidant-rich foods, and in some cases antioxidant supplements, is associated with a lower risk of these diseases.
2. Experimental evidence shows that oxidation of cells and structures (such as low-density lipoprotein, DNA, membranes, proteins, and mitochondria) is increased in individuals suffering from these disorders.
3. Experimental evidence shows that antioxidants protect protein, lipid, and DNA from oxidative damage.
4. Experimental evidence shows that biomarkers of oxidative damage to key structures are ameliorated by an increased intake of dietary antioxidants.

However, the following cautionary statements must be noted.

1. While there is a large body of observational evidence supporting a protective effect of dietary antioxidants, it has been suggested that the importance of this has been overstated, and recent studies are less supportive.
2. While phenomenological evidence is strong that oxidative damage does occur in aging and in chronic degenerative diseases, cause-and-effect relationships have not been confirmed.
3. While experimental evidence is quite strong, studies have generally been performed *in vitro* using very high concentrations of antioxidants, making their physiological relevance unclear.
4. Evidence from intervention trials is of variable quality and conflicting; animal studies have shown positive results but have often used very high does of antioxidants, and the relevance to human health is unclear; large human intervention trials completed to date, such as the α-Tocopherol β-Carotene Cancer Prevention Study, Gruppo Italiano per Io Studio Delia Sopravivenza nell'Infarto Miocardico, the Heart Protection Study, the Heart Outcomes Prevention Evaluation, and the Primary Prevention Project, have been largely disappointing in that they have not shown the expected benefits. These studies are summarized in **Table 4**.

Overall, observational data are supportive of beneficial effects of diets rich in antioxidants (**Figure 8**), and intervention trials have often used high-risk groups or individuals with established disease (**Table 4**). In addition, intervention trials have generally used antioxidant supplements (usually vitamin C or vitamin E) rather than antioxidant mixtures or antioxidant-rich foods. Therefore, while observational data support a role for antioxidant-rich food in health promotion, whether or not it is the antioxidants in the food that are responsible for the benefit remains to be confirmed.

Table 4 Summary of completed large antioxidant intervention trials

Name and aim of study	*Subjects*	*Supplementation*	*Results/comments*	*Reference*
The α-tocopherol β-carotene cancer prevention study (ATBC); primary prevention	29 133 high-risk subjects (male smokers, 50–69 years old; average of 20 cigarettes day^{-1} smoked for 36 years)	50 mg day^{-1} of α-tocopherol (synthetic) or 20 mg day^{-1} of β-carotene, or both, or placebo, for 5–8 years (median follow-up 6.1 years)	Supplementation with α-tocopherol had no effect on lung-cancer incidence; no evidence of interaction between α-tocopherol and β-carotene; significant increase in fatal coronary events in men with history of heart disease, and 18% increase in lung-cancer incidence in β-carotene supplemented men; follow-up showed significant (32%) decrease in risk of prostate cancer and nonsignificant (8%) decrease in fatal coronary heart disease in α-tocopherol supplemented subjects	The Alpha-tocopherol, Beta-carotene Cancer Prevention Study Group (1994) *Journal of the National Cancer Institute* 88: 1560–1570 Pryor WA (2000) *Free Radical Biology and Medicine* 28: 141–164
The Cambridge Heart Antioxidant Study (CHAOS); secondary prevention	2002 high-risk subjects (angiographically proven cardiovascular disease)	d-α-tocopherol 400 IU or 800 IU per day (median follow-up 510 days; results on different doses combined into one treatment group)	Significant decrease (77%) in non-fatal MI in treatment group; slight nonsignificant (18%) increase in fatal cardiovascular events in treatment group, but most (21/27) in noncompliant subjects	Stephens NG *et al.* (1996) *Lancet* 347: 781–786
Gruppo Italiano per lo Studio Delia Sopravivenza nell'Infarto Miocardico (GISSI); secondary prevention	11 324 survivors of MI within previous 3 months of enrollment	α-tocopherol (synthetic) 300 mg day^{-1} or omega 3 fatty acids (0.9 g day^{-1}) or both or neither for 3.5 years; subjects continued on normal medication (50% on statins)	Nonsignificant decrease (11%) in primary endpoints (death, nonfatal MI, and stroke) with vitamin E; high dropout rate (25%); open label study	Marchioli R (1999) *Lancet* 354: 447–455

Continued

Table 4 Continued

Name and aim of study	*Subjects*	*Supplementation*	*Results/comments*	*Reference*
Primary Prevention Project (PPP)	4495 subjects with $\geq$1 major cardiovascular risk factor	α-tocopherol (synthetic) 300 mg day^{-1} or aspirin 100 mg day^{-1} or both or neither, follow-up average of 3.6 years	Vitamin E had no significant effect on any primary endpoint (cardiovascular death, MI, or stroke)	Primary Prevention Project (2001) *Lancet* 357: 89–95
The Heart Outcomes Prevention Evaluation Study (HOPE); secondary prevention	9541 subjects (2545 women, 6996 men) aged $\geq$55 years, high-risk (cardiovascular disease or diabetes and $\geq$1 other CVD risk factor)	400 IU day^{-1} vitamin E (from 'natural sources') or Ramipril (angiotensin converting enzyme inhibitor) or both or neither; follow-up for 4–6 years	No significant effect of vitamin E on any primary endpoint (MI, stroke, or cardiovascular death)	The Heart Outcomes Prevention Evaluation Study Investigators (2000) *New England Journal of Medicine* 342: 154–160
Heart Protection Study	20 536 subjects, high-risk (diabetes, peripheral vascular disease or coronary heart disease)	Daily antioxidant cocktail (600 IU dl-α-tocopherol, 250 mg vitamin C, 20 mg β-carotene) or placebo for 5 years	No significant differences in hemorrhagic stroke or all-cause mortality between treatment and placebo groups	The Heart Protection Study Collaborative Group (2002) *Lancet* 360: 23–32

MI, myocardial infarction; CVD, cardiovascular disease.

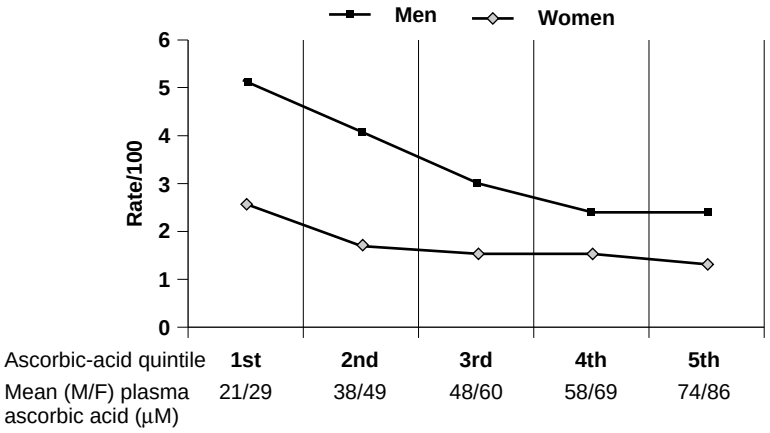


Figure 8 Age-adjusted rates of all-cause mortality by sex-specific ascorbic-acid quintiles in 8860 British men (squares) and 10 636 British women (diamonds).

Summary and Concluding Remarks

Our diet contains a multitude of antioxidants that we cannot synthesize, and most are plant-based. The available evidence supports a role for antioxidant-rich foods in the promotion of health, although it is not yet clear how many antioxidants and how much of each are needed to achieve an optimal status of antioxidant defense and minimize disease risk. Nor is it clear whether the benefit is of a threshold type or whether it continues to increase with the amount of antioxidant ingested. It is also not yet known whether those dietary antioxidants for which there is no absolute known requirement play a significant role in human antioxidant defense and health or whether they are merely coincidental co-travellers with other, as yet unknown, antioxidant or nonantioxidant dietary constituents that have beneficial effects. A reasonable recommendation is to eat a variety of antioxidant-rich foods on a regular basis. This is likely to be beneficial and is not associated with any harmful effects. However, further study is needed before firm conclusions can be drawn regarding the long-term health benefits of increasing antioxidant defense per se, whether through food or supplements. The challenge in nutritional and biomedical science remains to develop tools that will allow the measurement of biomarkers of functional and nutritional status and to clarify human requirements for dietary antioxidants, the goal being the design of nutritional strategies to promote health and functional longevity.

Acknowledgments

The authors thank The Hong Kong Polytechnic University and The University of Ulster for financially supporting this work.

See also: **Antioxidants**: Observational Studies; Intervention Studies. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Carotenoids**: Chemistry, Sources and Physiology; Epidemiology of Health Effects. **Copper**. **Folic Acid**. **Fruits and Vegetables**. **Iron**. **Riboflavin**. **Selenium**. **Vitamin E**: Physiology and Health Effects. **Zinc**: Physiology.

Further Reading

Ames BN and Wakimoto P (2002) Are vitamin and mineral deficiencies a major cancer risk? *Nature Reviews* **2**: 694–704.

Asplund K (2002) Antioxidant vitamins in the prevention of cardiovascular disease: a systematic review. *Journal of Internal Medicine* **251**: 372–392.

Benzie IFF (2003) Evolution of dietary antioxidants. *Journal of Comparative Biochemistry and Physiology* **136**: 113–126.

Block G, Norkus E, Hudes M, Mandel S, and Helzlsouer K (2001) Which plasma antioxidants are most related to fruit and vegetable consumption? *American Journal of Epidemiology* **154**: 1113–1118.

Chisholm GM and Steinberg D (2000) The oxidative modification hypothesis of atherogenesis: an overview. *Free Radical Biology and Medicine* **28**: 1815–1826.

Clarkson PM and Thompson HS (2000) Antioxidants: what role do they play in physical activity and health? *American Journal of Clinical Nutrition* **72**(Supplement 2): 637S–646S.

Cooke MS, Evans MD, Mistry N, and Lunec J (2002) Role of dietary antioxidants in the prevention of *in vivo* oxidative DNA damage. *Nutrition Research Reviews* **15**: 19–41.

Halliwell B and Gutteridge JMC (1999) *Free Radicals in Biology and Medicine*, 3rd edn. Oxford: Clarendon Press.

Khaw KT, Bingham S, Welch A *et al.* (2001) Relation between plasma ascorbic acid and mortality in men and women in EPIC-Norfolk prospective study: a prospective population study. *Lancet* **357**: 657–663.

Levine M, Wang Y, Padayatty SJ, and Morrow J (2001) A new recommended dietary allowance of vitamin C for healthy young women. *Proceedings of the National Academy of Science USA* **98**: 9842–9846.

Lindsay DG and Clifford MN (eds.) (2000) Critical reviews produced within the EU Concerted Action 'Nutritional

Enhancement of Plant-Based Food in European Trade' (NEODIET). *Journal of the Science of Food Agriculture* **80**: 793–1137.

McCall MR and Frei B (1999) Can antioxidant vitamins materially reduce oxidative damage in humans? *Free Radical Biology and Medicine* **26**: 1034–1053.

Polidori MC, Stahl W, Eichler O, Niestrol I, and Sies H (2001) Profiles of antioxidants in human plasma. *Free Radical Biology and Medicine* **30**: 456–462.

Pryor WA (2000) Vitamin E and heart disease: basic science to clinical intervention trials. *Free Radical Biology and Medicine* **28**: 141–164.

Szeto YT, Tomlinson B, and Benzie IFF (2002) Total antioxidant and ascorbic acid content of fresh fruits and vegetables: implications for dietary planning and food preservation. *British Journal of Nutrition* **87**: 55–59.

World Cancer Research Fund and the American Institute for Cancer Research (1997) *Food, Nutrition and the Prevention of Cancer: A Global Perspective*. Washington, DC: American Institute for Cancer Research.

Observational Studies

I F F Benzie, The Hong Kong Polytechnic University, Hong Kong, China

Introduction

The study of temporal and geographical variation in disease prevalence in association with differences in environment, diet, and lifestyle helps identify possible factors that may modulate the risk of disease within and across populations. As such, observational epidemiology is a powerful, albeit blunt, tool that serves to inform and guide experimental studies and intervention trials. In the case of dietary antioxidants and chronic age-related disease, there is a logical biochemical rationale for the protective effect of antioxidants, and there is strong, and consistent observational evidence supportive of this. The way in which dietary antioxidants are believed to act is described in a separate chapter. In this chapter, observational evidence relating to dietary antioxidants and the risk of disease states is discussed.

Epidemiology: Setting the Scene

The risk of developing a disease can be increased by exposure to a disease-promoting factor or decreased by a protective factor. In terms of antioxidants, high risk is generally assumed to be associated with low intakes, plasma levels, or tissue concentrations of antioxidants. Epidemiological studies often express results in terms of the relative risk (RR) of mortality or disease. The RR is generally given as the mean and 95% confidence interval (CI). In general, an RR of 0.80 indicates an average reduction in risk of 20%; however, RR values must be interpreted with caution and the CI must be considered. If the CI spans 1.0, the RR is not statistically significant, regardless of its magnitude.

Different approaches are used in observational epidemiology. Cross-cultural studies compare standardized mortality rates (from all causes or from a specific disease) or disease prevalence and the factor of interest ('exposure variable') in different populations within or between countries. These can be regarded as 'snapshot' observational surveys. Case–control studies compare the factor of interest in people who have a disease (the cases) with that in those who do not (the controls). Prospective trials are longitudinal studies of apparently disease-free subjects whose health is monitored over years or decades; the exposure variable of interest is compared, retrospectively, between those who develop the disease of interest and those who do not.

The Observational View of Dietary Antioxidants

Cancer and cardiovascular disease (CVD) are the two leading causes of death worldwide, diabetes mellitus is reaching epidemic proportions, and dementia and maculopathy are largely untreatable irreversible disorders that are increasingly common in our aging population. The prevalence and standardized mortality rates of these diseases vary considerably between and within populations. Mortality from CVD varies more than 10-fold amongst different populations, and incidences of specific cancers vary 20-fold or more across the globe. This enormous variation highlights the multiple factors at play in the etiology of chronic age-related diseases. These factors include smoking habit, socioeconomic status, exposure to infectious agents, cholesterol levels, certain genetic factors, and diet. Dietary factors have long been known to play an important role in determining disease risk. Indeed, 30–40% of overall cancer risk is reported to be diet-related, and there is a wealth of compelling observational evidence that a lower risk of cancer, CVD, diabetes, and other chronic age-related disorders is associated with diets that are rich in antioxidants.

In terms of dietary antioxidants, the major research focus to date has been on the water-soluble

vitamin C (ascorbic acid) and the lipophilic vitamin E. 'Vitamin E' is a group of eight lipid-soluble tocopherols and tocotrienols; however, the most widely studied form to date is α-tocopherol because it is the most abundant form in human plasma. Neither vitamin C nor vitamin E can be synthesized by humans, so they must be obtained in the diet, most coming from plant-based foods and oils. Deficiency of either of these vitamins is rare and can be prevented by the daily intake of a few milligrams of each. However, an adequate intake to prevent simple deficiency is unlikely to be sufficient for optimal health. Based on observational findings and experimental evidence that vitamin C and vitamin E protect key biological sites from oxidative damage *in vitro*, it has been suggested that there is a threshold of intake or plasma concentration for these antioxidants that confers minimum disease risk and promotes optimal health. The strength of the data supporting the health benefits of increased intakes of these vitamins was acknowledged in the US Food and Nutrition Board recommendation in 2000 to increase the daily intake of vitamin C to 75 mg day^{-1} for women and 90 mg day^{-1} for men and to increase that of vitamin E to 15 mg day^{-1} for both men and women. However, whether these new recommended intakes are 'optimal' is a contentious issue.

Supplementation trials with vitamin C or vitamin E have not to date shown the expected health benefits. The reasons for this mismatch between observational and supplementation data are not yet known, but some suggested reasons are outlined in **Table 1**. Nonetheless, despite the apparent lack of effect in supplementation trials, the variety and strength of observational findings, backed by a solid body of *in vitro* biochemical data, keep dietary antioxidants in the research spotlight, and in recent years attention has focused on the influence of 'non-nutrient' dietary antioxidants, such as polyphenolic compounds, in addition to the effects of vitamin C and vitamin E. The current evidence for vitamin C, vitamin E, and non-nutrient dietary antioxidants in relation to the major causes of morbidity and mortality in developed countries is discussed briefly below.

Vitamin C

Low plasma ascorbic-acid concentrations have been reported to be strongly predictive of mortality, particularly in men. Results of a prospective trial in the UK (EPIC-Norfolk Prospective Study), in which 19 496 men and women aged 45–79 years were followed for 4 years, showed that men and women in the highest quintile of plasma ascorbic-acid concentration in samples collected within 1 year of entry into the study had significantly ($p < 0.0001$) lower all-cause mortality than those in the lowest quintile. Highest-quintile concentrations of plasma ascorbic acid (mean $\pm$ standard deviation) were $72.6 \pm 11.5\,\mu mol\,l^{-1}$ for men and $85.1 \pm 13.7\,\mu mol\,l^{-1}$ for women; lowest quintiles were $20.8 \pm 7.1\,\mu mol\,l^{-1}$ and $30.3 \pm 10.1\,\mu mol\,l^{-1}$, respectively, for men and women. In men and women in the highest quintile, RRs (CI) for all-cause mortality were, respectively, 0.48 (0.33–0.70) and 0.50 (0.32–0.81), relative to those in the lowest quintile. Mortality from ischemic heart disease was also significantly ($p < 0.001$) lower in the highest quintiles: for men the RR (CI) was 0.32 (0.15–0.75), and for women it was 0.07 (0.01–0.67). The relationship held for CVD and cancer in men ($p < 0.001$), but no significant difference in cancer mortality was seen in women, and CVD rates in women were affected less than those in men. The mean ascorbic-acid level in

Table 1 Possible reasons for the conflict in results between observational epidemiological and supplementation trials

- Antioxidants are likely to work in cooperation with each other; more of one may increase the need for another
- The action of an antioxidant within a heterogeneous food matrix may be different from that in pure supplemental form
- A high intake of antioxidants may help to promote health when taken regularly over decades but may have little discernable effect over a few months or years
- A high intake of antioxidants may slow or even prevent some of the deleterious age-related changes that lead to chronic disease, but antioxidants are unlikely to reverse established pathological changes
- Benefits of increased antioxidant intake may be seen only in those with marginal or depleted antioxidant status at baseline
- The effect of antioxidant supplementation may be seen only in subgroups of the study population, e.g., in those individuals with certain single-nucleotide polymorphisms
- The key players may not be the most widely studied antioxidants; for example, γ-tocopherol, rather than α-tocopherol, may play an important role in modulation of cancer risk but has been little studied to date
- Antioxidants can act as pro-oxidants under certain conditions, and the net effect of a dietary antioxidant may well depend on dose and conditions at its site of action
- Antioxidant action per se may not be the key mechanism of action of protection; for example, immunomodulatory, anti-inflammatory, anti-proliferative, and pro-apoptotic effects of dietary agents (antioxidants or otherwise) may be more relevant to overall effects in terms of disease risk

each quintile in women was around 10 $\mu mol\,l^{-1}$ higher than that in men. Interestingly, the relationship between ascorbic-acid concentration and mortality was continuous throughout the range of plasma ascorbic-acid concentrations found. It was estimated that a 20 $\mu mol\,l^{-1}$ increase in plasma ascorbic acid (achievable by one or two additional servings of fruit and vegetables each day) was associated with a 20% decrease in all-cause mortality, independent of age, blood pressure, cholesterol, smoking habit, and diabetes. Interestingly, also, mortality was not associated with supplement use, indicating that dietary sources of vitamin C are crucial.

In the Third National Health and Nutrition Examination Survey (NHANES III) in the USA, the plasma ascorbic-acid concentrations of 7658 men and women were not found to be independently associated with a history of cardiovascular disease in participants who reported no alcohol consumption; however, in 3497 participants who consumed alcohol a significantly lower prevalence of pre-existing angina was found in those with high plasma ascorbic acid (>56 $\mu mol\,l^{-1}$) than in those with 'low to marginal' levels (<22 $\mu mol\,l^{-1}$). No significant association was seen between plasma ascorbic acid and previous myocardial infarction or stroke in this cross-sectional survey. In the NHANES II prospective study, a 43% decrease in mortality was associated with higher plasma ascorbic-acid levels in more than 3000 men followed for up to 16 years. Plasma ascorbic-acid levels in the highest and lowest quartiles in this study were more than 73 $\mu mol\,l^{-1}$ and less than 28.4 $\mu mol\,l^{-1}$, respectively. The corresponding values in women were again higher, at more than 85 $\mu mol\,l^{-1}$ and less than 39.7 $\mu mol\,l^{-1}$, respectively, and no significant relationship between plasma ascorbic-acid levels and mortality was seen in women.

The Kuopio IHD (ischemic heart disease) Risk Factor Study followed 1605 men for 5 years and reported an RR (CI) of 0.11 (0.04–0.30) for acute myocardial infarction in those men with higher plasma ascorbic-acid concentrations. The Medical Research Council Trial of Assessment and Management of Older People in the Community, a prospective trial in the UK of 1214 elderly subjects followed for a median of 4.4 years, showed that those in the highest quintile of plasma ascorbic-acid level (>66 $\mu mol\,l^{-1}$) at entry had less than half the risk of dying in the follow-up period compared with those in the lowest quintile (plasma ascorbic-acid level of <17 $\mu mol\,l^{-1}$). Data on men and women were not analyzed separately, but there were fewer men (27%) in the highest quintile of ascorbic-acid level. No relationship was seen between mortality and plasma levels or intake of β-carotene or lipid-standardized α-tocopherol. Interestingly, while the relationship between mortality and the concentration of plasma ascorbic acid was strong, there was no significant association between mortality and estimated dietary intake of vitamin C. This may reflect the difficulty in obtaining accurate dietary information, but it also suggests that different individuals may well need different intakes to achieve certain plasma levels of ascorbic acid.

There have been many case–control and cohort studies performed in Europe and the USA and published in the past 15 years, and some data from Asia have been gathered. In most case–control studies no significant relationship has been demonstrated between intake and/or plasma ascorbic-acid levels and the risk of cardiovascular events; however, the combination of findings from individual studies is revealing. In a detailed analysis of 11 cohort studies comparing high and low intakes of ascorbic acid in 50 000 subjects overall, with 2148 CVD events during follow-up, a Peto's Odds Ratio (95% CI) of 0.89 (0.79–0.99) for CVD was calculated, indicating a modest reduction in risk associated with a high intake of vitamin C. In an analysis of five cohort studies comparing high and low plasma ascorbic-acid levels, involving 13 018 subjects overall with 543 CVD events during follow-up, a Peto's Odds Ratio for CVD of 0.58 (0.47–0.72) was calculated. This was interpreted as showing high plasma ascorbic-acid levels to be a powerful predictor of freedom from CVD during follow-up.

The relationship between antioxidant-rich diets and protection from cancer is strong and clear; however, the influences of individual antioxidants are difficult to isolate. Cancer risk increases as total calorie intake increases, and this confounds prospective and retrospective dietary studies. Cancer causes many biochemical changes, and cancer treatment is harsh, and this confounds the results of studies comparing antioxidant levels in plasma in cases and controls unless the samples were collected and analysed before cancer developed (which may be a considerable time before diagnosis). Currently, the evidence for a cancer-opposing effect of high intakes or plasma concentrations of ascorbic acid is conflicting. To date, the strongest evidence of a role for vitamin C in lowering cancer risk is in relation to cancer of the stomach, with a low intake of vitamin C being associated with a two-to-three-fold increase in the risk of stomach cancer. A Spanish study showed a 69% lower risk of stomach cancer in those in the highest quintile of vitamin C intake, and low levels of ascorbic acid in gastric juice are

found in patients with chronic atrophic gastritis or *Helicobacter pylori* infection, both of which are associated with a greatly increased risk of gastric cancer. Whether the decrease in ascorbic acid is directly related to the development of gastric cancer is not known, but it is known that ascorbic acid inactivates carcinogenic nitrosamines within the stomach. There is also evidence of a decreased risk of cancer of the mouth, pharynx, pancreas, lung, cervix, and breast in association with increased vitamin C intake, though not all studies find this. It has been estimated that if the diets of postmenopausal women were enriched with vitamin C, a 16% decrease in breast cancer in these women would result. No significant association was reported between vitamin C intake and the incidence of ovarian cancer in 16 years of follow-up of 80 326 women in the Nurses' Health Study. A study of 100 children with brain tumours showed a three-fold increase in risk in those children whose mothers had a low intake of vitamin C during pregnancy, suggesting that the dietary intake of vitamin C by pregnant women may help to determine the future cancer risk in their children. In a prospective study of 19 496 British men and women aged 45–79 years and followed for 4 years (the EPIC (European Prospective Investigation into Cancer and Nutrition) study), the RR (CI) of mortality from cancer for a 20 $\mu mol\,l^{-1}$ increase in plasma ascorbic acid was 0.85 (0.74–0.99). In men there was a strong and continuous decrease in cancer risk with increasing plasma ascorbic-acid concentrations, with an RR (CI) in the highest quintile relative to the lowest quintile of 0.47 (0.27–0.88); i.e., the average risk in those with the highest plasma ascorbic-acid concentrations was less than half that of those in the lowest quintile. In women the decrease in RR did not reach statistical significance. The NHANES II study reported that men in the lowest quartile of ascorbic-acid level had a 62% higher risk of death from cancer during 12 years of follow-up than those in the highest quartile. However, this relationship was not seen in women. Of possible relevance here is the common finding in these studies that men, in general, had lower ascorbic-acid levels than women.

Vitamin C is concentrated in ocular tissues and fluids, particularly in the anterior aspect (cornea and lens). A case–control study in Spain reported a 64% reduction in the risk of cataract ($p < 0.0001$) in those with a plasma ascorbic-acid concentration of more than 49 $\mu mol\,l^{-1}$; however, no significant association with the dietary intake of vitamin C was seen. In a case–control study in the Netherlands, the prevalence of age-related maculopathy was reported to be twice as high in those with low antioxidant intake (from fruits and vegetables); however the data on vitamin C intake or plasma levels and maculopathy are conflicting. Lipid-soluble antioxidants, especially zeaxanthin and lutein (dietary-derived carotenoids that are highly concentrated in the lipid-rich fovea), may be more relevant in this condition than water-soluble vitamin C.

High plasma concentrations of ascorbic acid are reportedly associated with better memory performance, and lower plasma and cerebrospinal-fluid concentrations of ascorbic acid were found in patients with Alzheimer's disease than in non-demented controls. Individuals who took vitamin C supplements were reported to have a lower prevalence of Alzheimer's disease on follow-up after 4.3 years. However, not all studies have shown a significant association between vitamin C intake or plasma levels and cognitive decline or dementia.

Vitamin E

An extensive review noted that the data in relation to a connection between vitamin E and CVD risk are strong and convincing. In a large cross-cultural European (WHO/MONICA) observational study, a strong inverse relationship ($r^2 = 0.60$, $p < 0.005$) was found between plasma concentrations of lipid-standardized vitamin E and mortality from coronary heart disease (CHD) across 16 populations. In a detailed analysis, this relationship was found to be stronger than that between mortality and plasma cholesterol, smoking, and diastolic blood pressure combined ($r^2 = 0.44$, $p < 0.02$). In a case–control study in Scotland, patients with previously undiagnosed angina pectoris were found to have lower levels of plasma lipid-standardized vitamin E than controls. After adjustment for classical CHD risk factors, men in the highest quintile of lipid-standardized vitamin E level had an almost three-fold decrease in risk. Confusingly, some studies have reported a higher CVD risk in individuals with increased plasma total vitamin E, but these results are probably driven by elevated blood lipids. Vitamin E is carried in the lipoproteins, and it is important to lipid standardize plasma concentrations of this, and other, lipophilic antioxidants.

The Nurses' Health Study (women) and the Health Professionals Study (men) were initiated in the USA in 1980 and 1986, respectively, and recruited almost 200 000 subjects. It was found that women at the high end of vitamin E intake from diet alone had a small and non-significant decrease in CVD risk; however, those women in the highest quintile of vitamin E intake (more than 100 $IU\,day^{-1}$) had an RR (CI) for CVD of 0.54 (0.36–0.82). It should be noted that an intake of

100 IU day^{-1} of vitamin E is achievable only by using supplements: intake from food alone is unlikely be more than 15 IU day^{-1}. In this study, protection against CVD was seen only in those women who had taken vitamin E supplements for at least 2 years. In the Health Professionals Follow-up Study the findings were very similar. Supplemental, but not dietary, intake of vitamin E (more than 100 IU day^{-1}) in men was associated with a significant decrease in CVD risk, averaging over 30%, but again the effect was seen only if supplements had been taken for at least 2 years. A separate study in the USA of more than 11 000 elderly subjects showed that the use of vitamin E supplements was associated with a significant decrease in the risk of heart disease (RR (CI) of 0.53 (0.34–0.84)). The results also showed a significant decrease in all-cause mortality in users of vitamin E supplements and suggested that long-term use was beneficial. A study in Finland of more than 5000 men and women showed an average of 40% lower CVD risk in the highest versus the lowest tertile of vitamin E intake. Interestingly, most (97%) of subjects in this study did not take supplements, indicating that the protective effect was due to higher intake from food. An inverse association between dietary vitamin E intake and heart disease was also seen in The Women's Iowa Health Study, which involved almost 35 000 postmenopausal women. In this study, an RR (CI) of 0.38 (0.18–0.80) for CVD mortality was seen in women in the highest quintile relative to those in the lowest quintile of vitamin E intake from food alone. However, the Medical Research Council Trial of Assessment and Management of Older People in the Community (UK) found no relation between either dietary intake of vitamin E or plasma concentration of lipid-standardized α-tocopherol and all-cause mortality or death from CVD in 1214 elderly participants followed for a median of 4.4 years.

Cancer is caused by mutations in key genes. Anything that protects DNA will, in theory, help to prevent cancer-causing mutations. Lipid peroxide degradation products are reported to be carcinogenic, and vitamin E opposes lipid peroxidation, possibly conferring indirect protection against cancer. Furthermore, by interacting with reactive species elsewhere in the cell, vitamin E may spare other antioxidants, thereby also indirectly protecting DNA. Vitamin E reportedly protects against cancer of the upper digestive tract, skin cancer, including melanoma, and lung cancer. Follow-up analysis of the placebo group of the Finnish ATBC (Alpha Tocopherol Beta Carotene) study (incidentally, a study that showed no protection against lung cancer in a high-risk group supplemented with α-tocopherol and/or β-carotene) showed that there was a 36% higher incidence of lung cancer in those in the lowest quartile than in those in the highest quartile of diet-derived vitamin E. Vitamin E from dietary sources, but not supplements, has been reported to confer modest protection against breast cancer; however, as with vitamin C, no association was seen between vitamin E intake and the risk of ovarian cancer in the Nurses' Health Study follow-up.

Colorectal cancer is the second and third most common cancer in men and women, respectively. Dietary influences on the risk of colorectal cancer are currently unclear, and, based on recent findings of large prospective trials, it has been suggested that the influence of antioxidant-rich foods has been overstated. Nonetheless, there is evidence that vitamin E may be protective. In a case–control study in the USA of almost 1000 cases of rectal cancer, the risk was reported to be modestly decreased in women with a high vitamin E intake, but not in men. In a meta-analysis of five prospective nested case–control studies, there was a marginal decrease in the incidence of colorectal cancer in those in the highest quartile of plasma α-tocopherol, although no significant inverse association was seen in any of the studies individually. In the Iowa Women's Health Study, women with the highest risk of colon cancer were those with the lowest intake of vitamin E, although the relationship was significant only in women aged 55–59 years.

In addition to its antioxidant properties, vitamin E is reported to have immune-boosting and anti-inflammatory effects and to inhibit cell division, all of which may help explain the reported relationship between low intake or plasma concentrations of vitamin E and increased risk of various cancers. Currently, there is much interest in vitamin E in association with selenium in relation to the prevention of prostate, lung, and colon cancer. Indeed, the combination of vitamin E with other antioxidant micronutrients may be much more important than vitamin E alone. Furthermore, the different members of the vitamin E family may play cooperative or complementary roles in modulating the risk of disease. In terms of cancer prevention, γ-tocopherol is attracting much interest. Dietary intake of this form of vitamin E can be up to three times higher than that of α-tocopherol. Corn, canola, palm, soya bean, and peanut oils contain more γ-tocopherol than α-tocopherol. Despite a higher intake, however, our plasma levels of γ-tocopherol are only around 10% of those of α-tocopherol, owing to preferential placement of the α-form into very low-density lipoproteins. Interestingly, higher tissue levels of

α-tocopherol are reportedly found in animals fed both α-tocopherol and γ-tocopherol than in animals fed α-tocopherol alone, suggesting that intake of both forms may enhance the enrichment of tissues. Furthermore, the lower reaches of the gastrointestinal tract may contain high levels of γ-tocopherol, and this may help to destroy fecal mutagens. None of the epidemiological studies to date have estimated the dietary intake of γ-tocopherol, but the few studies that have measured plasma levels of γ-tocopherol show interesting results. In a nested case–control study of 6000 Japanese men, there was a statistically significant inverse relationship between the risk of cancer of the upper digestive tract and plasma levels of γ-tocopherol but not α-tocopherol. In a nested case–control study in the USA, a statistically significant protective effect against prostate cancer was found only when both plasma α-tocopherol and γ-tocopherol levels were high, with a five-fold decrease in prostate cancer in those in the highest quintile relative to those in the lowest quintile. Some of the putative effect of γ-tocopherol may be mediated through its antioxidant properties; however, γ-tocopherol has other properties relevant to cancer prevention, including effects on oncogenes and tumor suppressor genes and on cell cycle events, that the α-form does not have or demonstrates to a lesser extent. It is of interest that most vitamin E supplementation trials to date have used α-tocopherol. It may be that intake of both isomers is needed for optimal tissue uptake and effect. Furthermore, in view of the ability of α-tocopherol to displace bound γ-tocopherol, supplementation with the α-form alone may be counterproductive, in that it may deplete tissues of γ-tocopherol. Further studies are needed in this area.

The brain is rich in unsaturated fatty acids, and there is a reasonable rationale for the protection of lipid-rich neurones by vitamin E. Plasma and cerebrospinal α-tocopherol concentrations were found to be low in patients with Alzheimer's disease in some but not all studies. Cognitive function is reported to be directly correlated with plasma α-tocopherol levels. A high intake of vitamin E is associated with a decreased risk of the subsequent development of Alzheimer's disease, and an 8 month delay in significant worsening of Alzheimer's disease was reported in association with increased intakes of vitamin E. In the NHANES III study, better memory performance in elderly participants was reportedly found in those with higher plasma α-tocopherol levels. Based on data such as these, vitamin E (2000 IU day^{-1}) is currently being studied in relation to its possible ability to delay the onset of Alzheimer's disease in people with mild cognitive impairment.

'Non-Nutrient' Antioxidants

Plant-based foods contain a multitude of antioxidants other than vitamin C and vitamin E. The two major classes of these other dietary-derived antioxidants are the carotenoids and the polyphenolic flavonoids. There are hundreds of different carotenoids and thousands of flavonoids, and these compounds give fruits, vegetables, teas, and herbs their wonderful colors in shades of red, orange, yellow, and purple. These compounds are synthesized exclusively in plants and have no known function in human metabolism. No deficiency state for either class of compounds has been identified in humans. Consequently, there is no recommended daily intake or agreed requirement for any of these compounds, and they are regarded as 'non-nutrients.' Nonetheless, there is evidence that diets rich in carotenoids and flavonoids are beneficial to health. For example, in a study of 1299 elderly people in the USA, those with diets rich in carotenoid-containing fruits and vegetables were found to have a significantly decreased rate of CVD and fatal myocardial infarction: the RRs (CI) when the highest and lowest quartiles of intake were compared were 0.54 (0.34–0.86) for fatal CVD and 0.25 (0.09–0.67) for fatal myocardial infarction. The carotenoid lycopene has been reported to lower the risk of prostate cancer, but the evidence for a relationship between carotenoid intake and the risk of other cancers is conflicting. Increased intake of lutein and zeaxanthin may help to delay or prevent age-related maculopathy, because these carotenoids are concentrated in the macula and are likely to be very important in local protection of the lipid-rich retina. To date, however, epidemiological findings point to health benefits of foods containing carotenoids, and the influence, if any, of individual carotenoids remains to be established.

The same is true for the polyphenolic flavonoids, anthocyanins, and various other plant-based non-nutrient antioxidants in the diet. Many of these have antioxidant powers far higher than those of vitamin C and vitamin E when tested in *in vitro* systems. Dietary intake can be similar to that of vitamin C (100 mg day^{-1} or higher), but, as their bioavailability is low, plasma levels of individual flavonoids and other phenolic antioxidants are very low or undetectable. The major dietary polyphenolic compounds are quercetin, kaempferol, myricitin, and the catechins. These flavonoids are found in onions, apples, kale, broccoli, Brussels sprouts, teas, grapes, and wine. Moderate wine intake, especially of red wine (which is very rich in polyphenolic antioxidants), is associated with a significant

Table 2 Limitations of observational epidemiological studies of diet and disease

- Cross-cultural study has no power if rates of disease and/or population means of the exposure variable of interest do not vary significantly between the populations being compared
- Behavioral, genetic, and geographical, rather than dietary, variation may account for differences detected
- A 'snapshot' view of recent dietary habits or current status may not be representative of those in earlier or later life, and differences during these periods will confound and confuse the results
- In case–control studies, the disease process itself, drug treatment, or post-diagnosis changes in diet or lifestyle may cause or mask changes in the exposure variable
- Subclinical or undetected disease may be present in controls, decreasing contrast with cases
- Retrospective dietary recall may be unreliable, food tables may be out of date or incomplete, and analysis methods may be inaccurate
- In nested case–control studies, long-term follow-up is needed and may rely on a distant 'snapshot' measure of the exposure variable as a representative index of past and future levels
- Instability or inaccurate measurement of the exposure variable will lead to bias in the results
- Assessment methods and 'high' or 'low' thresholds may vary in different areas supplying data
- If protection is maximal above a 'threshold' level of the exposure variable, then no effect will be detectable if levels in most of the study population are below or above the threshold
- Prospective studies are very expensive, requiring a very large study group and years or decades of follow-up
- Prospective trials generally have disease or death as the measured outcome; this means that the participants in the trial cannot benefit from its findings

decrease in the risk of CHD. Tea consumption, especially a high intake of green tea, is associated with a lower risk of CVD and cancer. However, which of the myriad compounds contribute to the reported health benefits is not yet clear. It may be many; it may be none. It must be remembered that association does not prove causality. Equally, the lack of significant effects of supplementation trials in healthy subject does not mean that there is no effect. As outlined in **Table 1**, and further delineated in **Table 2**, observational studies have several limitations, and there are various reasons why a conflict may exist between what we observe and the outcome of supplementation trials.

Summary and Research Needs

Strong evidence from a variety of sources indicates that a high intake of vitamin C or something very closely associated with it in the diet is protective against cancer and CVD, the major causes of disability and death in our aging communities. Indeed, it may be that plasma ascorbic-acid concentration can predict overall mortality risk. This interesting concept remains to be confirmed. The evidence for the benefits of a high intake of vitamin E is also strong, but research is needed into which member(s) of the vitamin E family are most important. The evidence for the benefits of carotenoids and flavonoids stems largely from observational studies that show a decreased risk of disease in association with a high intake of foods or beverages rich in these non-nutrient antioxidants rather than the agents themselves. However, individuals who take these foods in large quantities are often more health conscious, take fewer total calories, do not smoke, exercise more, and eat less red meat and saturated fat. The relationship between diet and health is clear, but diet is complex and dynamic, the underlying mechanisms of chronic diseases are uncertain, and the influence of individual dietary antioxidants is difficult to discern within the heterogeneous framework of the human diet and lifestyle. Antioxidants do appear to play a role in protecting key biological sites, but further study is needed to establish which, how, and where and to establish the doses needed to achieve optimal effect. To date there is no evidence that high intakes of antioxidants in the diet are harmful. It is not yet known, however, whether intake above a threshold level brings additional benefit or whether the benefit of increased intake is limited to those with initially poor or marginal antioxidant status. Furthermore, antioxidants are likely to act within a coordinated system, and more of one may require more of others for beneficial effects to be achieved. Achieving 'target thresholds' of several antioxidants may be critical to achieving the optimal effect of each, and threshold plasma concentrations of $50\,\mu mol\,l^{-1}$ and $30\,\mu mol\,l^{-1}$ for vitamin C and vitamin E, respectively, with a ratio of more than 1.3, have been proposed for minimizing the risk of CHD.

To establish cause and effect and to make firm recommendations about the type and dose of antioxidants needed to achieve optimal health requires much in the way of further study. Of particular interest and value in such study is the growing field of orthomolecular nutrition, in which advances in genomics and proteomics are used to determine gene–nutrient interaction and the influence of diet

on epigenetic phenomena. Such molecular-based studies, guided by epidemiological data and incorporated into future supplementation trials, will help answer the questions about the mechanisms of action and which, if any, antioxidants are important, how much, and for whom. However, while many questions relating to dietary antioxidants and health remain unanswered, to understand how to obtain a mixture of antioxidants and promote health we need look only at the macro level of food rather than at the micro level of specific constituents or molecular level of response. Fruits, vegetables, teas, herbs, wines, juices, and some types of chocolate are rich in antioxidants. It is known that diets rich in a variety of such foods are beneficial to health. The results of molecular-based experimental studies will determine whether these two truths are linked in a cause-and-effect relationship.

Acknowledgments

The author thanks The Hong Kong Polytechnic University and The World Cancer Research Fund International for financially supporting this work.

See also: **Antioxidants**: Diet and Antioxidant Defense.

Further Reading

Ames BN and Wakimoto P (2002) Are vitamin and mineral deficiencies a major cancer risk? *Nature Reviews* **2**: 694–704.

Asplund K (2002) Antioxidant vitamins in the prevention of cardiovascular disease: a systematic review. *Journal of Internal Medicine* **251**: 372–392.

Benzie IFF (2003) Evolution of dietary antioxidants. *Journal of Comparative Biochemistry and Physiology* **136A**: 113–126.

Block G, Norkus E, Hudes M, Mandel S, and Helzlsouer K (2001) Which plasma antioxidants are most related to fruit and vegetable consumption? *American Journal of Epidemiology* **154**: 1113–1118.

Brigelius-Flohé R, Kelly FJ, Salonen JT *et al.* (2002) The European perspective on vitamin E: current knowledge and future research. *American Journal of Clinical Nutrition* **76**: 703–716.

Clarkson PM and Thompson HS (2000) Antioxidants: what role do they play in physical activity and health? *American Journal of Clinical Nutrition* **72**: 637S–646S.

Duthie GG, Gardner PT, and Kyle JAM (2003) Plant polyphenols: are they the new magic bullet? *Proceedings of the Nutrition Society* **62**: 599–603.

Gey KF (1998) Vitamins E plus C and interacting co-nutrients required for optimal health: a critical and constructive review of epidemiology and supplementation data regarding cardiovascular disease and cancer. *Biofactors* **7**: 113–175.

Grundman M and Delaney P (2002) Antioxidant strategies for Alzheimer's disease. *Proceedings of the Nutrition Society* **61**: 191–202.

Khaw KT, Bingham S, Welch A *et al.* (2001) Relation between plasma ascorbic acid and mortality in men and women in EPIC-Norfolk prospective study: a prospective population study. *Lancet* **357**: 657–663.

Lindsay DG and Clifford MN (eds.) (2000) Critical reviews within the EU Concerted Action 'Nutritional enhancement of plant-based food in European trade' ('NEODIET') *Journal of the Science of Food and Agriculture* **80**: 793–1137.

McCall MR and Frei B (1999) Can antioxidant vitamins materially reduce oxidative damage in humans? *Free Radical Biology and Medicine* **26**: 1034–1053.

Mensink RP and Plat J (2002) Post-genomic opportunities for understanding nutrition: the nutritionist's perspective. *Proceedings of the Nutrition Society* **61**: 404–463.

Padayatty SJ, Katz A, Wang Y *et al.* (2003) Vitamin C as an antioxidant: evaluation of its role in disease prevention. *Journal of the American College of Nutrition* **22**: 18–35.

Pryor WA (2000) Vitamin E and heart disease: basic science to clinical intervention trials. *Free Radical Biology and Medicine* **28**: 141–164.

World Cancer Research Fund and the American Institute for Cancer Research (1997) *Food, Nutrition and the Prevention of Cancer: A Global Perspective*. Washington DC: American Institute for Cancer Research.

Intervention Studies

S Stanner, British Nutrition Foundation, London, UK

A predominantly plant-based diet reduces the risk of developing several chronic diseases, including cancer and cardiovascular disease (CVD) coronary heart disease and stroke. It is often assumed that antioxidants, including vitamin C, vitamin E, the carotenoids (e.g., β-carotene, lycopene, and lutein), selenium, and the flavonoids (e.g., quercetin, kaempferol, myricetin, luteolin, and apigenin), contribute to this protection by interfering passively with oxidative damage to DNA, lipids, and proteins. This hypothesis is supported by numerous *in vitro* studies in animals and humans. A large number of descriptive, case–control, and cohort studies have also demonstrated an inverse association between high intakes and/or plasma levels of antioxidants and risk of CVD and cancer at numerous sites, as well as other conditions associated with oxidative damage, such as age-related macular degeneration, cataracts, and chronic obstructive pulmonary disease (COPD).

These findings provided a strong incentive for the initiation of intervention studies to investigate whether a lack of dietary antioxidants is causally related to chronic disease risk and if providing antioxidant supplements confers benefits for the prevention and treatment of these conditions. This article summarizes the

findings of the largest primary and secondary trials published to date and considers their implications for future research and current dietary advice.

Cardiovascular Disease

Of all the diseases in which excess oxidative stress has been implicated, CVD has the strongest supporting evidence. Oxidation of low-density lipoprotein (LDL) cholesterol appears to be a key step in the development of atherosclerosis, a known risk factor in the development of CVD. Small studies have demonstrated reductions in LDL oxidation (mostly *in vitro*) following supplementation with dietary antioxidants (particularly vitamin E, which is primarily carried in LDL-cholesterol), suggesting that they may provide protection against the development of heart disease. A number of large intervention trials using disease outcomes (rather than biomarkers such as LDL oxidation) have also been conducted to try to demonstrate a protective effect of vitamin E, β-carotene, and, to a lesser extent, vitamin C supplements on cardiovascular disease. Most have been carried out in high-risk groups (e.g., smokers) or those with established heart disease (i.e., people with angina or who have already suffered a heart attack).

Primary Prevention

The results of most primary prevention trials have not been encouraging (**Table 1**). For example, in the Finnish Alpha-Tocopherol Beta-Carotene Cancer Prevention (ATBC) study, approximately 30 000 male smokers received vitamin E (50 mg/day of α-tocopherol), β-carotene (20 mg/day), both, or an

Table 1 Summary of large intervention trials (>1000 subjects) investigating the role of antioxidants and CVD in primary prevention

Trial	*Characteristics of subjects*	*Sex*	*Length of follow-up (years)*	*Treatment*	*Effect of antioxidant supplementation*
ATBC	29 133 smokers, Finland	Male	6	50 mg α-tocopherol and/or 20 mg β-carotene	No significant effect on fatal or nonfatal-CHD or total strokes with either supplement Increase in deaths from hemorrhagic stroke in vitamin E group Increase in hemorrhagic stroke (+62%) and total mortality (+8%) in β-carotene group
CARET	14 254 smokers, 4 060 asbestos workers, United States	Male and Female	4	30 mg β-carotene and 25 000 IU retinol	Increase in deaths from CVD (+26%) (terminated early)
LCPS	29 584 poorly nourished, China	Male and Female	5	15 mg β-carotene, 30 mg α-tocopherol, and 50 μg selenium	Small decline in total mortality (+9%) Reduction in deaths from stroke in men (−55%) but not women
PHS	22 071 physicians, United States	Male	12	50 mg β-carotene and/or aspirin (alternate days)	No effect on fatal or nonfatal myocardial infarction or stroke
PPP	4 495 with one or more CVD risk factors, Italy	Male and Female	$3\frac{1}{2}$	Low-dose aspirin and/or 300 mg α-tocopherol	No effect on CVD deaths or events (but inadequate power due to premature interruption of trial)
SCPS	1720 with recent nonmelanoma skin cancer, Australia	Male and Female	8	50 mg β-carotene	No effect on CVD mortality
VACP II	1 204 former asbestos workers, Australia	Male and Female	5	30 mg β-carotene or 25 000 IU retinol (no placebo group)	No effect of β-carotene on CHD deaths
WHS	39 876, United States	Female	2	50 mg β-carotene (alternate days)	No effect on fatal or nonfatal CVD

ATBC, Alpha Tocopherol Beta Carotene Prevention Study; CARET, Beta Carotene and Retinol Efficacy Trial; LCPS, Linxian Cancer Prevention Study; PHS, Physicians Health Study; PPP, Primary Prevention Project; SCPS, Skin Cancer Prevention Study; VACP, Vitamin A and Cancer Prevention; WHS, Women's Health Study; CHD, Coronary Heart Disease; CVD, Cardiovascular disease.

inactive substance (placebo) for approximately 6 years. There was no reduction in risk of major coronary events with any of the treatments despite a 50% increase in blood vitamin E concentrations and a 17-fold increase in β-carotene levels. Moreover, with vitamin E supplementation, there was an unexpected increase in risk of death from hemorrhagic stroke and a small but significant increase in mortality from all causes with β-carotene supplementation (RR, 1.08; 95% confidence interval (CI), 1–16). An increase in CVD deaths was also observed in the Beta-Carotene and Retinol Efficacy Trial (CARET), which tested the effects of combined treatment with β-carotene (30 mg/day) and retinyl palmitate (25 000 IU/day) in 18 000 men and women with a history of cigarette smoking or occupational exposure to asbestos compared to the placebo group (RR, 1.26; 95% CI, 0.99–1.61).

Secondary Prevention

The most positive results from secondary prevention trials came from the Cambridge Heart Antioxidant Study (CHAOS), a controlled trial on 2002 heart disease patients with angiographically proven coronary atherosclerosis randomly assigned to receive a high dose of vitamin E (400 or 800 IU/day) or placebo (**Table 2**). Those receiving the supplements were 77% less likely to suffer from nonfatal heart disease over the $1\frac{1}{2}$-year trial period than those who did not receive vitamin E (RR, 0.23; 95% CI, 0.11–0.47), although there was no reduction in CVD deaths. However, other large secondary prevention trials with longer follow-up have been less encouraging. For example, in a further analysis of the ATBC study, the β-carotene supplementation was associated with an increased risk of coronary heart disease (CHD) deaths among men who had a previous heart attack and were thus at high risk of subsequent coronary events. There were significantly more deaths from fatal CHD in the β-carotene group (RR, 1.75; 95% CI, 1.16–2.64) and in the combined β-carotene and vitamin E group compared to the placebo group (RR, 1.58; 95% CI, 1.05–2.40). The Heart Outcomes Prevention Evaluation Study (HOPE) observed no benefit from vitamin E supplementation (400 IU/day) on CVD or all-cause mortality. The Heart Protection Study in the United Kingdom examined the effect of 5 years of

Table 2 Summary of large intervention trials (>1000 subjects) investigating the role of antioxidants and CVD in secondary prevention[a]

Trial	*Characteristics of subjects*	*Sex*	*Length of follow-up (years)*	*Treatment*	*Effect of antioxidant supplementation*
ATBC	1862 smokers with previous MI, Finland	Male	$5\frac{1}{2}$	50 mg α-tocopherol and/or 20 mg β-carotene	No effect on total coronary events (fatal and nonfatal) Increase in deaths from fatal CHD in β-carotene (+75%) and combined β-carotene/vitamin E group (+58%) vs placebo
	1795 heavy smokers with previous angina, Finland				No effect on symptoms or progression of angina or on total coronary events
CHAOS	2002 patients with coronary atherosclerosis, United Kingdom	Male and Female	$1\frac{1}{2}$	300 or 800 IU α-tocopherol	Reduction in nonfatal MI (−77%) but no effect on CVD mortality
GISSI	11324 patients with recent MI, Italy			300 mg α-tocopherol and/or 1 g n-3 PUFA	No benefit from vitamin E
HOPE	9541 known CVD or diabetes, Canada	Male and Female	4–6	400 IU α-tocopherol and/or ACE inhibitor	No effect on MI, stroke, or CVD death
HPS	20536 with known vascular disease or at high risk, United Kingdom	Male and Female	≥ 5	20 mg β-carotene, 600 mg α-tocopherol, and 250 mg vitamin C	No effect on fatal or nonfatal MI or stroke

[a]Secondary prevention is defined as including patients with known or documented vascular disease.

ACE, angiotensin converting enzyme; ATBC, Alpha Tocopherol Beta Carotene Prevention Study; CHAOS, Cambridge Heart Antioxidant Study; GISSI, GISSI Prevenzione Trial; HOPE, Heart Outcomes Prevention Evaluation Study; HPS, Heart Protection Study; CHD, Coronary Heart Disease; CVD, cardiovascular disease; MI, myocardial infarction; PUFA, polyunsatutated fatty acids.

supplementation with a cocktail of antioxidant vitamins (600 mg vitamin E, 250 mg vitamin C, and 20 mg β-carotene) alone or in combination with the lipid-lowering drug Simvastatin or placebo in more than 20 000 adults with CHD, other occlusive arterial disease, or diabetes mellitus. Although blood levels of antioxidant vitamins were substantially increased, no significant reduction in the 5-year mortality from vascular disease or any other major outcome was noted. In the Italian GISSI-Prevenzione Trial dietary fish oils reduced the risk of fatal or nonfatal CVD in men and women who had recently suffered from a heart attack but vitamin E supplementation (300 mg daily for $3\frac{1}{2}$ years) did not provide any benefit. In these three trials, no significant adverse effects of vitamin E were observed.

Systematic reviews and meta-analyses of the clinical trials to date have therefore concluded that despite evidence from observational studies, people with a high occurrence of CVD often have low intakes or plasma levels of antioxidant nutrients. Supplementation with any single antioxidant nutrient or combination of nutrients has not demonstrated any benefit for the treatment or prevention of CVD.

Cancer

The oxidative hypothesis of carcinogenesis asserts that carcinogens generate reactive oxygen species that damage RNA and DNA in cells, predisposing these cells to malignant changes and enhanced cancer risk. Most, but not all, damage is corrected by internal surveillance and repair systems involving dietary antioxidants, as well as endogenous antioxidant mechanisms. Antioxidants are therefore proposed to prevent cell damage by neutralizing free radicals and oxidants, thus preventing subsequent development of cancer.

β-Carotene

Many of the randomized controlled trials (RCTs) investigating a protective role for antioxidant nutrients in cancer prevention (**Table 3**) have focused on β-carotene. A study in Linxian, China, of a rural population with poor nutritional status found that supplementation with a combination of β-carotene, selenium, and vitamin E for 5 years provided a 21% reduction in stomach cancer mortality and a 13% reduction in all cancer deaths. Although interesting, the population studied was likely to have very low intakes of a number of micronutrients and this study does not contribute to knowledge about the effects of individual antioxidants or offer any insight into their effects on populations with good nutritional status.

The findings of a number of large double-blind RCTs in well-fed subjects using high-dose β-carotene supplements (either alone or in combination with other agents) have generally been unsupportive of any protective effect, although most have only focused on high-risk groups (e.g., smokers, asbestos workers, and older age groups). In the ATBC Cancer Prevention Trial, in which 29 000 male smokers were randomly assigned to receive β-carotene and/or α-tocopherol or placebo each day, β-carotene showed no protective effect on the incidence of any type of cancer after approximately 6 years. In fact, concern was raised following the publication of the findings of this trial because those randomized to receive this vitamin had an 18% higher risk of lung cancer (RR, 1.18; 95% CI, 3–36) as well as an 8% higher total mortality than nonrecipients. Subgroup analyses suggested that the adverse effect of β-carotene on lung cancer risk was restricted to heavy smokers and that the risk appeared to be transient, being lost at follow-up 4–6 years after cessation of supplementation.

The CARET was also terminated early because of similar findings; subjects receiving a combination of supplements (30 mg β-carotene and vitamin A daily) experienced a 28% increased risk of lung cancer incidence compared with the placebo group (RR, 1.28; 95% CI, 1.04–1.57). Subgroup analyses also suggested that the effect was found in current, but not former, smokers. In contrast, in the Physicians Health Study, supplementation of male physicians with 50 mg β-carotene on alternate days had no effect on cancer incidence (men who were smokers did not experience any benefit or harm). The Heart Protection Study also demonstrated no effect on 5-year cancer incidence or mortality from supplementation with 20 mg β-carotene in combination with vitamins E and C in individuals at high risk of CVD, despite increases in blood concentrations of these nutrients (plasma β-carotene concentrations increased 4-fold). They did not, however, find any harmful effects from these vitamins.

A number of trials have attempted to investigate the effect of β-carotene supplementation on nonmelanoma skin cancer, the most common forms of which are basal cell and squamous cell carcinomas (these types of cells are both found in the top layer of the skin). However, none have shown any significant effect on skin cancer prevention. For example, the Physicians Health Study found no effect after 12 years of β-carotene supplementation on the development of a first nonmelanoma skin cancer. The Nambour Skin Cancer Prevention Trial of 1621

Table 3 Summary of large intervention trials (>1000 subjects) investigating the role of antioxidants and cancer in primary prevention

Trial	*Characteristics of subjects*	*Sex*	*Length of follow-up (years)*	*Treatment*	*Effect of antioxidant supplementation*
ATBC	29 133 smokers, Finland	Male	5–8	50 mg α-tocopherol and/or 20 mg β-carotene	18% increase in lung cancer in β-carotene group (no effect in vitamin E group) 34% reduction in incidence of prostate cancer in vitamin E group No effect of either vitamin on colorectal, pancreatic, or urinary tract cancer
CARET	14 254 smokers, 4060 asbestos workers, United States	Male and Female	4	30 mg β-carotene and 25 000 IU retinol	Lung cancer increased by 28%
HPS	20 536 at high CVD risk, United Kingdom	Male and Female	$\geq$5	20 mg β-carotene, 600 mg α-tocopherol, and 250 mg vitamin C	No effect on cancer incidence or mortality
LCPS	29 584 poorly nourished, China	Male and Female	5	15 mg β-carotene, 30 mg α-tocopherol, and 50 μg selenium	Cancer deaths declined by 13% Stomach cancer declined by 21%
NSCPT	1621 (73% without skin cancer at baseline), Australia	Male and Female	$4\frac{1}{2}$	30 mg β-carotene with or without sunscreen application	No effect on basal cell or squamous cell carcinoma
PHS	22 071 physicians, United States	Male	12	50 mg β-carotene and/or aspirin (alternate days)	No effect on incidence of malignant neoplasms or nonmelanoma skin cancer
VACP II	1204 former asbestos workers, Australia	Male and Female	5	30 mg β-carotene or 25 000 IU retinol (no placebo group)	No effect of β-carotene on cancer mortality
WHS	39 876, United States	Female	2	50 mg β-carotene (alternate days)	No effect on cancer incidence

ATBC, Alpha Tocopherol Beta Carotene Prevention Study; CARET, Beta Carotene and Retinol Efficacy Trial; HPS, Heart Protection Study; LCPS, Linxian Cancer Prevention Study; NSCPT, Nambour Skin Cancer Prevention Trial; PHS, Physicians Health Study; VACP, Vitamin A and Cancer Prevention; WHS, Women's Health Study; CVD, Cardio Vascular disease.

men and women followed for nearly 5 years (most of whom had no history of skin cancer at baseline) showed that those supplemented with 30 mg β-carotene did not experience any reduction in risk of basal cell or squamous cell carcinoma or the occurrence of solar keratoses (precancerous skin growths that are a strong determinant of squamous cell carcinoma). A 5-year trial of 1805 men and women with recent nonmelanoma skin cancer (the Skin Cancer Prevention Study) also found that supplementation with 50 mg of β-carotene gave no protection against either type of skin cancer, although this may have been because these cancers have a long latency period of approximately 12 years (**Table 4**).

Together, these trials suggest that β-carotene supplements offer no protection against cancer at any site and, among smokers, may actually increase the risk of lung cancer. Investigators have sought to explain these findings by proposing that components of cigarette smoke may promote oxidation of β-carotene in the lungs, causing it to exert a prooxidant (rather than antioxidant) effect and act as a tumor promoter.

Vitamin C

There are no published RCTs of vitamin C alone in primary prevention, but data from the small number of trials of vitamin C in combination with other nutrients have not provided any support for a role for high-dose vitamin C supplementation in cancer prevention (**Table 3**). The Linxian trial found no significant effect of supplementing Chinese men and women with 120 mg vitamin C and 30 μg molybdenum daily for 5 years on the risk of cancers of the oesophagus or stomach. The Polyp Prevention Study, a trial of 864 patients with previous

Table 4 Summary of large intervention trials (>1000 subjects) investigating the role of antioxidants and cancer in secondary prevention[a]

Trial	*Characteristics of subjects*	*Sex*	*Length of follow-up (years)*	*Treatment*	*Effect of antioxidant supplementation*
NPCT	1312 with history of basal or squamous cell carcinoma, United States	Male and Female	$4\frac{1}{2}$	200 μg selenium	No effect on incidence of skin cancer Reduce cancer mortality (50%), cancer incidence (37%), prostate cancer (63%), colorectal cancer (58%), and lung cancer (46%)
SCPS	1805 with recent nonmelanoma skin cancer, United States	Male and Female	5	50 mg β-carotene	No effect on occurrence of new nonmelanoma skin cancer

[a]Secondary prevention defined as subjects with documented cancer including nonmelanoma skin cancer (although some of the primary prevention trials did not exclude those with nonmelanoma skin cancer at baseline).
NPCT, Nutritional Prevention of Cancer Trial; SCPS: Skin Cancer Prevention Study.

adenoma, found no effect of either β-carotene or a combination of vitamins E and C (1000 mg) on the incidence of subsequent colorectal adenomas. The Heart Protection Study also found no beneficial effects of supplementation with these three vitamins on cancer mortality. However, trials have generally been carried out on those with diets containing sufficient amounts of vitamin C and there is a need for further studies in people with low intakes.

Vitamin E

The ATBC trial showed no significant effect of α-tocopherol supplementation (50 mg/day) on risk of lung, pancreatic, colorectal, or urinary tract cancers among heavy smokers (**Table 3**). However, in a post hoc subgroup analysis a 34% reduction in the risk of prostate cancer was seen in men who received this supplement. Although interesting, prostate cancer was not a primary endpoint of this study, and no other studies have supported a preventative effect of vitamin E for prostate cancer. The Heart Protection Study found no effect of vitamin E in combination with vitamin C and β-carotene on cancer incidence or mortality. Two smaller, short-term intervention studies found no effect of α-tocopherol supplementation on mammary dysplasia or benign breast disease. Several trials have also been unable to demonstrate a protective effect of vitamin E supplementation on the risk or recurrence of colorectal adenomatous polyps.

Selenium

A few trials have suggested that selenium supplementation may have a protective effect on liver cancer in high-risk groups living in low-selenium areas. The provision of selenium-fortified salt to a town in Qidong, China, with high rates of primary liver cancer, reduced the incidence of this cancer by 35% compared with towns that did not receive this intervention (**Table 3**). Trials have also demonstrated the incidence of liver cancer to be significantly reduced in subjects with hepatitis B and among members of families with a history of liver cancer receiving a daily supplement of 200 μg of selenium for 4 and 2 years, respectively.

The Nutritional Prevention of Cancer Trial in the United States also supported a possible protective role of selenium (**Table 4**): 1312 patients (mostly men) with a previous history of skin cancer were supplemented with either placebo or 200 μg selenium per day for $4\frac{1}{2}$ years and those receiving selenium demonstrated significant reductions in the risk of cancer incidence (37%) and mortality (50%). Although selenium was not found to have a protective effect against recurrent skin cancer, the selenium-treated group had substantial reductions in the incidence of lung, colorectal, and prostate cancers of 46, 58, and 63%, respectively. Further analysis showed the protective effect on prostate cancer to be confined to those with lower baseline prostate-specific antigen and plasma selenium levels. Although these data need confirmation, they suggest that adequate selenium intake may be important for cancer prevention.

Other Diseases Associated with Oxidative Damage

Type 2 Diabetes

Type 2 diabetes is associated with elevated oxidative stress (especially lipid peroxidation) and declines in antioxidant defense. This is thought to be due in part

to elevated blood glucose levels (hyperglycemia), but severe oxidative stress may also precede and accelerate the development of type 2 diabetes and then of diabetic complications (CVD and microvascular complications such as retinopathy, neuropathy, and nephropathy).

Small-scale human trials have shown administration of high doses of vitamin E to reduce oxidative stress and improve some CVD risk factors, such as blood glycated hemoglobin, insulin, and triglyceride levels, in people with diabetes. Such trials have also indicated benefit from vitamin E in improving endothelial function, retinal blood flow, and renal dysfunction. However, the findings of large clinical trials investigating the role of individual or a combination of antioxidant nutrients in reducing the risk of CVD and microvascular complications in people with diabetes have generally been disappointing. For example, the Heart Outcomes Prevention Evaluation Trial investigated the effects of vitamin E and the drug Ramipril in patients at high risk for CVD events and included a large number of middle-aged and elderly people with diabetes (more than 3600). An average of $4\frac{1}{2}$ years of supplementation with 400 IU of vitamin E per day was found to exert no beneficial or harmful effect on CVD outcomes or on nephropathy. The Primary Prevention Project trial found no effect of vitamin E (300 mg/day) supplementation for 3 or 4 years in diabetic subjects, and the Heart Protection Study, which included a number of people with diabetes, also reported no benefit of a combination of antioxidant vitamins on mortality or incidence of vascular disease.

Chronic Obstructive Pulmonary Disease (COPD)

The generation of oxygen free radicals by activated inflammatory cells produces many of the pathophysiological changes associated with COPD. Common examples of COPD are asthma and bronchitis, each of which affects large numbers of children and adults. Antioxidant nutrients have therefore been suggested to play a role in the prevention and treatment of these conditions. A number of studies have demonstrated a beneficial effect of fruit and vegetable intake on lung function. For example, regular consumption of fresh fruit rich in vitamin C (citrus fruits and kiwi) has been found to have a beneficial effect on reducing wheezing and coughs in children.

Vitamin C is the major antioxidant present in extracellular fluid lining the lung, and intake in the general population has been inversely correlated with the incidence of asthma, bronchitis, and wheezing and with pulmonary problems. Although some trials have shown high-dose supplementation (1–2 g/day) to improve symptoms of asthma in adults and protect against airway responsiveness to viral infections, allergens, and irritants, this effect has been attributed to the antihistaminic action of the vitamin rather than to any antioxidant effect. The results of these trials have also been inconsistent, and a Cochrane review of eight RCTs concluded that there is insufficient evidence to recommend a specific role for the vitamin in the treatment of asthma. However, a need for further trials to address the question of the effectiveness of vitamin C in asthmatic children was highlighted.

Other dietary antioxidants have been positively associated with lung function in cohort studies but the findings of clinical trials have been mixed. In a study of 158 children with moderate to severe asthma, supplementation with vitamin E (50 mg/day) and vitamin C (250 mg/day) led to some improvement in lung function following ozone exposure. However, the much larger ATBC trial found no benefit from supplementation with α-tocopherol (50 mg/day) and β-carotene (20 mg/day) on symptoms of COPD, despite the fact that those with high dietary intakes and blood levels of these vitamins at baseline had a lower prevalence of chronic bronchitis and dyspnea. A small trial investigating the effects of selenium supplementation in asthmatics found that those receiving the supplements experienced a significant increase in glutathione peroxidase levels and reported improvement in their asthma symptoms. However, this improvement could not be validated by significant changes in the separate clinical parameters of lung function and airway hyperresponsiveness. Therefore, there is little evidence to support the role of other nutrients in COPD treatment.

Macular Degeneration and Cataracts

The eye is at particular risk of oxidative damage due to high oxygen concentrations, large amounts of oxidizable fatty acids in the retina, and exposure to ultraviolet rays. In Western countries, age-related macular degeneration (AMD) is the leading cause of blindness among older people. Cataracts are also widespread among the elderly and occur when the lens is unable to function properly due to the formation of opacities within the lens. These develop when proteins in the eye are damaged by photooxidation; these damaged proteins build up, clump, and precipitate. It has been proposed that antioxidants may prevent cellular damage in the eye by reacting with free radicals produced during the process of light absorption.

The results of intervention trials in this area have also been mixed. The Age-Related Eye Disease Study

in the United States investigating the effects of combined antioxidant vitamins C (500 mg), E (400 IU), and β-carotene (15 mg) with and without 80 mg zinc daily for 6 years showed some protective effect (a reduction in risk of approximately 25%) on the progression of moderately advanced AMD but no benefit on the incidence or progression or early AMD or cataracts. The Lutein Antioxidant Supplementation Trial, a 12-month study of 90 patients with AMD, found significant improvements in visual function with 10 mg/day lutein (one of the major carotenoids found in the pigment of a normal retina) alone or in combination with a number of other antioxidant nutrients. The Roche European Cataract Trial, providing a combined daily supplement of β-carotene, vitamin C, and vitamin E among adults with early signs of age-related cataract, showed a small deceleration in the progression of cataract after 3 years.

However, the Linxian trial found no influence of vitamin supplementation on risk of cataract; the ATBC trial found no reduction in the prevalence of cataracts with vitamin E, β-carotene, or both among male smokers; and the Health Physicians Study of more than 22 000 men showed no benefit from 12 years of supplementation with β-carotene (50 mg on alternate days) on cataract incidence. In fact, current smokers at the beginning of this trial who received the supplement experienced an increased risk of cataract (by approximately 25%) compared to the placebo group. The Vitamin E, Cataract and Age-Related Maculopathy Trial also reported no effect of supplementation with vitamin E for 4 years (500 IU/day) on the incidence or progression of cataracts or AMD.

Possible Explanations for the Disagreement between the Findings of Observational Studies and Clinical Trials

Various explanations have been given for the different findings of observational studies and intervention trials. Clearly, nonrandomized studies are unable to exclude the possibility that antioxidants are simply acting as a surrogate measure of a healthy diet or lifestyle and that the protective effect of certain dietary patterns, which has been presumed to be associated with dietary antioxidants, may in fact be due to other compounds in plant foods, substitution of these foods for others, or a reflection of other health behaviors common to people who have a high fruit and vegetable intake. However, although intervention studies provide a more rigorous source of evidence than observational studies, they are not without weaknesses from a nutritional perspective and the trials have been criticized for a number of reasons:

- The nature of the supplements used: It has been suggested that the synthetic forms used in most trials may have different biological activity or potency from natural forms of these vitamins, although trials using the natural forms have not found different clinical effects. The type of isomer used has also been questioned (e.g., β-carotene versus other carotenoids such as lycopene or lutein or α-tocopherol versus γ-tocopherol). Trials have not investigated other potentially beneficial antioxidants in foods, such as flavonoids and lycopenes.
- The use of high doses of one or two antioxidants: Mechanistic and epidemiological data suggest that antioxidants act not only individually but also cooperatively and in some cases synergistically. Single supplements may interfere with the uptake, transport, distribution, and metabolism of other antioxidant nutrients. An optimal effect would therefore be expected to be seen with a combination of nutrients at levels similar to those contained in the diet (corresponding to higher levels of intake associated with reduced risk in the observational studies). The findings of clinical trials testing the effect of a cocktail of antioxidant nutrients at low doses are awaited, but the Heart Protection Study did not demonstrate a protective effect of multiple antioxidants and a small RCT of 160 patients with coronary disease, using a combination of antioxidant nutrients (800 IU α-tocopherol, 1000 mg vitamin C, 25 mg β-carotene, and 100 μg selenium twice daily) for 3 years, showed no benefit for secondary prevention of vascular disease.
- Insufficient duration of treatment and follow-up: Most of the intervention trials published to date (except the Physicians Health Study, which found no effect despite 12 years of follow-up) had durations of treatment and follow-up lasting only approximately 4–6 years. Diseases such as cancer and CVD develop over a long period of time and trials may have been too short to demonstrate any benefit.
- The use of high-risk groups: Many of the supplementation trials have not been undertaken on normal 'healthy' individuals but on those with preexisting oxidative stress, either through smoking or through preexisting disease, among whom increasing antioxidant intake may not have been able to repair the oxidative damage process sufficiently to affect cancer or CVD risk.
- Lack of information about the impact of genetic variability: Unknown genetic factors (interacting

with nutrition) may explain some of the lack of effect in intervention studies. A greater understanding of the impact of factors such as genotype, age, and ill health on the interactions between antioxidants and reactive oxygen species would be helpful in designing future trials.

The Supplementation en Vitamines et Minéraux AntioXydants Study (SU.VI.MAX) has taken account of many of these issues in its design. This is a randomized, placebo-controlled trial testing the efficacy of supplementation among more than 12 000 healthy men and women over an 8-year period with a cocktail of antioxidant vitamins (120 mg vitamin C, 30 mg vitamin E, and 6 mg β-carotene) and minerals (100 μg selenium and 20 mg zinc) at doses achievable by diet (approximately one to three times the daily recommended dietary allowances) on premature death from CVD and cancer. Early reports suggest that this regime has not demonstrated an effect on CVD risk but has led to a 31% decrease in cancer incidence and a 37% reduction in total mortality among men but not women. This may reflect higher dietary intakes of these nutrients among the women in the trial compared to men, but publication of these results is still awaited. However, this is a good illustration of the type of nutritional approach that may be needed in the future.

Conclusion

Although there is a substantial body of evidence that diets rich in plant foods (particularly fruit and vegetables) convey health benefits, as do high plasma levels of several antioxidant nutrients found in these foods, a causal link between lack of antioxidants and disease occurrence or between antioxidant administration and disease prevention remains to be established. There is a lack of understanding of the mechanisms underpinning the apparent protective effect of plant foods and, as yet, no clear picture of which components are effective and hence no way of predicting whether all or just some plant foods are important in this respect.

If future trials do demonstrate a reduction in chronic disease risk with antioxidant supplementation, this cannot be definitively attributed to the antioxidant effect of these nutrients because other biological functions may also play a role. For example, in addition to retarding LDL oxidation, vitamin E may help to protect against CVD via its action on platelet aggregation and adhesion or by inhibition of the proliferation of smooth muscle cells. Furthermore, although vitamin C, vitamin E, and selenium have been shown to decrease the concentration of some of the biomarkers associated with oxidative stress, the relationship between many of these biomarkers and chronic disease remains to be elucidated.

The intervention studies highlight the lack of information on the safety of sustained intakes of moderate to high doses of micronutrient supplements and long-term harm cannot be ruled out, particularly in smokers. Further evidence is required regarding the efficacy, safety, and appropriate dosage of antioxidants in relation to chronic disease.

Currently, the most prudent public health advice continues to be to consume a variety of plant foods.

See also: **Antioxidants**: Diet and Antioxidant Defense; Observational Studies. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements; Deficiency States. **Cancer**: Epidemiology and Associations Between Diet and Cancer; Effects on Nutritional Status. **Carotenoids**: Chemistry, Sources and Physiology; Epidemiology of Health Effects. **Coronary Heart Disease**: Hemostatic Factors; Lipid Theory. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Lipoproteins**. **Lung Diseases**. **Selenium**. **Stroke, Nutritional Management**. **Vitamin E**: Metabolism and Requirements; Physiology and Health Effects.

Further Reading

Asplund K (2002) Antioxidant vitamins in the prevention of cardiovascular disease: A systematic review. *Journal of Internal Medicine* **251**: 372–392.

British Nutrition Foundation (2001) *Briefing Paper: Selenium and Health*. London: British Nutrition Foundation.

British Nutrition Foundation (2003) *Plants: Diet and Health. A Report of the British Nutrition Foundation Task Force* Goldberg G (ed.) Oxford: Blackwell Science.

Clarke R and Armitage J (2002) Antioxidant vitamins and risk of cardiovascular disease. Review of large-scale randomised trials. *Cardiovascular Drugs and Therapy* **16**: 411–415.

Evans J (2002) Antioxidant vitamin and mineral supplements for age-related macular degeneration. *Cochrane Database Systematic Review* **2**: CD000254.

Lawlor DA, Davey Smith G, Kundu D *et al.* (2004) Those confounded vitamins: What can we learn from the differences between observational versus randomised trial evidence? *Lancet* **363**: 1724–1727.

Lee I (1999) Antioxidant vitamins in the prevention of cancer. *Proceedings of the Association of American Physicians* **111**: 10–15.

Mares JA (2004) High-dose antioxidant supplementation and cataract risk. *Nutrition Review* **62**: 28–32.

Morris C and Carson S (2003) Routine vitamin supplementation to prevent cardiovascular disease: A summary of the evidence for the US Preventive Services Task Force. *Annals of Internal Medicine* **139**: 56–70.

National Academy of Sciences Food and Nutrition Board (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium and Carotenoids*. Washington, DC: National Academy Press.

Ram F, Rowe B, and Kaur B (2004) Vitamin C supplementation for asthma. *Cochrane Database Systematic Review* **3**: CD000993.

Stanner SA, Hughes J, Kelly CNM *et al.* (2004) A review of the epidemiological evidence for the 'antioxidant hypothesis.' *Public Health Nutrition* **7**: 407–422

Vivekananthan D, Penn MS, Sapp SK *et al.* (2003) Use of antioxidant vitamins for the prevention of cardiovascular disease: Meta-analysis of randomised trials. *Lancet* **361**: 2017–2023.

APPETITE

Contents

Physiological and Neurobiological Aspects

J C G Halford, University of Liverpool, Liverpool, UK
J E Blundell, University of Leeds, Leeds, UK

Appetite Regulation and Expression

Traditionally it has been thought that appetite is influenced solely by body components or by metabolism. These influences are commonly referred to as the glucostatic, aminostatic, thermostatic, or lipostatic hypotheses. Each suggests that a single variable such as glucose, amino acids, heat generation, or adipose tissue stores plays the major role in modulating the expression of appetite. It can be accepted that all four variables can be monitored and each can exert some influences over food consumption. In the last few years research has given renewed support to the lipostatic hypothesis, specifically, the identification of the adipose signal leptin. The short-term consequences of food ingestion generated by a meal also produce a powerful inhibition on further intake (satiety). We can draw a distinction between short-term satiety signals generated by the physiological consequences of meal intake (episodic), and the long-term signals generated by the body's constant metabolic need for energy (tonic). This distinction may be a useful starting point in our examination of the integration of the CNS systems responsible for the expression of appetite.

Episodic Events: Hunger, Satiety and the Appetite Cascade

A good place to start is the psychological experiences of hunger and satiety that underpin the pattern of eating behavior. Hunger can be defined as the motivation to seek and consume food initiating a period of feeding behavior. The process that brings this period to an end is termed satiation. Satiation processes ultimately lead to the state of satiety in which the hunger drive, and consequently eating behavior, is inhibited. The processes of satiation determine the meal size and the state of satiety determines the length of the post meal interval. The net effect of these systems can be considered before (preprandial or cephalic phase), during (prandial), and after (postprandial) a meal (see **Figure 1**).

Preconsumption physiological signals are generated by the sight and smell of the food, preparing the body for ingestion. Such afferent sensory information, carried to the brainstem via cranial nerves, stimulates hunger before eating and during the initial stages of consumption (the prandial phase). During the prandial phase the CNS receives postingestive sensory afferent input from the gut reflecting both the amount of food eaten and earliest representations of its nutrient content. Mechanoreceptors in the gut detect the distension of gut lining caused by the presence of food aiding the estimation of the volume of food consumed. Gut chemoreceptors detect the chemical presence of various nutrients in the gastrointestinal tract providing information on the composition (and possible energy content) of the food consumed. Prandial and postprandial signals are generated by the detection of nutrients that

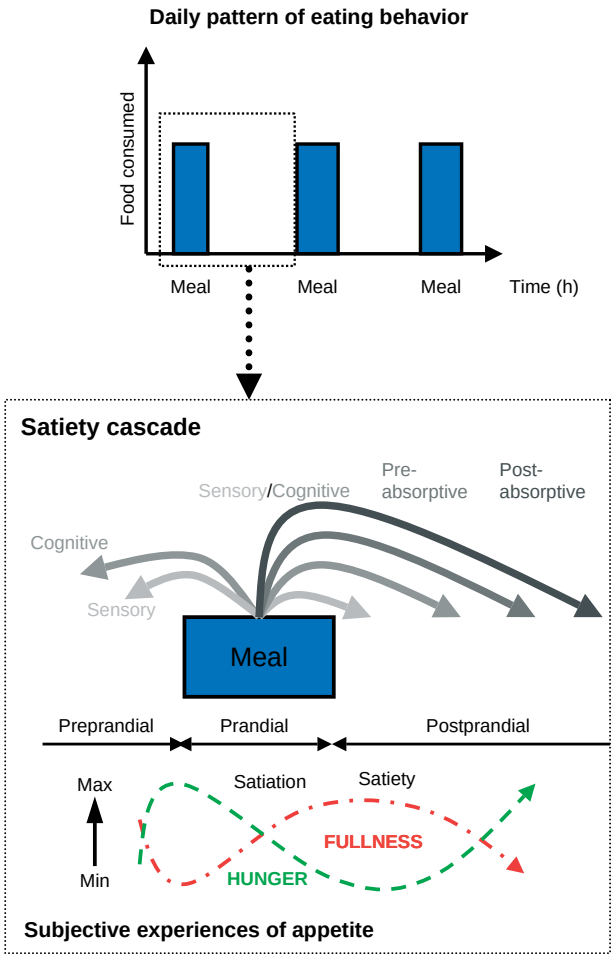


Figure 1 The satiety cascade. The signals generated prior to (preprandial), during (prandial), and after (postprandial) the consumption of a meal critical to short-term (episodic) meal-by-meal appetite regulation throughout the day.

have been absorbed from the gastrointestinal tract and have entered the circulation in the periphery (postabsorptive satiety signals). Circulating nutrients that are either metabolized in the periphery (e.g., liver) activate central nervous system (CNS) receptors (e.g., in brainstem), or they enter and affect the brain directly and act as postabsorptive metabolic satiety signals.

Neural Structures Critical to the Expression of Appetite

The CNS receives information generated by the sensory experience of eating, and from the periphery indicating the ingestion, absorption, metabolism, and storage of energy. To regulate appetite a variety of structures within the CNS integrate multiple signals, to assess the biological need for energy, to generate or inhibit conscious experiences of hunger, and subsequently to initiate the appropriate behavioral action. Information reaches the CNS via three main routes:

1. Signals from the periphery: peripheral receptors in the gut (distension and chemo-receptors) and metabolic changes in the liver (energy conversion and energy status) send afferent signals via the vagus nerve to the nucleus of the solitary tract/area postrema (NST/AP) complex in the brainstem.
2. Signals from specific receptors within the brain: receptors in the CNS, particularly in the brainstem detecting circulating levels of nutrients, their metabolites, and other factors within the periphery.
3. Substances crossing the blood–brain barrier entering the brain: factors such as neurotransmitter precursors, leptin or insulin cross the blood–brain barrier and directly alter CNS neurochemical activity, particularly in key hypothalamic nuclei and associated limbic areas.

Original theories of the neural control of appetite conceptualize food intake to be controlled by the opposing action of two hypothalamic centers (lateral hypothalamus, LH; and ventral medial hypothalamus, VMH). However, with later precision technologies numerous hypothalamic and nonhypothalamic nuclei have been implicated in the control of both hunger and satiety. For instance, infusions of various agents in or near the paraventricular nucleus (PVN), a key hypothalamic site, produce either marked increases or decreases in food intake and of specific macronutrients. Other key limbic sites identified as playing critical roles in appetite regulation include the arcuate nucleus (ARC), nucleus accumbens (NAc), the amygdala, posterior hypothalamus and the dorsal medial hypothalamus. Nonhypothalamic/limbic regions key to the expression of appetite include the NTS/AP adjacent areas in the hindbrain that relay vagal afferent satiety signals from the periphery (particularly receptors in the gastrointestinal tract and liver) to the hypothalamus. This area of the brainstem appears to possess receptors sensitive to levels of circulating nutrients and afferent sensory information from the mouth including taste (carried by cranial nerves).

Interrelated Levels of the System

Before we go on to consider the individual systems underpinning the expression of appetite it is useful to try and understand how neural, nutritional, and psychological events interact before, during, and after a meal. The biopsychological system

underlying the expression of appetite can be conceptualized as having three domains (**Figure 2**):

1. Psychological events (e.g., subjective sensations of hunger, satiety, hedonics, and cravings) accompanying observable behavioral operations (meal intake, snacking behavior, food choice) and their measurable consequences (energy intake and macronutrient composition of food consumed).
2. Peripheral physiology and metabolic events related to the effect of absorbed nutrients and their utilization or subsequent conversion for storage (i.e., the changes in the body due to either energy intake and/or energy deficit).
3. Neurochemical (classic neurotransmitters, neuropeptides, and hormones) and metabolic interactions within the CNS (i.e., how various signals of the body's energy status are detected in the brain).

The expression of appetite reflects the synchronous operation of events and processes in all three domains.

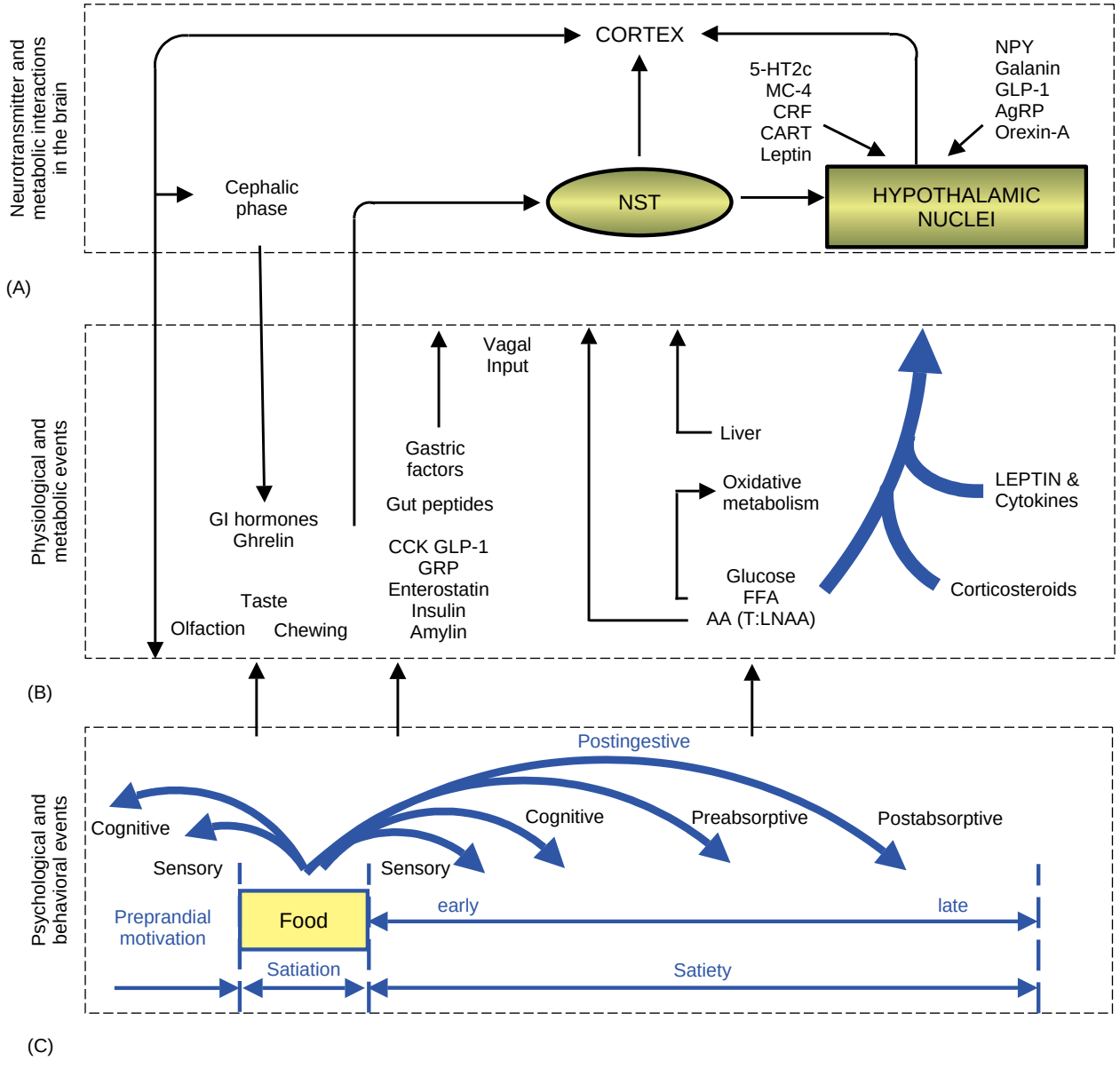


Figure 2 The psychobiological expression of appetite and the three levels of operation: (A) psychological and behavioral events; (B) physiological and metabolic operations; and (C) neurochemical and metabolic interactions within the CNS. *Abbreviations*: 5-HT, serotonin; AA, amino acids; AgRP, agouti-related peptide; CART, cocaine and amphetamine-regulated transcript; CCK, cholecystokinin; CRF, corticotropin releasing factor; FFA, free fatty acids; GI, gastrointestinal; GLP-1, glucagon-like peptide-1; GRP, gastric releasing peptide; MC, melanocortin; NPY, neuropeptide Y; T/LNAA, NTS, nucleus tractus solitarius; T/LNAA, tryptophan large neutral amino acids ratio.

Initiation and Stimulation of Eating: Mechanisms Underpinning Hunger

The intimate contact of mainly chemical, but also physical, stimuli with receptors in the mucosa of the nose and mouth set up orosensory effects of food stimuli. This is in turn transmitted to the brain by afferent fibers of primary olfactory, gustatory, and somatosensory neurons of cranial nerves 1, 5, 7, 9, and 10. These peripheral inputs appear to make contact with dopamine and opioid neurotransmitters in the brain. The cephalic phase of appetite control refers to physiological responses engendered by the sight or smell of food, which are anticipatory and serve to prepare the system for the imminent ingestion of food. Cephalic phase responses occur in the mouth (anticipatory secretion of saliva), stomach, and small intestine and represent preprandial changes that are precursors for the onset of a meal.

In addition it has been proposed that changes in blood glucose may serve as a signal for meal initiation. Recent evidence provides some support for a role for 'transient declines' in blood glucose in humans leading to increased expression of hunger and the initiation of eating. Potent feeding responses can also be obtained by microinjection of peptides to the brain of animals. A number of peptides, including β-endorphin, dynorphin, neuropeptide Y (NPY), orexins (OX-A and OX-B), galanin, agouti-related peptide (AgRP), and melanin-concentrating hormone (MCH) increase food intake.

Neuropeptide Y (NPY) is probably the most studied appetite stimulatory peptide. NPY is found throughout the CNS and in particular abundance in the PVN of the hypothalamus. Hypothalamic NPY neurons that are implicated in appetite regulation project from the ARC to the PVN. Infusing NPY directly into the CNS or increasing release of NPY within the PVN promotes meal initiation and produces an immediate and marked increase in food intake, delaying the onset of satiety. The hyperphagic effects of NPY appear to be mediated by both NPY Y1 and Y5 receptors. Endogenous NPY is sensitive to a variety of peripherally generated signals. It is stimulated by the gut factor ghrelin, but inhibited by the pancreatic hormone amylin, the adiposity signal leptin, and the satiety neurotransmitter serotonin (5-HT). Like NPY, galanin-induced hyperphagia has been well documented. Early studies demonstrated that direct infusion of galanin into the hypothalamus of rodents stimulated feeding behavior. Moreover, high concentrations of galanin and its receptors are found in the hypothalamus in areas associated with appetite regulation.

A more recently discovered CNS stimulatory system is that of the orexins. The endogenous orexin system is integrated with other critical hypothalamic energy regulatory systems. The system consists of two peptides, termed orexin-A and orexin-B, along with two orexin receptors, orexin-1 (OX1) and orexin-2 (OX2). The strongest and most reliable effect on food intake is produced by orexin-A. The endogenous orexin system responds to insulin-induced hypoglycemia and food restriction. Moreover, leptin reduces orexin-A concentration in the hypothalamus, and partially blocks orexin-A induced changes in feeding behavior.

Not all stimulatory peptide systems are within the brain. Ghrelin is a peripherally secreted hormone responsive to nutritional status. Human plasma ghrelin increases during fasting and decreases after food intake. Unlike the other gut-derived factors detailed in this review, ghrelin stimulates rather than inhibits feeding behavior. Both peripheral and central infusions of ghrelin have been shown to stimulate food intake in rats and mice, an effect in part mediated by central NPY. In lean humans, endogenous plasma ghrelin levels rise markedly before a meal and are suppressed by food intake. In lean healthy volunteers, ghrelin infusions increase food intake, premeal hunger, and prospective consumption.

Satiety: Peripheral Physiological Influences

Intake of energy can only be achieved (in mammals) through the gastrointestinal tract, and energy intake is limited by the capacity of the tract. Humans are periodic feeders and usually meals are separated by periods (intermeal intervals) of 3–5 h during which little food is eaten. It can be noted that the periodicity of meal eating is compatible with the time taken by the gastrointestinal tract to process a meal. After the ingestion of a meal the stomach mixes the meal with gastric secretions that aid its liquefaction and delivers the mixture at a steady rate into the small intestine for further chemical digestion and absorption. Four hours after ingestion most of the meal has left the stomach and the majority of nutrients have been absorbed. It seems obvious that the stomach must be involved in the termination of eating (satiation). Indeed stomach distension is regarded as being an important satiety signal. A good deal of experimental evidence indicates that gastric distension can arrest eating behavior, but the effect may be short lived. By itself, gastric distension does not appear to produce the sensation of satiety and cannot be the only factor controlling meal size. Indeed, chemicals released by gastric stimuli or by food processing in the gastrointestinal tract are critical to the episodic control of appetite. Many of these chemicals are peptide neurotransmitters, and many peripherally administered peptides cause changes in food consumption.

There is evidence for an endogenous role for cholecystokinin (CCK), pancreatic glucagon, gastrin releasing peptide (GRP), and somatostatin. Much recent research has confirmed the status of CCK as a hormone mediating meal termination (satiation) and possibly early phase satiety. Food consumption (mainly protein and fat) stimulates the release of CCK (from duodenal mucosal cells), which in turn activates CCK-A type receptors in the pyloric region of the stomach. This signal is transmitted via afferent fibers of the vagus nerve to the nucleus tractus solitarius (NTS) in the brainstem. From here the signal is relayed to the hypothalamic region where integration with other signals occurs. Direct infusions of CCK dose-dependently reduce food intake in mice, rats and monkeys, and in human volunteers the CCK octopeptide CCK-8 reduces food intake and enhances satiety. Peripheral CCK-8 administration has also been shown to increase the release of serotonin in the hypothalamus, a neurotransmitter that has been implicated in the integration of episodic satiety signals.

Other potential peripheral satiety signals include peptides such as enterostatin, neurotensin, and glucagon-like peptide (GLP-1). Researchers are continually searching for components of peripheral metabolism that could provide information to the brain concerning the pattern of eating behavior. There is considerable current interest in the peptide called enterostatin. It is formed by the cleavage of procolipase that produces colipase and this 5 amino acid activation peptide. The administration of enterostatin reduces food intake and, since it is increased after high-fat feeding, it has been suggested that enterostatin could be a specific fat-induced satiety signal. Another gut factor stimulated by the ingestion of dietary fat is intestinal glycoprotein apolipoprotein A-IV produced in the human small intestine and released into intestinal lymph in response to dietary lipids.

Glucagon-like peptide 1 (GLP-1) is a hormone that is released from the gut into the bloodstream in response to intestinal carbohydrate. Endogenous GLP-1 levels increase after meals with the largest increase in response to carbohydrate ingestion. In humans, infusions of glucose directly into the gut or the ingestion of carbohydrate produces a decrease in appetite and an increase in blood GLP-1. A series of studies by Meier and coworkers in both lean and obese human volunteers demonstrated that infusions of synthetic human GLP-1 enhanced ratings of fullness and satiety and reduced food intake and spontaneous eating behavior. Recent research has also focused on amylin, a pancreatic hormone, which also has a potent effect on both food intake and body weight. Peripheral administration of amylin reduces food intake in mice and rats, and meal size in rats.

One further source of biological information relevant to the control of appetite concerns fuel metabolism. The products of food digestion may be metabolized in peripheral tissues or organs, or may enter the brain directly. Most research has involved glucose metabolism and fatty acid oxidation in the hepatoportal area. The main hypothesis suggests that satiety is associated with an increase in fuel oxidation. Indirect evidence is provided by the use of antimetabolites that block oxidation pathways or impair fuel availability and lead to increases in food intake. It is argued that membranes or tissues sensitive to this metabolic activity modulate afferent discharges that are relayed to the brain via the vagus nerve. Pathways in the CNS that are sensitive to this metabolic signaling have begun to be mapped out. It is difficult to identify CNS mechanisms that specifically integrate short-term satiety signals alone into appetite regulation. Generally, the peripheral release and detection of various gut peptides could account for their satiety function. However, direct entry into, and action on receptors in the CNS, may also contribute to their satiety action. The one central factor clearly associated with episodic satiety, rather than tonic energy status, is serotonin (see below).

Tonic Signals: The Moderating Effects of Energy Status

Appetite is not only derived from the daily flux of physiology associated with meals and eating behavior but also must respond to the long-term (tonic) energy status of the organism. Factors derived from the processes of energy storage and the status of the body's energy stores must also contribute to appetite and its expression (e.g., indicators of glucose metabolism and fat storage). Blood carries various substances (other than nutrients) generated in organs implicated in nutrient metabolism and energy storage such as the liver, the pancreas and in adipose tissue depots that reflect the body's energy status and that have been shown to have potent effects on food intake (insulin, glucagons, and leptin). The number of potentially active metabolites and by-products produced by energy metabolism of differing nutrients is vast providing a wide range of potential indicator substances.

As noted earlier one of the classical theories of appetite control has involved the notion of a so-called long-term regulation involving a signal that informs the brain about the state of adipose tissue stores. This idea has given rise to the notion of a lipostatic or ponderostatic mechanism. Indeed, this is a specific example of a more general class of peripheral appetite (satiety) signals believed to circulate in the blood reflecting the state of depletion or repletion of energy reserves that directly

modulate brain mechanisms. Levels of substances such as satietin and adipsin, or cytokine signals such as interleukin-6 (IL-6) tumor-necrosing factors (e.g., TNFα) may all be influenced by adipose tissue. Levels of other circulating hormones, for example, gonadal steroids (androgens, estrogens, and progesterone) also reflect the body fat mass, and so its energy status. Gonadal steroids have potent effects on (both increasing and reducing) food intake, meal size and frequency, body weight, and per cent body fat.

Leptin (ob Protein)

For 40 years, scientists searched for a mechanism by which the brain could monitor body fat deposition in order to keep an animal's body weight constant. In 1994, a gene that controlled the expression of a protein produced by adipose tissue was identified. Circulating levels of this protein (the ob protein) could be measured in normal weight mice. However, in obese ob/ob mice, which display marked overeating, this protein was absent due to a mutation of the ob gene. A series of studies demonstrated that the absence of this protein was responsible for overconsumption and obesity in the obese ob/ob. As the ob protein reduces food intake and also increases metabolic energy expenditure, both of which would result in weight loss, it was named leptin from the Greek 'leptos' meaning thin. In general, circulating levels of leptin appear to reflect the current status of body fat deposition and increase with the level of adiposity demonstrating the responsiveness of endogenous leptin to weight gain and energy status.

Since the identification of leptin and its receptor researchers have isolated numerous CNS hypothalamic neuropeptide systems, which mediate the hypophagic action of leptin. NPY, the melanocortins, corticotropin releasing factor (CRF), cocaine and amphetamine regulated transcript (CART), and the orexins may all be part of the circuit linking adipose tissue with central appetite regulatory mechanisms. Certainly the fact that leptin has multiple effects on complex hypothalamic appetite systems, consisting of wide ranging but integrated regulatory neuropeptides, would appear to support the view that leptin is a major factor in body weight regulation. There is evidence of synergy between leptin and the short-term meal-generated satiety factor CCK. Systemic administration of CCK enhances leptin-induced decreases in food intake and augments weight loss in rodents.

Integration of Episodic and Tonic Signals within the CNS

As stated previously, certain brainstem, hypothalamic, and other limbic sites appear critical in the regulation of food intake and feeding behavior. Within these sites numerous neurochemicals (first neurotransmitters and then neuropeptides) have been identified as potent inhibitors and stimulators of feeding behavior. 5-HT has been implicated as a critical CNS satiety factor in the short-term regulation of food intake. Specifically, the 5-HT system appears to be sensitive to meal-generated satiety factors such as CCK, enterostatin, and ingested macronutrients. Moreover, 5-HT drugs appear to enhance satiety, suppress CNS NPY release, and inhibit hunger. 5-HT appears to mediate the effects of episodic meal-generated satiety on appetite. The second CNS system to be involved is that of the melanocortins, which appear integral in the action of circulating leptin on intake and (like 5-HT) its agonists also inhibit NPY functioning. Thus, the melanocortins may mediate the effects of tonic energy status on appetite.

Of all the monoamines, 5-HT has been most closely linked with the episodic process of satiation and the state of satiety. Moreover, it has been known for a long time that serotoninergic drugs reliably reduce food intake and body weight, both in animals and humans. A number of researchers have identified the critical role of 5-HT_{1B} and 5-HT_{2C} in mediating the satiety effect of 5-HT. Direct agonists of 5-HT_{1B} and 5-HT_{2C} have been shown to reduce intake and to produce changes in feeding behavior consistent with the operation of satiety. Studies with selective 5-HT_{1B} and 5-HT_{2C} agonists in humans confirm their satiety-enhancing properties. Hypothalamic areas, including the PVN, have been implicated in 5-HT hypophagia. 5-HT activation suppresses the levels of the appetite stimulatory peptide NPY within the PVN. Conversely, blocking 5-HT synthesis or antagonizing 5-HT receptors increases NPY functioning in the PVN. This interaction between 5-HT and NPY may be one of the critical mechanisms in short-term (episodic) appetite, determining the generation of either hunger or satiety. As mentioned previously, 5-HT function has been linked to peripheral signals triggered by fat ingestion, such as CCK and enterostatin. Moreover, CNS levels of the 5-HT precursor tryptophan are directly affected by dietary carbohydrate, through its action on the tryptophan/large neutral amino acid competition to cross the blood–brain barrier. Thus, CNS 5-HT is sensitive to both fat and carbohydrate ingestion.

The melanocortins are one of the inhibitory systems through which the tonic adiposity signal leptin inhibits food intake. Like leptin, the role of CNS melanocortins was revealed through the investigation of a genetic mouse model of obesity (agouti (A y/a)) syndrome. These animals were obese, displayed marked hyperphagia, and produced excessive amounts of agouti, an endogenous antagonist of melanocortin receptors. The

hyperphagia was linked specifically to blockade of the melanocortin MC4R receptor. Children who are unable to synthesize endogenous melanocortin receptor agonists also display abnormal eating patterns and obesity. It soon became apparent that the melanocortin receptor MC4R, and a number of its endogenous agonists (such as αMSH), and antagonists (such as AgRP) were part of the endogenous body weight regulation system. MC4R receptors are expressed widely throughout the CNS and in the hypothalamus and these systems appear to mediate the effects of a number of factors such as leptin and insulin. For instance, CNS administration of MC4R receptor agonists inhibits NPY-induced hyperphagia.

Even if tonic signals such as leptin are generated independently from episodic signals they must feedback to reduce intake by altering subjective experiences of hunger and satiety. Increases and decreases in endogenous circulating leptin do have an effect on the modulation of subjective experiences of hunger. Moreover, endogenous leptin levels may fluctuate across the day, in part as a consequence of meal consumption and the effects this has on metabolism. Individuals with an inability to produce leptin experience constant hunger and without the tonic leptin signal the constant drive for energy is unleashed and leads to continuous and voracious food-seeking behavior. Similarly, specific deficits in a system that responds to circulating leptin, in this case the absence of functional melanocortin MC4R receptors, produces similar effects on appetite and intake. These extreme examples demonstrate the inability of short-term meal-generated episodic signals alone to block the override demand for energy (expressed as continuous hunger) generated by tonic basal metabolism.

Summary: Episodic and Tonic Factors in the Regulation of Appetite

Endogenous 5-HT and leptin represent two aspects of negative feedback integral to the appetite control

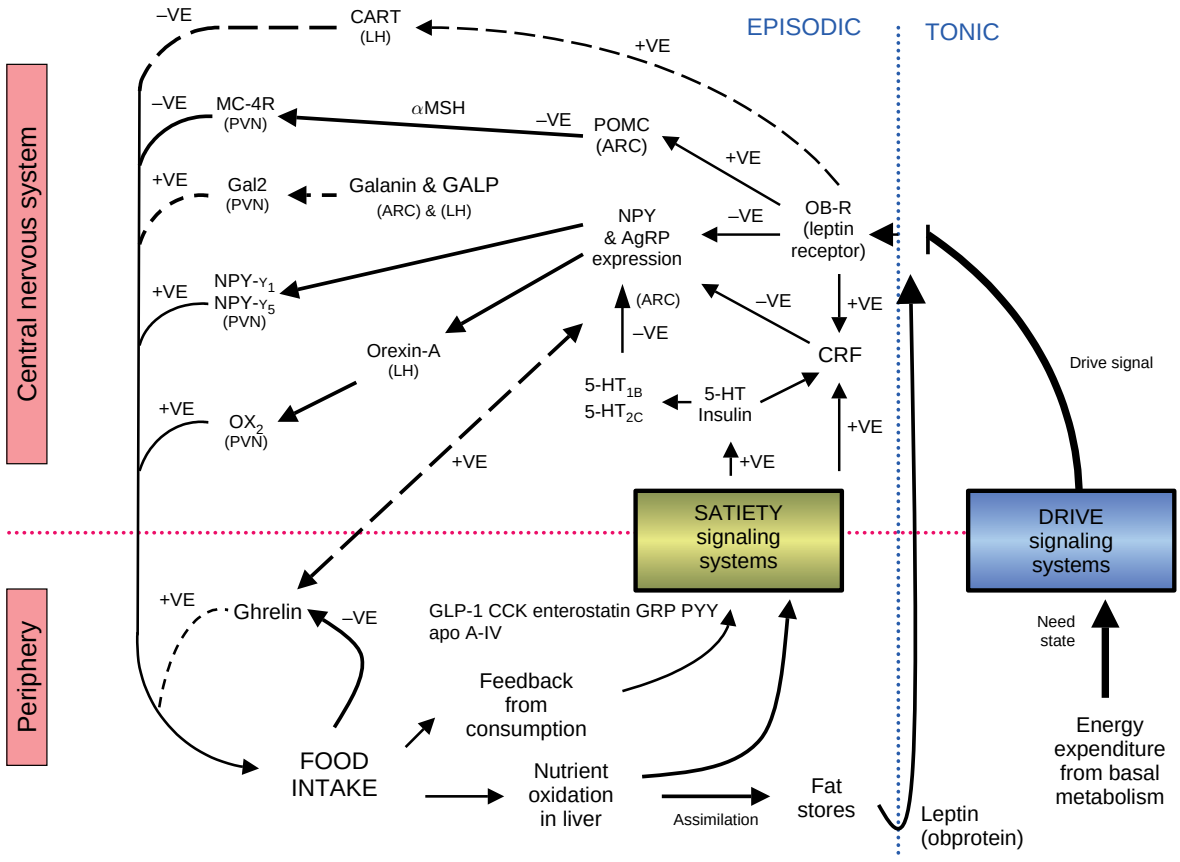


Figure 3 The integration of peripherally generated episodic and tonic signals critical to the expression of appetite. Signals generated by both meal consumption and fat deposition are integrated into a complex hypothalamic system of neuropeptides, which in turn either stimulate or inhibit subsequent food intake. Abbreviations: 5-HT, serotonin; αMSH, alpha melanocortin stimulating hormone; AgRP, agouti-related peptide; A-IV, apoliproprotein-IV; ARC, arcuate nucleus; CART, cocaine and amphetamine-regulated transcript; CCK, cholecystokinin; CRF, corticotropin releasing factor; GAL, galanin; GLP-1, glucagon-like peptide-1; GRP, gastric releasing peptide; LH, lateral hypothalamus; MC, melanocortin; NPY, neuropeptide Y; OX, orexin; PVN, paraventricular hypothalamus; POMC, pre-pro-opiomelanocortin; PYY, peptide YY.

system (see **Figure 3**). Both systems appear to inhibit NPY functioning, the effect of leptin being partly mediated by melanocortins and other excitatory and inhibitory neuropeptides. 5-HT mediates the effect of meal-derived satiety factors derived from pre- and postingestive processes (such as CCK and enterostatin relsease) and promotes meal termination, prolonging the intermeal interval. By such a mechanism the body deals with the daily physiological fluxes that result from meal intake ensuring an approximately appropriate daily energy intake. Circulating leptin accurately reflects the current status of the body's energy store. Leptin levels continually modify total daily and meal food intake to maintain a sufficient but not excessive level of energy deposition. Thus, 5-HT and leptin represent two classes of signals: short-term episodic and long-term tonic feedback, respectively. The net result of the action both episodic and tonic signals will be an adjustment in the expression of appetite, adjusting subsequent feeding behavior to compensate for previous intake and energy stored.

See also: **Appetite**: Psychobiological and Behavioral Aspects. **Brain and Nervous System**. **Hunger**. **Lipoproteins**.

Further Reading

Blundell JE (1991) Pharmacological approaches to appetite suppression. *Trends in Pharmacological Sciences* **12**: 147–157.

Blundell JE and Halford JCG (1998) Serotonin and appetite regulation: Implications for the treatment of obesity. *CNS Drugs* **9**(6): 473–495.

Gehlert DR (1999) Role of hypothalamic neuropeptide Y in feeding and obesity. *Neuropeptides* **33**: 329–338.

Halford JCG, Cooper GD, Dovey TM, Ishii Y, Rodgers RJ, and Blundell JCG (2003) Pharmacological Approaches to Obesity treatment, Current Medical Chemistry. *Central Nervous System Agents* **3**: 283–310.

MacNeil DJ, Howard AD, Guan XM, Fong TM, Nargaund RP, Bednarek MA, Goulet MT, Weinberg DH, Strack AM, Marsh DJ, Chen HY, Shen CP, Chen AS, Rosenblun Cl, Macneil T, Tota M, Maclntyre ED, and Van der Ploeg LHT (2002) The role of melanocortins in body weight regulation: opportunities for the treatment of obesity. *European Journal of Pharmacology* **440**(2–3): 141–157.

Meier JJ, Gallwitz B, Schmidt WE, and Nauck MA (2002) Glucagon-like peptide 1 as a regulator of food intake and body weight: therapeutic perspectives. *Eur. J. Pharmacol.* **440**: 267–279.

Moran TH (2000) Cholecystokinin and satiety: current perspectives. *Nutrition* **16**: 858–865.

Morton GJ and Schwartz MW (2001) The NPY/AgRP neuron and energy homeostasis. *International Journal of Obesity* **25**(supplement 5): s56–62.

Rodgers RJ, Ishii Y, Halford JCG, and Blundell JE (2002) Orexins and appetite regulation. *Neuropeptides* **36**(5): 303–326.

Woods SC and Seeley RJ (2000) Adiposity signals and the control of energy homeostasis. *Nutrition* **16**: 894–902.

Zhang Y, Proenca R, Maffei M, Barone M, Leopold L, and Friedman JM (1994) Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**: 425–432.

Psychobiological and Behavioral Aspects

R J Stubbs and S Whybrow, The Rowett Research Institute, Aberdeen, UK
J E Blundell, University of Leeds, Leeds, UK

The Nature of Feeding Behavior and Appetite 'Regulation'

Mammalian feeding occurs regularly and intermittently and despite a general lack of conscious nutritional knowledge on the part of the animal, usually appears to match energy intake (EI) and nutrient intakes with requirements. How is this achieved? The common explanation is that appetite, EI, or feeding behavior are regulated to ensure that physiological requirements are met. However, there is a lack of direct evidence for this regulation. It may be that neither feeding behavior nor appetite are regulated in a strictly physiological sense since: (1) neither are held constant within certain narrow limits; and (2) feeding responses are not an inevitable response to an altered physiological signal or need. Feeding behavior is responsive to a number of induced states such as pregnancy, cold exposure, growth and development, and weight loss. These responses have often been cited as evidence of a system that is regulated. It is probable that aspects of body size and composition are regulated and that changes in feeding behavior are functionally coupled to those regulatory processes. Indeed, feeding behavior might be said to be adaptive rather than regulated since patterns of food intake are flexible, responsive, and anticipatory, enabling the animal to adapt to changes in the state of the internal and external environment.

Hunger and satiety often have a large learned, anticipatory component rather than being the direct consequences of unconditioned physiological signals *per se*, such as reduced gastrointestinal content. Such physiological events can act as important cues for feeding but they do not necessarily directly determine that behavior.

The mechanism by which feeding behavior is coupled to physiological (and other events) is the process of learning. To understand feeding behavior, hunger, and satiety processes, the mechanism by which learning links feeding behavior to physiological, sensory, nutritional, situational, and other learning cues must be appreciated. This is true for mice and men. This mechanism is termed 'associative conditioning' of preferences, appetites, and satieties.

Learned Appetites, Satieties, and Feeding Behavior

Animals and humans learn (or become conditioned) to associate a given food with the physiological consequences of having ingested it. They associate certain proximal stimuli such as the smell, color, taste, or texture of a food (the conditioning stimulus) with a set of sensations that are directly felt (sensory afferent inputs), in relation to the external stimulus and to the endogenous changes such as physiological and neuroendocrine responses to food. The physiological changes that occur as a result of ingesting the food are termed the 'unconditioned stimulus.' The subject forms a learned or 'conditioned association' between the conditioning stimulus and the unconditioned stimulus (the detectable consequence of eating), which informs them of the sensory and physiological consequences of ingesting that food. This process is summarized in **Figure 1**. Conditioned or learned associations are most efficiently established if the food is sensorially distinct, if there is a significant detectable postingestive consequence of ingesting the food and if a training or learning schedule is encountered (e.g., by repeated exposure to the food under similar conditions). Learning is facilitated by social interaction.

As regards appetite 'regulation' a problem arises when foods are constructed to look and taste like foods with a different composition. For some time after the initial exposure to the food subjects will respond to it in a manner that is determined not by immediate exposure to the food but by what they have learned during the previous period of exposure to the similar foods upon which the learning was originally based. Only if the food produces a very large unconditioned stimulus will this previously learned response be instantly over-ridden. This raises the possibility that the use of food mimetics (e.g., artificial sweeteners) may disrupt stable patterns of learned feeding behavior in consumers at large.

The above view of the nature of feeding behavior has implications for the way the appetite system functions in lean and overweight people.

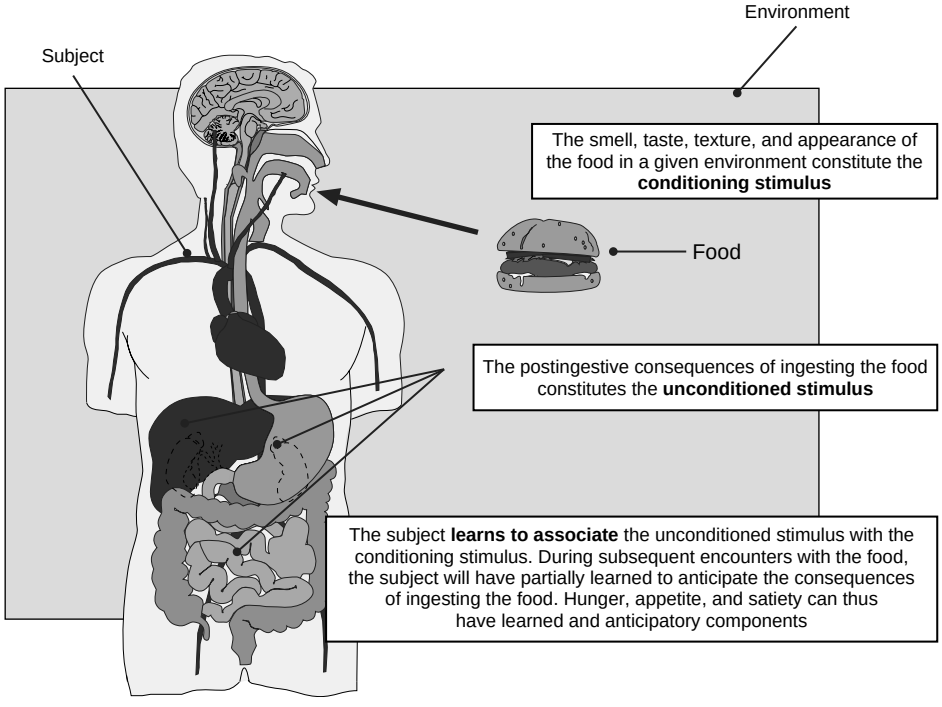


Figure 1 The process by which the subject learns to associate the postingestive consequences of eating with the food eaten and the environment in which it was eaten. Environmental influences can vary in strength from negligible effect to, in extremis, influences so strong that they can constitute the major factor determining a subject's subsequent response to that food.

Physiologists have expended considerable time and effort in attempting to understand how feeding behavior is geared to the regulation of a stable body weight. Obesity is therefore seen as a consequence of defects in this regulation. The evidence from behavioral studies suggests that feeding behavior is inherently more responsive to decreases rather than increases in body weight. Second, current secular trends in body weight suggest that, over time, it is relatively easy to increase body weight, which infers body weight is not tightly regulated, at least with reference to weight gain. For instance, according to the National Health and Nutrition Examination Survey (NHANES) ongoing dietary surveys of American adults, average weight change amounts to a gain of 0.2–2.0 kg year^{-1}. Third, while obese subjects exhibit differences in their feeding behavior and physiology, the literature is remarkably short of clear lean–obese differences in feeding behavior of a type that suggest defects in a regulatory system. For example, evidence suggests that the obese tend to select a diet rich in fat, which in itself facilitates over consumption. However, the tendency to select fat cannot be viewed as a defect in physiological regulation. It may be far more profitable to attempt to understand how feeding responds to environmental and endogenous stimuli, and which of these responses are functional and/or adaptive and which are not. This consideration influences the methodological approaches that attempt to investigate how feeding behavior responds to endogenous and environmental influences.

Methodological Issues

Measuring Hunger

Hunger is a subjectively expressed construct that people use to express a motivation to eat. The most appropriate measure of hunger is its subjective expression at a given time. This is achieved by asking subjects to mark a visual analog scale, which takes the form of a straight line with two extreme representations of hunger anchored at either end. It is most useful to track changes in subjective hunger over time and in relation to feeding events, diet composition, or physiological parameters. Hunger itself exhibits a large learned component (see above) as reflected by the fact that most of the variation in the subjectively expressed hunger of human subjects is accounted for by time. If hunger is plotted against time in Western subjects feeding ad libitum then it generally exhibits three peaks and troughs, which broadly correspond to the three main meal times of a Western feeding schedule (**Figure 2**). While subjective hunger is a relatively poor proxy for the amount eaten it is a reasonably good predictor of when eating will occur. It is important to recognize that hunger can be influenced by a large number of factors and so a search for 'the hunger signal' is likely to prove fruitless. Thus, a large survey of over 600 men, women, boys, and girls could find no clear constellation of traits, sensations, or characteristics that typified hunger. Several laboratories have found inverse correlations between indices of postprandial carbohydrate utilization and hunger. While the postprandial utilization of carbohydrate is likely to influence hunger and indeed may act as a learned cue that conditions hunger it is not the exclusive physiological signal that determines hunger. A variety of hormones and drugs, the sight and smell of food, its perceived palatability, timing, and social situation can all influence hunger.

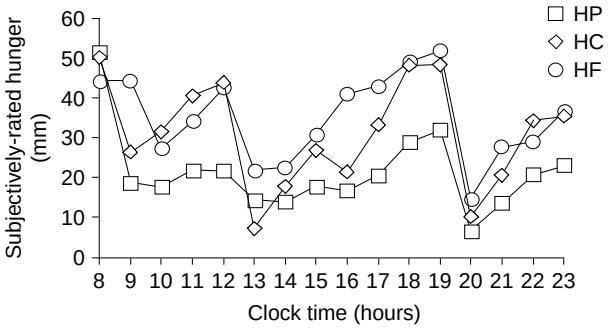


Figure 2 Subjective hunger tracked during waking hours in six subjects feeding on isoenergetically dense high-protein (HP), high-fat (HF), and high-carbohydrate (HC) diets. Subjects exhibit the three peaks and troughs of hunger that typify the Western feeding schedule.

A number of proxy measures of hunger, such as salivation, have been made in an attempt to characterize hunger more objectively. These approaches have had relatively limited success and are difficult to compare across environmental circumstances. Satiety (or postingestive satiety (PI satiety)) is reciprocally related to hunger and can therefore be measured as such.

Measurement of Appetite

Appetite is specific to foods, exhibits wide intersubject variability, and tends to decline for a specific food as that food is eaten, leading to selection of other foods. Appetite is therefore said by Le Magnen to be sensory specific. The sensory specificity of appetite has been shown to relate *inter alia* to the postingestive consequences (satiation and satiety) of having ingested a food. The most objective measure

of appetite for a given food in a specific experimental situation is therefore the amount of that food that a subject chooses to eat. Appetite is not rigidly determined by physiological signals *per se*, although they may greatly influence it. Both the palatability of a food and the appetite for it tend to co-vary and are often increased subsequent to a period of negative energy balance. Two examples are dieting and illness, both of which lead to lowered intake and a subsequent rebound in appetite. As discussed above the appetite for a food will be learned on the basis of the consequences of having ingested that food on previous eating occasions. Because of this it is possible to use covertly manipulated foods to deceive subjects into behaving in a manner largely determined by prior learned experience. If this were not so such deception would be impossible because the physiological signals produced by the sensorially similar yet nutritionally different food would immediately translate (through physiological signals) into behavioral compensation. This fact has implications for consumers since food technology can now dissociate the sensory and nutritional properties of foods, which may undermine the learned basis of food intake control in some people.

Measures of Feeding Behavior

There are many different techniques that can be used to measure feeding behavior in a number of different environmental situations (see 00000 and 00000). These techniques are used in real-life and/or laboratory studies of feeding. Generally, in the laboratory, specific aspects of the system are manipulated at the cognitive, sensory, gastrointestinal, or even the metabolic level by, for example, deceiving subjects about the energy content of foods, altering the sensory variety of the diet, and administering nasogastric or parenteral infusions, respectively. A number of other manipulations can be achieved. Because the environment can have such a large influence on feeding behavior it is important to consider a particular influence on feeding in several contexts. For instance, the effects of fat on energy intake have been considered in several laboratory and real-life contexts. Under most sets of circumstances dietary fat appears to be a risk for excess intake. The relationship between the experimental context and how it influences the investigations made is depicted in **Figure 3**.

Sensory Stimulation and Palatability

Palatability can be measured as the subjective preference for a food, its subjective pleasantness, or indeed the amount (in grams) of a food a subject eats. The relative palatability of a food can be determined by choice tests or taste tests relative to other standard ingestants (e.g., the 5% sugar solution). However, there has been much more controversy over the actual definition of palatability. In general the palatability of a food can be thought of as: (1) the momentary subjective orosensory pleasantness of a food; or (2) its sensory capacity to stimulate ingestion of that

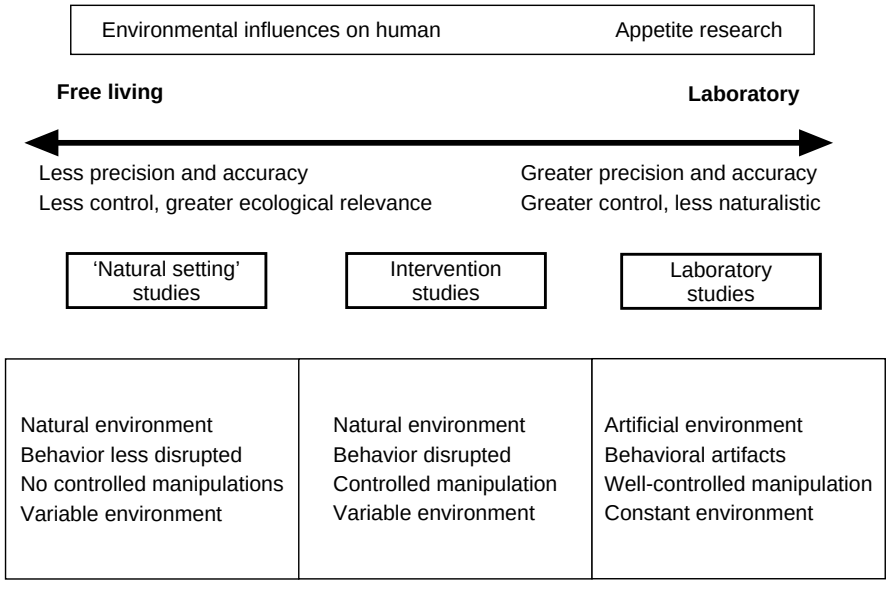


Figure 3 The constraints and limitations that the experimental environment places on studies of human feeding. In general the environment ranges from totally free living, which is realistic but very difficult to make measurements in, to the laboratory where measurements are easy but may be contaminated by artifacts due to the artificiality of the laboratory surroundings.

food. However, the second definition should not be taken to indicate that there is a direct correlation between the perceived palatability of a food and the amount of that food which is ingested. As with hunger, the coupling between the expressed sensation and the amount of food or energy ingested is loose. This definition takes account of the fact that the palatability of the food is jointly determined by the nature of the food (smell, taste, texture, and state), the sensory capabilities and metabolic state of the subject, and the environment in which the food and subject interact. Palatability is therefore not stable; indeed, the palatability of a food typically declines as its own ingestion proceeds. Work on military personnel suggests that the decline in preference for highly preferred foods (e.g., chocolate) is greater than that for staple foods such as bread, which exhibit more stable preference profiles. Palatability can be dissociated from sensory intensity since sensory intensity increases with the concentration of the compound or food being tasted; palatability generally shows a parabolic n-shaped curve with increasing sensory intensity of the food. Palatability can be conditioned, as can aversions. Palatability of a food generally increases with food deprivation.

Because of the mutable nature of palatability and sensory preference the role that they play in determining EI and degree of overweight are unclear. Short-term experiments suggest that more preferred foods stimulate hunger and food intake. Recent work in American and French consumers suggest that on a palatability scale of 1–7 subjects rarely select any food below a score of 3 and that palatability does increase the size of meals. But these works do not address energy balance. Indeed, virtually no published evidence supports the notion that altering dietary palatability or sensory variety *per se* will influence longer term energy balance of human subjects, despite the common perception that increasing dietary palatability will increase intake. This perception is so strong that the food industry feels unable to sacrifice the palatability of their products in developing healthier options, as the risk of decreased consumer acceptance is believed to be too high.

Certain combinations of the sensory and nutritional profiles of foods (e.g., sweet, high-fat foods) are conducive to overconsumption. This effect is often due to the combination of sensory stimuli and the postingestive effects of the food, which re-enforce each other. The individual sensory and nutritional stimuli in isolation are often less effective. Thus, cafeteria regimes, which can produce obesity in rats, typically alter the composition of the diet by increasing its fat content. When rats are given petroleum jelly in chow as a fat mimetic, they initially prefer this diet to normal chow. This preference soon becomes extinct, suggesting that sensory factors alone do not maintain preference.

Sensory Stimuli and Body Weight

It was originally proposed that obese subjects are more susceptible to external stimuli such as sensory stimuli than lean subjects who were more reliant upon internal physiological cues. This predicted that in a Western context the numerous external food stimuli would promote excess EI in susceptible individuals. As can be appreciated from the above, the interplay between 'external' and 'internal' signal is much more complex and interactive than was initially supposed. Nevertheless, given the multiplicity of afferent inputs that can influence feeding it is possible that some subjects have learned to base feeding predominantly on some cues rather than others. However, dividing these cues into simply internal and external sources is perhaps oversimplified. Current questionnaires that attempt to characterize a subject's responsiveness to food do so by dissociating 'externality,' 'restraint' and 'emotionality.'

Systematic differences have been found in sensory preference profiles, but not perceptions of intensity of various tastes between lean and obese subjects. Recent work has suggested that subjects with a history of weight fluctuation have an enhanced sensory preference for high-fat stimuli that are sweet. It has also recently been shown that sensory preference for fats is associated with degree of fatness in adults and children. This may be important. While there is little evidence that sensory stimuli alone promote positive energy balances, there is evidence that people select foods they prefer. Since high-fat foods are usually more energy dense than lower fat foods, preferential selection of these foods can lead to excess EIs without any apparent change in the amount of food eaten. Thus, sensory factors are likely to play a role in the selection of foods that are conducive to weight gain.

Sensory Versus Nutritional Determinants of Intake

The major problem with the concept of sensory preference or palatability as determinants of hyperphagia and obesity in humans is: (1) that there is little direct evidence for this effect *per se* (because

the appropriate experiments are very difficult to carry out); and (2) both animals and humans appear to acquire sensory preferences for foods dense in readily available energy. Dissociation of the sensory characteristics and postingestive consequences of ingesting a food becomes difficult and perhaps artificial. It seems that at the present time the data from animal studies tends to suggest that maximal sensory preference for a food or diet is achieved when the sensory stimulus is reinforced by the metabolic consequences that form part of the satiety sequence. Indeed, there is controversial evidence that ingestion of one sensory stimulus (sweetness) without the associated nutrient (carbohydrate) promotes ingestion of energy shortly afterwards. The contribution of sensory and nutritional determinants of feeding is still poorly understood in humans.

Meal Patterns, Appetite, and Energy Balance

The effect of meal patterns on appetite and energy balance is also an unresolved issue. It has been noted that snacking and commercially available snack food are often believed to elevate EI. However, there is considerably less evidence that meal or snack patterns contribute to the development of obesity. It is important to note at this point that the relationship between a meal and a snack relates to timing and size of ingestive events in meal-feeding animals. In nonhuman species (and indeed humans) who engage in numerous small feeding bouts throughout their diurnal cycle there is little if any distinction between a meal and a snack. Meal-feeding animals are conditioned to ingest the majority of their EI in a few large ingestive events in their diurnal cycle, at approximately the same time points. Under these conditions, a snack can be defined as an energetically small, intermeal ingestive event (SIMIE). To avoid confusion with a common use of the word to describe a certain type of commercially available food, we use the phrase 'commercially available snack foods' to describe those specific foods. Commercially available snack foods tend to differ from the rest of the diet as they are more energy dense, high in fat and carbohydrate and low in protein, and usually contain a large fraction of their edible mass as dry matter. They are by no means the only food eaten as a SIMIE in many people at large.

There are two alternative hypotheses about how snacking may influence EI and body weight: (1) snacking helps 'fine tune' meal-time EI to match intake with requirements; or (2) habitual consumption of calorific drinks and snacks between meals is a major factor driving EI up and predisposing people to weight gain and obesity.

The evidence in relation to meal patterns, appetite, EI, and body weight is, however, indirect and fragmentary. On aggregate, cross-sectional studies tend to support no or a negative relationship between meal frequency and BMI. However, Bellisle *et al.* (1991) convincingly argue that examinations of the relationship between snacking and energy balance in free-living subjects are extensively flawed by misreporting, misclassification of meals and snacks, and potentially by reverse causality. Under these conditions it is difficult to draw clear conclusions about the effects of snacking in cross-sectional studies. It is therefore important to conduct controlled laboratory interventions over a number of days in humans. These studies suggest that in the short to medium term adding mandatory snacks to the diet leads to overconsumption. This effect is most pronounced in those who do not habitually snack and least pronounced in those who do. It is also of note that rats tend to be 'snackers' and Western humans tend to be meal feeders. The rat tends to adjust EI by varying meal frequency, the human by varying meal size. However, if rats are meal fed, they learn to adjust EI by varying meal size. Humans placed in time isolation begin to adjust intake by varying meal frequency. These comparisons illustrate the fact that adjustment of intake to energy or nutrient requirements occurs within a conditioned time framework, which itself is variable depending on the conditioning environment. Despite large changes in the pattern of feeding, EI can still be adjusted to satisfy requirements.

Social and Situational Influences on Feeding Behavior

There are a number of social and situational influences on food intake in humans. In general, the shorter the time period of measurement the greater the effect of situational and social influences. Thus, there are a large number of factors that can influence single meal size in humans. These factors are summarized in **Figure 4**.

Time of day appears to influence meal size in that the amount eaten and the EI increases on going from breakfast, to lunch to the evening meal. Meal size also increases across the feeding period in rats. It has been suggested that this occurs in learned anticipation of the energy requirements in the fasting period (night for humans, day for rats)[6]. Meal size and EI

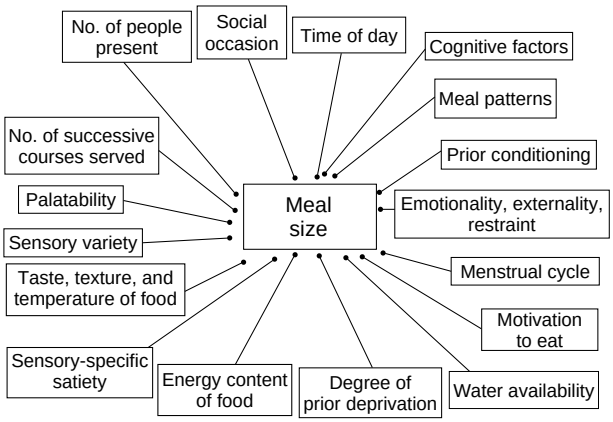


Figure 4 Major factors known to affect single meal size in humans. In general, the shorter the time interval of measurement the greater the influence of these factors on feeding behavior.

tend to be greater at weekends than on weekdays in Western adults. Meal size also varies as a power function of the number of people present at a meal. DeCastro has termed this effect 'social facilitation' of feeding. Social facilitation and daily routine account for much of this effect.

Seasonality can influence feeding. A number of studies suggest that EI, meal size, and eating rate are all elevated in the autumn. In one particular study hunger was associated with meal size in winter and spring, but not so clearly in the autumn.

Cognitive and Social Cues

Throughout the 1960s and 1970s a large number of behavioral studies examined the effect of cognitive and social cues (perceived energy content of foods, salience of cues, eating behavior of others present) on feeding in relation to the externality hypothesis. While a large number of studies found that so-called external cues do relate to short-term feeding behavior, a large number of others did not. However, the presence of external cues alone does not reliably predict how much food people will eat. Neither does the presence of external cues always relate to lean–obese differences in feeding patterns. Some of these differences in relation to cognitive and social cues are better explained in relation to dietary restraint. 'Restraint' is a term used to describe people who are attempting to limit or reduce their body weight by means of cognitive energy restriction (dieting). In doing so it is proposed that they are placing their motivation in relation to feeding at odds with physiological feeding stimuli. Placing cognition at odds with physiological drives can result in pathologies of eating since the normal 'regulatory' processes are cognitively undermined. According to Herman and Polivy, these aberrations can extend into disturbances of emotion and cognition, which, *in extremis*, may partially underlie the increase in the prevalence of eating disorders. Furthermore, it is argued that restraint will increase the probability that a person will break a diet. It has been shown that an intervention (usually a preload) that breaks the rules of restraint, almost paradoxically induces a greater intake. This phenomenon has been termed 'counter-regulation.' This effect is cognitive, since it can be induced by deceiving a restrained eater into believing that a preload was high in calories. Because the concept of restraint has predictable behavioral outcomes it is a useful tool in characterizing different people with respect to their feeding behavior.

Macronutrients and Appetite

Food and nutrient ingestion influence human appetite through multiple feedbacks at several levels, which can be traced through the processes of food location, ingestion, digestion, absorption, and metabolism. Satiety is therefore maintained by a functional sequence or cascade of sequential physiological events that reinforce each other. Removing parts of a food or nutrient's effects from this sequence will therefore diminish its impact on satiety.

Protein

Protein suppresses EI to a greater extent than any of the other macronutrients. This effect is apparent in free-living subjects and in the laboratory. There may be a critical threshold in the amount of protein required to suppress subsequent EI since studies that have found little effect of protein relative to other macronutrient preloads have only used small amounts of energy in the preloads. Thus, protein appears to be particularly satiating when given in moderate and large amounts (**Figure 5**).

The mechanisms responsible for the apparent appetite-restraining effect of protein have not yet been determined. Essential amino acids when ingested in excess of requirement form a physiological stress that must be disposed of by oxidation. It is known that animals will alter feeding behavior in order to alleviate a physiological stress. Pigs, in particular, appear capable of learning to select a protein:energy ratio in the diet that is optimal for growth, as can rats.

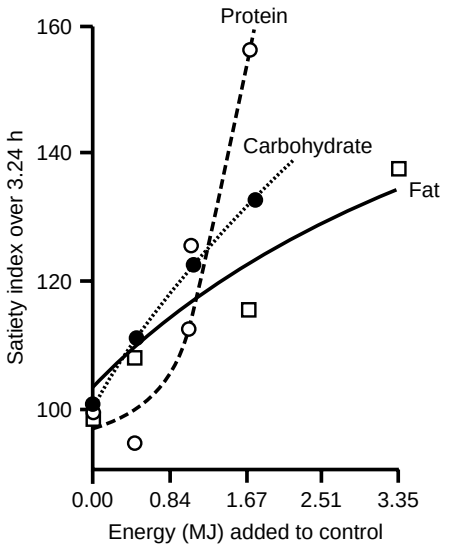


Figure 5 Effect of increasing energy content of macronutrient loads on satiety index subjectively expressed over 3.25 h. (Reproduced from Weststrate JL (1992) Effect of nutrients on the regulation of food intake. Unilever Research, Vlaardingen, The Netherlands.

Carbohydrates

By the mid 1990s it was generally accepted that carbohydrates are absorbed, metabolized, and stored with less energetic efficiency than dietary fat and were protective against weight gain. Indeed, a general perception was developing that because *de novo* lipogenesis appears limited when humans feed on Western diets, carbohydrate ingestion does not promote fat storage. At the same time the notion that carbohydrate metabolism or stores exert powerful negative feedback on EI became quite firmly established. By the same reasoning, diets high in carbohydrates were deemed to be more satiating, specifically because they were high in carbohydrates. High-fat diets were seen to promote overconsumption because they are relatively low in carbohydrate.

Recently, doubts have surfaced about the paramount role of carbohydrates as the central nutrient around which energy balance is regulated and body weight controlled. Several rigorous tests of carbohydrate-specific models of feeding have suggested that carbohydrate oxidation or stores do not exert powerful negative feedback on EI. Rather, as macronutrients come in the diet (where fat is disproportionately energy dense) there appears to be a hierarchy in the satiating efficiency of the macronutrients protein, carbohydrate, and fat. Per megajoule of energy ingested protein induces supercaloric compensation, carbohydrate generates approximately caloric compensation, and fat precipitates subcaloric compensation, and hence often excess EI. When energy density is controlled protein is still far more satiating than carbohydrates or fats (at least when ingested in excess of 1–1.5 MJ loads). Under these conditions differences in the satiating efficiency of carbohydrates and fats become subtle. Some studies are now showing that it is possible to overeat when consuming a high-carbohydrate, energy-dense diet. Furthermore, in recent studies where both fats and sugars are added to the diet, there is no evidence that increasing sugar intake levers fat out of the diet and protects against weight gain.

The foods most capable of limiting EI (both voluntary and metabolizable) are those rich in unavailable complex carbohydrates. However, humans are not too fond of these foods. The average Western adult's fiber intake is spectacularly low.

It has recently been suggested that carbohydrates with a high glycemic index are especially conducive to weight gain. The evidence relating to the glycemic index of carbohydrates and appetite control is currently very inconclusive. It is likely several factors associated with readily absorbed carbohydrates can promote higher energy intakes. These include their sweetness, ready solubility, and ease with which they can be added to foods and absorbed across the gut wall. It may be a coincidence that these traits also determine the high glycemic load of these carbohydrates. Thus, while certain high-glycemic-index carbohydrates may promote higher energy intakes the effect may not be due to their glycemic index *per se*.

Fat

Numerous studies have now shown that when humans or animals are allowed to feed ad libitum on high-fat (HF) energy-dense diets, they consume similar amounts (weight) of food but more energy (which is usually accompanied by weight gain) than when they feed ad libitum on lower fat, less energy-dense diets. However, fat is not likely to be the only risk factor for over consumption and few analyses take account of how fat may interact with other nutrients. For instance, sweet high-fat foods have a potent effect on stimulating EI.

We are beginning to gain insights into the effects of types of fat on appetite control, due to the search for forms of fat that do not predispose the general population to weight gain. There is already some evidence that certain subtypes of fat limit the excess EI that occurs as a consequence of ingesting a high-fat diet. In the future specific nutrients could be tailored to exert quantitatively significant effects on appetite control, tissue deposition, and energy balance. In this context it is of note that certain isomers of conjugated linoleic acid (CLA) can be used to suppress appetite and fat deposition in animals and perhaps humans.

The Combined Effects of Macronutrients and Energy Density on Energy Intake

There has recently been considerable debate as to whether the effects of diet composition on EI can be simply explained in terms of dietary energy density. Here energy density is defined as the metabolizable energy per unit weight of ready to eat food. The major determinants of dietary energy density are water and fat, with water having the greatest effect. In general the energy density of ready to eat foods is largely determined by a fat–water seesaw, with energy density falling as the water content of food rises and as the fat content of foods falls. Protein and carbohydrate contribute relatively little to dietary energy density. There is considerable scope for technological developments that can alter the energy density of foods without compromising palatability.

Energy density exerts profound effects in constraining EI in short-to-medium term studies. Subjects behave differently in longer-term interventions because they learn to adjust their feeding behavior. Energy density is a factor, which at high levels can facilitate excess EI, and at low levels constrains EI. However, the effects that dietary energy density may exert on appetite and EI should be considered in the context of other nutritional and non-nutritional determinants of EI rather than as a substitute for those considerations.

Multifactor models appear more appropriate to explain nutritional determinants of feeding since they explain a far greater proportion of the variance in EI than single nutrient-based models (see **Figure 6**).

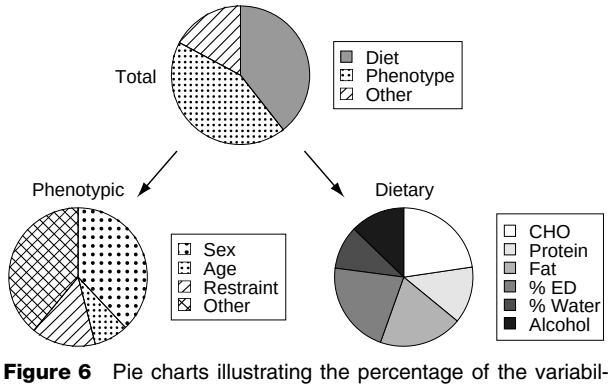


Figure 6 Pie charts illustrating the percentage of the variability in energy intake ascribable to different sources in 102 subjects self recording their food intake for 7 consecutive days. Approximately 39% of the variability was due to diet and ~40% was due to intersubject variability. These two major sources of variation are subdivided further. These charts clearly illustrate that the determinants of energy intake in human adults is multifactorial. ED, energy density.

Micronutrients

There is currently very little data on the effect of micronutrients on feeding behavior and body weight under normal feeding conditions. There is evidence that rodents will learn to select a diet that alleviates a micronutrient deficiency. It may also be supposed that the administration of a micronutrient that will, for instance, improve a deficiency-related defect in nutrient metabolism will also improve appetite for that nutrient, and perhaps appetite in general. For instance, heat-stressed chickens will learn to consume more of a food rich in ascorbic acid, which apparently helps alleviate the stress.

See also: **Appetite**: Physiological and Neurobiological Aspects. **Dietary Intake Measurement**: Methodology; Validation. **Dietary Surveys**. **Energy**: Balance. **Hunger**. **Meal Size and Frequency**. **Obesity**: Definition, Etiology and Assessment.

Further Reading

Blundell JE (1979) Hunger, appetite and satiety – Constructs in search of identities. In: Turner M (ed.) *Nutrition and Lifestyles*, pp. 21–42. London: Applied Science Publishers.

Blundell JE and Stubbs RJ. Diet and food intake in humans. In: Bray GA, Bouchard C, and James WPT (eds.) *International Handbook of Obesity*, 2nd edn. Dekker Inc.(in press).

Booth DA, Lee M, and Macleavey C (1976) Acquired sensory control of satiation in man. *British Journal of Psychology* **67**: 137–147.

De Castro JM (1997) How can energy balance be achieved by free-living human subjects? *Proc Nutr Soc* **56**: 1–14.

Forbes JM (1995) *Diet, Voluntary Food Intake and Selection in Farm Animals*. Oxford: CAB International.

Friedman MI and Tordoff MG (1986) Fatty acid oxidation and glucose utilisation interact to control food intake in rats. *American Journal of Physiology* **251**: R840–R845.

Herman P and Polivy J (1991) Fat is a psychological issue. *New Scientist* 19th November: 41–45.

Hill AJ, Rogers PJ, and Blundell JE (1995) Techniques for the experimental measurement of human eating behaviour and food intake: a practical guide. *International Journal of Obesity* **19**: 361–375.

Langhans W and Scharrer E (1992) The metabolic control of food intake. *World Review of Nutrition and Diet* **70**: 1–68.

Le Magnen J (1992) *Neurobiology of Feeding and Nutrition*. California Academic Press, Millbrae, CA.

Mattes R (1990) Hunger ratings are not a valid proxy measure of food intake in humans. *Appetite* **15**: 103–113.

Mattes RD (1985) Gustation as a determinant of ingestion: methodological issues. *American Journal of Clinical Nutrition* **41**: 672–683.

Mayer J (1955) The regulation of energy intake and the body weight. *Annals of the New York Academy of Science* **63**: 15–43.

Monello LF and Mayer J (1967) The Hunger and satiety sensations of men, women, boys and girls. *American Journal of Clinical Nutrition* **20**: 253–261.
Ramirez I (1990) What do we mean when we say palatable food? *Appetite* **14**: 159–161.
Schachter S and Rodin J (1974) *Obese Humans and Rats*. Washington DC: Erlbaum/Halsted.
Spitzer L and Rodin J (1981) Human eating behaviour: a critical review of studies in normal weight and overweight individuals. *Appetite* **2**: 293–329.
Stubbs J, Ferres S, and Horgan G (2000) Energy density of foods: effects on energy intake. *Critical Reviews in Food Science and Nutrition* **40**(6): 481–515.

ARTHRITIS

L A Coleman, Marshfield Clinic Research Foundation, Marshfield, WI, USA
R Roubenoff, Millennium Pharmaceuticals, Inc., Cambridge, MA, USA and Tufts University, Boston, MA, USA

Introduction

There are many different types of arthritis, which, broadly defined, can be categorized into two major groups: degenerative and inflammatory. Osteoarthritis (OA) is the most common type of degenerative arthritis. In an inflammatory arthritis such as rheumatoid arthritis (RA) there is a systemic illness with inflammation of many joints. There is evidence of a systemic immune response; there is also activation of the acute-phase response. The systemic inflammation leads to altered energy and protein metabolism and wasting of body cell mass and muscle mass, described as 'rheumatoid cachexia.'

Dietary management of these two major types of arthritis differs substantially. In OA, the primary goal of dietary management is prevention via weight loss. Very little is known about the prevention of RA; the primary goal is treatment of inflammatory symptoms and prevention of joint damage and disability. We discuss dietary therapies including modification of dietary fatty acids and supplementation, in addition to vitamin and mineral supplementation and elimination diets; and make recommendations for optimizing dietary intake in order to attempt to alleviate disease symptoms.

Definitions and Etiology

Over a hundred types of arthritis are currently recognized. Among the degenerative arthritides, OA is the most common form, and the prototype of this group. In OA there is inflammation within the joint, but there is no evidence of whole-body inflammation, a key feature in distinguishing OA from the inflammatory arthritides. In general, OA affects a few joints, usually the large weight-bearing joints of the lower extremities, such as the knees and hips. Osteoarthritis can also affect the hands, especially in women, but without the systemic illness that characterizes inflammatory diseases such as RA. The etiology of OA is unknown, but the primary pathological problem is degradation of the cartilage leading to loss of joint space and bony overgrowth, causing pain first with weight-bearing, then with passive motion, and finally at rest.

In contrast, in an inflammatory arthritis such as RA there is a systemic illness with inflammation of many joints, usually the small joints of the hands, wrists and feet, often spreading to include the knees and hips. There is evidence of a systemic immune response, with activation of clones of autoreactive T cells and increased production of many cytokines, including interleukin (IL)-1β, tumor necrosis factor (TNF)-α, IL-6, and others. There is also activation of the acute-phase response, with reduced albumin synthesis and increased production of fibrinogen, C reactive protein, and other acute-phase reactants. The systemic inflammation leads to altered energy and protein metabolism and wasting of body cell mass and muscle mass, described as 'rheumatoid cachexia.'

Prevalence

Osteoarthritis is the most common joint affliction, and its prevalence increases dramatically with age. Radiographic evidence of OA is seen in 70% of people aged over 65 years, but symptoms do not necessarily correspond with X-ray changes. Osteoarthritis of the hip has been reported in 7–25% of adults aged 55 years and older, while knee OA is approximately twice as common, and OA of the hands three times as common. Rheumatoid arthritis,

on the other hand, affects 1–2% of the population, but generally attacks many more joints than OA and is associated with a twofold or higher increased risk of death. Other inflammatory arthritides, such as the seronegative spondyloarthropathies, are much less common.

Clinical Features

The term 'arthritis' simply means the presence of pain and inflammation (heat, swelling, redness) in a joint. Joint pain without inflammation is 'arthralgia', and may be due to disease within the joint or in the surrounding soft tissues, ligaments, and tendons. Degenerative arthritis such as OA is generally a disease of the large weight-bearing joints of the lower extremities, such as the knees and hips. In addition, OA commonly strikes the distal interphalangeal (DIP) and first carpometacarpal joints of the hands, especially in women. The affected joints have pain on motion, mild swelling, and sometimes intra-articular effusions or swelling. As the disease progresses, bony overgrowth becomes clinically apparent, coinciding with the development of osteophytes on radiographic examination. These osteophytes, together with loss of joint space, are the radiographic hallmarks of OA, and reflect new bone formation at the joint margins. Over time, the range of motion in the joint is restricted, first by pain, later by loss of joint space, and finally by the osteophytes. Treatment of OA is essentially symptomatic, using analgesics and nonsteroidal anti-inflammatory drugs (NSAIDs) to reduce pain and limit the intra-articular inflammation. However, this treatment is seldom completely satisfactory, and progression of the disease is usually seen. Joint replacement surgery has revolutionized the care of end-stage OA, allowing return of function of joints that are otherwise immobile.

In inflammatory arthritis the situation is quite different. Rheumatoid arthritis is a symmetric, additive polyarthritis involving up to several dozen small joints of the hands, wrists, and feet, often with involvement of the knees, hips and ankles, and sometimes the elbows, shoulders and cervical spine. There is pain, swelling, and warmth in the affected joints and stiffness upon awakening or after prolonged immobility that can last for several hours. Unlike OA, in RA there is evidence of whole-body inflammation with activation of the acute-phase response. This leads to suppression of albumin gene expression and upregulation of the production of acute-phase proteins such as C reactive protein, transferrin, and fibrinogen. In addition, there is suppression of serum iron, increased zinc, and increased whole-body protein breakdown and resting metabolic rate. Treatment begins with rest, physical therapy, and use of NSAIDs to reduce pain. Low-dose oral corticosteroids, equivalent to 5–10 mg day^{-1} of prednisone, are often necessary to control symptoms. However, these therapies do not alter the natural history of the disease. The best chance of doing so rests with the so-called 'slow-acting antirheumatic drugs' (SAARDs), such as methotrexate, TNF-α inhibitors, and other medications that have been shown to prevent erosions. It should be noted that some of these medications may also affect the nutritional status of individuals with RA via either altered appetite, blood sugar, plasma lipids, absorption, or protein metabolism (**Table 1**).

Table 1 Examples of drug side effects on nutritional status

Effect	*Drugs*
Appetite increased	Alcohol, insulin, steroids, thyroid hormone, sulfonylureas, some psychoactive drugs, antihistamines
Appetite decreased	Bulk agents (methylcellulose, guar gum), glucagon, indometacin, morphine, cyclophosphamide, digitalis
Malabsorption	Neomycin, kanamycin, chlortetracycline, phenindione, *p*-aminosalicylic acid, indometacin, methotrexate
Hyperglycemia	Narcotic analgesics, phenothiazines, thiazide diuretics, probenecid, phenytoin, coumarin
Hypoglycemia	Sulfonamides, aspirin, phenacetin, β-blockers, monoamine oxidase inhibitors, phenylbutazone, barbiturates
Plasma lipids reduced	Aspirin and *p*-aminosalicylic acid, L-asparaginase, chlortetracycline, colchicine, dextrans, fenfluramine, glucagon, phenindione, sulfinpyrazone, trifluperidol
Plasma lipids increased	Oral contraceptives (estrogen-progestogen type), adrenal corticosteroids, chlorpromazine, ethanol, thiouracil, growth hormone, vitamin D
Protein metabolism decreased	Tetracycline, chloramphenicol

From The Merck Manual of Diagnosis and Therapy, Edition 17, p. 22, edited by Mark H. Beers and Robert Berkow. Copyright 1999 by Merck & Co., Inc., Whitehouse Station, NJ.

Role of Diet in the Management of Inflammatory Arthritis

Nutritional Assessment in Rheumatoid Arthritis

It is important to recognize that patients with RA do not have normal nutritional status. Compared with

healthy persons of the same weight, age, race, gender, and height, patients with RA have lower body cell mass (especially muscle mass) and increased fat mass. This condition has been termed 'rheumatoid cachexia,' and it occurs despite adequate and even excessive dietary intake. This cachexia is generally seen in the presence of hypermetabolism (elevated resting energy expenditure) and hypercatabolism (elevated protein breakdown), along with reduced physical activity. These metabolic abnormalities are linked to increased production of the catabolic cytokines IL-1β and TNF-α. The problems are further exacerbated by reduced physical activity, which reduces the anabolic stimulus to muscle, and disordered growth hormone kinetics. In addition, patients with RA have lower concentrations of serum albumin and other markers of visceral protein status, and often have anemia of chronic disease with disordered iron metabolism.

Although many foods or food components have been considered as possible treatments for RA, most studies have focused on either supplementation (particularly the use of fish oil) or the use of an elimination diet, especially fasting or a vegetarian regimen.

Supplementation with Dietary Fatty Acids

Various dietary fatty acids have been shown to have numerous immunomodulatory effects. Arachidonic acid (AA, 20:4 *n*-6) is synthesized in mammalian tissues from the essential fatty acid linoleic acid (18:2 n-6), found in many plant products. The release of AA from cell membrane phospholipids via the action of phospholipase A_2 results in the subsequent production of AA-derived eicosanoids, such as prostaglandin (PG) E_2 and leukotriene (LT) B_4, which have potent proinflammatory and chemotactic effects. Alternatively, when AA is replaced with an *n*-3 fatty acid in the diet, such as eicosapentaenoic acid (EPA, 20:5 *n*-3) or docosahexaenoic acid (DHA, 22:6 *n*-3), there is competitive inhibition of the use of AA as a substrate, and eicosanoids with different biological activity (PGE_3 and LTB_5) are produced through the cyclooxygenase and 5-lipoxygenase cellular metabolic pathways (**Figure 1**). More specifically, EPA-derived eicosanoids result in decreased platelet aggregation, reduced neutrophil chemotaxis, and anti-inflammatory effects. Omega-3 fatty acids are derived primarily from marine sources, including fish and shellfish. Because modulation of dietary fatty acids can alter cellular eicosanoid production, it has been hypothesized that increased consumption of *n*-3 fatty acids can affect the immunologic and inflammatory responses accompanying RA.

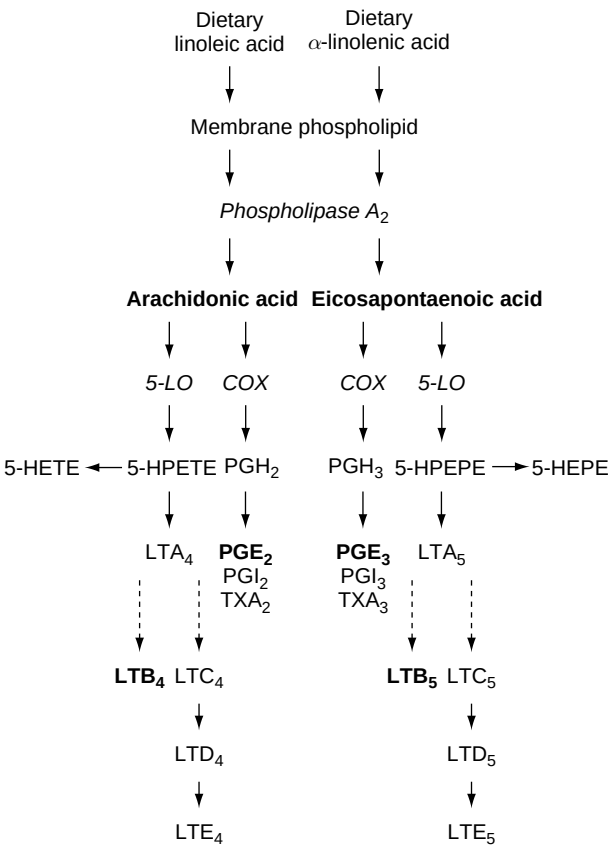


Figure 1 Simplified diagram of eicosanoid formation. Enzymes are italicized. Key intermediates are in bold. *COX*, cyclooxygenase; 5-HEPE, 5-hydroxyeicosapentaenoic acid; 5-HETE, 5-hydroxyeicosatetraenoic acid; 5-HPETE, 5-hydroperoxyeicosatetraenoic acid; 5-HPEPE, 5-hydroperoxyeicosapentaenoic acid; LT, leukotriene; PG, prostaglandin; TX, thromboxane.

Eicosapentaenoic acid supplementation causes modest improvement in the number of tender joints and fatigue among patients with RA although clinical benefits have generally been small, subjective, and transient. Possible mechanisms for this improvement in clinical symptoms of inflammation include decreased LTB_4 production, altered neutrophil membrane lipid composition, reduced interleukin-1 (IL-1) production, or a change in the α-tocopherol content of the diet. Overall, findings suggest that clinical benefits of dietary supplementation with *n*-3 polyunsaturated fatty acids (PUFAs) are more commonly observed among patients consuming higher dosages of fish oil, for longer periods than those previously studied. Indeed, beneficial clinical effects have been observed for as long as 1 year among patients with RA ingesting 2.6 g daily of *n*-3 PUFA supplements. In terms of optimal dosage of supplementation, however, a level of $130\,mg\,kg^{-1}\,day^{-1}$ (9 g of *n*-3 PUFAs in a person weighing 70 kg) has been shown to result in no

additional improvement compared with patients receiving doses ranging from 3 to 6 g daily. Therefore, although the optimal level of fish oil supplementation is yet to be determined, there does appear to be an upper limit beyond which no additional benefit exists for patients.

Although some studies seem to suggest modest clinical improvements as a result of dietary fish oil supplementation in patients with RA, the question of the effects of a patient's medical regimen on the efficacy of fish oil supplementation remains. Nonsteroidal anti-inflammatory agents are known to inhibit the cyclooxygenase enzyme system, which is the same pathway that seems to be inhibited by EPA and DHA. It is therefore possible that in studies of fish oil supplementation where patients are simultaneously maintained on NSAIDs, the effect of EPA is diminished, since the cyclooxygenase pathway is already inhibited by concurrent treatment with NSAIDs. Several studies have attempted to address this issue and have demonstrated a modest effect of *n*-3 fatty acid supplementation in both patients who are treated with NSAIDs and those who are not, suggesting that concurrent treatment with NSAIDs does not seem to diminish the effect of *n*-3 fatty acids. On the other hand, a clinically important NSAID-sparing effect of fish oil among patients who discontinue NSAIDs has not been demonstrated, suggesting that the benefits of fish oil supplementation are modest at best, relative to the effects of medication.

Although the majority of studies regarding manipulation of dietary fatty acids have focused on fish oil supplementation, other fatty acids have also been studied. The use of α-linolenic acid, the precursor of EPA and DHA, has not been shown to be of any benefit in RA. However, γ-linolenic acid, found in blackcurrent seed, evening primrose, and borage seed oils, has resulted in clinically important reductions in the signs and symptoms of disease activity in patients with RA, perhaps via a reduction in PGE_2, IL-1, and IL-6. Because these oils do not cause an unpleasant fishy taste and odor in recipients, they may be preferred to fish oils for chronic treatment.

In summary, most studies of dietary supplementation with *n*-3 fatty acids suggest a modest improvement in clinical symptoms associated with RA, which are to some extent dose-and time-dependent. The most consistent clinical benefits have been reductions in tender joint counts and morning stiffness. Studies do not suggest that benefits are great enough to warrant discontinuing patients' other medications. However, the use of fish oil supplements, or diets rich in marine fish, may further improve clinical symptoms among patients with RA. Beyond the possible benefits in terms of controlling inflammatory symptoms of RA, increases in *n*-3 PUFAs are also associated with reduced risk of cardiovascular disease and other health benefits. Thus, there is reason to promote fish consumption among patients with RA consistent with general healthy eating recommendations. Both the American Heart Association and World Health Organization suggest consuming a minimum of two servings of fish per week, with one or more of the servings as oily fish. Those individuals who do not consume fish regularly may consider supplementation with modest levels of fish oils; however, until more is known about optimal dosages, caution should be taken with the use of concentrated high-dose fish oil supplements. Of note, there has been concern raised over environmental contamination of some types of marine fish with methylmercury. Therefore, eating a variety of fish will help to reduce any potentially negative health effects due to environmental contaminants.

Vitamin and Mineral Supplementation

Most studies involving vitamin or mineral supplementation in RA have focused on either the antioxidant nutrients (vitamin C, vitamin E, beta carotene, selenium) or B vitamins. Various studies have examined the effects of vitamins C, E, and selenium supplementation on the management of RA. In general, results from randomized controlled trials of vitamin E supplementation have been of relatively short duration and have led to conflicting results so that there continues to be a lack of concrete evidence to support vitamin E supplementation at a particular dosage. Nonetheless, patients with RA could certainly be encouraged to increase their intake of vitamin E-rich foods, including edible vegetable oils (sunflower, safflower, canola, olive), unprocessed cereal grains, and nuts. Similarly, the effect of dietary sources of other antioxidant nutrients, such as selenium and vitamin C, on inflammatory symptoms in RA has also been ambiguous. It should be emphasized that providing individual nutrient supplements does not necessarily offer the same overall benefit as when nutrients are obtained from whole foods. It is possible that the combination of nutrients that are present in whole foods, or even some unidentified components of a food, are responsible for any observed beneficial effects, and that supplementing a typical diet with individual nutrients will not provide the same benefit.

We have studied vitamin B_6 levels in patients with RA and healthy controls, and found that plasma levels of pyridoxal-5-phosphate (PLP), the metabolically active form of vitamin B_6, were lower in

patients with RA compared to control subjects. Furthermore, plasma levels of PLP were inversely associated with TNF-α production by peripheral blood mononuclear cells, suggesting that abnormal vitamin B_6 status may be contributing to inflammation in RA. However, there is no evidence to support the efficacy of oral vitamin B_6 supplements for treating the symptoms of RA at this time. Furthermore, large doses of vitamin B_6 can be toxic; therefore, as with the anti-oxidant nutrients, patients with RA would obtain the greatest benefit by increasing dietary sources of vitamin B_6, consistent with the dietary reference intake (DRI) for this nutrient. If supplementation is considered, it should not exceed twice the DRI level.

Fasting and Vegetarian Diets

An alternative approach to alleviating the symptoms associated with chronic inflammation is elimination of various foods or food components, most often by fasting or a vegetarian diet. Some studies have demonstrated a significant improvement in various objective and subjective measures of disease activity, including number of tender and swollen joints, Ritchie articular index, duration of morning stiffness, erythrocyte sedimentation rate (ESR), C reactive protein (CRP), grip strength and score on health assessment questionnaires, among patients with RA 6 weeks to 2 years after initiating a vegetarian diet. Furthermore, these clinical improvements were accompanied by changes in biochemical and immunological parameters consistent with a substantial reduction in inflammatory activity. Other studies, however, have demonstrated no clinical improvement among patients with RA following a vegetarian diet.

Several possible mechanisms have been proposed to explain the impact of elimination diets on clinical symptoms in RA. One possibility is that RA might be the result of hypersensitivity to environmental toxins or specifically to foods or food-related products, resulting in a food allergy of sorts that exacerbates symptoms of RA. However, true food intolerance, involving a systemic humoral immune response against food items, appears to be relatively uncommon among patients with RA. Another possible mechanism that has been proposed includes an alteration in the fatty acid content of the diet. Vegetarian diets contain more linoleic acid, but less AA, EPA, and DHA than omnivorous diets. Therefore, the eicosanoid precursors (AA, EPA, and DHA) must be produced endogenously from linoleic and α-linolenic acid (see **Figure 1**). It has been hypothesized that if this endogenous production cannot compensate for the absence of AA in the diet, then the precursor of the proinflammatory eicosanoids would be reduced, perhaps explaining the beneficial effect of vegetarian diets in patients with RA. Furthermore, it has also been demonstrated that fasting for 7 days resulted in decreased release of LTB_4 from neutrophils, in addition to reductions in morning stiffness, articular index, and ESR, but that this reduction in LTB_4 occurred despite an increased AA content of the serum, platelets, and neutrophils. These findings suggest that perhaps fasting may impair a metabolic step of AA conversion.

Other potential mechanisms include the possible effect of a vegetarian diet on antioxidant status, or on other dietary practices frequently associated with vegetarianism. Plant-based foods are naturally high in antioxidant nutrients (vitamin C, vitamin E, and beta-carotene) and low serum antioxidant levels have been associated with an increased risk of developing RA, although the specific mechanism involved remains unknown. Certainly, RA is associated with increased production of reactive oxygen species; these compounds seem to contribute to the inflammatory process, so a diet high in antioxidants could limit damage via their anti-inflammatory properties. While changes in fatty acid composition or antioxidant status seem to be the most plausible explanations for the potential benefit of adhering to a vegetarian diet, there are other possible mechanisms as well. Fasting, for example, suppresses inflammation and frequently a period of fasting is recommended prior to initiating an elimination or vegetarian diet; it is possible that this fasting period contributes to the reduction in inflammation among patients with RA following a vegetarian diet.

In summary, the notion that food sensitivity reactions contribute significantly to clinical symptoms associated with RA remains controversial. However, it seems that at least a small subgroup of patients with RA may benefit from individualized dietary manipulation involving elimination of specific foods or food components, in combination with other medical therapies. However, fasting and other elimination diets should be used with caution in light of the prevalence of rheumatoid cachexia in this population. Such patients are prone to further loss of cell mass during restrictive diets, and the net effect may be to do more harm than good.

Conclusions

Of the two primary approaches to the dietary management of inflammatory arthritis – supplementation and elimination diets – it appears that dietary supplementation with fish oil may result in the most

consistent clinical benefits, although improvements still remain modest. Elimination diets, including fasting and vegetarian regimens, may provide some benefit for a limited number of patients with RA, but consistent alleviation of disease activity by objective clinical measures has not been demonstrated.

In neither case does the use of dietary management warrant discontinuing a patient's medical regimen; rather, diet may be useful as an adjunct to other more substantiated therapies. Perhaps the most prudent approach for patients with RA interested in attempting to control their disease activity through diet is to recommend a diet consistent with current recommendations for all individuals, including an intake high in fresh fruits, vegetables, and grains, with moderate amounts of lean meats and poultry, and an emphasis on fish, particularly marine fish high in *n*-3 fatty acid content. More definitive research demonstrating consistent, objective clinical benefits is needed before specific dietary manipulations for patients with RA can be recommended. In addition, there is growing evidence that exercise, both resistance and aerobic, has important beneficial effects on both inflammatory and degenerative arthritis.

Role of Diet in the Management and Prevention of Degenerative Arthritis

Much less is known about the role of diet in the treatment of OA and other degenerative arthritides. The above discussion regarding *n*-3 PUFAs in RA also may pertain to OA, although the strength of the effect has not been studied as thoroughly. However, the same eicosanoid metabolism occurs in OA as in RA, with the exception that the disorder is limited to the joint rather than involving the whole body. Thus, fish oils may well be of benefit in OA. Antioxidant intervention with vitamin E may also be effective in OA, with several studies showing an effect comparable with NSAIDs. Although not strictly nutrients, glucosamine and chondroitin sulfate, which are two of the constituents of normal cartilage that decline with arthritis, have been shown to be useful when given as an oral supplement, especially in patients with early OA.

In contrast to RA, where diet's main role is in the treatment and little is known about prevention, there is more known about dietary components that lead to OA than about nutritional management of OA. It is clear that OA of the lower extremities is largely a problem brought on by obesity, especially OA of the knee (and hip, to a much lesser extent), suggesting that obesity seems to be a mechanical rather than systemic risk factor. Thus, maintaining body weight within the recommended ranges is probably the most important nutritional intervention to prevent OA. Weight loss leads to reduction in joint stress, and often reduces symptoms. In fact, recent studies have suggested that if all overweight and obese individuals reduced their body weight by 5 kg, or until their body mass index (BMI) was within the desirable range, 24% of surgeries for knee OA could be avoided. Furthermore, studies have demonstrated that exercise can improve OA symptoms even independently of weight loss, presumably by increasing muscle strength and thus improving the shock-absorbing power of the muscles, hence sparing the cartilage and joint. However, patients with OA have a great deal of difficulty with exercise, and their sedentary life style is reinforced by their joint pain, generally leading to weight gain after the onset of OA, which in turn exacerbates the disease, creating a vicious cycle. Exercise programs that increase physical activity and strengthen the muscles surrounding afflicted joints clearly improve symptoms in OA. Thus, OA can be thought of as a disease of overnutrition, while RA is generally a disease of undernutrition. Interestingly, recent twin studies have examined the role of genetic versus environmental factors as mediators of the obesity–OA relationship, and have suggested that shared genetic factors are not as important as environmental factors in mediating the obesity–OA relationship. Dietary modification leading to weight loss is a critical component of the management of OA.

See also: **Cytokines**. **Fatty Acids**: Omega-3 Polyunsaturated. **Obesity**: Complications. **Starvation and Fasting**. **Supplementation**: Dietary Supplements; Role of Micronutrient Supplementation. **Vegetarian Diets**.

Further Reading

Adam O, Beringer C, Kless T *et al.* (2003) Anti-inflammatory effects of a low arachidonic acid diet and fish oil in patients with rheumatoid arthritis. *Rheumatology International* **1**: 27–36.

Anderson RJ (2001) Rheumatoid arthritis: clinical and laboratory features. In: Klippel JH (ed.) *Primer on the Rheumatic Diseases*, 12th edn, pp. 218–225. Atlanta, GA: Arthritis Foundation.

Baker K, Nelson M, Felson DT *et al.* (2001) The efficacy of home-based progressive strength training in older adults with knee osteoarthritis: a randomized controlled trial. *Journal of Rheumatology* **28**: 1655–1665.

Belch JJF and Hill A (2000) Evening primrose oil and borage oil in rheumatologic conditions. *American Journal of Clinical Nutrition* **71**: 352S–356S.

Coggon D, Reading I, Croft P *et al.* (2001) Knee osteoarthritis and obesity. *International Journal of Obesity and Related Metabolic Disorders* **25**: 622–627.

Goronzy JJ and Weyand CM (2001) Rheumatoid arthritis: epidemiology, pathology, and pathogenesis. In: Klippel JH (ed.) *Primer on the Rheumatic Diseases*, 12th edn, pp. 209–217. Atlanta, GA: Arthritis Foundation.

Hafstrom I, Ringertz B, Spangberg A *et al.* (2001) A vegan diet free of gluten improves the signs and symptoms of rheumatoid arthritis: the effects on arthritis correlate with a reduction in antibodies to food antigens. *Rheumatology* **40**: 1175–1179.

McCabe BJ, Frankel EH, and Wolfe JJ (eds.) (2003) *Handbook of Food-Drug Interactions*. Boca Raton, FL: CRC Press.

Manek NJ, Hart D, Spector TD, and MacGregor AJ (2003) The association of body mass index and osteoarthritis of the knee joint: an examination of genetic and environmental influences. *Arthritis and Rheumatism* **48**: 1024–1029.

Miller GD, Rejeski WJ, Williamson JD *et al.* and ADAPT Investigators (2003) The Arthritis Diet and Activity Promotion Trial (ADAPT): design, rationale, and baseline results. *Controlled Clinical Trials* **24**: 462–480.

Nelson M, Baker K, Roubenoff R, and Lindner L (2002) *Strong Women and Men Beat Arthritis*. New York: Putnam.

Panush RS, Carter RL, Katz P, Kowsari B, Longley S, and Finnie S (1983) Diet therapy for rheumatoid arthritis. *Arthritis and Rheumatism* **26**: 462–471.

Rall LC, Rosen CJ, Dolnikowski *et al.* (1996) Protein metabolism in rheumatoid arthritis and aging: effects of muscle strength training and tumor necrosis factor α. *Arthritis and Rheumatism* **39**: 1115–1124.

Reginster JY, Deroisy R, Rovati LC *et al.* (2001) Long-term effects of glucosamine sulphate on osteoarthritis progression: a randomized, placebo-controlled clinical trial. *Lancet* **357**: 251–256.

Rennie KL, Hughes J, Lang R, and Jebb SA (2003) Nutritional management of rheumatoid arthritis: a review of the evidence. *Journal of Human Nutrition and Dietetics* **16**: 97–109.

Roubenoff R, Roubenoff RA, Selhub J *et al.* (1995) Abnormal vitamin B6 status in rheumatoid cachexia. Association with spontaneous tumor necrosis factor alpha production and markers of inflammation. *Arthritis and Rheumatism* **38**: 105–109.

Roubenoff R, Walsmith J, Lundgren N *et al.* (2002) Low physical activity reduces total energy expenditure in women with rheumatoid arthritis: implications for dietary intake recommendations. *American Journal of Clinical Nutrition* **76**: 774–779.

Sarzi-Puttini P, Comi D, Boccassini L *et al.* (2000) Diet therapy for rheumatoid arthritis: a controlled double-blind study of two different dietary regimens. *Scandanavian Journal of Rheumatology* **29**: 302–307.

Skoldstam L, Hagfors L, and Johansson G (2003) An experimental study of a Mediterranean diet intervention for patients with rheumatoid arthritis. *Annals of Rheumatic Disease* **62**: 208–214.

Walsmith J and Roubenoff R (2002) Cachexia in rheumatoid arthritis. *International Journal of Cardiology* **85**: 89–99.

Relevant Website

http://www.euro.who.int/nutrition – World Health Organization website: CINDI (Countrywide Integrated Noncommunicable Disease Intervention) Dietary Guide. Provides a guide for healthy eating and healthy lifestyles, as suggested by the World Health Organization.

ASCORBIC ACID

Contents

Physiology, Dietary Sources and Requirements

D A Bender, University College London, London, UK

Ascorbic acid is a vitamin (vitamin C) for only a limited number of species: man and the other primates, bats, the guinea pig, and a number of birds and fishes.

In other species ascorbic acid is not a vitamin, but is an intermediate in glucuronic acid catabolism, and its rate of synthesis bears no relation to physiological requirements for ascorbate. Species for which ascorbate is a vitamin lack the enzyme gulonolactone oxidase (EC 1.11.3.8) and have an alternative pathway for glucuronic acid metabolism.

Ascorbic acid functions as a relatively nonspecific, radical-trapping antioxidant and also reduces the tocopheroxyl radical formed by oxidation of vitamin E. It has a specific metabolic function as the redox coenzyme for dopamine β-hydroxylase and peptidyl glycine hydroxylase, and it is required to maintain the iron of 2-oxoglutarate-dependent hydroxylases in the reduced state.

Absorption, Transport, and Storage

In species for which ascorbate is not a vitamin, intestinal absorption is passive, while in human

beings and guinea pigs there is sodium-dependent active transport of the vitamin at the brush border membrane, with a sodium-independent mechanism at the basolateral membrane. Dehydroascorbate is absorbed passively in the intestinal mucosa and is reduced to ascorbate before transport across the basolateral membrane.

At intakes up to about 100 mg per day, 80–95% of dietary ascorbate is absorbed, falling from 50% of a 1 g dose to 25% of a 6 g and 16% of a 12 g dose. Unabsorbed ascorbate is a substrate for intestinal bacterial metabolism.

Ascorbate and dehydroascorbate circulate in the bloodstream both in free solution and bound to albumin. About 5% of plasma vitamin C is normally in the form of dehydroascorbate. Ascorbate enters cells by sodium-dependent active transport; dehydroascorbate is transported by the insulin-dependent glucose transporter and is accumulated intracellularly by reduction to ascorbate. In poorly controlled diabetes mellitus, tissue uptake of dehydroascorbate is impaired because of competition by glucose, and there may be functional deficiency of vitamin C despite an apparently adequate intake.

About 70% of blood-borne ascorbate is in plasma and erythrocytes (which do not concentrate the vitamin from plasma). The remainder is in white cells, which have a marked ability to concentrate ascorbate; mononuclear leukocytes achieve 80-fold concentration, platelets 40-fold, and granulocytes 25-fold, compared with the plasma concentration.

There is no specific storage organ for ascorbate; apart from leukocytes (which account for 10% of total blood ascorbate), the only tissues showing a significant concentration of the vitamin are the adrenal and pituitary glands. Although the concentration of ascorbate in muscle is relatively low, skeletal muscle contains much of the body pool of 5–8.5 mmol (900–1500 mg) of ascorbate.

Metabolism and Excretion

As shown in **Figure 1**, oxidation of ascorbic acid proceeds by a one-electron process, forming monodehydroascorbate, which disproportionates to ascorbate and dehydroascorbate. Most tissues also contain monodehydroascorbate reductase (EC 1.6.5.4), a flavoprotein that reduces the radical back to ascorbate. Dehydroascorbate is reduced to ascorbate by dehydroascorbate reductase (EC 1.8.5.1), a glutathione-dependent enzyme; little is oxidized to diketogulonic acid in human beings.

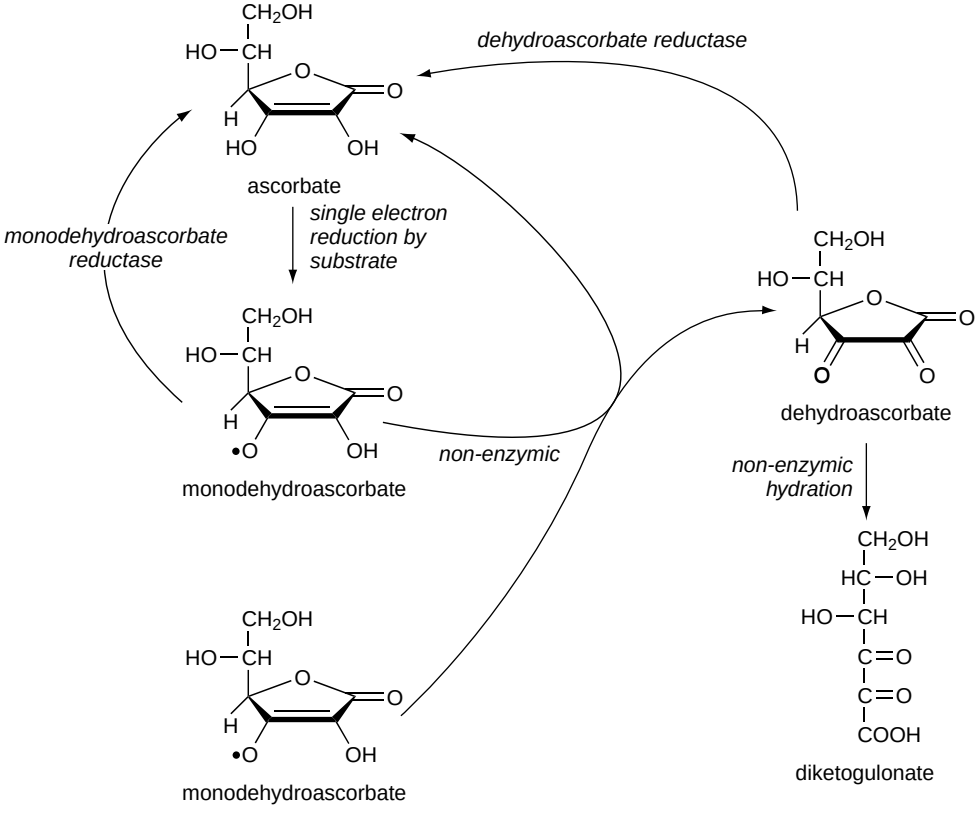


Figure 1 The metabolism of ascorbate. Monodehydroascorbate reductase, EC 1.6.5.4; dehydroascorbate reductase, EC 1.8.5.1.

Both ascorbate and dehydroascorbate are filtered at the glomerulus, then reabsorbed by facilitated diffusion. When glomerular filtration exceeds the capacity of the transport systems, at a plasma concentration of ascorbate above about 85 μmol/l, the vitamin is excreted in the urine in amounts proportional to intake.

It has been reported that approximately 25% of the dietary intake of ascorbate is excreted as oxalate; this would account for about 40% of the total urinary excretion of oxalate. However, there is no known metabolic pathway for the synthesis of oxalate from ascorbate, and it is likely that all or most of the oxalate found in urine after loading doses of ascorbate is formed nonenzymically, after the urine has been collected. Even in people at risk of forming oxalate renal stones it is unlikely that normal or high intakes of ascorbate pose any additional hazard.

Metabolic Functions of Ascorbic Acid

Ascorbic acid has specific and well-defined roles in two classes of enzymes: copper-containing hydroxylases and the 2-oxoglutarate-linked, iron-containing hydroxylases. It also increases the activity of a number of other enzymes *in vitro*—a nonspecific reducing action rather than reflecting a metabolic function of the vitamin. In addition, ascorbic acid has a number of less specific effects due to its action as a reducing agent and oxygen radical quencher. There is also evidence that ascorbate has a role in regulating the expression of connective tissue protein (and some other) genes; its mechanism of action is unknown.

Copper-Containing Hydroxylases

Dopamine β-hydroxylase (EC 1.14.17.1) is a copper-containing enzyme involved in the synthesis of the catecholamines noradrenaline and adrenaline from tyrosine in the adrenal medulla and central nervous system. The active enzyme contains Cu^{+}, which is oxidized to Cu^{2+} during the hydroxylation of the substrate; reduction back to Cu^{+} specifically requires ascorbate, which is oxidized to monodehydroascorbate.

A number of peptide hormones have a terminal amide, and amidation is essential for biological activity. The amide group is derived from a glycine residue in the precursor peptide, by proteolysis to leave a carboxy terminal glycine. This is hydroxylated on the α-carbon; the hydroxyglycine decomposes nonenzymically to yield the amidated peptide and glyoxylate. This reaction is catalyzed by peptidyl glycine hydroxylase (peptidyl α-amidase, EC 1.14.17.3); like dopamine β-hydroxylase, it is a copper-containing enzyme, and it requires ascorbate as the electron donor.

2-Oxoglutarate-Linked, Iron-Containing Hydroxylases

A number of iron-containing hydroxylases (**Table 1**) share a common reaction mechanism, in which hydroxylation of the substrate is linked to decarboxylation of 2-oxoglutarate. Ascorbate is required for the activity of all of these enzymes, but it does not function as either a stoichiometric substrate or a conventional coenzyme (which would not be consumed in the reaction).

Proline and lysine hydroxylases are required for the postsynthetic modification of collagen, and proline hydroxylase also for the postsynthetic modification of osteocalcin in bone and the Cl_q component of complement. Aspartate β-hydroxylase is required for the postsynthetic modification of protein C, the vitamin K-dependent protease which hydrolyzes activated factor V in the blood-clotting cascade. Trimethyllysine and γ-butyrobetaine hydroxylases are required for the synthesis of carnitine.

The best studied of this class of enzymes is procollagen proline hydroxylase; it is assumed that the others follow essentially the same mechanism. As shown in **Figure 2**, the first step is binding of oxygen to the enzyme-bound iron, followed by attack on the 2-oxoglutarate substrate, resulting in decarboxylation to succinate, leaving a ferryl radical at the active site of the enzyme. This catalyzes the hydroxylation of proline, restoring the free iron to undergo further reaction with oxygen.

It has long been known that ascorbate is oxidized during the reaction, but not stoichiometrically with hydroxylation of proline and decarboxylation of 2-oxoglutarate. The purified enzyme is active in the absence of ascorbate, but after about 5–10 s (about 15–30 cycles of enzyme action) the rate of reaction falls. The loss of activity is due a side reaction of the highly reactive ferryl radical in which the iron is oxidized to Fe^{3+}, which is catalytically inactive—so-called uncoupled decarboxylation of

Table 1 Vitamin C-dependent, 2-oxoglutarate-linked hydroxylases

Aspartate β-hydroxylase	EC 1.14.11.16
γ-Butyrobetaine hydroxylase	EC 1.14.11.1
p-Hydroxyphenylpyruvate hydroxylase	EC 1.14.11.27
Procollagen lysine hydroxylase	EC 1.14.11.4
Procollagen proline 3-hydroxylase	EC 1.14.11.7
Procollagen proline 4-hydroxylase	EC 1.14.11.2
Pyrimidine deoxynucleotide dioxygenase	EC 1.14.11.3
Thymidine dioxygenase	EC 1.14.11.10
Thymine dioxygenase	EC 1.14.11.6
Trimethyllysine hydroxylase	EC 1.14.11.8

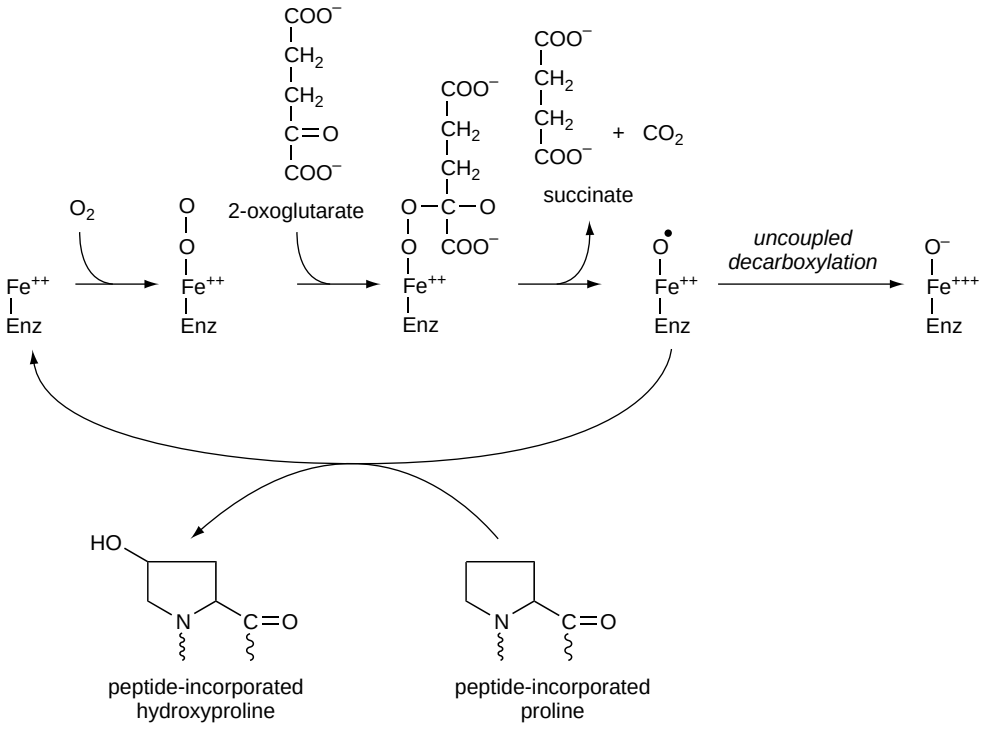


Figure 2 The reaction of procollagen proline hydroxylase.

2-oxoglutarate. Activity is only restored by ascorbate, which reduces the iron back to Fe^{2+}.

The Role of Ascorbate in Iron Absorption

Inorganic dietary iron is absorbed as Fe^{2+} and not as Fe^{3+}; ascorbic acid in the intestinal lumen not only maintains iron in the reduced state but also chelates it, increasing absorption considerably. A dose of 25 mg of vitamin C taken together with a meal increases the absorption of iron approximately 65%, while a 1 g dose gives a 9-fold increase. This is an effect of ascorbic acid present together with the test meal; neither intravenous administration of vitamin C nor supplements several hours before the test meal affects iron absorption, although the ascorbate secreted in gastric juice should be effective. This is not a specific effect of ascorbate; a variety of other reducing agents including alcohol and fructose also enhance the absorption of inorganic iron.

Inhibition of Nitrosamine Formation

Oral bacteria can reduce nitrate to nitrite which, under the acidic conditions of the stomach, can react with amines in foods to form carcinogenic *N*-nitrosamines. In addition to dietary sources, a significant amount of nitrate is formed endogenously by the metabolism of nitric oxide—1 mg/kg body weight/day (about the same as the average dietary intake), increasing 20-fold in response to inflammation and immune stimulation, and nitrate is secreted in saliva.

Ascorbate reacts with nitrite forming NO, NO_2, and N_2, so preventing the formation of nitrosamines. In addition to ascorbate in foods, there is considerable secretion of ascorbate in the gastric juice, and inhibition of gastric secretion for treatment of gastric ulcers, as well as reducing vitamin B_{12} absorption, also inhibits this presumably protective gastric secretion of ascorbate.

However, while ascorbate can deplete nitrosating compounds under anaerobic conditions, the situation may be reversed in the presence of oxygen. Nitric oxide reacts with oxygen to form N_2O_3 and N_2O_4, both of which are nitrosating reagents, and can also react with ascorbate to form NO and monodehydroascorbate. It is thus possible for ascorbate to be depleted, with no significant effect on the total concentration of nitrosating species. It remains to be determined whether or not ascorbate has any significant effect in reducing the risk of nitrosamine formation and carcinogenesis.

Antioxidant and Prooxidant Actions of Ascorbate

Chemically, ascorbate is a potent reducing agent, both reducing hydrogen peroxide and also acting as a radical trapping antioxidant, reacting with superoxide and a proton to yield hydrogen peroxide

or with the hydroxy radical to yield water. In each case the product is monodehydroascorbate, which, as shown in **Figure 1**, undergoes dismutation to ascorbate and dehydroascorbate. In studies of ascorbate depletion in men there is a significant increase in abnormalities of sperm DNA, suggesting that vitamin C may have a general, nonspecific radical-trapping antioxidant function.

Ascorbate also acts to reduce the tocopheroxyl radical formed by oxidation of vitamin E in cell membranes and plasma lipoproteins. It thus has a vitamin E sparing antioxidant action, coupling lipophilic and hydrophilic antioxidant reactions.

The antioxidant efficiency of ascorbate is variable. From the chemistry involved, it would be expected that overall 2 mol of tocopheroxyl radical would be reduced per mole of ascorbate because of the reaction of 2 mol of monodehydroascorbate to yield ascorbate and dehydroascorbate. However, as the concentration of ascorbate increases, so the molar ratio decreases, and it is only at very low concentrations of ascorbate that it tends toward the theoretical ratio. This is because, as well as its antioxidant role, ascorbate can be a source of hydroxyl and superoxide radicals.

At high concentrations, ascorbate can reduce molecular oxygen to superoxide, being oxidized to monodehydroascorbate. Both Fe^{3+} and Cu^{2+} ions are reduced by ascorbate, again yielding monodehydroascorbate; the resultant Fe^{2+} and Cu^{+} are reoxidized by reaction with hydrogen peroxide to yield hydroxide ions and hydroxyl radicals. Thus, as well as its antioxidant role, ascorbate has prooxidant action; the net result will depend on the relative rates of formation of superoxide and hydroxyl radicals by autooxidation and metal-catalyzed reactions of ascorbate, and the trapping of these radicals by ascorbate.

It seems likely that the prooxidant actions of ascorbate are of relatively little importance *in vivo*. Except in cases of iron overload there are almost no transition metal ions in free solution, they are all bound to proteins, and because the renal transport system is readily saturated, plasma and tissue concentrations of ascorbate are unlikely to rise to a sufficient extent to lead to significant radical formation.

Assessment of Vitamin C Status

The early method of assessing vitamin C nutritional status was by testing the extent of saturation of the body's reserves by giving a test dose of 500 mg (2.8 mmol) and measuring the amount excreted in the urine. In a subject with high status, more or less all of the test dose is recovered over a period of 5 or 6 h.

More sensitive assessment of status is achieved by measuring the concentration of the vitamin in whole blood, plasma, or leukocytes. Criteria of adequacy are shown in **Table 2**. The determination of ascorbate in whole blood is complicated by nonenzymic oxidation of the vitamin by hemoglobin, and most studies rely on plasma or leukocyte concentrations of ascorbate.

A problem arises in the interpretation of leukocyte ascorbate concentrations because of the different capacity of different classes of leukocytes to accumulate the vitamin. Granulocytes are saturated at a concentration of about 530 pmol/10^6 cells, while mononuclear leukocytes can accumulate 2.5 times more ascorbate. A considerable mythology has developed to the effect that vitamin C requirements are increased in response to infection, inflammation, and trauma, based on reduced leukocyte concentrations of ascorbate in these conditions. However, the fall in leukocyte ascorbate can be accounted for by an increase in the proportion of granulocytes in response to trauma and infection (and hence a fall in the proportion of mononuclear leukocytes). Total leukocyte ascorbate is not a useful index of vitamin C status without a differential white cell count.

There is increased formation of 8-hydroxyguanine (a marker of oxidative radical damage) in DNA during (short-term) vitamin C depletion, and the rate of removal of 8-hydroxyguanine from DNA by excision repair, and hence its urinary excretion, is affected by vitamin C status. This suggests that measurement of urinary excretion of

Table 2 Plasma and leukocyte ascorbate concentrations as criteria of vitamin C nutritional status

		Deficient	*Marginal*	*Adequate*
Whole blood	mmol/l	<17	17–28	>28
	mg/l	<3.0	3.0–5.0	>5.0
Plasma	mmol/l	<11	11–17	>17
	mg/l	<2.0	2.0–3.0	>3.0
Leukocytes	pmol/10^6 cells	<1.1	1.1–2.8	>2.8
	μg/10^6 cells	<0.2	0.2–0.5	>0.5

8-hydroxyguanine may provide a biomarker of optimum status, as a basis for estimating requirements.

Requirements

While the minimum requirement for ascorbate is firmly established, there are considerable differences between the reference intakes published by different national and international authorities. Depending on the chosen criteria of adequacy, and assumptions made in interpreting experimental results, it is possible to produce arguments in support of reference intakes ranging from 30 to 100 mg/day. Studies of intakes associated with reduced risks of cancer and cardiovascular disease suggest an average requirement of 90–100 mg/day and a reference intake of 120 mg/day.

Minimum Requirement

The minimum requirement for vitamin C was established in the 1940s in a depletion/repletion study, which showed that an intake of less than 10 mg per day was adequate to prevent the development of scurvy, or to cure the clinical signs. At this level of intake, wound healing is impaired, and optimum wound healing requires a mean intake of 20 mg per day. Allowing for individual variation, this gives reference intake of 30 mg/day, which was the UK figure until 1991 and the WHO/FAO figure until 2001.

Requirements Estimated from the Plasma and Leukocyte Concentrations of Ascorbate

The plasma concentration of ascorbate shows a sigmoidal relationship with intake. Below about 30 mg/day it is extremely low and does not reflect increasing intake to any significant extent. As the intake rises above 30 mg/day, so the plasma concentration begins to increase sharply, reaching a plateau of 70–85 μmol/l, at intakes between 70 and 100 mg/day, when the renal threshold is reached and the vitamin is excreted quantitatively with increasing intake.

The point at which the plasma concentration increases more or less linearly with increasing intake represents a state where reserves are adequate and ascorbate is available for transfer between tissues. This corresponds to an intake of 40 mg/day and is the basis of the UK, EU, and FAO figures. At this level of intake the total body pool is about 5.1 mmol (900 mg). It has been argued that setting requirements and reference intakes on the basis of the steep part of a sigmoidal curve is undesirable, and a more appropriate point would be the intake at which the plasma concentration reaches a plateau, at an intake of around 100–200 mg/day.

The US/Canadian reference intakes of 75 mg for women and 90 mg for men are based on studies of leukocyte saturation during depletion/repletion studies.

Requirements Estimated from Maintenance of the Body Pool of Ascorbate

An alternative approach to estimating requirements is to determine the fractional rate of catabolism of total body ascorbate; an appropriate intake would then be that required to replace losses and maintain the body pool.

Clinical signs of scurvy are seen when the total body pool of ascorbate is below 1.7 mmol (300 mg). The pool increases with intake, reaching a maximum of about 8.5 mmol (1500 mg) in adults—114 μmol (20 mg)/kg body weight. The basis for the 1989 US RDA of 60 mg was the observed mean fractional turnover rate of 3.2% of a body pool of 20 mg/kg body weight/day, with allowances for incomplete absorption of dietary ascorbate and individual variation.

It has been argued that a total body pool of 5.1 mmol (900 mg) is adequate; it is threefold higher than the minimum required to prevent scurvy, and there is no evidence that there are any health benefits from a body pool greater than 600 mg. The observed body pool of 8.5 mmol in depletion/repletion studies was found in subjects previously consuming a self-selected diet, with a relatively high intake of vitamin C, and therefore might not represent any index of requirement. Assuming a total body pool of 5.1 mmol and catabolism of 2.7%/day, allowing for efficiency of absorption and individual variation gives a reference intake of 40 mg/day.

Because the fractional turnover rate was determined during a depletion study, and the rate of ascorbate catabolism varies with intake, it has been suggested that this implies a rate of 3.6%/day before depletion. On this basis, and allowing for incomplete absorption and individual variation, various national authorities arrive at a recommended intake of 80 mg.

The rate of ascorbate catabolism is affected by intake, and the requirement to maintain the body pool cannot be estimated as an absolute value. A habitual low intake, with a consequent low rate of catabolism, will maintain the same body pool as a habitual higher intake with a higher rate of catabolism.

Dietary Sources and High Intakes

It is apparent from the list of rich sources of vitamin C in **Table 3** that the major determinant of vitamin C intake is the consumption of fruit and

Table 3 Rich sources of vitamin C

	Portion (g)	*mg/portion*
Black currants	80	160
Oranges	250	125
Orange juice	200	100
Strawberries	100	60
Grapefruit	140	56
Melon	200	50
Green peppers	45	45
Sweet potato	150	38
Loganberries	85	34
Spinach	130	33
Red currants	80	32
White currants	80	32
Pineapple	125	31
Brussels sprouts	75	30
Mangoes	100	30
Satsumas	100	30
Tangerines	100	30
Turnips	120	30
Gooseberries	70	28
Potato chips	265	27
Broccoli	75	26
Swedes	120	24
Spring greens	75	23
Artichokes, globe	220	22
Potatoes	140	21
Avocados	130	20
Leeks	125	20
Lemons	25	20
Okra	80	20
Peas	75	20
Raspberries	80	20
Tomato juice	100	20
Plantain, green	85	17
Bilberries	80	16
Blackberries	80	16
Kidney	150	15
Tomatoes	75	15
Bananas	135	14
Cauliflower	65	13
Beans, broad	75	11
Cabbage	75	11
Nectarines	110	11
Parsnips	110	11
Rhubarb	100	10

vegetables; deficiency is likely in people whose habitual intake of fruit and vegetables is very low. However, clinical signs of deficiency are rarely seen in developed countries. The range of intakes by healthy adults in Britain reflects fruit and vegetable consumption: the 2.5 percentile intake is 19 mg per day (men) and 14 mg per day (women), while the 97.5 percentile intake from foods (excluding supplements) is 170 mg per day (men) and 160 mg per day (women). Smokers may be at increased risk of deficiency; there is some evidence that the rate of ascorbate catabolism is 2-fold higher in smokers than in nonsmokers.

There is a school of thought that human requirements for vitamin C are considerably higher than those discussed above. The evidence is largely based on observation of the vitamin C intake of gorillas in captivity, assuming that this is the same as their intake in the wild (where they eat considerably less fruit than under zoo conditions), and then assuming that because they have this intake, it is their requirement—an unjustified assumption. Scaling this to human beings suggests a requirement of 1–2 g per day.

Intakes in excess of about 80–100 mg per day lead to a quantitative increase in urinary excretion of unmetabolized ascorbate, suggesting saturation of tissue reserves. It is difficult to justify a requirement in excess of tissue storage capacity.

A number of studies have reported low ascorbate status in patients with advanced cancer—perhaps an unsurprising finding in seriously ill patients. One study has suggested, on the basis of an uncontrolled open trial, that 10 g daily doses of vitamin C resulted in increased survival. Controlled studies have not demonstrated any beneficial effects of high-dose ascorbic acid in the treatment of advanced cancer.

High doses of ascorbate are popularly recommended for the prevention and treatment of the common cold. The evidence from controlled trials is unconvincing, and meta-analysis shows no evidence of a protective effect against the incidence of colds. There is, however, consistent evidence of a beneficial effect in reducing the severity and duration of symptoms. This may be due to the antioxidant actions of ascorbate against the oxidizing agents produced by, and released from, activated phagocytes, and hence a decreased inflammatory response.

Scorbutic guinea pigs develop hypercholesterolemia. While there is no evidence that high intakes of vitamin C result in increased cholesterol catabolism, there is evidence that monodehydroascorbate inhibits hydroxymethylglutaryl CoA reductase, resulting in reduced synthesis of cholesterol, and high intakes of ascorbate may have some hypocholesterolaemic action. There is limited evidence of benefits of high intakes of vitamin C in reducing the incidence of stroke, but inconsistent evidence with respect to coronary heart disease.

Regardless of whether or not high intakes of ascorbate have any beneficial effects, large numbers of people habitually take between 1 and 5 g per day of vitamin C supplements. There is little evidence of any significant toxicity from these high intakes. Once the plasma concentration of ascorbate reaches the renal threshold, it is excreted more or less quantitatively with increasing intake.

Because the rate of ascorbate catabolism increases with increasing intake, it has been suggested that abrupt cessation of high intakes of ascorbate may result in rebound scurvy because of 'metabolic conditioning' and a greatly increased rate of catabolism. While there have been a number of anecdotal reports, there is no evidence that this occurs.

See also: **Antioxidants**: Diet and Antioxidant Defense. **Ascorbic Acid**: Deficiency States. **Cholesterol**: Factors Determining Blood Levels. **Diabetes Mellitus**: Dietary Management. **Fruits and Vegetables**. **Iron**. **Vitamin E**: Physiology and Health Effects.

Further Reading

Basu TK and Schorah CI (1981) *Vitamin C in Health and Disease*. p. 160. London: Croom Helm.

Bender DA (2003) Ascorbic acid (vitamin C). In: *Nutritional Biochemistry of the Vitamins*, 2nd edn., pp. 357–384. New York: Cambridge University Press.

Benzie IF (1999) Vitamin C: Prospective functional markers for defining optimal nutritional status. *Proceedings of the Nutrition Society* **58**: 469–476.

Chatterjee IB (1978) Ascorbic acid metabolism. *World Review of Nutrition and Dietetics* **30**: 69–87.

Englard S and Seifter S (1986) The biochemical functions of ascorbic acid. *Annual Review of Nutrition* **6**: 365–406.

Ginter E (1978) Marginal vitamin C deficiency, lipid metabolism and atherogenesis. *Advances in Lipid Research* **16**: 167–220.

Rivers JM (1987) Safety of high-level vitamin C ingestion. *Annals of the New York Academy of Sciences* **498**: 445–451.

Sato P and Udenfriend S (1978) Studies on vitamin C related to the genetic basis of scurvy. *Vitamins and Hormones* **36**: 33–52.

Sauberlich HE (1994) Pharmacology of vitamin C. *Annual Review of Nutrition* **14**: 371–391.

Smirnoff N (2000) Ascorbic acid: Metabolism and functions of a multi-facetted molecule. *Current Opinion in Plant Biology* **3**: 229–235.

Deficiency States

C J Bates, MRC Human Nutrition Research, Cambridge, UK

Scurvy: The History, and Discovery of Vitamin C

Scurvy is traditionally associated with long sea voyages (**Table 1**), which typically lasted for several years; the seamen's diets were confined to whatever could be stored at room temperature for long periods. In the absence of refrigeration, their diets typically consisted of dried biscuits and other dry cereal foods (wheat flour and oatmeal), salted meat, dried peas, cheese, butter, and ale, i.e., whatever could be dried and preserved, often for long periods in adverse tropical climates. The signs and symptoms that were commonly described in classical accounts of scurvy, written long before its cause was understood, included lassitude, swollen joints, putrid and bleeding gums, failure of wound healing and the opening of old wounds and sores, intradermal bleeding due to capillary fragility, heart failure, and sudden death (**Table 2**). Although nowadays we carefully distinguish the symptoms of true scurvy (now known to be produced specifically by vitamin C deficiency) from conditions such as beriberi (thiamine deficiency, which is associated with oedema of the lower limbs), vitamin A deficiency (associated with night-blindness and corneal lesions), and rickets (caused mainly by a lack of exposure to sunlight in children), in the older literature these conditions were often not recognized as distinct. Signs and symptoms of scurvy occurred on land in times of siege or during prolonged military campaigns where dietary variety and access to fresh foods were severely restricted. While some medical practitioners and leaders became convinced that the cause of scurvy was dietary and that cure and prevention were possible by including fresh plant food such as scurvy grass, decoctions of evergreen needles etc. in the diet (**Table 1**), many remained convinced, right up to the beginning of the twentieth century, that other factors such as 'foul vapours' or infections were to blame. Indeed, before the recognition and discovery of essential micronutrients and experimental animal studies in the early years of the twentieth century, which confirmed that small amounts of certain complex organic molecules are needed in the diet to maintain health, the idea that food needed to supply anything other than energy, protein, certain minerals, and water was not generally accepted.

Although ad hoc treatments for scurvy had been successfully applied in many situations before the eighteenth century, the definitive proof of a dietary cure is attributed to James Lind, whose controlled trial of several different treatments of the disease on board HMS Salisbury, followed by his *Treatise of the Scurvy* published in 1753, provided the decisive evidence that persuaded the British Admiralty to insist on the inclusion of citrus fruit regularly in the naval diet (**Table 1**). Lind showed that a 'rob' or decoction of oranges could rapidly cure the disease, and the application of his discovery rapidly brought about dramatic reductions in the incidence of and mortality due to scurvy. However, the labile component in fruit that was responsible for protection against scurvy was not isolated until 1928. Its

Table 1 Scurvy and vitamin C; selected historical milestones, 1500–1950

1536	Jacques Cartier used a leaf extract of evergreen-tree needles to cure scurvy in explorers in Newfoundland
1593	Sir Richard Hawkins successfully used citrus fruit (amongst other cures and precautions) as a cure for scurvy in sailors
1753	James Lind's *Treatise of the Scurvy* described the first controlled nutrition experiment, in which oranges and lemons, but not other treatments, cured scurvy in sailors
Late eighteenth century	Successful public-health practices were introduced to reduce scurvy in the British navy based on citrus-fruit rations (Sir Gilbert Blane); Captain James Cook's long sea voyages used, and benefited from, this new knowledge
Late nineteenth century	Outbreaks of 'Barlow's disease' (infantile scurvy) in young children receiving mainly condensed cow's milk
Early twentieth century	Guinea-pig model of human scurvy developed by Holst and Frohlich in Norway; this enabled antiscorbutic foods and extracts to be studied in the laboratory
1928	Isolation of crystalline 'hexuronic acid' by Albert Szent-Gyorgyi in Cambridge, followed by its recognition as 'the antiscorbutic vitamin', alias vitamin C, by Charles King in Pittsburgh
1933	Norman Haworth (UK) and Tadeus Reichstein (Switzerland) separately synthesized vitamin C (L-ascorbic acid) and resolved its structure
Mid twentieth century	Human studies showed 10 mg vitamin C per day to be sufficient to prevent or cure scurvy; animal studies showed that vitamin C is essential for normal collagen and hence connective-tissue formation, thereby resolving the biochemical basis for many of the clinical signs of scurvy

chemical structure was proved by *de novo* synthesis from common sugars a few years later (**Table 1**). Paradoxically, crystalline ascorbic acid (vitamin C) was first isolated, not from a plant source such as fruit or green leaves, but from an animal source, namely adrenal glands, where high concentrations are also found. Indeed, the original motivation for the isolation of the crystalline material by Albert Szent-Gyorgyi in Hopkins' laboratory in Cambridge arose from his attempt to isolate a new adrenal hormone. The 'hexuronic acid' that he crystallized was not immediately equated with the antiscorbutic vitamin. Fortunately, studies by Charles Glen King in Pittsburgh led to the recognition of this unstable easily oxidized sugar derivative as the long-sought antiscorbutic principle, vitamin C or L-ascorbic acid.

Table 2 The signs and symptoms of scurvy, and evidence of inadequate tissue levels of vitamin C

Clinical signs and symptoms of scurvy in adults
Petechiae (small hemorrhagic spots), perifollicular hemorrhages, and larger sheet hemorrhages, especially of the skin of the limbs and trunk
Positive 'Hess test' (pressure or suction test) for increased capillary fragility
Hyperkeratosis of hair follicles; hairs abnormally coiled (ecchymoses)
Swollen and bleeding gums
Swollen painful joints, with effusions and arthralgia
Failure of wound healing and breakdown (reopening) of old wounds
Oedema, dyspnoea, dry eyes and mouth
Lassitude and impaired mental state
Sudden heart failure, often leading to sudden death, in severe cases
Clinical signs and symptoms in young infants
Painful joints leading to frog-like lying appearance; bleeding into the joints
Swelling ('beading') of rib cage and swelling of long-bone joints
Changes in x-ray appearance of the epiphyses of the long bones: 'ground glass' appearance, differing from that of rickets (with which infantile scurvy was often confused)
Intracranial hemorrhages, 'bulging eyes', and anemia in some cases; gums affected only if teeth already erupted
Biochemical evidence of inadequate vitamin C status
Serum or plasma vitamin C concentration $<0.2\,mg\,dl^{-1}$ ($<11\,\mu mol\,l^{-1}$)
Buffy-coat vitamin C concentration $<15\,\mu g$ (<85 nmol) per 10^8 white cells
Untreated subjects with clinical scurvy invariably have very low biochemical vitamin C levels, but people may have very low biochemical levels without having clinical evidence of scurvy; very low intakes for periods of months to years, plus additional stresses, increase the likelihood of clinical scurvy

Degradation, Turnover, and Factors that Induce Increased Requirements for Vitamin C

The instability of vitamin C in air, and especially in neutral or alkaline aqueous solution, is attributable to the fact that in the presence of oxygen or other oxidizing agents it readily undergoes two successive one-electron oxidation steps to produce dehydroascorbate. Since the oxidation products are also unstable and undergo an irreversible lactone ring opening to diketogulonic acid, the vitamin is very easily destroyed, both in foods and (to a lesser extent because of efficient recycling mechanisms) in the body. Diketogulonic acid is one of several degradation products of vitamin C that cannot be reconverted to the vitamin and are further degraded to stable excretory products, such as oxalic acid, by oxidative metabolism. Of all the micronutrients that are essential for human health and survival, vitamin C is the most easily destroyed during drying and other traditional methods of preserving food. Citrus fruits contain other organic acids that inhibit this process of oxidation by lowering the pH of the fruit juice. This enables them, and extracts of them, to preserve at least some of their vitamin content for several weeks and even months of storage and thereby helps them to prevent and cure scurvy.

It remains largely a mystery why some people succumb to classical scurvy after a short period of virtually zero intake, whereas others survive for much longer. It has been speculated that some people may be able to produce all of the enzymes of the vitamin C synthetic pathway, including gulonolactone oxidase, which is normally absent from humans. However, this now seems unlikely, and it is more probable that the retention and recycling mechanisms for the vitamin are more efficient in some people than in others. We now know, for example, that smokers have a higher turnover of endogenous vitamin C than non-smokers, presumably because of the free-radical oxidant species in cigarette smoke. People with infections also have increased vitamin C turnover, which is associated with the liberation of pro-oxidant substances (such as hypochlorous acid) that are used by the body to kill bacteria. Some people have isoforms of certain blood proteins such as haptoglobins that are associated with relatively low levels of vitamin C in the blood. Very occasionally, there arise non-lethal mutations of vitamin C-dependent pathways whose abnormalities can be treated with high vitamin C intakes. A well-characterized example is Ehlers–Danlos syndrome, type VI, which is associated with impaired collagen lysyl hydroxylation and presents with a variety of clinical and biochemical connective-tissue (collagen-related) defects. However, much more research is needed to determine which of many possible genetic and environmental factors modulate the turnover of vitamin C in the body and to determine individual requirements and hence relative resistance to scurvy. Although 100–200 mg of the vitamin per day is needed to approach saturation of the tissues of humans, the amount needed to prevent or cure scurvy is less than 10 mg day^{-1}, as was shown by experiments involving prolonged periods of feeding with depleted diets in the middle of the twentieth century (**Table 1**). Today, overt clinical scurvy is rare. It is occasionally seen in refugee camps or in elderly people with poor diets that are devoid of the usual sources of the vitamin. The latter high-risk group contains many individuals who are unable to chew fresh fruit and vegetables because of bad dentures or poor gastric tolerance of acidic or fibrous foods (see below).

An essential dietary requirement for vitamin C (L-ascorbic acid) is shared with humans by only a small number of other vertebrates, including primates, guinea pigs and agoutis, and some birds and fishes. Most mammals synthesize the vitamin in their livers from hexose sugars; birds synthesize it in their kidneys. The final enzyme in the pathway, L-gulonolactone oxidase, has been lost in several unrelated species, suggesting a vulnerable and easily mutated locus on the genome. Presumably this mutation was neutral or advantageous during the natural selection of man's ancestors, when human and related-primate diets were rich in plant-derived sources of the vitamin.

Well-Established Metabolic Functions of Vitamin C that are Impaired in Deficiency

Studies of guinea pigs (and other species requiring a dietary source of vitamin C) have revealed that, when deprived of the vitamin, characteristic lesions of growing bones, failure of wound-healing of skin and bones, capillary defects, and other lesions arise, all of which point to a failure of the new synthesis of, or repair processes for, connective tissues and especially the protein collagen, which is the major extracellular protein and comprises a third of all the protein in the body (**Table 1**). As the biochemical pathway of collagen biosynthesis became better understood, during the middle years of the twentieth century, it became clear that certain unusual and characteristic hydroxylated amino-acids, comprising two different hydroxylated forms of proline and one

of lysine, occurred uniquely in collagen. These were not coded for by the genome or inserted by the amino-acid-assembly machinery of the cell but instead were created by 'post-translational' amino-acid hydroxylation processes that took place after the nascent pro-collagen polypeptide chain had been synthesized on the polysomal messenger RNA. Some of the prolyl residues of the pro-collagen molecule were then hydroxylated to hydroxyprolyl residues, and some of the lysyl residues were hydroxylated to hydroxylysyl residues. The hydroxylated prolyl residues are essential for subsequent collagen triple-helix formation and hence for the secretion of nascent collagen; the hydroxylated lysyl residues form part of the essential pyridinoline-type crosslinks that stabilize the collagen fibers, especially those in bone. In the absence of sufficient vitamin C, these hydroxylation reactions rapidly fail, because the iron at the active center of the 'mixed function oxidase' enzymes that catalyze them is rapidly inactivated by oxidation. Vitamin C, specifically, is needed to keep the essential ferrous residues at the hydroxylase-enzyme active centers in the reduced, active, form. In the absence of the vitamin, these enzymes are inactivated after only a few cycles of hydroxylation.

The essential function of vitamin C in collagen maturation can go a long way towards explaining many of the clinical lesions of scurvy (**Table 2**). However, recent evidence indicates that the vitamin may also act directly on the transcription and translation of collagen mRNA and on the synthesis of other parts of the cell machinery that are needed for the formation of normal connective tissues. Parts of this process have yet to be clarified.

Vitamin C also plays a cofactor-like role in the reactions of several other enzymes that split molecular oxygen, notably members of the group of enzymes that are classified as 'mixed-function oxidases.' Two enzymes containing ferrous iron that are involved in carnitine biosynthesis (trimethyl lysine hydroxylase and γ-butyrobetaine hydroxylase) fall into this category. Aspartate β-hydoxylase, which is needed for the post-synthetic modification of protein kinase C, also requires vitamin C. Another enzyme that requires vitamin C is the copper enzyme dopamine β-hydroxylase, and, in this reaction, ascorbic acid is needed to reduce cupric copper to the cuprous form at the active site. Peptidyl glycine hydroxylase (peptidyl α-amidase) is also a copper enzyme that requires vitamin C as cosubstrate. Vitamin C can increase the activities of several other enzymes, by a non-specific reducing or protective action that is shared by other cellular reductants. This action is, however, distinct from the functions described above, which are more specific to vitamin C.

In the course of its functional roles, ascorbic acid is oxidized in two successive one-electron reversible steps, and it is thought that most, if not all, of its essential biological actions are centred around this key redox cycle. The first oxidation product is the free-radical form of the vitamin, which is known variously as 'monodehydroascorbate,' 'semidehydroascorbate,' or 'ascorbate free radical' (AFR). Although this intermediate shares with most other free radicals the properties of having a relatively short half life and a high degree of chemical reactivity, it is, nevertheless, more stable than many other free radicals, contrasting with the highly reactive and damaging radicals such as hydroxyl or superoxide radical that are derived from molecular oxygen. By reacting with, and thus quenching, these damaging oxygen free radicals, ascorbate can act as a free-radical chain terminator and can thereby protect vulnerable macromolecules such as DNA, lipids, and proteins from oxidative damage by free-radical chain reactions. Such reactions would otherwise cause extensive damage, including genetic damage (to DNA), the formation of potentially atherogenic oxidized lipids, and oxidative inactivation of enzymes. For this reason, ascorbic acid is thought to possess important 'protective' antioxidant properties that are not directly connected with its other cofactor-like or cosubstrate-like roles in enzyme reactions. Ascorbate probably also protects host tissues against damage by oxidants such as hypochlorous acid that are produced in the normal course of bacterial killing by white cells.

The second one-electron oxidation step in ascorbate oxidation produces dehydroascorbate from the free-radical intermediate AFR. Both of these oxidized forms can be recycled to ascorbate either by non-enzymatic reactions with glutathione as the reductant (electron acceptor) or by pyridine nucleotide-dependent enzymatically catalyzed reactions. Thus, the two sequential one-electron oxidation steps from ascorbate to dehydroascorbate are fully reversible *in vivo*. However, the subsequent spontaneous non-enzymatic reaction comprising hydrolysis of the 1,4-lactone ring is not reversible, so that the product of this reaction, diketogulonic acid, has no provitamin activity. Normally, about 3% of the vitamin C in the body is degraded every day, and this loss must be replaced from the diet. Nevertheless, many weeks at or near zero intake are usually needed to reach scorbutic levels, if the tissues are reasonably well supplied to begin with.

Measurement of Vitamin C Status; Biochemical Tests for Adequacy and Deficiency

In species (such as humans) that cannot synthesize vitamin C in their bodies, the vitamin concentration in tissues and blood compartments (plasma, erythrocytes, and white blood cells) varies characteristically with the dietary intake of the vitamin. Since the blood-compartment concentrations mirror the concentrations in most other cells and tissue compartments, tissue vitamin C status can be monitored by measuring the concentration in plasma or blood, even though the blood concentrations are generally lower than those in most tissues. The concentration ratios between extracellular and various intracellular compartments are determined by active transport systems that concentrate the vitamin inside many cell types. At high intakes of the vitamin, the intestinal absorption process is overwhelmed, so that some of the ingested vitamin remains unabsorbed and is destroyed in the lower intestine by intestinal bacteria. The maximum steady-state level in plasma can be temporarily exceeded following a high bolus intake, but the excess vitamin is rapidly excreted in the urine once the renal threshold for filtration and reabsorption is exceeded. These safety mechanisms limit the maximum concentration of the vitamin to which the tissues are exposed.

For many years, the best biochemical measure of vitamin C status was considered to be the buffy coat, or total white-cell concentration of the vitamin (**Table 2**), expressed as micrograms per 10^8 white cells, the cell count in the assay sample usually being estimated by an electronic cell counter. This status index varied predictably with the magnitude of the total body vitamin C stores during controlled (animal) depletion studies. However, in practice it has proved to be a difficult test to use in human studies and especially in surveys, as it requires complex laboratory operations to be performed immediately after collecting the blood. It is also difficult to harmonize between laboratories, and, since it measures the average vitamin C content across several different white-cell types, whose individual proportions and relative vitamin C contents may vary considerably, its interpretation was not always straightforward. In addition, infection has rather unpredictable effects on the values obtained. For all of these reasons, this assay has fallen out of favor and is now rarely used. The concentration of the vitamin in erythrocytes or whole blood is not an ideal alternative, partly because hemoglobin can catalyze the oxidative destruction of the vitamin *in vitro* and partly because erythrocyte concentrations do not mirror other body compartments in a simple manner.

Serum or plasma vitamin C has therefore become the most commonly used status assay. In order to avoid short-term fluctuations caused by recent bolus intakes from food or supplements, it is preferable to collect an overnight-fasting blood sample. Since the vitamin is extremely easily oxidized, the sample must be carefully preserved unless the assay is to be performed immediately. The usual approach is to add freshly prepared metaphosphoric acid, usually at between 2 and 5% w/v, which precipitates plasma proteins, chelates transition-metal ions, and provides a protective acidic environment of a suitable pH. If stored, the samples must be kept at a low temperature, e.g., at $-25\,^{\circ}C$ for not more than a week or two or at $-80\,^{\circ}C$ for up to 1–2 years. There are many alternative physicochemical and chemical assay methods for measuring vitamin C in extracts of plasma or serum. These include (a) the measurement of its chemical reducing action on reducible dyes such as dichlorophenol indophenol and (b) the formation of either a colored osazone or a fluorescent derivative with orthophenylene diamine, after conversion to dehydroascorbate. Quantitation by absorbance or by electrochemical detection after separation by high-performance liquid chromatography is favored by many workers. This procedure has the advantage of being relatively specific (i.e., free from most forms of interference) and highly sensitive, but it is more time-consuming than the simpler nonchromatographic methods. Different methods may differ with respect to their specificity and their sensitivity to problems of interference as well as in the precautions that are needed to avoid oxidative destruction of the vitamin during the assay. Careful validation and robust quality-control procedures are essential.

Plasma or serum levels below $11\,\mu mol\,l^{-1}$ (<0.2 mg per 100 ml) are considered to be evidence of biochemical deficiency, and if this is severe and prolonged, the risk of clinical deficiency, i.e., scorbutic signs and symptoms, gradually increases. Intakes below around $20\,mg\,day^{-1}$ are likely to result in plasma levels in this range. Studies of human volunteers in the middle of the twentieth century showed clearly that an intake of 10 mg vitamin C per day in a healthy adult is sufficient to prevent clinical scurvy, and this small amount is also sufficient to cure scorbutic signs and symptoms (**Table 1**).

Assay methods based on urinary excretion of vitamin C have been used to study status, but they are too cumbersome and difficult to interpret to be useful in population studies. There are no well-established functional assays available to define vitamin C status and requirements at present. An older

method known as the 'Hess test,' which measures relative capillary fragility under pressure or suction (**Table 2**), is useful only if subclinical scurvy is present and is rarely attempted today. Studies of collagen crosslinks or oxidative damage to macromolecules such as DNA or lipids may yield evidence about functional status in the future, but this remains a research challenge and is not yet an available option for routine studies or surveys.

Occurrence of Low Intakes and Poor Biochemical Status in Present-Day Societies

Although scurvy is rare, biochemical evidence of poor vitamin C status is not uncommon in certain high-risk groups in different human populations. Studies in The Gambia in West Africa, for instance, have shown that there is a regular seasonal cycle of availability of foods rich in vitamin C, with a good availability in the dry season alternating with a severe shortage during the rainy season. Plasma, buffy-coat, and breast-milk concentrations are all, on average, adequate in the dry season but are severely reduced during the rains. Functional and health-related parameters also deteriorate during the rains, but it has so far proved difficult to reverse this deterioration by vitamin C supplements alone. Therefore, robust evidence of health consequences of this seasonal availability cycle has not yet been obtained.

From recent surveys in the UK, **Table 3** shows the prevalence of low intakes of vitamin C (estimated from the proportion of participants receiving less than the lower reference nutrient intake (LRNI), which is the amount deemed to be sufficient for only a few people in a population group, namely the 2.5% with the lowest requirements). Also shown in **Table 3** is the prevalence of plasma concentrations below the lower cut-off of normality, set at $0.2\,mg\,dl^{-1}$ or $11\,\mu mol\,l^{-1}$. This is shown for several subgroups of the British population of different ages, from data collected in three nationally representative population surveys during the decade 1990s. It is clear from these results that very few people were getting less than the LRNI for vitamin C over a 4 day or 7 day period of weighed-intake estimates of their diets. Low plasma levels were likewise relatively uncommon in the younger age groups; however, they were more common in older people and were especially prevalent, at almost 40%, in older people living in institutions such as nursing homes. These relatively low plasma levels seen in frail older people are likely to be caused by factors other than very low intakes of the vitamin. In the UK, unlike The Gambia, there was relatively little evidence of a major seasonal variation in vitamin C intake or status at the end of the twentieth century.

Other studies have shown that vitamin C absorption does not appear to be abnormally low in healthy older people. However, there is growing evidence that the multiple pathologies associated with old age (and with debility at any age) are associated with increased turnover of the vitamin. Older people with very low levels of vitamin C are at higher risk of dying sooner than those with high levels, although short-term vitamin supplements generally fail to reverse this increased risk. It thus appears that vitamin C status can act as a barometer of health as well as being a marker of adequacy of vitamin C intake. Further research is needed to determine the key mechanisms that affect the rate of vitamin C turnover and its control in different age

Table 3 Prevalence of low vitamin C intakes and low plasma vitamin C concentrations in Britain at the end of the twentieth century: data from the National Diet and Nutrition Surveys[a]

Age group	*LRNI[b] ($mg\,day^{-1}$)*	*Intake less than LRNI[b]*	*C Less than $11\,\mu mol\,l^{-1}$ plasma vitamin*
Pre-school 1.5–4.5 years	8	8/723 = 1.1%	24/723 = 3.3%
Young people 4–18 years			
4–10 years	8	1/423 = 0.2%	6/422 = 1.4%
11–14 years	9	0/307 = 0%	4/307 = 1.3%
15–18 years	10	1/271 = 0.4%	8/271 = 3.0%
Adults 65 years and over			
Free-living 65–79 years	10	8/606 = 1.3%	88/606 = 14.5%
Free-living 80+ years	10	7/274 = 2.5%	45/274 = 16.4%
Institution-living	10	2/248 = 0.8%	98/248 = 39.5%

[a]Confined to those participants who provided both a weighed-intake record for 4 or 7 days and a blood sample for the biochemical analysis. Source: National Diet and Nutrition Survey Series, commissioned jointly by the Department of Health and MAFF, whose responsibility has since been transferred to the Food Standards Agency. The National Diet and Nutrition Survey Reports are published by The Stationery Office.

[b]Lower reference nutrient intake, deemed to be sufficient for only 2.5% of the population who have the smallest requirements.

groups and different metabolic states. Since frail older people are at high risk of developing pressure sores and of needing surgery for a variety of ailments, there seems to be a potential public-health advantage in avoiding the development of very low vitamin C stores in this vulnerable age group, as a sensible precautionary measure.

See also: **Antioxidants**: Diet and Antioxidant Defense. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Fruits and Vegetables**. **Nutritional Assessment**: Biochemical Indices. **Older People**: Nutritional-Related Problems. **Supplementation**: Role of Micronutrient Supplementation.

Further Reading

Barnes MJ and Kodicek E (1972) Biological hydroxylations and ascorbic acid with special regard to collagen metabolism. *Vitamins and Hormones* **30**: 1–43.

Bates CJ, Prentice AM, and Paul AA (1994) Seasonal variations in vitamins A, C, riboflavin and folate intakes and status of pregnant and lactating women in a rural Gambian community. *European Journal of Clinical Nutrition* **48**: 660–668.

Carpenter KJ (1986) *The History of Scurvy and Vitamin C*. Cambridge: Cambridge University Press.

Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium and Carotenoids*. Washington, DC: National Academy Press.

Kivirikko KI and Myllyla R (1982) Posttranslational enzymes in the biosynthesis of collagen: intracellular enzymes. *Methods in Enzymology* **82**: 245–304.

Langlois MR, Delanghe JR, de Buyzere ML, Bernard DR, and Ouyang J (1997) Effect of haptoglobin on the metabolism of vitamin C. *American Journal of Clinical Nutrition* **66**: 606–610.

Packer L and Fuchs J eds. (1997) *Vitamin C in Health and Disease*. New York: Marcel Dekker Inc.

Pinnell SR, Krane SM, Kenzora JE, and Glimcher MJ (1972) A heritable disorder of connective tissue. Hydroxylysine-deficient collagen disease. *New England Journal of Medicine* **286**: 1013–1020.

Sato P and Udenfriend S (1978) Studies on vitamin C related to the genetic basis of scurvy. *Vitamins and Hormones* **36**: 33–52.

Atherosclerosis *see* **Cholesterol**: Sources, Absorption, Function and Metabolism. **Coronary Heart Disease**: Prevention

B

B Vitamins *see* **Cobalamins**. **Niacin**. **Pantothenic Acid**. **Riboflavin**. **Thiamin**: Physiology; Beriberi. **Vitamin B_6**

Bacteria *see* **Infection**: Nutritional Interactions; Nutritional Management in Adults

Bases *see* **Electrolytes**: Acid-Base Balance

Beer *see* **Alcohol**: Absorption, Metabolism and Physiological Effects; Disease Risk and Beneficial Effects; Effects of Consumption on Diet and Nutritional Status

BEHAVIOR

E L Gibson, University College London, London, UK
M W Green, Aston University, Birmingham, UK

Introduction

The effects of diet on behavior have long been a topic of folklore, superstition, and popular mythology and more recently the subject of rigorous, and not so rigorous, scientific study. Most research into dietary effects on human behavior has assessed changes in mood or mental function after eating (or drinking) or after fasting. Typical measures of mental function include reaction time, attention, memory, problem solving, and intelligence. In addition, research has addressed the effects of diet on disturbed behavior, including attention-deficit–hyperactivity disorder (ADHD) in children, antisocial behavior and aggression, mental illness – for example depression – and dementia (**Table 1**).

Clearly, chronic malnutrition can seriously affect behavior by impairing brain development, and acute malnutrition may result in insufficient nutrients being available for optimal cognitive function. However, this article concentrates on dietary effects on behavior that are not the result of chronic malnutrition or of pharmacologically active ingredients of the diet such as alcohol or caffeine. Rather, the behaviors arise from more subtle effects of variation in nutrient intake within the normally nourished population.

Effects of Meals

The commonest way in which food can affect behavior is the change in mood and arousal that occurs after eating a meal. This might sound trite, but it is not trivial: this general meal effect is probably the most reliable example of an effect of diet on behavior. Many animals, including

Table 1 Examples of nutritional variables known or suspected to affect behavior, mood, and cognition

Food restriction
Early-life undernutrition
Chronic semi-starvation
Dieting to lose weight
Short-term fasting (e.g., missing a meal)
Meal effects
Pre- to post-meal changes
Meal timing (e.g., morning, afternoon, night)
Meal size
Macronutrient composition (acute and chronic effects)
Amino-acids
Neurotransmitter precursors (e.g., tryptophan, tyrosine, phenylalanine)
Phenylketonuria
Sugars
Sucrose (dietary intake)
Glucose (supplement, tolerance)
Micronutrients
Iodine
Iron
Selenium
B-vitamins: B_1, B_6, B_{12}, folate
Vitamin C
Vitamin E
Diabetes
Acute effects of hypoglycemia
Chronic effects
Pharmacological
Caffeine
Alcohol
Nutraceuticals (e.g., plant compounds)

humans, tend to be aroused, alert, and even irritable when hungry. This encourages their search for food. However, their mental processes become distracted by this task, to the detriment of other behaviors. After eating a satiating meal, we and other animals become calm, lethargic, and may even sleep.

Nevertheless, even this seemingly straightforward phenomenon can be distorted and can vary across individuals and situations. The impact of a food or drink will depend on the person's initial state. For example, thirsty people improved their vigilance when allowed to drink water, whereas when people were asked to drink when not thirsty, their performance deteriorated. Numerous experiments have shown that manipulation of the structure of meals results in variation in postprandial changes in mood and mental function. One obvious facet of meals that has been investigated is what is eaten, i.e., nutrient composition; the other two main aspects of meal structure that have been studied are meal timing and meal size. Of course, the effect of a meal on appetite also represents a behavioral effect, but this aspect is covered elsewhere in this encyclopedia.

Besides any nutritional effects, two other influences on behavior are known to interact with attempts to measure dietary effects on behavior. First, most people are very habitual in their choice of food and in the size and timing of their meals. As a result, they have learned a set of beliefs and expectations about the impact of their habitual dietary regime. Therefore, particularly in short-term tests, these expectations may override or mitigate physiological changes. Dietary experiences that differ from a person's habitual eating pattern could lead their behavior to change through cognitive rather than (or as well as) physiological influences.

Second, there are circadian rhythms and sleep–wake cycles in arousal and performance, which complicate the interpretation of meal effects, as we discuss in the next section.

Meal Timing

Does the timing of a meal in the day make a difference to any effects on behavior? In other words, do any behavioral effects differ between breakfast, midday, and evening meals or between mid-morning and afternoon snacks?

Breakfast The potential effects of breakfast on performance and well-being continue to attract much interest, not least from industry, especially concerning the performance of schoolchildren. Pollitt and colleagues have argued that children are likely to be more susceptible than adults to the effects of fasting, owing to their greater brain metabolic demands relative to their glycogenic and gluconeogenic capacity. The numerous studies in this area have produced inconsistent results, which is partly attributable to variation in the populations studied, their nutritional status, and the designs used. There is a consensus that breakfast is more likely than not to benefit schoolchildren's performance, particularly if the children are already nutritionally vulnerable and have mental abilities with room for improvement.

In all of us, there is a tendency for levels of arousal and alertness to rise during the morning, reaching a peak near midday. Some evidence suggests that breakfast may help to control this arousal, so that attention can be successfully focused on the task in hand. Conversely, omitting breakfast may increase autonomic reactivity, leading

to less-focused attention. This effect could explain the finding that children without breakfast showed better recall of objects to which they had not been asked to attend; such attention to irrelevant stimuli is also known to occur with increased anxiety. Furthermore, increasing hunger is likely to be distracting.

Less attention has been paid to the effects of breakfast in adults. However, there are several studies of the effects of giving breakfast to students that show a benefit in spatial and verbal-recall tasks 1–2 h later, compared with missing breakfast. Interestingly, attention-based and reaction-time tasks were not improved by breakfast, and a logical-reasoning task was even slightly impaired. Perhaps performance in those tests benefits more from mild arousal, which could be acutely reduced by some breakfasts. These studies did not determine whether performance later in the morning is affected by breakfast. Differential effects of breakfast content and size will be discussed below.

Midday meal Several studies have demonstrated a drop in performance after the midday meal, particularly in vigilance tasks requiring sustained attention. However, this 'post-lunch dip' may not simply be an effect of eating, because vigilance has also been found to decline from late morning to early afternoon in subjects not eating lunch. That is, there is an underlying circadian rhythm in performance that is confounded by the effect of a midday meal. In fact, using noise stress to arouse subjects during a midday meal prevented any decline in performance due to the meal. It has also been shown that the more anxious one is feeling prior to lunch, the less one will experience a post-lunch dip in performance. In support of this, another study found that subjects scoring highly on a personality measure of extraversion and low on neuroticism were more likely to be affected by a post-lunch dip. These are examples of the importance of individual differences and context for meal effects.

Evening meal There are few studies of the effects of eating later in the day, although there has been some interest in the effects of meals during night shifts. Accuracy of performance declines with eating during a night shift, but, unlike the effects of lunch, pre-meal anxiety levels had no effect. One study in students on the effects of eating a large freely chosen evening meal found little evidence for consistent changes in performance relative to missing the meal. Despite this, the students who omitted the meal reported feeling more feeble and incompetent and less outgoing than those who had eaten.

Snacks One study specifically addressed whether an afternoon snack (approximately 1–1.2 MJ (240–290 kcal) of yoghurt or confectionery) eaten 3 h after lunch (or no lunch) affected task performance. A beneficial effect of the snack was found on memory, arithmetic reasoning, and reaction time 15–60 min later. The comparison was with performance after a 'placebo' zero-energy drink (participants were unaware of the energy content). This rather different placebo does not preclude effects due to differences in sensory experience and expectations. Moreover, whether or not lunch had been eaten beforehand had little effect on the outcome, suggesting that any nutritional effects must be isolated to the acute impact of the snack. It is known that snacks of this size eaten after a meal have only a small effect on blood glucose, although insulin rises sufficiently to inhibit lipolysis and suppress the release of plasma free fatty acids later in the post-prandial period.

The authors reported that these performance benefits from an afternoon snack were not found with a snack eaten in the late morning. The most likely reason is that the beneficial effect depends on the decline in alertness that normally occurs during the afternoon.

Other studies have found differential effects of the macronutrient content of snacks; these are discussed below.

Meal Size

The effect of meal size on behavior has been little studied, perhaps because there are a number of methodological difficulties and an absence of theory. For example, what counts as a large or small meal? Should the difference be measured in terms of absorbed energy, or weight or volume eaten, or even consumption time? If absorbed energy is used as the measure, then behavioral outcomes would need to be measured with a sufficient delay for differences in energy absorption to be discriminable. Moreover, the influence of expectations and habit might confound experimental nutritional differences.

Two studies in adults found that eating large lunches (at least 4 MJ (1000 kcal)) impaired vigilance relative to eating small or medium-sized lunches. There was also evidence that this effect depended on the meal size being different from that habitually consumed. In adolescents, a larger breakfast (2.6 MJ (634 kcal), on average) resulted in poorer vigilance but better short-term memory 3 h later,

compared with a smaller breakfast (1.6 MJ (389 kcal), on average). Thus, there is some evidence that vigilance is adversely affected by a large meal.

Meal Composition

Carbohydrate versus protein The effects of varying the nutrient composition of meals have been studied extensively, rather more for mood than performance. This is largely because of evidence that plasma and brain levels of precursor amino-acids for the synthesis of monoamine neurotransmitters (chemicals responsible for signalling between nerve cells), strongly implicated in mood disorders, can depend on the ratio between carbohydrate and protein in the diet. Synthesis of the neurotransmitter serotonin (or 5-hydroxytryptamine (5-HT)) depends on the dietary availability of the precursor essential amino-acid, tryptophan, owing to a lack of saturation of the rate-limiting enzyme, tryptophan hydroxylase, which converts tryptophan to 5-hydroxytryptophan (see **Figure 1**). An important complication is that tryptophan competes with several other amino-acids, the large neutral, primarily branched-chain, amino-acids (LNAA), for the same transport system from blood to brain. If the protein content of a meal is sufficiently low, for example less than 5% of the total energy as protein, then relatively few amino-acids will be absorbed from the food in the gut. At the same time, insulin will stimulate tissue uptake of competing amino-acids from the circulation, and the plasma ratio of tryptophan to those amino-acids (tryptophan/LNAA) will rise, favoring more tryptophan entry to the brain. Conversely, a high-protein meal, which would be less insulinogenic, results in the absorption of large amounts of competing amino-acids into the blood, especially the branched-chain amino-acids (leucine, isoleucine, and valine). On the other hand, tryptophan is scarce in most protein sources and is readily metabolized on passage through the liver: thus, the plasma ratio of tryptophan to competing amino-acids falls after a protein-rich meal. Indeed, the protein-induced reduction in plasma tryptophan ratio often seems to be more marked than any carbohydrate-induced rise. Such effects also depend on the interval since, and nutrient content of, the last meal.

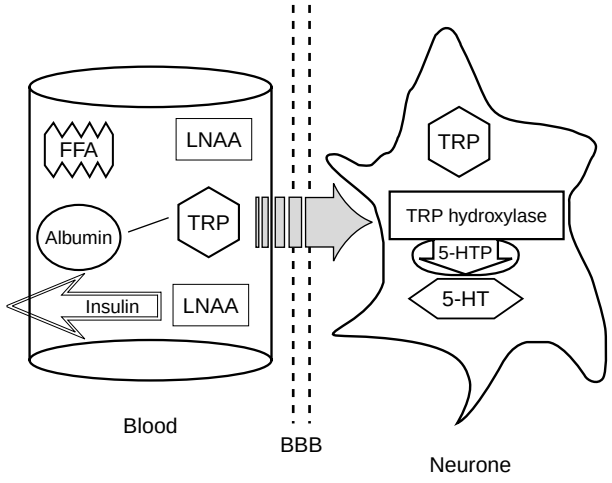


Figure 1 The pathways involved in the synthesis of the neurotransmitter 5-hydroxytryptamine (5-HT; serotonin) from the precursor essential amino-acid, tryptophan (TRP). Tryptophan is taken up by neurones from the blood, but its passage across the blood–brain barrier (BBB) is in competition with that of another group of essential amino-acids known as the large neutral amino-acids (LNAA). Thus, the ratio of tryptophan to total LNAA determines how much tryptophan enters the brain. Most tryptophan is normally bound to albumin in plasma, so it is not available for uptake into the brain. However, after a carbohydrate-rich low-protein meal, increased release of insulin raises levels of free fatty acids (FFA) in plasma, and these displace tryptophan from albumin. In addition, insulin promotes tissue uptake of the LNAA from plasma. Hence, the ratio of tryptophan to total LNAA increases and more tryptophan enters the brain. Increased availability of tryptophan in neurones drives greater synthesis of 5-HT because the rate-limiting enzyme, tryptophan hydroxylase, which converts tryptophan to the intermediate 5-hydroxytryptophan (5-HTP), is not fully saturated.

This evidence is particularly relevant to dietary effects on mood and arousal, because 5-HT has long been implicated in sleep and in affective disorders such as depression and anxiety. However, cognitive performance might also be affected, given the known role of 5-HT in responsiveness to environmental stimuli and stressors, impulsivity, and information processing. Importantly, there is evidence that the dietary availability of tryptophan can influence brain function in humans: for instance, feeding a tryptophan-free diet, which considerably reduced plasma tryptophan (and so could be expected to impair 5-HT function), induced depression in previously recovered depressives and in people with a genetic predisposition to depression. Furthermore, a tryptophan-free drink has been shown to impair performance in tests of visuospatial and visual-discrimination learning and memory.

There is evidence that people feel calmer and sleepier after snacks or meals rich in carbohydrate but virtually free of protein (an unusual situation) than after protein-rich meals containing little carbohydrate. This is compatible with changes in 5-HT function, but these studies did not determine whether this is due to an increase in 5-HT after the carbohydrate-rich meal or a decrease after the protein-rich meal, which could prevent the postprandial sleepiness. Furthermore, adding more than 5–6% protein (of total energy) to the carbohydrate meal has been shown to prevent the increased synthesis of central

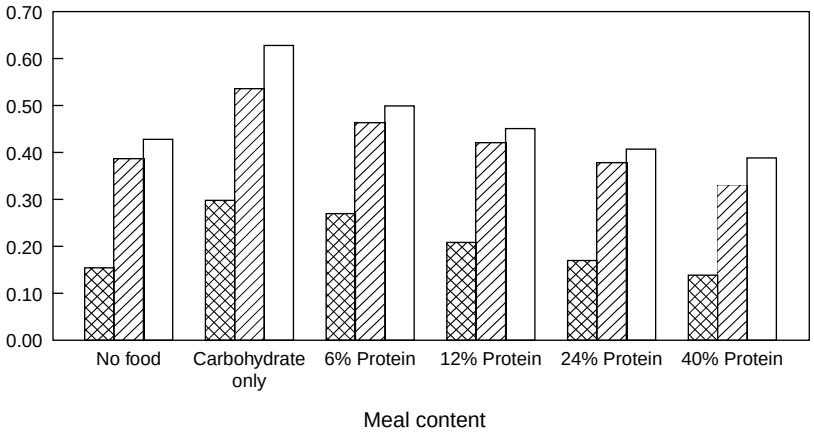


Figure 2 The effect in rats of no meal, a carbohydrate meal with no protein, and meals with increasing amounts of protein on the plasma ratio of tryptophan to the large neutral amino-acids with which tryptophan competes for entry across the blood–brain barrier (cross-hatched bars), levels of tryptophan in the hypothalamus of the rat brain (hatched bars, expressed in $\mu mol\,g^{-1}$), and levels of 5-hydroxytryptophan, an intermediate precursor of serotonin synthesis, in the hypothalamus (open bars, expressed in $0.1\,\mu g\,g^{-1}$). The rise in tryptophan entering the brain after a carbohydrate meal drives increased serotonin synthesis, but this effect is progressively inhibited by increasing protein content.

5-HT, relative to fasted levels, in both rats and people (see **Figure 2**). Also, even pure carbohydrate does not appear to induce sleepiness in everyone.

Another difficulty in comparing the effects of carbohydrate and protein intakes is that relative changes in mood and performance might be due to a protein-induced increase in plasma tyrosine, the precursor amino-acid for synthesis of the catecholamine neurotransmitters (adrenaline, noradrenaline, dopamine), which also competes with LNAA for entry into the brain. In catecholamine systems where the neurones are firing rapidly, acute physiological increases in brain tyrosine (e.g., by feeding a high-protein diet) can raise the tyrosine hydroxylation rate and catecholamine turnover. Such systems include dopaminergic neurones involved in arousal, attention, and motivation. Nevertheless, high-protein meals in humans do not always raise the plasma tyrosine–LNAA ratio; the effect depends on nutritional status and time of day, for example.

Differential effects on performance have been seen with less extreme variations in protein and carbohydrate intakes. For example, a lunch of 55% energy as protein and 15% as carbohydrate produced faster responses to peripheral stimuli, but greater susceptibility to distraction, than eating the reverse proportions of protein and carbohydrate. Sleepiness was not affected by macronutrient composition in that study. However, with these protein–carbohydrate ratios, the plasma tryptophan–LNAA ratio could still be lowered by the protein-rich meal relative to the ratio after the carbohydrate-rich meal, even if tryptophan/LNAA does not rise from pre-meal levels after a carbohydrate-rich meal with much more than 5% protein (**Figure 2**).

A delay of at least 2 h after eating may be necessary for changes in neurotransmitter precursors to influence behavior. Earlier effects may be related to changes in glucose availability and levels of insulin and counter-regulatory hormones such as adrenaline, glucagon, and cortisol. These changes could underlie recent results after breakfasts of 20:80, 50:50, and 80:20 protein–carbohydrate ratios (1.67 MJ, 400 kcal). A measure of central attention initially improved after the carbohydrate-rich breakfast, but later improved after the protein-rich ones; the opposite was found for peripheral attention. This study also found that the 80% protein breakfast produced the best short-term memory performance about 1–2 h after eating, but not at 3.5 h.

Effects of dietary fat Most studies of the effects of fat have varied its level together with that of carbohydrate, while keeping protein constant and so allowing equicaloric meals. Comparisons have been made of low-fat (e.g., 11–29% of energy as fat), medium-fat (e.g., 45% of energy as fat), and high-fat (e.g., 56–74% of energy as fat) breakfasts, mid-morning and midday meals, and intraduodenal infusions of lipid or saline. On balance, high-fat meals appear to increase subsequent fatigue and reduce alertness and attention, relative to high-carbohydrate/low-fat meals. However, there are inconsistencies in changes in specific moods and the effects of meal timing: for instance, feelings of drowsiness, confusion, and uncertainty were found

to increase after both low- and high-fat lunches but not after a medium-fat lunch. One possibility is that mood may be adversely affected by meals that differ substantially from habitual ones in macronutrient composition. An alternative is that similar mood effects could be induced (albeit by different mechanisms) by high carbohydrate in one meal and high fat in the other: for example, 1.67 MJ (400 kcal) drinks of pure fat or carbohydrate taken in the morning both increased an objective measure of fatigue relative to a mixed-macronutrient drink, although the two single-nutrient drinks had opposite effects on plasma tryptophan/LNAA ratios.

In many of these studies, the meals were designed to disguise variation in fat level from the participants. It is therefore possible that effects on mood may have resulted from discrepancies between subjects' expectations of certain post-ingestive effects and the actual effects that resulted from neurohormonal responses to the detection of specific nutrients in the duodenum and liver. A case in point may be the increase in tension, 90 min after lunch, with increasing fat intake reported by predominately female subjects: this might reflect an aversive reaction to (unexpected) fat-related post-ingestive sensations.

Postprandial declines in arousal can be quite noticeable 2.5–3 h after high-fat meals, but fat in mid-morning meals seems to be more sedating than fat ingested at lunchtime, which might relate to expectations. By comparison, when lipid was infused directly into the duodenum, a decline in alertness was apparent much sooner, by 30–90 min after the meal. These effects of fat may result from increased release of the gastric regulatory hormone cholecystokinin. However, in a study comparing ingestion of pure fat, carbohydrate, and protein (1.67 MJ (400 kcal) at breakfast), measures of memory, attention, and reaction time deteriorated more after carbohydrate and protein than after fat. This beneficial effect of fat was attributed to the demonstrated relative absence of glycemic and hormonal (insulin, glucagon, and cortisol) perturbations in the 3 h following fat ingestion.

Carbohydrates, Stress, Mood, and Mental Function

Susceptibility to Mood Enhancement by Diet

The possibility that a carbohydrate-rich low-protein meal could raise 5-HT function gave rise to the proposal that some depressed people may self-medicate by eating carbohydrate, so leading to increased 5-HT release in a manner reminiscent of the effects of antidepressant drugs, which enhance aspects of 5-HT function by inhibiting the removal of 5-HT from the synaptic cleft between nerve cells. For the most part, however, early behavioral and pharmacological evidence for such a phenomenon was not very convincing.

Nevertheless, recent research provides some further support for beneficial effects of carbohydrate-rich protein-poor meals on mood and emotion in some people. When participants were divided into high and low stress-prone groups, as defined by a questionnaire, carbohydrate-rich protein-poor meals prior to a stressful task were found to block task-induced depressive feelings and the release of the glucocorticoid stress hormone cortisol, but only in the high stress-prone group. This finding was replicated using high- and low-tryptophan-containing proteins (α-lactalbumin and casein, respectively). It was argued that, because stress increases 5-HT activity, the poor response to stress of the sensitive group might indicate a deficit in 5-HT synthesis that is improved by this dietary intervention.

There is another link between macronutrient intake, stress, and mood. Chronic dysfunction of the stress-sensitive hormone cortisol and its controlling hypothalamic pituitary adrenal (HPA) axis is associated with depression and anxiety and with abdominal obesity. Moreover, protein-rich meals that prevent a meal-induced fall in arousal also stimulate the release of cortisol in unstressed people, and the degree of this effect is positively correlated with the probability of poor psychological well-being. Chronically, a carbohydrate-rich diet is associated with better overall mood state and lower average plasma cortisol than a high-protein diet. Acutely, a carbohydrate preload, but not protein or fat load, enhances cortisol release during stress. This may be related to findings from both human and animal research that suggest that eating carbohydrate-rich and perhaps high-fat foods can help restore normal HPA axis function and glucocorticoid stress responses. Raised levels of cortisol in stressed people contribute to insulin resistance, which in turn promotes abdominal obesity. However, insulin resistance may increase the likelihood of high-carbohydrate low-protein foods raising brain tryptophan and 5-HT levels, because of increased levels of plasma fatty acids, which result in more unbound tryptophan in plasma. Conversely, it has also been found that high baseline cortisol predicts induction of depression by dietary depletion of tryptophan. This might underlie recent findings that insulin-resistant people are less prone to suicide and depression, both of which are believed to be increased by low 5-HT function. Similarly, patients with seasonal affective disorder

show increased insulin resistance in the winter, together with a greater predilection for sugar-rich foods. Unfortunately, despite this protective effect, insulin resistance is a substantial risk to health because of its association with cardiovascular disease.

Sugars and Opioids

Endogenous opioids are released during stress and are known to be important for adaptive effects such as resistance to pain. They are also involved in motivational and reward processes in eating behavior, such as the stimulation of appetite by palatable foods. Perhaps the best evidence for opioid involvement in an interaction between stress and eating is the finding that, in animals and human infants, the ingestion of sweet and fatty foods, including milk, alleviates crying and other behavioral signs of stress. Recently, this effect was shown to depend on sweet taste rather than calories, as non-nutritive sweeteners also reduce crying. This stress-reducing effect can be blocked by opioid antagonists. The conclusion that adults select sweet fatty foods for opioid-mediated relief of stress is tempting, but remains speculative. Also, such behavior would need to be explained in the context of stress itself enhancing endogenous opioid release.

Glucose, Mood, and Mental Function

The possibility that ingesting glucose can alter mood and improve mental function has generated considerable research interest. However, there is space here only to summarize and interpret the key findings and controversies. The interest in glucose arises from two observations: first, that the primary source of energy for brain function is glucose, and, second, that mental function and mood deteriorate when blood glucose concentration falls below basal physiological levels (hypoglycemia; $<3.6\,\mathrm{mmol\,l^{-1}}$). The first observation must be qualified by recent evidence that, first, in times of metabolic demand, the brain can also use lactate very effectively as an energy source, and, second, the brain contains significant stores of glycogen in specialized cells called astrocytes, which can be metabolized for energy by neighboring neurones. Nevertheless, in rats, extracellular glucose levels in a specific region of the brain critical for memory, the hippocampus, decline to a greater extent during more demanding memory tasks, and this decline is prevented by a systemic glucose load.

Hypoglycemia is rarely induced by normal food, although large quantities of sugar-rich drinks taken on an empty stomach might do so in some people. Yet, many studies of the effects of glucose use a method similar to the oral glucose tolerance test (OGTT), in which fasted patients drink aqueous solutions containing 50–75 g of some form of glucose. This is meant to be not a normal nutritional manipulation but a test of glucoregulation. Associations have been reported between rapid and substantial declines in blood glucose after OGTTs and aggressive thoughts and behavior. However, this might be mediated by greater counter-regulatory hormone release.

In studies comparing sugar-rich drinks with zero-energy sweet placebos, many participants report no effect on mood, but some report a rise in subjective energy within an hour, followed by increased calmness. In children, controlled studies failed to support the popular myth that sugar is excitatory: again, it either had no effect or was calming. However, it is worth noting that some adults, and probably children, are especially sensitive to rapid drops in blood glucose, showing counter-regulatory hormone release and 'hypoglycemic symptoms' even though actual hypoglycemic levels of glucose are not reached.

It may be that beneficial effects of glucose ingestion become consistently apparent only when demands are placed on mental function. The findings on glucose and cognitive performance can be summarized as follows.

- The majority of studies that administered a glucose drink found subsequent improvements in performance compared with administration of a placebo, particularly in tests of short-term memory or vigilance tasks that require a large component of 'working memory'.
- Improvements in performance can be associated with rising or falling blood glucose levels, even independently of consuming a glucose load.
- Young healthy subjects require more demanding tasks than the elderly to detect a beneficial effect of glucose load.
- Associations between performance and glucose may be mediated by individual glucoregulatory efficiency.
- Both glucoregulation and performance are influenced by hormones, such as adrenaline and cortisol, that are sensitive to stressful or arousing cognitive tasks.
- Personality, stress-sensitivity, and task involvement can influence glucose uptake and disposal and, hence, the effects of glucose on cognition.

However, one important pattern does emerge: memory performance is worse in poorer than in better glucoregulators (**Figure 3**). This is true not just for elderly patients but also among a healthy student population, especially if the task is sufficiently demanding. High peak blood glucose predicts poor memory performance in elderly patients, whether or

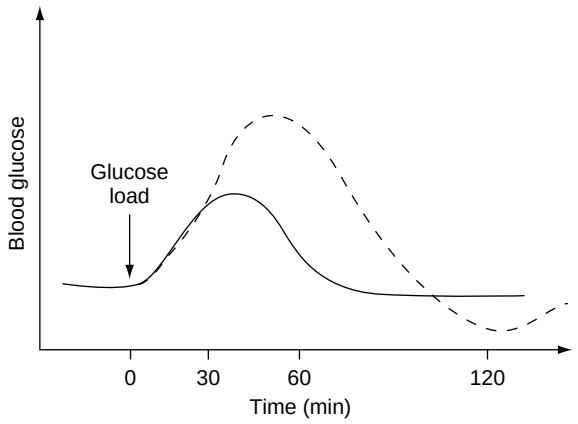


Figure 3 A model of changes in blood glucose levels after a glucose load in people who are either good regulators (solid line) or poor regulators (broken line) of blood glucose. The increased peak and delayed recovery of blood glucose levels in poor regulators suggests glucose intolerance and insulin resistance. In such people, there may be a brief period of mild hypoglycemia prior to a return to baseline levels. This difference in glucoregulation has been demonstrated in both young and elderly people without diabetes. Poor glucoregulation predicts poor cognitive performance in challenging tests, especially those involving memory.

not a glucose load has been given prior to testing. This relationship between raised glucose levels and poor memory performance could underlie a recent finding that a snack with a high glycemic index (greater plasma glucose rise) resulted in poorer memory performance 2–3 h later than a snack with a low glycemic index. Even so, it seems that a glucose load can lessen the memory deficit present in young and old poor glucoregulators (with little consistency or no effect in good glucoregulators).

One reason why poor glucoregulation predicts poor memory performance may be that glucose intolerance is associated with higher basal and stress-induced cortisol secretion. Cortisol is known to impair memory, probably by an action on hippocampal neurones that includes inhibition of glucose uptake. However, the substantial rise in insulin induced by a glucose load in poor glucoregulators may overcome the negative impact of cortisol in some cases: hyperinsulinemia induced independently of hyperglycemia has been shown to ameliorate memory deficits in patients with Alzheimer-type dementia.

Two other mechanisms might explain the ability of a glucose load to improve performance in subsequent challenging tasks. One is an increase in sympathetic activation by the glucose load: adrenaline is known to enhance memory. The other is increased synthesis and release of the neurotransmitter acetylcholine during challenging tasks: acetylcholine is known to be critically involved in learning and memory and is synthesized from dietary choline and from acetyl CoA, which is a by-product of glucose metabolism.

Hyperactivity and Antisocial Behavior

In children, there is an increasing frequency of the diagnosis of ADHD, a condition characterized by inattention, impulsive and disruptive behavior, learning difficulties, and increased levels of gross motor activity and fidgeting. Also, the prevalence of food allergies and intolerances has been increasing. Perhaps it is not surprising that dietary explanations and treatments for ADHD have been sought regularly for several decades, given theories of allergic reactions or intolerance to food additives, ingredients in chocolate, and even refined sugar (often grouped as the 'Feingold theory', after an early instigator of unproven dietary intervention). There has also been a long-standing interest in the possibility that antisocial behavior in children and adults might in part result from poor nutrition, although early studies were poorly designed. Behavioral effects of sugar and of many additives have by and large not been supported by controlled studies; however, determining unequivocally whether the behavior of young children is affected by specific dietary components is difficult. ADHD may be associated with disrupted eating behavior and poor nutrition, so that removal of a number of nutrient deficiencies might improve behavior. In addition, parents or unqualified health professionals may devise unsuitable dietary regimes that can increase the risk of undernutrition. As a result, there is little consensus as to what in the diet may or may not provoke disturbed behavior in children, other than that only a small minority of children are likely to be affected. Nevertheless, a recent British study, in which children were given a collection of food colorings and preservatives, or a placebo, in drinks, found a deterioration in the behavior reported by parents for both hyperactive and normal children given the additives, which seemed unrelated to allergic history. This effect was not detectable in a clinical setting. Clearly, a definitive answer will require more research.

Nevertheless, one promising line of research has involved supplementation with *n*-3 and *n*-6 highly unsaturated essential fatty acids: a recent study of ADHD children found improvement in the supplemented group on several behavioral measures after 3 months. Furthermore, a recent randomized placebo-controlled trial of dietary supplementation in young adult prisoners in the UK found a substantial reduction in antisocial behavior. The supplement used was a multivitamin and mineral preparation taken together with a combined *n*-3 and *n*-6 fatty-acid supplement, for a period varying from 2 weeks

to 9 months. Other effects of essential fatty acids on cognition are discussed below.

Micronutrients and Mental Function

There has been, over the years, an increasing body of evidence suggesting that vitamin and mineral status is significantly related to both brain development in childhood and the degree of cognitive decline experienced as we age. Indeed, it is certainly the case that deficiencies of some vitamins are associated with negative neurological symptoms such as neural-tube defects. The work examining vitamin and mineral supplementation comprises both cohort studies and nutritional interventions and has generated much confusing and contradictory data. To a large degree, this confusion and contradiction result from variation in a number of factors such as the methodological rigour of each study, the measures of cognitive function used, and the precise nutrient being studied.

Early work in this area concentrated on the notion that supplementing the diet of schoolchildren with multivitamin supplements would improve both their IQ scores and their academic achievement. This work was controversial and marked by a number of deficiencies, such as any clear indication of whether the participants were nutritionally compromised prior to treatment, difficulties in determining which, if any, of the vitamins in the cocktail were producing effects, and the lack of any clear hypotheses regarding the mechanisms responsible. The consensus now is that supplementation will benefit cognitive development and IQ (especially non-verbal) in a minority of children who are not otherwise adequately nourished. In Britain and the USA at least, there is particular concern that a significant proportion of adolescents, especially girls, are deficient in iron. There is good evidence that iron deficiency contributes to poor cognitive ability, perhaps in association with low vitamin C status, which has also been linked to reduced cognitive function. Iron is known to be essential for the synthesis and function of neurotransmitters, such as dopamine, noradrenaline, and serotonin (5-HT). Selenium is another mineral that may be important for brain function, and low levels of selenium have been associated with cognitive decline and depressed mood in the European population.

Much recent work, however, has concentrated on the use of vitamins in the treatment of age-related cognitive decline and dementia and, to varying degrees, is more scientifically rigorous than the earlier work. The overwhelming majority of the experimental work has targeted the action of two groups of micronutrients: antioxidants and B-complex vitamins. Research into the effects of antioxidant vitamins, whilst showing some promise in that correlational studies show that levels of these vitamins (vitamin E most consistently) are associated with function in a range of cognitive domains, is more contradictory when one considers the clinical intervention trials. The work on B-complex vitamins is, however, more consistent and supported by a strong hypothetical basis. This relies on the roles of vitamins B_{12} and folate in methylation of membrane phospholipids (see below) and neurotransmitters and in breaking down the toxic sulfur amino-acid homocysteine. High levels of homocysteine are now considered by some to be a far greater risk factor for the development of coronary and vascular problems than are high levels of cholesterol. Elevated levels of homocysteine may be a cause of minor ischemic events, which, cumulatively, lead to a degradation of cognitive function due to either subclinical deficiencies of or problems with the absorption of B-complex vitamins. Indeed, a large number of studies have consistently demonstrated relationships between homocysteine levels, B-complex vitamin levels, and neuropsychological task performance. The number of direct intervention trials that have supplemented the diets of the elderly with B-complex vitamins is, however, small. Whilst some studies have shown no net benefit of supplementation on the cognitive function of the elderly, a larger number of studies have shown a stabilization of cognitive function and a reduction in homocysteine levels resulting from B-complex vitamin supplementation. As with the studies of antioxidant vitamins, however, these studies must be interpreted with a degree of caution since they use differing dosages, periods of supplementation, and measures of neuropsychological function.

Lipids

Lipids have attracted a good deal of research interest in terms of their possible effects on psychological function. The main nutrients studied fall into three groups: cholesterol, *n*-3 and *n*-6 essential fatty acids, and phospholipids. In general, the theoretical basis underpinning the effects (or lack of effects) of these nutrients on psychological function relates to how their relative concentrations affect cell-membrane fluidity. The rigidity of lipid bilayers of cell membranes is thought to be essential for neurotransmitter function because it maintains maximum exposure of receptors at the synaptic cleft between neurones.

Cholesterol

The interest in cholesterol as a substance that is related to psychological well-being stems back to the 1980s. During this period, a number of epidemiological studies found that individuals with low cholesterol levels were more prone to aggressive behavior and at greater risk of suicide and violent death. In addition, it was found that non-human primates increased their incidence of aggressive behavior when kept on a low-cholesterol diet. In terms of neuropsychological function, a number of studies have found associations between cholesterol levels and choice reaction time and/or memory function. Two of these studies sought actively to reduce cholesterol levels by pharmacological or dietary means: both found that lowering cholesterol produced small but statistically significant impairments in memory and attention. Conversely, however, a number of studies have demonstrated that high cholesterol levels are a significant risk factor for the development of Alzheimer's dementia. One mechanism for the negative effects of cholesterol lowering may be that neural cell walls lose their rigidity, thereby decreasing the relative exposure of serotonin (5-HT) membrane receptors at the synaptic cleft and impeding 5-HT signal transmission.

Essential Fatty Acids

There is a body of evidence suggesting that dietary supplementation with docosahexanoic acid (DHA; 22:6, *n*-3) and arachidonic acid (AA; 20:4, *n*-6) is effective in reducing the symptoms of clinical depression and schizophrenia. These fatty acids are also important for the development of the central nervous system in mammals. Recent work has demonstrated that maternal supplementation with these particular fatty acids during pregnancy significantly improves children's IQ at age 4 compared with children whose mothers took corn oil during pregnancy. There is no convincing evidence yet concerning the effects of essential-fatty-acid supplementation on the cognitive function of the elderly, although one study has found a correlational link between *n*-3 fatty-acid intake and the risk of developing Alzheimer's dementia. The proposed modes of action of these fatty acids involve their antithrombotic and anti-inflammatory properties in addition to their being a primary component of membrane phospholipids in the brain.

Phospholipids

The final group of lipids that has been investigated in terms of possible psychological benefits is the phospholipids and, in particular, phosphotidylserine. As with cholesterol, the proposed mode of action whereby this substance affects psychological function is via alterations in cell-membrane fluidity. A number of studies have claimed that dietary supplementation with phosphotidylserine can arrest the symptoms of both age-related decline and full dementia. In general, however, these studies have been poorly carried out and suffer from procedural difficulties such as a lack of appropriate control populations, alterations in dosage during the trial, and use of unvalidated or inappropriate neuropsychological assessment measures. Indeed, some of the more methodologically rigorous trials have specifically shown that phosphotidylserine supplementation exerts no significant effect on psychological function. The evidence for the efficacy of phosphotidylserine is, therefore, unclear.

Food Deprivation

There is evidence to suggest psychological effects of undernutrition. In severe cases, such as anorexia nervosa, neuropsychological function is impaired primarily as a result of structural changes in brain anatomy resulting from starvation. Evidence that undernutrition is associated with psychological problems in those not suffering from eating disorders was first hinted at in the Minnesota Study of Semi-Starvation in the 1950s. Volunteers who were kept on a half-calorie-intake diet for a period of months reported mood swings, increased irritability, poorer memory, and an inability to concentrate. These self-reported effects were not supported by objective testing, but the lack of a non-deprived control group means that there may have been an affect that could have been masked by a practice effect.

Dieting to lose weight is one of the most common food-choice-related behaviors in the Developed World, and it has been consistently associated with negative psychological consequences such as preoccupation with body shape and depression. In addition, a number of investigators have found that dieting to lose weight is associated with impairments in cognitive function, with dieters performing more poorly than non-dieters on measures of reaction time, immediate memory, and the ability to sustain attention. This is unlikely to be due to pre-existing differences between individuals who happen to be dieting and those who are not dieting at the time of testing since, within individuals, performance is poorer when dieting than when not dieting. It is unlikely that these effects are due to the gross physical effects of food deprivation since experimentally induced food deprivation of varying lengths

fails to produce a comparable impairment in task performance; in addition, poorer task performance is found amongst dieters who claim not to have lost any weight over the course of the diet.

Rather than being a function of food deprivation per se, the poorer task performance amongst current dieters appears to be a function of the preoccupying concerns with hunger and body shape that are characteristic of dieters. Indeed, the impairments in task performance amongst dieters appear to be comparable in both structure and magnitude to those that result from the preoccupying concerns that are characteristic of clinical depression and anxiety disorders. Specifically, the primary deficit appears to be a reduction in the amount of available working-memory capacity, working memory being the primary cognitive system that allocates processing capacity to ongoing cognitive operations. A threshold hypothesis has been formulated to account for this phenomenon. Non-dieting highly restrained eaters are characterized by an enduring trait concern with body shape, which consumes a certain amount of working-memory capacity (explaining why non-dieting restrained eaters perform at a level intermediate between the levels of current dieters and unrestrained eaters). When they decide to diet, they then experience preoccupations with food and an increased desire to eat; this extra drain on working-memory capacity reaches a point where insufficient capacity is available to maintain task performance. Support for this hypothesis can be seen in the results of a study in which highly restrained non-dieters were instructed to imagine eating their favorite food or to imagine their favorite holiday whilst performing a reaction-time task. When imagining their favorite food, but not their favorite holiday, restrained non-dieters performed as poorly as current dieters on the reaction-time task.

Although evidence seems to be mounting that the poor cognitive function of current dieters is due to psychological and not biological factors, work continues to examine some of the more subtle possible biological mechanisms that may underlie the effects. One possible mechanism is that a low dietary intake of the amino-acid tryptophan (the precursor for 5-HT) leads dieters to have impaired serotonergic function. However, analysis of the urine of dieters for the 5-HT metabolite 5-hydroxyindoleacetic acid found no evidence for this. Another possibility (not yet investigated) is that, by avoiding red meat, dieters experience mild iron deficiency, with deleterious consequences for hemoglobin status, brain oxygen supply, and neurotransmitter function.

The types of dieter studied so far are those who attempt to lose weight in an unsupported and unsupervised manner. Comparisons between this type of dieter and those who attempt to lose weight in the context of an organized weight-loss group reveal dramatic differences. Those who diet as part of a group do not show the impairments in task performance typical of unsupported dieters. In addition, unsupervised dieters display an elevated stress response after 1 week of attempted weight loss (as measured by salivary cortisol levels), whereas supported dieters do not. It would appear, therefore, that the poor performance characteristic of unsupervised dieting is a result of the stress associated with this type of weight-loss attempt and that the psychological manifestation of this stress is the preoccupying thoughts outlined above.

Functional and Pharmacological Components of Foods and Drinks

There is growing interest, particularly in the food and beverage industry, in developing foods and drinks with functional properties (nutraceuticals) attractive to the consumer. These include effects on behavior, such as improvements in cognitive function, mood, and physical performance. Components of interest include caffeine, herbal extracts such as ginkgo biloba and panax ginseng, micronutrients, essential fatty acids, amino-acids, and carbohydrates. There is some support for beneficial effects of these components, but they are not reviewed further here.

Conclusion

The scientific understanding of dietary effects on behavior has moved in from the fringes of respectability sufficiently to attract substantial commercial interest. Advances in nutritional and neuropsychological knowledge, experimental design, and sensitivity of measures of behavior and brain function have produced replicable findings in some areas to mollify earlier scepticism. New understanding of the impact of nutrition on brain function and of predictors of individual susceptibility has allowed reinterpretation of old data. Promising areas with encouraging developments in understanding include the interactions between macronutrients, stress, and mood disorders, and the effects of vitamins, minerals, and lipids on cognition, dementia, and psychiatric disorders. Some findings, including recent discoveries of poorly nourished sectors of the population, suggest useful interventions. Nevertheless, research in this field is at an early stage, and the coming years should bring further revelations relating to the link between diet and behavior. With industrial backing, few may escape the consequences.

See also: **Appetite**: Psychobiological and Behavioral Aspects. **Brain and Nervous System**. **Caffeine**. **Children**: Nutritional Problems. **Diabetes Mellitus**: Etiology and Epidemiology. **Eating Disorders**: Anorexia Nervosa. **Exercise**: Diet and Exercise. **Fatty Acids**: Omega-3 Polyunsaturated; Omega-6 Polyunsaturated. **Food Choice, Influencing Factors**. **Food Folklore**. **Food Intolerance**. **Glucose**: Metabolism and Maintenance of Blood Glucose Level; Glucose Tolerance. **Glycemic Index**. **Homocysteine**. **Hunger**. **Hyperactivity**. **Hypoglycemia**. **Iodine**: Deficiency Disorders. **Iron**. **Lipids**: Composition and Role of Phospholipids. **Meal Size and Frequency**. **Older People**: Nutrition-Related Problems. **Premenstrual Syndrome**. **Sports Nutrition**. **Supplementation**: Dietary Supplements; Role of Micronutrient Supplementation. **Vitamin E**: Metabolism and Requirements.

Further Reading

Awad N, Gagnon M, Desrochers A, Tsiakas M, and Messier C (2002) Impact of peripheral glucoregulation on memory. *Behavioral Neuroscience* **116**: 691–702.

Benton D (2002) Carbohydrate ingestion, blood glucose and mood. *Neuroscience and Biobehavioral Reviews* **26**: 293–308.

Berr C (2002) Oxidative stress and cognitive impairment in the elderly. *Journal of Nutrition Health and Aging* **6**: 261–266.

Bruinsma KA and Taren DL (2000) Dieting, essential fatty acid intake and depression. *Nutrition Reviews* **58**: 98–108.

Dallman MF, Pecoraro N, Akana SF *et al.* (2003) Chronic stress and obesity: a new view of "comfort food". *Proceedings of the National Academy of Sciences of the United States of America* **100**: 11696–11701.

Ernst E (1999) Diet and dementia: is there a link? A systematic review. *Nutritional Neuroscience* **2**: 1–6.

Fernstrom JD (1994) Dietary amino acids and brain function. *Journal of the American Dietetic Association* **94**: 71–77.

Gesch CB, Hammond SM, Hampson SE, Eves A, and Crowder MJ (2002) Influence of supplementary vitamins, minerals and essential fatty acids on the antisocial behavior of young adult prisoners: randomised, placebo controlled trial. *British Journal of Psychiatry* **181**: 22–28.

Gibson EL and Green MW (2002) Nutritional influences on cognitive function: mechanisms of susceptibility. *Nutrition Research Reviews* **15**: 169–206.

Green MW (2001) Dietary restraint and craving. In: Hetherington MM (ed.) *Food Cravings and Addiction*. pp. 521–548. Leatherhead: Leatherhead Food RA.

Hillbrand M and Spitz RT (1996) *Lipids, Health and Behavior*. Washington, DC: American Psychological Society.

Kalmijn S, Feskens EJM, Launer LJ, and Kromhout D (1997) Polyunsaturated fatty acids, antioxidants and cognitive function in very old men. *American Journal of Epidemiology* **145**: 33–41.

Krummel DA, Seligson FH, and Guthrie HA (1996) Hyperactivity: is candy causal? *Critical Reviews in Food Science and Nutrition* **36**: 31–47.

Markus CR, Olivier B, Panhuysen GEM *et al.* (2000) The bovine protein alpha-lactalbumin increases the plasma ratio of tryptophan to the other large neutral amino acids, and in vulnerable subjects raises brain serotonin activity, reduces cortisol concentration, and improves mood under stress. *American Journal of Clinical Nutrition* **71**: 1536–1544.

Pollitt E and Mathews R (1998) Breakfast and cognition: an integrative summary. *American Journal of Clinical Nutrition* **67**: 804S–813S.

Reynolds EH (2002) Folic acid, ageing, depression, and dementia. *BMJ* **324**: 1512–1515.

Riggs K, Spiro A, Tucker K, and Rush D (1996) Relations of vitamin B_{12}, vitamin B_6, folate, and homocysteine to cognitive performance in the Normative Aging Study. *American Journal of Clinical Nutrition* **63**: 306–314.

Rogers PJ (2001) A healthy body, a healthy mind: long-term impact of diet on mood and cognitive function. *Proceedings of the Nutrition Society* **60**: 135–143.

Sher L (2001) Role of thyroid hormones in the effects of selenium on mood, behavior, and cognitive function. *Medical Hypotheses* **57**: 480–483.

Smith AP and Jones DM (eds.) (1992) *Handbook of Human Performance*, vol. 2, *Health and Performance*. London: Academic Press.

Story M and Neumark-Sztainer D (1998) Diet and adolescent behavior: is there a relationship? *Adolescent Medicine* **9**: 283–298.

Tucker DM, Penland JG, Sandstead HH *et al.* (1990) Nutrition status and brain function in aging. *American Journal of Clinical Nutrition* **52**: 93–102.

Wolraich ML, Wilson DB, and White W (1995) The effect of sugar on behavior or cognition in children: a meta-analysis. *JAMA,* **274**: 1617–1621.

Beriberi *see* **Thiamin**: Beriberi

Beverages *see* **Alcohol**: Absorption, Metabolism and Physiological Effects; Disease Risk and Beneficial Effects; Effects of Consumption on Diet and Nutritional Status. **Tea**

BIOAVAILABILITY

R J Wood, Tufts University, Boston, MA, USA

Introduction

Bioavailability is an important consideration in a number of areas of nutrition, including the derivation of Dietary Reference Intakes, estimation of potential impact of changes in dietary pattern, the selection of specific food fortificants, and the formulation of whole meal products, such as infant formulas or meal replacements. In this article, the role of food processing in nutrient bioavailability, specific determinants of nutrient bioavailability, methods for measuring bioavailability, and bioavailability of food fortificants are discussed.

Definition

Nutrient bioavailability is defined as the fraction of a nutrient in a food that is absorbed and utilized. In practice, however, measurements of bioavailability have focused on either direct measurement of absorption or determination of the change in some functional or biochemical endpoint reflecting absorption and utilization of the nutrient. In general, the bioavailability of all nutrients can be estimated by measuring absorption alone because, once absorbed, nutrients are freely available for biological utilization, irrespective of their original dietary source. For example, consider the case of iron bioavailability. Iron absorption can be measured directly from a food by a variety of methods (described in more detail later). In addition, absorbed iron enters the plasma iron pool carried on the protein transferrin. In turn, this absorbed iron will be used in large part (about 80%) immediately in the synthesis of hemoglobin by erythrocyte precursor cells in the bone marrow. The fraction of iron that is utilized for hemoglobin synthesis is not dependent in any way on the food source of that iron. Thus, food iron bioavailability can be conveniently measured and compared in relative terms among various sources by determining the change in blood hemoglobin after consumption of various forms of iron in iron-deficient subjects. Thus, nutrient bioavailability can be estimated by measuring these appropriate endpoints, such as hemoglobin incorporation for iron, hepatic tissue or mineral content of bone for various bone-seeking minerals, or more generally as growth stimulation under nutrient limiting conditions, etc.

An exception to the 'absorption equals bioavailability' rule is selenium. One form of dietary selenium is selenomethionine. This selenium-containing compound is handled by the body exactly like the amino acid methionine and gets readily incorporated into methionine-containing proteins. However, the selenium found in selenium-dependent enzymes is in the form of a special amino acid called selenocysteine, which must be synthesized in the body during the process of incorporation of selenocysteine into these selenoproteins. Selenomethionine catabolism will result in the release of this selenium into an active endogenous selenium pool, which serves as the source of selenium for selenocysteine synthesis. Thus, selenium in selenomethionine is not immediately available to support selenium-dependent functions in the body.

Effects of Food Processing

Food processing can have positive or negative effects on nutrient bioavailability. For example, milling of grains removes all or part of the external covering of the grain (bran) that contains high amounts of phytic acid, an important inhibitor of bioavailability of divalent minerals, such as iron, zinc, and copper. One disadvantage of this form of food processing is that much of the mineral content of grains is in the bran fraction and is lost in the process of milling. To compensate for this loss of mineral (and some vitamins as well) grain products can be 'enriched' by fortifying the flour made from the milled grain with specific micronutrients. Simple food processing techniques, such as sprouting or fermentation, are also effective in lowering the phytate content and increasing mineral bioavailability of grains. Other techniques used in food manufacturing, such as browning, which produces the Maillard reaction, and extrusion, can have negative effects on bioavailability of certain nutrients. Processing practices that affect the polyphenol content of cereals and legumes can influence nutrient bioavailability. Polyphenols interact with plant proteins and form tannin–protein complexes that can inactivate enzymes or lead to protein insolubility adversely influencing amino acid and protein bioavailability. The antinutritional properties of polyphenols can be decreased by removing them from grain with chemical treatments (such as alkaline treatment and ammonia) or removing from grain the polyphenol-rich pericarp and testa by pearling.

Importance of Nutrient Bioavailability to Human Nutrition

Assessment of bioavailability of nutrients is an essential component of deriving dietary reference intakes (DRIs), guidelines for optimal intake of individual nutrients established for North American populations. Many DRIs are based on evaluation of available physiological data to determine the obligatory daily needs for a nutrient to replace losses, or the amount needed for optimal growth of tissues, and an estimate of overall dietary bioavailability of the nutrient in question. In many populations, the content of a nutrient in the diet (e.g., iron or zinc) may be sufficient to meet recommended intake, but bioavailability is suboptimal due to the presence of high levels of inhibitory substances (such as phytate) in the diet leading to a high risk of developing this nutrient deficiency.

Determinants of Nutrient Bioavailability

Speciation

Speciation refers to the form of the nutrient found in food, which may in turn influence the absorption of the nutrient in the gastrointestinal tract. For example, this term could be applied to *cis*- or *trans*-forms of unsaturated fatty acids found in partially hydrogenated oils; individual coenzyme forms of vitamins, such as free thiamin, thiamin pyrophosphate (TPP), thiamin monophosphate (TMP) and thiamin triphosphate (TTP), or the various coenzyme forms of riboflavin – flavin mononucleotide (FMN) and adenine dinucleotide (FAD), or for vitamin B_6 – pyridoxine, pyridoxamine, and pyridoxal; or minerals, such as ferrous or ferric forms of nonheme iron or heme iron; or the selenomethionine, selenite, or selenate form of selenium.

Digestion and Metabolism

During the transit of digested food material through the gastrointestinal tract many changes occur in the intestinal lumen that could influence nutrient bioavailability. Digestion of food constituents is an important aspect of nutrient bioavailability. The secretion of acid into the stomach following food ingestion activates certain digestive enzymes, as well as creating an acidic environment that influences mineral solubility and extraction from food. In this regard, the choice of a mineral to use for supplementation purposes might be influenced by certain physiological conditions, such as achlorhydria. Aging is associated with a decrease in gastric acid secretion in many elderly persons leading to either hypochlorhydria or achlorhydria, characterized by either low or complete absence of acid secretion, respectively, leading to a neutral or slightly alkaline gastric pH. This raised pH condition in the stomach can have detrimental effects on micronutrient bioavailability. For example, elderly persons with achlorhydria are at risk of vitamin B_{12} deficiency because of an inability to remove properly protein-bound vitamin B_{12} from food when it enters the stomach. In these subjects, the capacity to absorb vitamin B_{12} is normal, however, because they can readily absorb crystalline (nonprotein-bound) vitamin B_{12} from a supplemental vitamin B_{12} dose. Lowering the gastric pH towards normal by administering hydrochloric acid restores vitamin B_{12} absorption from food.

In achlorhydric elderly persons the absorption of calcium carbonate from a dietary supplement after an overnight fast is very poor, presumably because calcium carbonate is a relatively insoluble calcium salt and needs gastric acid to be solubilized. Again, these subjects have normal calcium absorption if the calcium is delivered in a more soluble form, such as calcium citrate. In addition to elderly persons who develop achlorhydria, a large number of people regularly use gastric acid-lowering medications, such as the gastric proton pump inhibitor omeprazole, for antiulcer therapy. These medications can reduce zinc absorption and presumably may affect the absorption of other divalent minerals. Acidification of the gastric contents and solubilization of minerals in the gastric juice is important because many mineral nutrients are preferentially absorbed in the duodenum (upper small intestine) and need to be available in free or low-molecular-weight complexes when they leave the stomach to facilitate contact with intestinal nutrient transporters on the apical (luminal) surface of the absorptive enterocytes.

The lower gastrointestinal tract can also be a potentially important site affecting the bioavailability of bioactive substances found in food. In this regard, intestinal bacteria found in the large intestine can influence bioavailability. Bacteria are instrumental in the metabolic conversion of certain phytonutrients into forms that are more readily absorbed. In addition, bacteria likely play an important role in the enhancing effects of prebiotics on mineral bioavailability. Consumption of nonabsorbable carbohydrates, such as inulin, can have a positive effect on mineral absorption. A possible mechanism of this effect is that the nonabsorbable carbohydrates pass the small intestine and enter into the large intestine where they serve as a food substrate for bacteria. The metabolism of these prebiotic substances by intestinal bacteria lowers the pH of the lumen of the large intestine and may thereby

serve to solubilize some insoluble mineral complexes that have passed through the small intestine. The freed mineral would then be available for absorption from the large intestine, thereby increasing overall mineral bioavailability. Other scenarios are possible as well, such as the release of short-chain fatty acids by the bacteria, that may facilitate mineral absorption.

Chelation

Chelation is the process whereby an organic moiety acts as a ligand to bind a metal ion through two or more coordination bonds. Some low-molecular-weight compounds that may be released during the digestion of food can act as metal chelators and increase metal solubility in the intestinal lumen. In some circumstances, chelated forms of metals are naturally present in food, such as heme iron (part of hemoglobin or myoglobin protein) found in meat. Heme is a stable protoporphyrin ring-containing compound that protects a central iron atom from interacting with other potentially deleterious compounds, such as phytic acid, that would reduce its availability and inhibit iron absorption. Nonheme iron bioavailability is affected by various enhancing and inhibitory substances found in food. In contrast, heme iron bioavailability is not. The heme moiety is absorbed intact by the enterocyte. Inside the enterocyte a cytosolic enzyme heme oxygenase breaks apart the protoporphyrin ring and releases the caged iron atom, which can then be transferred out into the blood.

Bioavailability Enhancers and Inhibitors

Various food substances have been identified that act as enhancers or inhibitors of divalent mineral absorption. In general, these food factors influence nutrient bioavailability by either forming relatively insoluble complexes with nutrients or preventing them from interacting with their respective nutrient transporter, or by protecting the nutrient from such untoward interaction maintaining it in a state that can be absorbed or as an absorbable chelated complex (e.g., heme iron).

A list of factors known to influence mineral bioavailability is given in **Table 1**.

Table 1 Dietary enhancers and inhibitors of mineral bioavailability

Enhancers	*Inhibitors*
Ascorbic acid	Polyphenols (especially galloyl groups)
Organic acids	Phytic acid
Meat factor	Myricetin
Alcohol	Chlorogenic acid (coffee)
Inulin	Insoluble dietary fiber

General Physiology of Nutrient Absorption

Nutrients enter the blood by passing through the intestinal mucosa. Intestinal nutrient transport can occur via two distinct pathways. One is termed the paracellular pathway and represents the movement of a nutrient between the absorptive enterocytes on the intestinal villi. This transport pathway is an energy-independent diffusional process and depends on the electrochemical gradient across the mucosa and its permeability characteristics to the nutrient in question. The characteristics of the diffusional pathway are not regulated in response to nutrient deficiency or excess. A second transport pathway represents the transcellular movement of a nutrient across the intestine. The transcellular transport rate of the nutrient is composed of both diffusional and carrier-mediated transport pathways. Often in response to changes in nutrient status the number of nutrient carriers will be changed to facilitate appropriate increases or decreases in intestinal absorption to help maintain nutrient homeostasis. Within a class of nutrients there can be substantial differences in absorption rates. For example, in the case of minerals, monovalent minerals (such as sodium, potassium, and iodine) are absorbed at very high efficiency approximating complete absorption, while multivalent elements (such as chromium, heavy metals, and iron) are relatively poorly absorbed (1–20%). In addition, because some nutrient transporters, e.g., the divalent metal transporter DMT-1 that is responsible for intestinal iron transport, can also transport more than one type of mineral, unintended alterations in the absorption rate of one mineral may occur by a process of co-adaptation in response to changes in the status or physiologic need for another mineral. For example, in iron deficiency, iron absorption is increased, but the absorption of cadmium and lead are also inadvertently increased. However, these effects (homeostatic regulation and co-adaptation) are responses to a change in physiological state and should not be confused with alterations in nutrient 'bioavailability' *per se*, a characteristic of an individual food, a complex mixed meal, or a longer term characteristic of a particular dietary pattern.

Methods for Measuring Nutrient Bioavailability

In Vitro Bioavailability Technique

Nutrient bioavailability, estimated as absorbability alone, can be measured by various *in vitro* methods. *In vitro* methods have obvious distinct advantages

in that they are less expensive, rapid, and amenable to high throughput analyses. Often, experimental *in vitro* methods involve an initial 'digestion phase' where the food is treated with acid and digestive enzymes to simulate the initial steps of food breakdown. The digestion phase is then followed by a second phase wherein the goal is to estimate the potential relative availability of a nutrient. This usually involves the measurement of the concentration of the soluble nutrient of interest in a supernatant of the digested food following centrifugation or after dialysis of the digested food products across a semi-permeable membrane designed to select only low-molecular-weight complexes. Variations on this theme include the addition of radioactive isotopes following the digestion phase and the *in vitro* measurement of cellular uptake of the nutrient in a cell culture preparation or some appropriate index of nutrient uptake. In the case of iron, for example, cellular synthesis of ferritin, an iron storage protein, has been used. Similar applications of this *in vitro* technique are now appearing in the scientific literature for the measurement of phytochemical bioavailability, such as beta-carotene, lycopene, lutein, etc. However, although promising at the moment, there is little confidence that these methods can adequately replace *in vivo* methods of measuring nutrient bioavailability.

In Vivo Bioavailability Techniques

Various animal models and techniques have been used to estimate nutrient bioavailability. A primary concern that must be initially addressed in using animal models to estimate nutrient bioavailability is whether these various model systems accurately reflect nutrient bioavailability in humans. For example, usual experimental animal models, such as rats and mice, cannot be used to assess beta-carotene bioavailability because the absorptive mechanism in these rodents for this nutrient is quite different from that in humans. In contrast, the ferret appears to be a suitable animal model to at least mimic this carotenoid's intestinal absorptive pathway. Similarly, poor correlation of iron bioavailability between chicks and humans for elemental iron powders has raised questions about the suitability of that species for estimating nutrient bioavailability of iron.

Measuring Nutrient Bioavailability in Humans

The first consideration is often whether it is necessary to have an accurate quantitative estimate of nutrient bioavailability, or whether an estimate of relative bioavailability compared to a known standard nutrient source will suffice. An accurate quantitative measure of bioavailability might be necessary when the intention is to provide data to derive a recommendation for dietary intake to meet a nutrient requirement. In this case, it is important to have a reasonably good estimate of the true fraction of a given dose of the ingested nutrient that could be absorbed and utilized, for example, to replace endogenous losses of the nutrient.

A common application of bioavailability measurements is to compare relative bioavailability between two or more sources of a nutrient. For example, one might be concerned with determining the calcium bioavailability from milk compared to calcium derived from a vegetable source such as broccoli or calcium-fortified orange juice. There are many techniques available to measure relative nutrient bioavailability under *in vivo* conditions based on a comparison of the rise in plasma level (or urinary excretion) of the nutrient or rate of appearance in plasma of a radioactively labeled nutrient after an oral test dose. An important technical advance in measuring food mineral bioavailability in humans was the validation of an extrinsic tag method. Extrinsic tag studies were validated by measuring the extent of absorption of a mineral isotope mixed exogenously (the 'extrinsic tag') with a food compared to that of an intrinsic tag where the absorption of the isotope is determined from an intrinsically labeled food source. The intrinsic tag is often achieved by growing plants hydroponically in a solution enriched in a radioactive or stable mineral isotope to label the plant food of interest during growth, or by supplying the mineral isotope tag to a growing animal used for meat, or one that was used for milk production, for example. These studies have shown that in most cases the ratio of absorption of the extrinsic to the intrinsic isotope was approximately one, indicating that the extrinsically added isotope tag became homogenously incorporated into the pool of absorbed mineral found endogenously in the food of interest. The use of the extrinsic tag method has greatly facilitated the study of relative bioavailability of minerals from food in human subjects.

A large and growing number of people are consuming dietary supplements. However, due to the relative difficulty of labeling these supplements, in most cases, little information on the bioavailability of the nutrients in the supplements is available. A study of vitamin and mineral bioavailability from a popular multinutrient supplement found good absorption of the water-soluble vitamins (B vitamins and vitamin C) from the tablet but relatively poor absorption of copper and zinc.

Food Fortification

In the summer of 1941 a National Nutrition Conference for Defense was held that led to the recommendation that there should be improvement of the nutritive value of certain low-cost stable food products (e.g., flour and bread) by nutrient enrichment to replace nutrients lost during the milling and refining process. This led to recommendations to fortify milk with vitamin D, margarine with vitamin A, and salt with iodine using the new recommended dietary allowances (RDAs) established by the Committee on Food and Nutrition of the National Research Council (currently the Food and Nutrition Board) as a yardstick to judge the appropriate levels of fortification. Standards of identity for 'enriched' flour' were initially established that allowed for the addition of the 'basic four': iron, thiamin, niacin, and riboflavin (with optional calcium). In subsequent years, the standard of identity concept was expanded to include some other enriched foods. As of 1998, the Food and Drug Administration mandated that folate be added to the standards of identity for enriched breads, flours, corn meals, pastas, rice, and other grain products.

Micronutrient Fortification in the UK

In the UK, fortification of foods is subject to the Food Safety Act 1990. Fortification of certain micronutrients to margarine and most types of flour is mandatory. Calcium, iron, thiamin, and niacin are required to be added to both white and brown flours, but not to wholemeal flours. The level of required fortification is shown in **Table 2**.

Margarine is required to be fortified with vitamin A and D to levels comparable with or exceeding those found in butter. Additional mandatory fortification requirements determine the nutrient content of infant formulas and follow-on formulas, weaning foods, and foods intended to be used in energy-restricted diets. In the UK, voluntary fortification is allowed for certain products, such as breakfast cereal, soft drinks, and milks. In most cases the level of fortification per serving is between 15% and 33% of the relevant RDA.

Table 2 Nutrients required to be added to white and brown flours in the United Kingdom

Nutrient	*Amount of nutrient (mg) per 100 g flour*
Calcium	235–390
Iron	Not less than 1.65
Thiamin	Not less than 0.24
Niacin	Not less than 1.6

Bioavailability of Food Fortificants

The bioavailability of nutrients added to food can be determined by a number of factors. For example, in some cases the reactivity of added nutrients can cause untoward reactions that adversely affect the organoleptic properties of food. In these cases, there must be a trade-off of some kind and it may be necessary to intentionally select a somewhat less bioavailable form of a nutrient to provide an acceptable consumer product or to provide an acceptable shelf life to the product under given field conditions. Moreover, once added to a food, the bioavailability of a fortificant can be altered by various food manufacturing processes, such as those that demand high heat and pressure. Normal home food preparation techniques can also affect nutrient bioavailability. In addition, plant breeding and horticultural practices can contribute to the development and use of superior plant varieties supplying additional or more bioavailable micronutrients. For example, genetic engineering of plants has led to the development of rice and other grain products that have lower phytate content and higher mineral bioavailability. The development of 'Golden Rice,' which is rich in β-carotene, a dietary precursor of vitamin A, represents a well-known example of genetic plant engineering to enhance nutrient intakes. There is increasing interest in genetic manipulation of plant stocks to achieve higher content of potentially healthful phytonutrients, such as lycopene and lutein. Internationally, the traditional focus of fortification has been directed at the 'Big 3' – deficiencies of vitamin A, iodine, and iron – due to the widespread prevalence of deficiencies of these particular micronutrients and well-known adverse health effects of these nutrient deficiencies.

Bioavailability of Fortified Iron

A list of iron sources that are generally recognized as safe (GRAS) by the US Food and Drug Administration (FDA) is given in **Table 3**. However, as shown in **Table 4**, the bioavailability of different iron sources varies widely. Moreover, even within a given source of iron, such as the elemental iron powders commonly used to fortify various ready-to-eat breakfast cereals and other products, a significant disparity (5–148%) in relative bioavailability (compared to ferrous sulfate) can be observed. To some extent, these differing bioavailability estimates reflect the influence of the characteristics of the fortified product in terms of its contribution of various enhancers or inhibitors (**Table 1**) on iron bioavailability. In addition, other factors also can affect the bioavailability of elemental iron powders, such as the particle size of the fortificant

Table 3 Iron and zinc compounds listed as generally recognized as safe by the US Food and Drug Administration

Iron compounds	*Zinc compounds*
Elemental iron	Zinc sulfate
Ferrous ascorbate	Zinc chloride
Ferrous carbonate	Zinc gluconate
Ferrous citrate	Zinc oxide
Ferrous fumarate	Zinc stearate
Ferrous gluconate	
Ferrous lactate	
Ferrous sulfate	
Ferric ammonium citrate	
Ferric chloride	
Ferric citrate	
Ferric pyrophosphate	
Ferric sulfate	

compound – a finer particle size is associated with greater iron bioavailability.

There is very limited information available concerning the bioavailability of calcium or other nutrients added as a fortificant to various products. One study in elderly women found that calcium citrate malate used to fortify orange juice had equivalent bioavailability to calcium from milk or calcium from a calcium carbonate supplement. In a study with adult subjects, the bioavailability of a single 25 000 IU dose of vitamin D_2 was assessed from whole milk, skim milk, and vitamin D-fortified oil given with toast. No difference in peak serum vitamin D_2 was found following these three treatments, suggesting that the fat content of whole milk does not influence vitamin D bioavailability. It has also been shown that consumption of vitamin D-fortified orange juice (1000 IU/240 ml) for 12 weeks significantly increased serum 25-hydroxyvitamin D concentrations.

Fat content of a meal may have an important effect on carotenoid bioavailability. The absorption of carotenoids (α-carotene, β-carotene, and lycopene) from salad vegetables was found to be undetectable if a fat-free salad dressing was used, but substantially greater absorption occurred with a full-fat salad dressing. The amount of fat needed to promote optimal absorption of vitamin E and carotenoids may be rather limited. No difference in absorption of vitamin E and α- or β-carotene was observed when supplements were administered with either 3 g or 36 g of dietary fat. In contrast, lutein ester absorption was more than twice as great when consumed with the higher fat level.

Bioavailability of food sources of folate are usually only about 50% of synthetic folic acid. This systematic difference may be due to the occurrence of polyglutamyl folic acid in foods that reduce folate absorption.

Table 4 Average relative bioavailability in humans of various iron sources used as iron fortification compounds

Average relative bioavailability	*Iron compound*	*Approximate iron content (%)*
>90% group		
106[a]	Ferrous lactate	19
100	Ferrous sulfate $7H_2O$	20
100	Ferrous fumarate	33
92	Ferrous succinate	35
>60–<90% group		
89	Ferrous gluconate	12
75	Electrolytic elemental Fe powders	97
74	Ferric saccharate	10
74	Ferrous citrate	24
62	Ferrous tartrate	22
Variable (%)		
21–74	Ferric pyrophosphate	25
25–32	Ferric orthophosphate	28
13–148	H-reduced elemental Fe powder	97
5–20	Carbonyl elemental Fe powder	99

[a]Relative to absorption of iron from ferrous sulfate = 100%
Adapted from Lynch S (2002) Food iron absorption and its importance for the design of food fortification strategies. *Nutrition Reviews* **60**: S3–S15.

See also: **Calcium**. **Carotenoids**: Chemistry, Sources and Physiology. **Cobalamins**. **Copper**. **Food Fortification**: Developed Countries; Developing Countries. **Iodine**: Physiology, Dietary Sources and Requirements. **Microbiota of the Intestine**: Prebiotics. **Osteoporosis**. **Selenium**. **Vitamin A**: Biochemistry and Physiological Role. **Zinc**: Physiology.

Further Reading

Backstrand J (2002) The history and future of food fortification in the United States: a public health perspective. *Nutrition Reviews* **60**: 15–26.

Bouis HE (2003) Micronutrient fortification of plants through plant breeding: can it improve nutrition in man at low cost? *Proceedings of the Nutrition Society* **62**(2): 403–411.

Calvo MS and Whiting SJ (2003) Prevalence of vitamin D insufficiency in Canada and the United States: importance to health status and efficacy of current food fortification and dietary supplement use. *Nutrition Reviews* **61**(3): 107–113.

Chavasit V and Nopburabutr P (2003) Combating iodine and iron deficiencies through the double fortification of fish sauce, mixed fish sauce, and salt brine. *Food and Nutrition Bulletin* **24**(2): 200–207.

Dary O and Mora JO (2002) Food fortification to reduce vitamin A deficiency: International Vitamin A Consultative Group recommendations. *Journal of Nutrition* **132**(supplement 9): 2927S–2933S.

Delange FM (2003) Control of iodine deficiency in Western and Central Europe. *Central Europe Journal of Public Health* **11**(3): 120–123.

Fairweather-Tait SJ and Teucher B (2002) Iron and calcium bioavailability of fortified foods and dietary supplements. *Nutrition Review* **60**(11): 360–367.

Johnson-Down L, L'Abbe MR, Lee NS, and Gray-Donald K (2003) Appropriate calcium fortification of the food supply presents a challenge. *Journal of Nutrition* **133**(7): 2232–2238.

Lutter CK and Dewey KG (2003) Proposed nutrient composition for fortified complementary foods. *Journal of Nutrition* **133**(9): 3011S–3020S.

Lynch S (2002) Food iron absorption and its importance for the design of food fortification strategies. *Nutrition Reviews* **60**: S3–S6.

Meltzer HM, Aro A, Andersen NL, Koch B, and Alexander J (2003) Risk analysis applied to food fortification. *Public Health and Nutrition* **6**(3): 281–291.

Penniston KL and Tanumihardjo SA (2003) Vitamin A in dietary supplements and fortified food: too much of a good thing? *Journal of the American Dietetic Association* **103**(9): 1185–1187.

Quinlivan EP and Gregory JK 3rd (2003) Effect of food fortification on folic acid intake in the United States. *American Journal of Clinical Nutrition* **77**(1): 8–9.

Rosado JL (2003) Zinc and copper: proposed fortification levels and recommended zinc compounds. *Journal of Nutrition* **133**(9): 2985S–2989S.

BIOTIN

D M Mock, University of Arkansas for Medical Sciences, Little Rock, AR, USA

Biotin is a water-soluble vitamin that is generally classified in the B complex group. Biotin was discovered in nutritional experiments that demonstrated a factor in many foodstuffs capable of curing the scaly dermatitis, hair loss, and neurologic signs induced in rats fed dried egg white. Avidin, a glycoprotein found in egg white, binds biotin very specifically and tightly. From an evolutionary standpoint, avidin probably serves as a bacteriostat in egg white; consistent with this hypothesis is the observation that avidin is resistant to a broad range of bacterial proteases in both the free and biotin-bound form. Because avidin is also resistant to pancreatic proteases, dietary avidin binds to dietary biotin (and probably any biotin from intestinal microbes) and prevents absorption, carrying the biotin through the gastrointestinal tract. Biotin is synthesized by many intestinal microbes; however, the contribution of microbial biotin to absorbed biotin, if any, remains unknown. Cooking denatures avidin, rendering this protein susceptible to digestion and unable to interfere with absorption of biotin.

Absorption and Transport

Digestion of Protein-Bound Biotin

The content of free biotin and protein-bound biotin in foods is variable, but the majority of biotin in meats and cereals appears to be protein-bound via an amide bond between biotin and lysine. Neither the mechanisms of intestinal hydrolysis of protein-bound biotin nor the determinants of bioavailability have been clearly delineated. Because this bond is not hydrolyzed by cellular proteases, release is likely mediated by a specific biotin—amide hydrolase (biotinidase, EC 3.5.1.12). Biotinidase mRNA is present in pancreas and, in lesser amounts, in intestinal mucosa. Biotinidase is also present in many other tissues, including heart, brain, liver, lung, skeletal muscle, kidney, plasma, and placenta. Biotinidase also likely plays a critical role in intracellular recycling of biotin by releasing biotin from intracellular proteins such as carboxylases during protein turnover.

Intestinal Absorption and Transport into Somatic Cells

At physiologic pH, the carboxylate group of biotin is negatively charged. Thus, biotin is at least modestly water-soluble and requires a transporter to cross cell membranes such as enterocytes for intestinal absorption, somatic cells for utilization, and renal tubule cells for reclamation from the glomerular filtrate. In intact intestinal preparations such as loops and everted gut sacks, biotin transport exhibits two components. One component is saturable at a k_m of approximately 10 μM biotin; the other is not saturable even at very large concentrations of biotin. This observation is consistent with passive diffusion. Absorption of biocytin, the biotinyl-lysine product of intraluminal protein digestion, is inefficient relative to biotin, suggesting that biotinidase releases

biotin from dietary protein. The transporter is present in the intestinal brush border membrane. Transport is highly structurally specific, temperature dependent, Na^+ coupled, and electroneutral. In the presence of a sodium ion gradient, biotin transport occurs against a concentration gradient.

In rats, biotin transport is upregulated with maturation and by biotin deficiency. Although carrier-mediated transport of biotin is most active in the proximal small bowel of the rat, the absorption of biotin from the proximal colon is still significant, supporting the potential nutritional significance of biotin synthesized and released by enteric flora. Clinical studies have provided evidence that biotin is absorbed from the human colon, but studies in swine indicate that absorption of biotin from the hindgut is much less efficient than from the upper intestine; furthermore, biotin synthesized by enteric flora is probably not present at a location or in a form in which bacterial biotin contributes importantly to absorbed biotin. Exit of biotin from the enterocyte (i.e., transport across the basolateral membrane) is also carrier mediated. However, basolateral transport is independent of Na^+, electrogenic, and does not accumulate biotin against a concentration gradient.

Based on a study in which biotin was administered orally in pharmacologic amounts, the bioavailability of biotin is approximately 100%. Thus, the pharmacologic doses of biotin given to treat biotin-dependent inborn errors of metabolism are likely to be well absorbed. Moreover, the finding of high bioavailability of biotin at pharmacologic doses provides at least some basis for predicting that bioavailability will also be high at the physiologic doses at which the biotin transporter mediates uptake.

Studies of a variety of hepatic cell lines indicate that uptake of free biotin is similar to intestinal uptake; transport is mediated by a specialized carrier system that is Na^+ dependent, electroneutral, and structurally specific for a free carboxyl group. At large concentrations, transport is mediated by diffusion. Metabolic trapping (e.g., biotin bound covalently to intracellular proteins) is also important. After entering the hepatocyte, biotin diffuses into the mitochondria via a pH-dependent process.

Two biotin transporters have been described: a multivitamin transporter present in many tissues and a biotin transporter identified in human lymphocytes. In 1997, Prasad and coworkers discovered a Na^+-coupled, saturable, structurally specific transporter present in human placental choriocarcinoma cells that can transport pantothenic acid, lipoic acid, and biotin. This sodium-dependent multivitamin transporter has been named SMVT and is widely expressed in human tissues. Studies by Said and coworkers using RNA interference specific for SMVT provide strong evidence that biotin uptake by Caco-2 and HepG2 cells occurs via SMVT; thus, intestinal absorption and hepatic uptake are likely mediated by SMVT. The biotin transporter identified in lymphocytes is also Na^+ coupled, saturable, and structurally specific. Studies by Zempleni and coworkers provide evidence in favor of monocarboxylate transporter-1 as the lymphocyte biotin transporter.

A child with biotin dependence due to a defect in the lymphocyte biotin transporter has been reported. The SMVT gene sequence was normal. The investigators speculate that lymphocyte biotin transporter is expressed in other tissues and mediates some critical aspect of biotin homeostasis.

Ozand and collaborators described several patients in Saudi Arabia with biotin-responsive basal ganglia disease. Symptoms include confusion, lethargy, vomiting, seizures, dystonia, dysarthria, dysphagia, seventh nerve paralysis, quadriparesis, ataxia, hypertension, chorea, and coma. A defect in the biotin transporter system across the blood–brain barrier was postulated. Additional work by Gusella and coworkers has suggested that SLC19A3 may be responsible for the reported defect.

The relationship of these putative biotin transporters to each other and their relative roles in intestinal absorption, transport into various organs, and renal reclamation remain to be elucidated.

Transport of Biotin from the Intestine to Peripheral Tissues

Biotin concentrations in plasma are small relative those of other water-soluble vitamins. Most biotin in plasma is free, dissolved in the aqueous phase of plasma. However, small amounts are reversibly bound and covalently bound to plasma protein (approximately 7 and 12%, respectively); binding to human serum albumin likely accounts for the reversible binding. Biotinidase has been proposed as a biotin binding protein or biotin carrier protein for the transport into cells. A biotin binding plasma glycoprotein has been observed in pregnant rats. Although the importance of protein binding in the transport of biotin from the intestine to the peripheral tissues is not clear, the immunoneutralization of this protein led to decreased transport of biotin to the fetus and early death of the embryo.

Transport of Biotin into the Central Nervous System

Biotin is transported across the blood–brain barrier. The transporter is saturable and structurally specific

for the free carboxylate group on the valeric acid side chain. Transport into the neuron also appears to involve a specific transport system with subsequent trapping of biotin by covalent binding to brain proteins, presumably carboxylases.

Placental Transport of Biotin

Biotin concentrations are 3- to 17-fold greater in plasma from human fetuses compared to those in their mothers in the second trimester, consistent with active placental transport. The microvillus membrane of the placenta contains a saturable transport system for biotin that is Na^+ dependent and actively accumulates biotin within the placenta, consistent with SMVT.

Transport of Biotin into Human Milk

More than 95% of the biotin in human milk is free in the skim fraction. The concentration of biotin varies substantially in some women and exceeds the concentration in serum by one or two orders of magnitude, suggesting that there is a system for transport into milk. Metabolites account for more than half of the total biotin plus metabolites in early and transitional human milk. With postpartum maturation, the biotin concentration increases, but inactive metabolites still account for approximately one-third of the total biotin plus metabolites at 5 weeks postpartum. Studies have not detected a soluble biotin binding protein.

Metabolism and Urinary Excretion of Biotin and Metabolites

Biotin is a bicyclic compound (**Figure 1**). One of the rings contains an ureido group (—N—CO—N—). The tetrahydrothiophene ring contains sulfur and has a valeric acid side chain. A significant proportion of biotin undergoes catabolism before excretion (**Figure 1**). Two principal pathways of biotin catabolism have been identified in mammals. In the first pathway, the valeric acid side chain of biotin is degraded by β-oxidation. β-Oxidation of biotin leads to the formation of bisnorbiotin, tetranorbiotin, and related intermediates that are known to result from β-oxidation of fatty acids. The cellular site of this β-oxidation of biotin is uncertain. Spontaneous (nonenzymatic) decarboxylation of the unstable β-keto acids (β-keto-biotin and β-keto-bisnorbiotin) leads to formation of bisnorbiotin methylketone and tetranorbiotin methylketone; these catabolites appear in urine.

In the second pathway, the sulfur in the thiophane ring of biotin is oxidized, leading to the formation of biotin-L-sulfoxide, biotin-D-sulfoxide, and biotin sulfone. Sulfur oxidation may be catalyzed by a NADPH-dependent process in the smooth endoplasmic reticulum. Combined oxidation of the ring sulfur and β-oxidation of the side chain lead to metabolites such as bisnorbiotin sulfone. In mammals, degradation of the biotin ring to release carbon dioxide and urea is quantitatively minor. Biotin metabolism is accelerated in some individuals by anticonvulsants and during pregnancy, thereby increasing in urine the ratio of biotin metabolites to biotin.

Animal studies and studies using brush border membrane vesicles from human kidney cortex indicate that biotin is reclaimed from the glomerular filtrate against a concentration gradient by a saturable, Na^+-dependent, structurally specific system, but biocytin does not inhibit tubular reabsorption of biotin. Subsequent egress of biotin from the tubular cells occurs via a basolateral membrane transport system that is not dependent on Na^+. Studies of patients with biotinidase deficiency suggest that there may be a role for biotinidase in the renal handling of biotin.

On a molar basis, biotin accounts for approximately half of the total avidin-binding substances in human serum and urine (**Table 1**). Biocytin, bisnorbiotin, bisnorbiotin methylketone, biotin-D,L-sulfoxide, and biotin sulfone account for most of the balance.

Biliary Excretion of Biotin and Metabolites

Biliary excretion of biotin and metabolites is quantitatively negligible based on animal studies. When [^{14}C]biotin was injected intravenously into rats, biotin, bisnorbiotin, biotin-D,L-sulfoxide, and bisnorbiotin methylketone accounted for less than 2% of the administered ^{14}C, but urinary excretion accounted for 60%. Although the concentrations of biotin, bisnorbiotin, and biotin-D,L-sulfoxide were approximately 10-fold greater in bile than in serum of pigs, the bile-to-serum ratios of biotin and metabolites were more than 10-fold less than those of bilirubin, which is actively excreted in bile.

Metabolic Functions

In mammals, biotin serves as an essential cofactor for five carboxylases, each of which catalyses a critical step in intermediary metabolism. All five of the mammalian carboxylases catalyze the incorporation of bicarbonate as a carboxyl group into a substrate and employ a similar catalytic mechanism.

Biotin is attached to the apocarboxylase by a condensation reaction catalyzed by holocarboxylase synthetase (**Figure 1**). An amide bond is formed

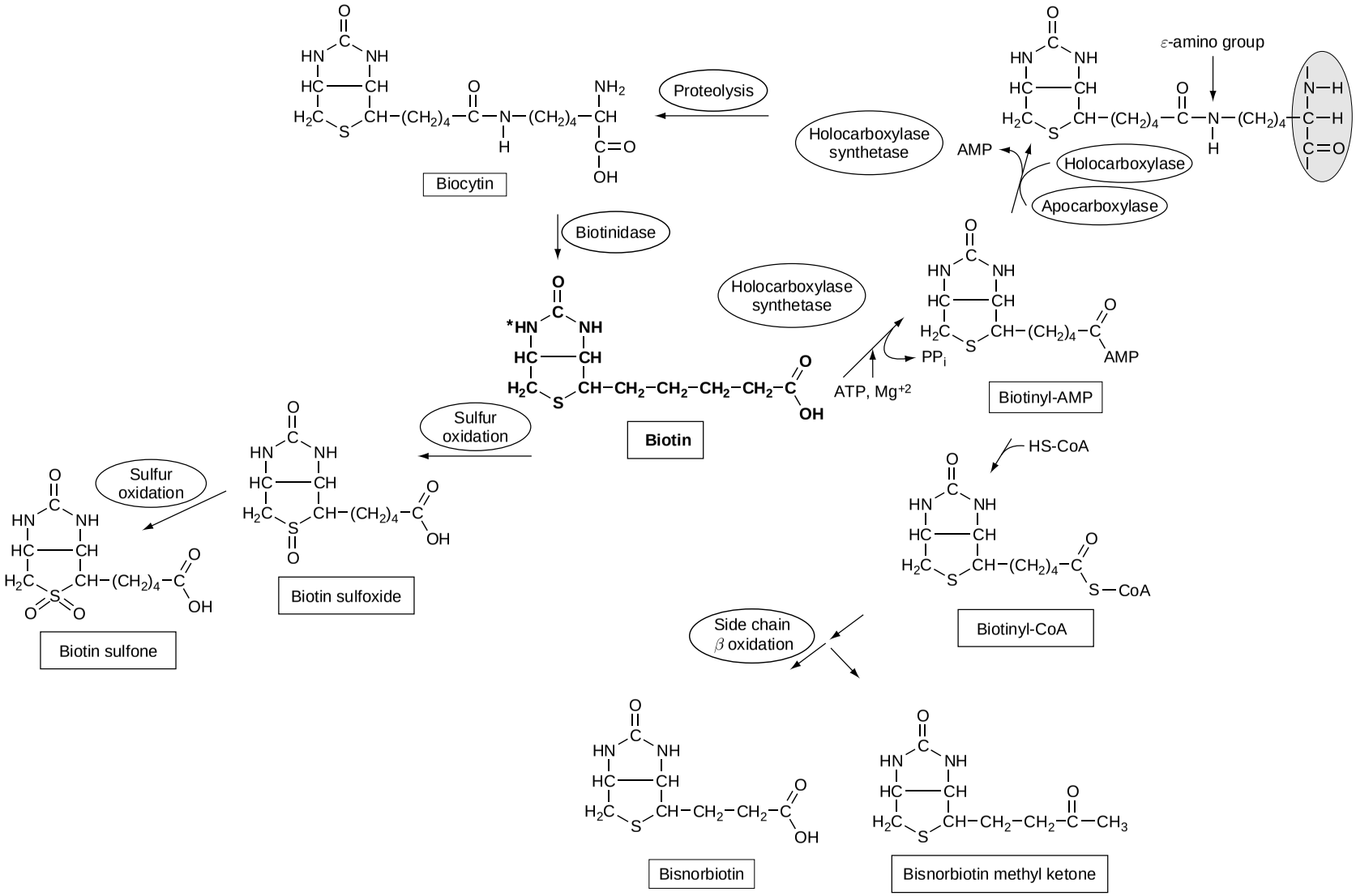


Figure 1 Biotin metabolism and degradation. Ovals denote enzymes or enzyme systems; rectangles denote biotin, intermediates, and metabolites. AMP, adenosine monophosphae; ATP, adenosine triphosphate; CoA, coenzyme A; Pp_i, pyrophosphate; *, site of attachment of carboxyl moiety.

Table 1 Normal range for biotin and metabolites in human serum and urine[a]

Compound	*Serum (pmol/l)*	*Urine (nmol/24 h)*
Biotin	133–329	18–127
Bisnorbiotin	21–563	6–39
Biotin-D,L-sulfoxide	0–120	5–19
Bisnorbiotin methylketone	0–120	2–13
Biotin sulfone	ND	1–8
Biocytin	0–26	1–13
Total biotinyl compounds	294–1021[b]	46–128

[a]Normal ranges are reported ($n = 15$ for serum; $n = 16$ for urine, except biocytin, $n = 10$).
[b]Including unidentified biotin metabolites.
ND, not determined.

between the carboxyl group of the valeric acid side chain of biotin and the ε-amino group of a specific lysyl residue in the apocarboxylase; these regions contain sequences of amino acids that are highly conserved for the individual carboxylases both within and between species.

In the carboxylase reaction, the carboxyl moiety is first attached to biotin at the ureido nitrogen opposite the side chain; then the carboxyl group is transferred to the substrate. The reaction is driven by the hydrolysis of ATP to ADP and inorganic phosphate. Subsequent reactions in the pathways of the mammalian carboxylases release carbon dioxide from the product of the carboxylase reaction. Thus, these reaction sequences rearrange the substrates into more useful intermediates but do not violate the classic observation that mammalian metabolism does not result in the net fixation of carbon dioxide.

Regulation of intracellular mammalian carboxylase activity by biotin remains to be elucidated. However, the interaction of biotin synthesis and production of holoacetyl-CoA carboxylase in *Escherichia coli* has been extensively studied. In the bacterial system, the apocarboxylase protein and biotin (as the intermediate biotinyl-AMP) act together to control the rate of biotin synthesis by direct interaction with promoter regions of the biotin operon, which in turn controls a cluster of genes that encode enzymes that catalyze the synthesis of biotin.

The five biotin-dependent mammalian carboxylases are acetyl-CoA carboxylase isoforms I and II (also known as α-ACC (EC 6.4.1.2) and β-ACC (EC 6.4.1.2)), pyruvate carboxylase (EC 6.4.1.1), methylcrotonyl-CoA carboxylase (EC 6.4.1.4), and propionyl-CoA carboxylase (EC 6.4.1.3). ACC catalyzes the incorporation of bicarbonate into acetyl-CoA to form malonyl-CoA (**Figure 2**). There are two isoforms of ACC. Isoform I is located in the cytosol and produces malonyl-CoA, which is rate limiting in fatty acid synthesis (elongation). Isoform II is located on the outer mitochondrial membrane and controls fatty acid oxidation in mitochondria through the inhibitory effect of malonyl-CoA on fatty acid transport into mitochondria. An inactive mitochondrial form of ACC may serve as storage for biotin.

The three remaining carboxylases are mitochondrial. Pyruvate carboxylase (PC) catalyzes the incorporation of bicarbonate into pyruvate to form oxaloacetate, an intermediate in the Krebs tricarboxylic acid cycle (**Figure 2**). Thus, PC catalyzes an anaplerotic reaction. In gluconeogenic tissues (i.e., liver and kidney), the oxaloacetate can be converted to glucose. Deficiency of PC is probably the cause of the lactic acidemia, central nervous system lactic acidosis, and abnormalities in glucose regulation observed in biotin deficiency and biotinidase deficiency. β-Methylcrotonyl-CoA carboxylase (MCC) catalyzes an essential step in the degradation of the branched-chain amino acid leucine (**Figure 2**). Deficient activity of MCC leads to metabolism of

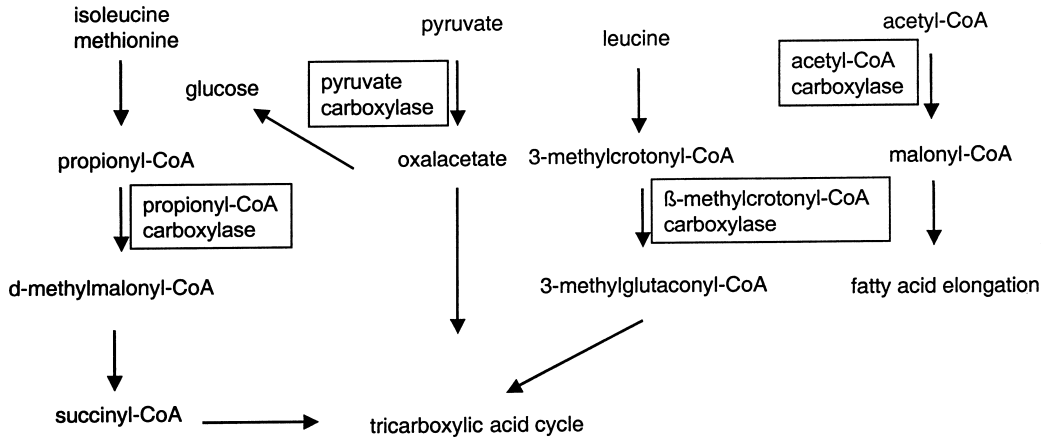


Figure 2 Interrelationship of pathways catalyzed by biotin-dependent enzymes (shown in boxes).

3-methylcrotonyl-CoA to 3-hydroxyisovaleric acid and 3-methylcrotonylglycine by an alternate pathway. Thus, increased urinary excretion of these abnormal metabolites reflects deficient activity of MCC.

Propionyl-CoA carboxylase (PCC) catalyzes the incorporation of bicarbonate into propionyl-CoA to form methylmalonyl-CoA; methylmalonyl-CoA undergoes isomerization to succinyl-CoA and enters the tricarboxylic acid cycle (**Figure 2**). In a manner analogous to MCC deficiency, deficiency of PCC leads to increased urinary excretion of 3-hydroxypropionic acid and 3-methylcitric acid.

In the normal turnover of cellular proteins, holocarboxylases are degraded to biocytin or biotin linked to an oligopeptide containing at most a few amino acid residues (**Figure 1**). Biotinidase releases biotin for recycling. Genetic deficiencies of holocarboxylase synthetase and biotinidase cause the two types of multiple carboxylase deficiency that were previously designated the neonatal and juvenile forms.

A Potential Role for Biotin in Gene Expression

In 1995, Hymes and Wolf discovered that biotinidase can act as a biotinyl-transferase; biocytin serves as the source of biotin, and histones are specifically biotinylated. Approximately 25% of total cellular biotinidase activity occurs in the nucleus. Zempleni and coworkers demonstrated that the abundance of biotinylated histones varies with the cell cycle, that biotinylated histones are increased approximately twofold compared to quiescent lymphocytes, and that histones are debiotinylated enzymatically in a process that is at least partially catalyzed by biotinidase. These observations suggest that biotin plays a role in regulating DNA transcription and regulation.

Although the mechanisms remain to be elucidated, biotin status has been shown to clearly effect gene expression. Cell culture studies suggest that cell proliferation generates an increased demand for biotin, perhaps mediated by increased synthesis of biotin-dependent carboxylases. Solozano-Vargas and coworkers reported that biotin deficiency reduces messenger RNA levels of holocarboxylase synthetase, α-ACC, and PCC and postulated that a cyclic GMP-dependent signaling pathway is involved in the pathogenesis.

Studies have been conducted on diabetic humans and rats that support an effect of biotin status on carbohydrate metabolism. Genes studied include glucokinase, phosphoenolpyruvate carboxykinase (PEPCK), and expression of the asialoglycoprotein receptor on the surface of hepatocytes. The effect of biotin status on PEPCK expression was particularly striking when diabetic rats were compared to nondiabetic rats. However, most studies have been performed on rats in which metabolic pathways have been perturbed prior to administration of biotin. Thus, the role of biotin in regulation of these genes during normal biotin status remains to be elucidated.

Hyperammonemia is a finding in biotin deficiency. Maeda and colleges have reported that ornithine transcarbamoylase (an enzyme in the urea cycle) is significantly reduced in biotin-deficient rats.

Assessment of Biotin Status

Measurement of Biotin

For measuring biotin at physiological concentrations (i.e., $100\,\text{pmol}\,\text{l}^{-1}$ to $100\,\text{nmol}\,\text{l}^{-1}$), a variety of assays have been proposed, and a limited number have been used to study biotin nutritional status. Most published studies of biotin nutritional status have used one of two basic types of biotin assays: bioassays (most studies) or avidin-binding assays (several recent studies).

Bioassays are generally sensitive enough to measure biotin in blood and urine. However, the bacterial bioassays (and perhaps the eukaryotic bioassays as well) suffer interference from unrelated substances and variable growth response to biotin analogues. Bioassays give conflicting results if biotin is bound to protein.

Avidin-binding assays generally measure the ability of biotin (i) to compete with radiolabeled biotin for binding to avidin (isotope dilution assays), (ii) to bind to avidin coupled to a reporter and thus prevent the avidin from binding to a biotin linked to solid phase, or (iii) to prevent inhibition of a biotinylated enzyme by avidin. Avidin-binding assays generally detect all avidin-binding substances, although the relative detectabilities of biotin and analogues vary between analogues and between assays, depending on how the assay is conducted. Chromatographic separation of biotin analogues with subsequent avidin-binding assay of the chromatographic fractions appears to be both sensitive and chemically specific.

Laboratory Findings of Biotin Deficiency

Although various indices have been used to assess biotin status, these have been validated in humans only twice during progressive biotin deficiency. In both studies, marginal biotin deficiency was induced in normal adults by feeding egg white. The urinary excretion of biotin declined dramatically with time on the egg-white diet, reaching frankly abnormal values in 17 of 21 subjects by day 20 of egg-white

feeding. Bisnorbiotin excretion declined in parallel, providing evidence for regulated catabolism of biotin. In most subjects, urinary excretion of 3-hydroxyisovaleric acid increased steadily. By day 14 of egg-white feeding, 3-hydroxyisovaleric acid excretion was abnormally increased in 18 of 21 subjects, providing evidence that biotin depletion decreases the activity of MCC and alters leucine metabolism early in progressive biotin deficiency. Based on a study of only 5 subjects, 3-hydroxyisovaleric acid excretion in response to a leucine challenge may be even more sensitive than 3-hydroxyisovaleric acid excretion. Urinary excretions of 3-methylcrotonylglycine, 3-hydroxypropionic acid, and 3-methylcitric acid are not sensitive indicators of biotin deficiency compared to 3-hydroxyisovaleric acid excretion.

In a single study, plasma concentrations of free biotin decreased to abnormal values in only half of the subjects. This observation provides confirmation of the impression that blood biotin concentration is not an early or sensitive indicator of marginal biotin deficiency.

Lymphocyte PCC activity is an early and sensitive indicator of marginal biotin deficiency. In 11 of 11 subjects, lymphocyte PCC activity decreased to abnormal values by day 28 of egg-white feeding and returned to normal in 8 of 11 within 3 weeks of resuming a general diet with or without biotin supplement.

Odd-chain fatty acid accumulation is also a marker of biotin deficiency. The accumulation of odd-chain fatty acid is thought to result from PCC deficiency (**Figure 3**); the accumulation of propionyl-CoA likely leads to the substitution of a propionyl-CoA moiety for acetyl-CoA in the ACC reaction and to the incorporation of a three- (rather than two-) carbon moiety during fatty acid elongation. However, in comparison to lymphocyte PCC activity and urinary excretion of 3-hydroxyisovaleric acid, odd-chain fatty acids accumulate in blood lipids more slowly during biotin deficiency and return to normal more gradually after biotin repletion.

Requirements and Allowances

Data providing an accurate estimate of the dietary and parenteral biotin requirements for infants, children, and adults are lacking. However, recommendations for biotin supplementation have been formulated for oral and parenteral intake for preterm infants, term infants, children, and adults (**Table 2**).

Table 2 Adequate intake of biotin

Life-stage group	*Adequate intake (μg/day)*
Infants (months)	
0–6	5
7–12	6
Children (years)	
1–3	8
4–8	12
Males and females (years)	
9–13	20
14–18	25
≥19	30
Pregnancy	30
Lactation	35

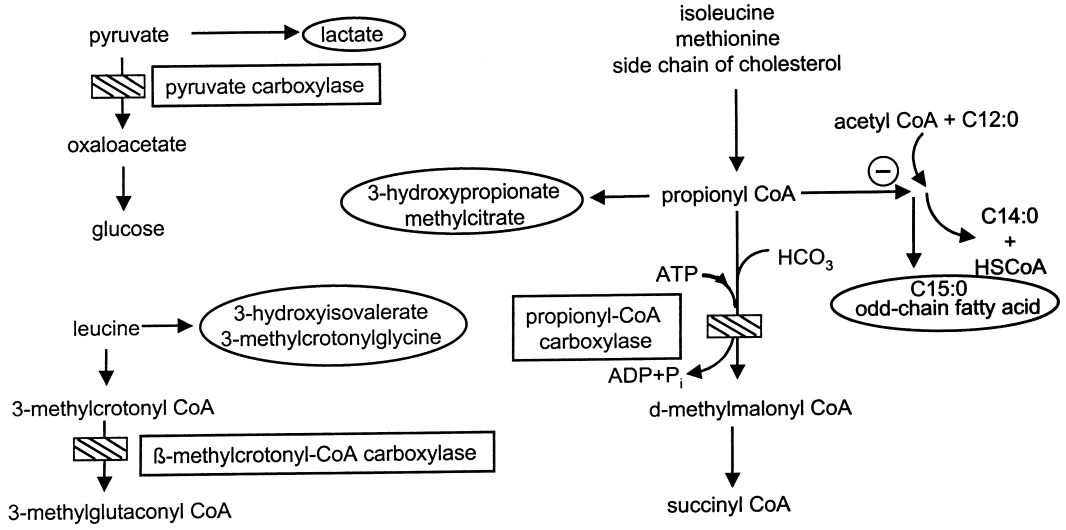


Figure 3 Organic acids and odd-chain fatty acids accumulate because biotin deficiency causes reduced activity of biotin-dependent enzymes. Hatched bars denote metabolic blocks at deficient carboxylases; ovals denote accumulation of products from alternative pathways.

Dietary Sources, Deficiency, and High Intakes

Dietary Sources

There is no published evidence that biotin can be synthesized by mammals; thus, the higher animals must derive biotin from other sources. The ultimate source of biotin appears to be de novo synthesis by bacteria, primitive eukaryotic organisms such as yeast, moulds, and algae, and some plant species.

The great majority of measurements of biotin content of foods have used bioassays. Recent publications provide evidence that the values are likely to contain substantial errors. However, some worthwhile generalizations can be made. Biotin is widely distributed in natural foodstuffs, but the absolute content of even the richest sources is low compared to the content of most other water-soluble vitamins. Foods relatively rich in biotin are listed in **Table 3**. The average daily dietary biotin intake has been estimated to be approximately 35–70 μg.

Circumstances Leading to Deficiency

The fact that normal humans have a requirement for biotin has been clearly documented in two situations: prolonged consumption of raw egg white and parenteral nutrition without biotin supplementation in patients with short bowel syndrome and other causes of malabsorption. Based on lymphocyte carboxylase activities and plasma biotin levels, some children with severe protein–energy malnutrition are biotin deficient. Investigators have speculated that the effects of biotin deficiency may be responsible for part of the clinical syndrome of protein–energy malnutrition.

Biotin deficiency has also been reported or inferred in several other clinical circumstances, including long-term anticonvulsant therapy, Leiner's disease, sudden infant death syndrome, renal dialysis, gastrointestinal diseases, and alcoholism. Studies of biotin status during pregnancy and of biotin supplementation during pregnancy provide evidence that a marginal degree of biotin deficiency develops in at least one-third of women during normal pregnancy. Although the degree of biotin deficiency is not severe enough to produce overt manifestations of biotin deficiency, the deficiency is sufficiently severe to produce metabolic derangements. A similar marginal degree of biotin deficiency causes high rates of fetal malformations in some mammals. Moreover, data from a multivitamin supplementation study provide significant albeit indirect evidence that the marginal degree of biotin deficiency that occurs spontaneously in normal human gestation is teratogenic.

Clinical Findings of Frank Deficiency

The clinical findings of frank biotin deficiency in adults, older children, and infants are similar. Typically, the findings appear gradually after weeks to several years of egg-white feeding or parenteral nutrition. Thinning of hair and progression to loss of all hair, including eyebrows and lashes, has been reported. A scaly (seborrheic), red (eczematous) skin rash was present in the majority; in several, the rash was distributed around the eyes, nose, mouth, and perineal orifices. These cutaneous manifestations, in conjunction with an unusual distribution of facial fat, have been termed 'biotin deficiency facies.' Depression, lethargy, hallucinations, and paraesthesia of the extremities were prominent neurologic symptoms in the majority of adults. The most striking neurologic findings in infants were hypotonia, lethargy, and developmental delay.

The clinical response to administration of biotin has been dramatic in all well-documented cases of

Table 3 Foods relatively rich in biotin

Food	*ng biotin/g food*	*serving size (g)*	*μg biotin/serving*
Chicken liver, cooked	1872.00	74	138.00
Beef liver, cooked	416.00	74	30.80
Egg, whole, cooked	214.00	47	10.00
Peanuts, roasted, salted	175.00	28	4.91
Egg, yolk, cooked	272.00	15	4.08
Salmon, pink, canned in water	59.00	63	3.69
Pork chop, cooked	45.00	80	3.57
Mushrooms, canned	21.60	120	2.59
Sunflower seeds, roasted, salted	78.00	31	2.42
Chili	5.20	441	2.29
Hot dog, chicken and pork, cooked	37.00	56	2.06
Egg, white, cooked	58.00	35	2.02
Banana pudding	10.20	170	1.73
Strawberries, fresh	15.00	111	1.67

biotin deficiency. Healing of the rash was striking within a few weeks, and growth of healthy hair was generally present by 1 or 2 months. Hypotonia, lethargy, and depression generally resolved within 1 or 2 weeks, followed by accelerated mental and motor development in infants. Pharmacological doses of biotin (e.g., 1–10 mg) have been used to treat most patients.

High Intakes

Daily doses up to 200 mg orally and up to 20 mg intravenously have been given to treat biotin-responsive inborn errors of metabolism and acquired biotin deficiency. Toxicity has not been reported.

See also: **Brain and Nervous System**. **Breast Feeding**. **Cofactors**: Organic. **Meat, Poultry and Meat Products**. **Microbiota of the Intestine**: Prebiotics. **Pregnancy**: Role of Placenta in Nutrient Transfer.

Further Reading

Bender DA (1999) Optimum nutrition: Thiamin, biotin, and pantothenate. *Proceedings of the Nutritional Society* **58**: 427–433.

Cronan JE Jr (2002) Interchangeable enzyme modules. Functional replacement of the essential linker of the biotinylated subunit of acetyl-CoA carboxylase with a linker from the lipoylated subunit of pyruvate dehydrogenase. *Journal of Biological Chemistry* **277**: 22520–22527.

Flume MZ (2001) Final report on the safety assessment of biotin. *International Journal of Toxicology* **20**(supplement 4): 1–12.

McCormick DB (2001) Bioorganic mechanisms important to coenzyme functions. In: Rucker RB, Suttie JW, McCormick DB, and Machlin LJ (eds.) *Handbook of Vitamins*, 3rd edn, pp. 199–212. New York: Marcel Dekker.

McMahon RJ (2002) Biotin in metabolism and molecular biology. *Annual Reviews in Nutrition* **22**: 221–239.

Mock DM (1992) Biotin in human milk: When, where, and in what form? In: Picciano MF and Lonnerdal B (eds.) *Mechanisms Regulating Lactation and Infant Nutrient Utilization*. New York: John Wiley.

National Research Council, Food and Nutrition Board, and Institute of Medicine (1998) Dietary reference intakes for thiamin, riboflavin, niacin, vitamin B-6, folate, vitamin B-12, pantothenic acid, biotin, and choline. In *Recommended Dietary Allowances*., pp. 374–389. Washington, DC. National Academy Press.

Pacheco-Alvarez D, Solorzano-Vargas RS, and Del Rio AL (2002) Biotin in metabolism and its relationship to human disease. *Archives of Medical Research* **33**: 439–447.

Solbiati J, Chapman-Smith A, and Cronan JE Jr (2002) Stabilization of the biotinoyl domain of *Escherichia coli* acetyl-CoA carboxylase by interactions between the attached biotin and the protruding "thumb" structure. *Journal of Biological Chemistry* **277**: 21604–21609.

Wolf B (2001) Disorders of biotin metabolism. In: Scriver CR, Beaudet AL, Sly WS, and Valle D (eds.) *The Metabolic and Molecular Basis of Inherited Disease*, vol. 3, pp. 3151–3177. New York: McGraw-Hill.

Zempleni J and Mock DM (2001) Biotin. In: Song WO and Beecher GR (eds.) *Modern Analytical Methodologies in Fat and Water-Soluble Vitamins*, pp. 459–466. Baltimore: John Wiley.

Blood Lipids/Fats *see* **Hyperlipidemia**: Overview. **Lipoproteins**

Blood Pressure *see* **Hypertension**: Etiology

BODY COMPOSITION

D Gallagher and S Chung, Columbia University, New York, NY, USA

Introduction

Historically, the measurement of the body and its components centered around cadaver analyses where specific tissues and organs were extracted from the body for inspection. The extraction of tissue samples from the living body was a step forward in allowing for the analyses of tissue morphology in a state more closely resembling the *in vivo* state. However, both cadaver and *in vitro* tissue analyses are subject to inaccuracies when extrapolations are being made to the living body. Nevertheless, much of our understanding of human body composition in both children and adults has roots in these approaches. During the twentieth century, significant advances were made in the development of *in vivo* methods of body composition analysis thanks to the disciplines of physics, engineering, and medicine. Methodologies with minimal or no risk to the participant have allowed for the assessment of body composition in growth and development, aging, and disease.

The physiological significance of knowing the composition of the body greatly depends on the question of interest. Common applications involving medical/clinical diagnoses include osteopenia/osteoporosis; muscle wasting; sarcopenia; lipodystrophy; altered states of hydration; malnutrition; and obesity. There are also metabolic consequences (e.g., insulin resistance) associated with high and low levels of body fat and where the fat is distributed. From a nutritional perspective, the interest in body composition has increased multifold with the global increase in the prevalence of obesity and its complications. This chapter will focus on our current state of body composition knowledge and how this knowledge was determined with the available most advanced methodologies.

Body Composition Determination

There is no single gold standard for body composition measurements *in vivo*. All methods incorporate assumptions that do not apply in all individuals and the more accurate models are derived by using a combination of measurements, thereby reducing the importance of each assumption. The most commonly used technique today with good reproducibility in children and adults is dual-energy X-ray absorptiometry.

Dual-Energy X-Ray Absorptiometry (DXA)

The DXA method evolved from earlier single and dual photon absorptiometry methods for evaluating bone mineral. DXA systems share in common an x-ray source that, after appropriate filtration, emits two photon energy peaks. The attenuation of the two energy peaks relative to each other depends on the elemental content of tissues through which the photons pass. Bone, fat, and lean soft tissues are relatively rich in calcium/phosphorus, carbon, and oxygen, respectively. DXA systems are designed to separate pixels, based on appropriate models and relative attenuation, into these three components. There are no known factors, including hydration effects that significantly influence the validity of DXA fat and bone mineral estimates. Excessive or reduced fluid volume would be interpreted as changes in lean soft tissue. The radiation exposure is minimal and can be used in children and adults of all ages. DXA measures in persons who fit within the DXA field-of-view have good reproducibility for total body and regional components.

Hydrodensitometry/Air Plethysmography

One of the oldest methods of measuring body composition, the determination of body volume by water displacement (Archimedes principle) allows for the estimation of fat-free mass (FFM) density (where an assumption is made that densities of fat and FFM are constant) from which percent body fat is calculated using a two-compartment body composition model. Today there are a number of additional methods for measuring body volume, including air displacement plethysmography. Limitations with this approach include the assumptions of stable densities of fat and FFM across the age range where this may not be true in older individuals and across race/ethnic groups where it is now known that the density of FFM in Blacks is higher.

Dilution Techniques

Since fat is relatively anhydrous, the body's water is found primarily in the body's FFM compartment where approximately 73% of a healthy non-obese adults FFM compartment is water. The body's water pool can be measured using tracers which after administration, dilute throughout the body.

Basic assumptions involved with tracer dilution for body composition determination include equal distribution though out the pool of interest and dilution is complete within a specific period of time without any loss. Examples of commonly used tracers include deuterium oxide for total body water and sodium bromide for extracellular water.

Whole-Body Counting

A small constant percentage of total body potassium (TBK) is radioactive (^{40}K) and emits a γ-ray. With appropriate shielding from background, this γ-ray can be counted using scintillation detectors. As the ratio of ^{40}K to ^{39}K is known and constant, ^{39}K or "total body potassium" can be estimated accordingly. All of the body's potassium is within the FFM compartment and the proportion of the body FFM compartment TBK/FFM ratio is relatively stable in the same subject over time and between different subjects. However, with increasing age or when comparing young versus elderly, the TBK to FFM ratio decreases.

Magnetic Resonance Imaging/Computed Tomography

The use of computed tomography (CT) has had limited application in body composition research due primarily to radiation exposure. Its use has primarily been limited to single slice acquisitions in the abdomen and mid-thigh whereby information on adipose tissue distribution and muscle cross-sectional area have been derived. The use of magnetic resonance imaging (MRI) has resulted in important advances in body composition phenotyping. MRI studies are safe and instruments are available in most hospital or related facilities. Expense is a limiting factor. The importance of both CT and MRI is that both methods acquire cross-sectional images of the body at pre-defined anatomic locations. Image analysis software then allows estimation of the adipose tissue, skeletal muscle, and organs based on pixel intensity. Acquiring images at predefined intervals and integrating the area between slices allows reconstruction of an entire organ of interest such as skeletal muscle mass. A significant advancement made possible by these imaging methods has been the characterization of a tissue distribution, such as adipose tissue where it is now possible to quantify visceral, subcutaneous, and intermuscular depots at the regional and whole-body level.

Bioimpedance Analysis (BIA)

BIA is a simple, low-expense, noninvasive body composition measurement method. BIA is based on the electrical conductive properties of the human body. Measures of bioelectrical conductivity are proportional to total body water and the body compartments with high water concentrations such as fat free and skeletal muscle mass. BIA assumes that the body consists of two compartments, fat and FFM (Body weight = Fat + FFM). BIA is best known as a technique for the measurement of percent body fat although it has more recently been used for estimating skeletal muscle mass too.

Anthropometry

For routine clinical use, anthropometric measurements (circumference measures and skinfold thickness) have been preferred due to ease of measurement and low cost. Waist circumference and the waist-hip ratio measurements are commonly used surrogates of fat distribution, especially in epidemiology studies. Waist circumference is highly correlated with visceral fat and was recently included as a clinical risk factor in the definition of the metabolic syndrome. Specifically, waist circumferences greater than 102 cm (40 in) in men and greater than 88 cm (35 in) in women are suggestive of elevated risk.

Skinfold thicknesses which estimate the thickness of the subcutaneous fat layer are highly correlated with percent body fat. Since the subcutaneous fat layer varies in thickness throughout the body, a combination of site measures is recommended, reflecting upper and lower body distribution. Predictive percent body fat equations based on skinfold measures are age and sex specific in adults and children.

Body Mass Index (BMI)

The body mass index (BMI = weight kg/height m^2) continues to be the most commonly used index of weight status, where normal weight is a BMI 18.5–25.9 kg/m^2; overweight is a BMI 25.0–29.9 kg/m^2; and obese a BMI >30.0 kg/m^2. BMI is a commonly used index of fatness due to the high correlation between BMI and percent body fat in children and adults. The prediction of percent body is dependent on age (higher in older persons), sex (higher in males), and race (higher in Asian compared to African American and Caucasian).

In Vivo Neutron Activation

Nitrogen, carbon, hydrogen, phosphorus, sodium, chlorine, calcium, and oxygen are all measurable *in vivo* by methods known as neutron activation analysis. A source emits a neutron stream that interacts with body tissues. The resulting decay products of activated elements can be counted by detectors and elemental mass established. Carbon, nitrogen, and

calcium can be used to estimate total body fat, protein, and bone mineral mass using established equations. Neutron activation analysis is uniquely valuable in body composition research as there are no known age or sex effects of currently applied equations, however, facilities that provide these techniques are limited.

Models in Body Composition

The use of models in the assessment of body composition allows for the indirect assessment of compartments in the body. Typically, a compartment is homogenous in composition (e.g., fat), however, the simpler the model the greater the assumptions made and the greater the likelihood of error. The sum of components in each model is equivalent to body weight (**Figure 1**). These models make assessments at the whole-body level and do not provide for regional or specific organ/tissue assessments.

The basic two-compartment (2C) model (**Table 1**) is derived from measuring the density of fat-free mass (FFM) by hydrodensitometry and subtracting FFM from total body weight thereby deriving fat mass (body weight – FFM = fat mass). FFM is a heterogeneous compartment consisting of numerous tissues and organs. A 2C approach becomes inadequate when the tissue of interest in included within the FFM compartment. Nevertheless, the 2C model is routinely and regularly used to calculate fat mass from hydrodensitometry, total body water, and total body potassium.

A three-compartment (3C) model consists of fat, fat-free solids, and water. The water content of FFM is assumed to be between 70% and 76% for most species and results from cross-sectional studies in adult humans show no evidence of differences in the hydration of FFM with age. The fat-free solids component of FFM refers to minerals (including bone) and proteins. The 3C approach involves the measurement of body density (usually by hydrodensitometry) and total body water by an isotope dilution technique. Assumptions

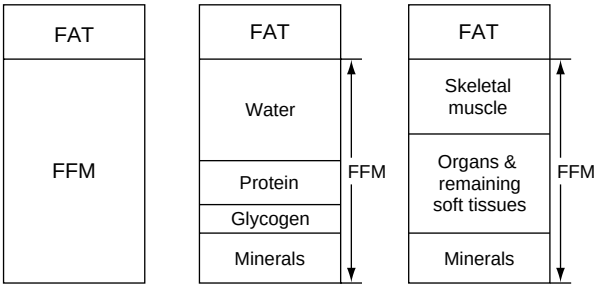


Figure 1 Three different models for characterizing body composition compartments. Components are as labeled: FFM, fat-free body mass.

Table 1 Multicompartment body composition models

Model	*Equations for % fat*	*Reference*
2C	$100 \times (4.971/D_b - 4.519)$	a
3C	$100 \times (2.118/D_b - 0.78 \times (TBW/W) - 1.354)$	b
4C	$100 \times (2.747/D_b - 0.727 \times (TBW/W) + 1.146 \times (BMC/W) - 2.0503)$	c
6C	$100 \times (2.513/D_b - 0.739 \times (TBW/W) + 0.947 \times (TBBM/W) - 1.79)$	d

[a]Behnke AR Jr, Feen BG, and Welham WC (1942) The specific gravity of healthy men. *Journal of the American Medical Association* **118**: 495–498.
[b]Siri WE (1961) Body composition from fluid spaces and density: analysis of methods. In: Brozek J and Hensch el A (eds.) *Techniques for Measuring Body Composition*, pp. 223–224. Washington, DC: National Academy of Science.
[c]Boileau RA, Lohman TG, and Slaughter MH (1985) Exercise and body composition of children and youth. *Scandinavian Journal of Sports Sciences* **7**: 17–27.
[d]Heymsfield SB, Wang ZM, and Withers RT (1996) Multicomponent molecular level models of body composition analysis. In: Roche AF, Heymsfield SB, and Lohman TG (eds.) *Human Body Composition*, pp. 129–147. Champaign: Human Kinetics.
D_b, body density; TBW, total body water; W, body weight; BMC, bone mineral content; TBBM, total body bone mineral.

are made that both the hydration of FFM and the solids portion of FFM are constant. Since bone mineral content is known to decrease with age, the 3C approach is limited in its accuracy in persons or populations where these assumptions are incorrect.

A four-compartment (4C) model involves the measurement of body density (for fat), total body water, bone mineral content by dual-energy X-ray absorptiometry (DXA), and residual (residual = body weight – (fat + water + bone)). This model allows for the assessment of several assumptions that are central to the 2C model. The 4C approach is frequently used as the criterion method against which new body composition methods are compared in both children and adults.

The more complex 4C model involves neutron activation methods for the measurement of total body nitrogen and total body calcium, where total body fat = body weight – total body protein (from total body nitrogen) + total body water (dilution volume) + total body ash (from total body calcium). A six-compartment model is calculated as follows: fat mass (measured from total body carbon) = body weight – (total body protein + total body water + bone mineral + soft tissue mineral (from a combination of total body potassium, total body nitrogen, total body chloride, total body calcium) + glycogen (total body nitrogen) + unmeasured residuals). However, the availability of neutron activation facilities is limited and therefore the latter models are not readily obtainable by most researchers.

At the organizational level, a five-level model was developed where the body can be characterized at five levels. The following are the levels and their constituents: atomic = oxygen, carbon, hydrogen, and other (level 1); molecular = water, lipid, protein, and other (level 2); cellular = cell mass, extracellular fluid, and extracellular solids (level 3); tissue-system level = skeletal muscle, adipose tissue, bone, blood, and other (level 4); and whole body (level 5).

Tissues and Organs

The aforementioned models do not allow for subregion and/or specific organ and tissue measurements. For example, skeletal muscle mass (SM) is contained within the FFM compartment. SM represents the single largest tissue in the adult body and is equivalent to ~40% of body weight in young adults, decreasing to ~30% of young values at elderly ages. SM is one of the more difficult components to quantify. Estimates of SM are commonly derived from anthropometry, total body potassium, and DXA using modeling approaches previously described. The use of magnetic resonance imaging (MRI) in body composition research has allowed for a good estimation of SM, adipose tissue (AT), and select organs *in vivo*, in all age groups with no risk to the participant (**Figure 2**). Moreover, AT distribution, including subcutaneous adipose tissue (SAT), visceral adipose tissue (VAT), and intermuscular adipose tissue (IMAT) is also measurable using a whole-body multislice MRI protocol (**Figure 3**). In studies relating body composition to energy expenditure, high metabolic rate organs including liver, kidneys, heart, spleen, and brain are also measurable using MRI.

Bone mineral content and bone mineral density of specific body sites (e.g., radius, hip, lumbar spine) are most commonly measured using DXA. Bone mass and microarchitecture are important determinants of bone strength, with microarchitectural deterioration being one of the specific changes associated with osteoporosis. Using high-resolution microcomputed tomography (micro-CT) and computer software, detailed analysis of three-dimensional (3D) architecture is feasible and allows microstructural 3D bone information to be collected.

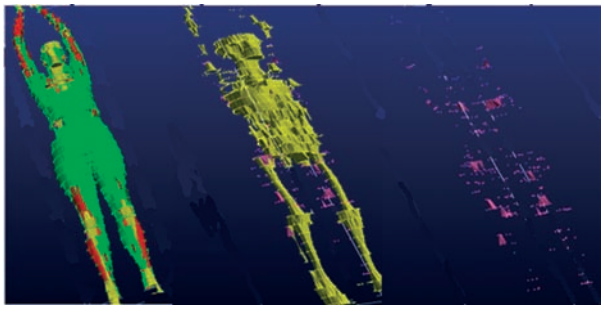


Figure 2 3D reconstructed image of whole-body scan (from MRI). Skeletal muscle (red); adipose tissue (green); bone, organs, and residual (yellow); intermuscular adipose tissue (pink).

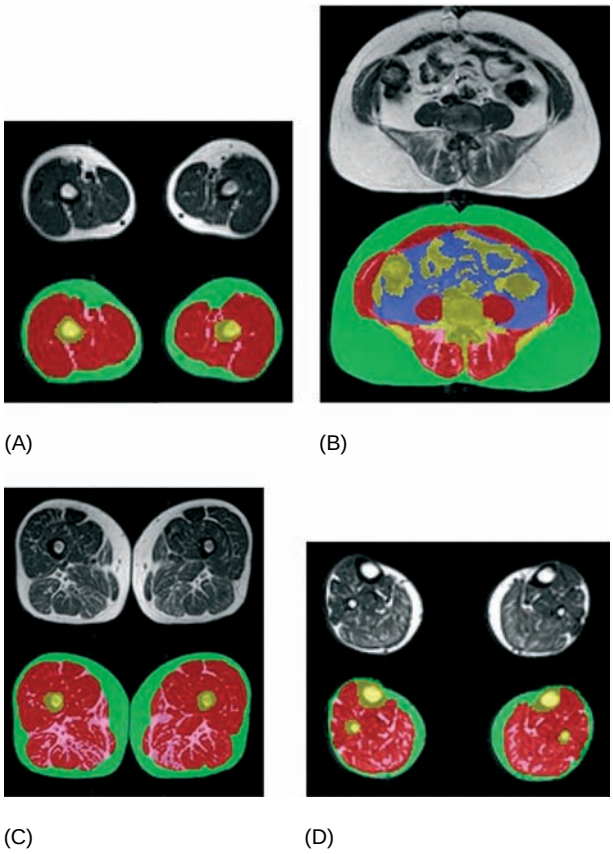


Figure 3 Cross-sectional images from (A) upper arm, (B) trunk (L4-L5 level), (C) mid-thigh, and (D) mid-calf in an elderly female volunteer. IMAT, intermuscular adipose tissue (pink); SM, skeletal muscle (red); SAT, subcutaneous adipose tissue (green); VAT, visceral adipose tissue (blue).

Body Composition Applications During Growth

Skeletal muscle mass has a central role in intermediary metabolism, aerobic power, and strength. Its mass increases as a portion of body weight during growth, accounting for 21% at birth and 36% at adolescence. The essential role of skeletal muscle in many physiologic processes throughout the lifespan makes understanding of factors affecting it significant. The greater incidence of type 2 diabetes mellitus in adolescents in the US (particularly in girls from minority populations) and in Japan makes evaluation of race and sex differences in pediatric skeletal muscle mass (and adipose tissue or fat mass) especially important. Identification and

characterization of differences could form the basis for further investigation of the associated metabolic implications.

Race differences in SM are known to exist as early as prepuberty. African-Americans have greater limb lean tissue mass compared to Asian and Caucasian children, while Caucasian children have greater amounts than Asians throughout Tanner stages 1 to 5. Race differences in total body bone mineral content adjusted for total body bone area, age, height, and weight have been reported in prepubertal African-American, Asian, and Caucasian females and males. African-American children had greater total body bone mineral content than Asian and Caucasian children, while differences between Asian and Caucasian children are less clear. Collectively, these findings suggest that the proportions of specific FFM subcomponents may differ by race. Although mechanisms leading to bone and skeletal muscle differences between races are not well understood, endocrine factors may be involved.

Sex differences in FFM have been reported from birth throughout childhood with females having smaller amounts than males. Total body bone mineral content is less in Tanner 1 females compared to males in African-Americans, Asians, and Caucasians. The mechanism for this sex difference in unclear. Gonadal steroids are significant mediators of adult sexual dimorphism of body composition, including fat-free soft tissues. Prepubertal females have higher concentrations of circulating estradiol than prepubertal males, and gonadotropin and gonadal steroids increase gradually in both males and females from the age of 5 years. Thus, prepuberty is a period with sex differences in circulating concentrations of sex steroids and of changes in these concentrations with advancing age. The earlier skeletal maturation of females, for example, has been attributed to the greater estradiol level in females compared to males. However, non-hormonal (possibly genetic) mechanisms may also play a role.

Fat or adipose tissue distribution is recognized as a risk factor for cardiovascular disease in both adults and children. An android or male fat pattern, with relatively greater fat in the upper body region, is associated with negative metabolic predictors whereas a gynoid or female fat pattern, with relatively greater fat in the hip and thigh areas, is associated with less metabolic risk. More and more studies are showing that the syndrome develops during childhood and is highly prevalent among overweight children and adolescents. While the concept of the metabolic syndrome referred initially to the presence of combined risk factors including VAT, dyslipidemia, hypertension, and insulin resistance in adults, it is now known to exist in children, especially where obesity and/or higher levels of VAT are present. Although sex-specific patterns of fat distribution had previously been thought to emerge during puberty, sex and race differences in fat distribution are now known to exist in prepubertal children. The implications are that a specific body composition pattern may differ by sex and race. An example is the relationship of blood pressure to central fat distribution in boys compared to girls where a significant positive relationship between trunk fat and blood pressure was reported in boys but not girls, and was independent of race, height, weight, and total body fat. Understanding the predictors of blood pressure in children is important since childhood blood pressure has been shown to track into adulthood in longitudinal studies. Children whose blood pressure levels were in the highest quintile, were two times more likely to be in the highest quintile 15 years later. Identification of clinically useful body composition measures would allow for the identification of children at increased risk for hypertension, who could benefit from monitoring.

Race differences in fat distribution among prepubertal Asians, African-Americans, and Caucasians also exist. Previous reports in adolescents have suggested significantly smaller hip circumferences in Asian females at all pubertal stages compared to Caucasians and Hispanics and greater trunk subcutaneous fat in Asian females compared to Caucasians. Differences in subcutaneous fat mass and fat distribution in Asian compared to Caucasian adults have also been described. Understanding the sex- and race-specific effects of puberty on regional body composition may help delineate the developmental timing of specific health risk associations.

Race difference in blood pressure has been reported in many studies of adults, where a higher prevalence of hypertension has been found among African-American women, placing this group at a higher risk for cardiovascular-related morbidities and mortality. Previous studies attempting to determine whether this race difference appears in childhood or early in adulthood have produced inconclusive findings.

Body Composition Applications During Aging

During the adult life span, body weight generally increases slowly and progressively until about the seventh decade of life, and thereafter, declines into old age. An increased incidence of physical disabilities and comorbidities is likely linked to aging-associated body composition changes. Characterization of the

aging processes has identified losses in muscle mass, force, and strength, which collectively are defined as 'sarcopenia.' Little is known about the overall rate at which sarcopenia develops in otherwise healthy elderly subjects, if this rate of progression differs between women and men, and the underlying mechanisms responsible for age-related sarcopenia. Peak SM mass is attained in the young adulthood years and slowly declines thereafter. During the latter adult years, SM decreases more rapidly as body fat becomes more centralized. Anthropometric equations have been developed for predicting appendicular skeletal muscle (ASM = SM of the limbs) in the elderly where sarcopenia was defined as ASM (kg)/height^2 (m^2) less than two standard deviations below the mean of the young reference group. In the elderly men, the mean ASM/height^2 was approximately 87% of the young group. The corresponding value in women was approximately 80%. **Table 2** shows the estimated prevalence's of sarcopenia in the same survey sample for each ethnic group, by age and sex. The same authors have reported that obese and sarcopenic persons have worse outcomes than those who are nonobese and sarcopenic.

Even in healthy, weight-stable elderly persons, changes in body composition over a 2-year period can include decreases in SM mass and bone mineral content with corresponding increases in IMAT and VAT, after adjusting for their baseline values, despite no detectable changes in physical function or food intake.

In adults, excess abdominal or VAT is recognized as an important risk factor in the development of coronary heart disease and non-insulin dependent diabetes mellitus. Waist circumference and the waist:hip ratio are commonly used to predict visceral fat accumulation in epidemiological studies. However, waist circumference is unable to differentiate VAT from SAT. As a result, persons with similar waist circumferences could have markedly different quantities of VAT and abdominal SAT. Skinfold thickness has been used as a continuous variable grading adiposity or adipose tissue distribution within study populations.

The most accurate measurement of VAT requires imaging techniques (MRI and computed tomography (CT)), which are expensive and not readily available in many clinical settings. **Figure 3B** shows an MRI-derived cross-sectional image at the L4-L5 level with adipose tissue depots identified. The AT located between muscle bundles (IMAT; **Figure 3**) and visible by MRI and CT may be negatively associated with insulin sensitivity. In the elderly, greater IMAT (as suggested by lower skeletal muscle attenuation by CT) is associated with lower specific force production. Currently, there is no simple or clinic-based method to measure adipose tissue located between the muscle groups, defined in our laboratory as intermuscular adipose tissue (IMAT). IMAT has been reported to be significantly negatively correlated with insulin sensitivity and higher in type 2 diabetic women compared to nondiabetic women.

Sex and race differences in body composition are well established in adults. Men acquire higher peak SM mass than women and some evidence exists suggesting that men may lose SM faster than women with age. Moreover, it is well established that women have a larger amount of total body fat or total adipose tissue than men. Among races, African-American adult men and women have larger amounts of SM than Asian and Caucasians even after adjusting for differences in body weight, height, age, and skeletal limb lengths.

Efforts are ongoing to better understand variations in IMAT as a function of age, race, and level of fatness. IMAT deposits appear comparable in size in adult African-Americans, Asians, and Caucasians at low levels of adiposity but accumulate as a greater proportion of TAT in African-Americans compared to Caucasians and Asians subjects (58 g IMAT/kg TAT in African-Americans; 46 g IMAT/kg TAT in

Table 2 Prevalance (%) of sarcopenia[a] in the New Mexico Elder Health Survey, by age, sex, and ethnicity, 1993–1995

Age group (years)	*Men*		*Women*	
	Hispanic (n = 221)	*Non-Hispanic whites (n = 205)*	*Hispanics (n = 209)*	*Non-Hispanic whites (n = 173)*
<70	16.9	13.5	24.1	23.1
70–74	18.3	19.8	35.1	33.3
75–80	36.4	26.7	35.3	35.9
>80	57.5	52.6	60.0	43.2

[a]Appendicular skeletal muscle mass/height^2 (kg/m^2) less than two standard deviations below the mean value for the young adults from Gallagher D, Visser M, De Meersman RE *et al.* (1997) Appendicular skeletal muscle mass: effects of age, gender, and ethnicity. *Journal of Applied Physiology* **83**: 229–239.

Adapted from Baumgartner RN, Koehler KM, Gallagher D *et al.* (1998) Epidemiology of sarcopenia among the elderly in New Mexico. *American Journal of Epidemiology* **147**: 755–763.

Caucasians; 44 g IMAT/kg TAT in Asians). Across race groups, VAT deposits also appear comparable in size at low levels of adiposity but with increasing adiposity VAT accumulates more in Asians and Caucasians compared to IMAT, although accumulation rates for IMAT and VAT do not differ in African-Americans. While the association between greater amounts of abdominal or VAT and increased insulin resistance and the metabolic syndrome is well established compared to the peripherally located SAT, the role of the IMAT compartment in the metabolic alterations leading to the development of insulin resistance warrants further investigation, especially as it may influence race/ethnicity differences in dysglycemia. Collectively, sex and race differences exist in body composition in children and adults.

Physiological Application: Two Examples

Example 1

Expressing heat production relative to body mass is required when comparing energy expenditure rates between individuals that differ in size. Age and gender-specific resting energy expenditure (REE) norms based on body weight and stature-derived were developed in the early 1900s by Kleiber and showed that adult mammals differing widely in body size had similar metabolic rates relative to body weight raised to the 0.75 power. Two components are usually considered as representative of whole-body metabolically active tissue, body cell mass (BCM), and FFM. BCM is typically estimated as the exchangeable potassium space that can be measured by total body potassium. The FFM component can be measured using two-component body composition methods.

In studies assessing REE, FFM is considered the principal contributor to energy requirements, and is commonly used as a surrogate for metabolically active tissue. However, this practice is inherently flawed as it pools together numerous organs and tissues that differ significantly in metabolic rate. The brain, liver, heart, and kidneys alone account for approximately 60% of REE in adults while their combined weight is less than 6% of total body weight or 7% of FFM. The skeletal muscle component of FFM comprises 40–50% of total body weight (or 51% of FFM) and accounts for only 18–25% of REE. REE varies in relation to body size across mammalian species. Within humans, REE per kg of body weight or FFM is highest in newborns ($\sim$56 kcal kg^{-1} day^{-1}), declines sharply until 4 years, and slowly thereafter reaching adult values ($\sim$25 kcal kg^{-1} day^{-1}). Among adults, REE is lower in the later adult years, to an extent beyond that explained by changes in body composition. That is, the loss of FFM cannot fully explain the decrease (5–25%) in REE in healthy elderly persons.

Recent attention has been given to modeling REE based on available information on organ- and tissue-specific metabolic rates combined (**Table 3**) with the mass of these tissues as determined by MRI. Whole-body REE can be calculated from organ- tissue mass (REE_c) and then compared to REE measured using indirect calorimetry (REE_m) for individuals or groups. REE (in kJ day^{-1}) of each organ- tissue component (subscript i) can be calculated using the following equation:

$$REE_i = OMR_i \times M_i \quad [1]$$

where OMR (organ metabolic rate) is the metabolic rate constant (in kJ per kg per day) for each organ-tissue component (**Table 3**) and M is the mass of the corresponding organ/tissue (in kg). Whole-body REE (in kJ per day) is calculated as the sum of the seven individual organ-tissue REE

$$REE_c = \sum_{i=1}^{7} (REE_i) \quad [2]$$

The whole-body REE equation is:

$$REE_C = 1008 \times M_{brain} + 840 \times M_{liver} + 1848 \times M_{heart} + 1848 \times M_{kidneys} + 55 \times M_{SM} + 19 \times M_{AT} + 50 \times M_{residual} \quad [3]$$

This approach has allowed for the hypothesis to be tested that the proportion of FFM as certain

Table 3 Organ and tissue coefficients used in developing models

	Weight (kg)[a]	*Density (kg l^{-1})*[a]	*Metabolic rate (kJ kg^{-1} day^{-1})*[b]
Skeletal muscle	28.0	1.04	55
Adipose tissue	15.0	0.92	19
Liver	1.8	1.05	840
Brain	1.4	1.03	1008
Heart	0.3	1.03	1848
Kidneys	0.3	1.05	1848
Residual	23.2	*	50

[a]Adapted from Snyder WS, Cook MJ, Nasset ES *et al.* (1975) Report of the task group on reference men. *International Commission on Radiological Protection* 23. Oxford: Pergamon.
[b]Adapted from Elia M (1992) Organ and tissue contribution to metabolic rate. In: Kinney JM and Tucker HN (eds.) *Energy Metabolism. Tissue Determinants and Cellular Corollaries*, pp. 61–77. New York: Raven Press.
*Residual mass was not assigned a density but was calculated as body mass minus sum of other measured mass components.

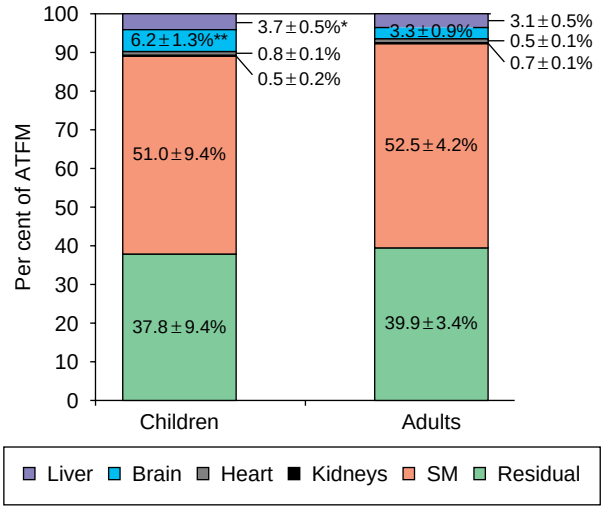


Figure 4 Proportional contribution of each organ/tissue to Adipose Tissue Free Mass (ATFM). Liver (▭), brain (▭), heart (▭), kidneys (▭), skeletal muscle mass (▭), residual mass (▭), * $p < 0.01$ and ** $p < 0.001$ for children vs. adults. Reproduced with permission from Hsu A, Heshka S, Janumala I, Song MY, Horlick M, Krasnow N, and Gallagher D (2003) Larger mass of high-metabolic-rate organs does not explain higher resting energy expenditure in children. *American Journal of Clinical Nutrition* **77**: 1506–11.

high metabolic rate organs, specifically liver and brain, is greater in children compared to young adults (**Figure 4**). Findings thus far have shown that after accounting for this disproportion, the specific organ/tissue metabolic constants available in the literature (**Table 3**) are not adequate to account for REE in children. These results therefore imply that the decline in REE per kilogram body weight (or per kilogram FFM) during the growth years is likely due to both changes in body composition and changes in the metabolic rate of individual organs/tissues. When this approach was applied to young adults (31.2 ± 7.2 years), REE_c and REE_m were highly correlated, with no significant differences between them. When this approach was applied to persons over 70 years, both older men and women had significantly lower REE_m compared to REE_c, and the magnitude of the differences were 13% and 9.5%, respectively, for men and women. These findings suggest that even after adjustment for age-related organ and tissue atrophy in the elderly, whole body REE by indirect calorimetry continues to be lower than expected. The latter suggests that the metabolic rate constants used (**Table 3**) for specific organs and tissues may not be appropriate in the elderly.

At the individual or clinic level, the measurement of REE by indirect calorimetry is frequently unavailable. An alternate approach has been to estimate REE based on body weight, height, age, and sex. Many studies have examined the association between these basic and easily acquired measures and REE. A small number of studies have included FFM in their REE prediction equations. **Table 4** lists published equations for the prediction of REE in healthy individuals.

Example 2

The global increase in the prevalence of childhood overweight and obesity and their associations with disease during childhood and adulthood is now alarming public health officials. One approach to understanding the pathways between overweight/obesity and disease is identifying the factors that cause excess weight gain. The time *in utero* is considered a critical period. During the growth and development years, the periods known as 'adiposity rebound' and adolescence are considered critical periods in the development and persistence of overweight in the pediatric age group. Children born small or large for gestational age appear to be at increased risk for cardiovascular disease and diabetes in adulthood. To what extent the growth trajectory between birth and adulthood influences the risk of disease burden is unclear. Important questions that need to be answered include the role that adiposity or fat accretion and adipose tissue distribution has on the development of disease. To answer such questions, the measurement of body composition needs to occur at the organ/tissue level beginning as early as birth, if not earlier. The current MRI methodology allows for such measurements after birth although no data exists thus far where infants have been followed longitudinally into adolescence or adulthood.

Conclusion

The measurement of body composition allows for the estimation of body tissues, organs, and their distributions in living persons without inflicting harm. It is important to recognize that there is no single measurement method in existence that allows for the measurement of all tissues and organs and no method is error free. Furthermore, bias can be introduced if a measurement method makes assumptions related to body composition proportions and characteristics that are inaccurate across different populations. The clinical significance of the body compartment to be measured should first be determined before a measurement method is selected since the more advanced techniques are less accessible and more costly.

Table 4 REE prediction equations based on anthropometrics or body composition

Authors	*Subjects/gender/nation*	*Weight status*	*Age (years)*	*Equation*
Harris & Benedict (1919)	239/M-F/USA	NW	29 ± 14 (X ± SD)	F: **BMR** = 9.5 wt (kg) + 1.9 ht (cm) − 4.7 age (years) + 655 M: **BMR** = 13.8 wt (kg) + 5.0 ht (cm) − 6.8 age (years) + 66
Robertson & Reid (1952)	2310/M-F/UK	NS	Range (3–80)	**RMR** = BSA (m^2) × 24 × age-specific value
Altman & Dittmer (1968)	>200/M-F/USA	NW	Range (3–16)	F: **REE** = 0.778 wt (kg) + 24.11 M: **REE** = 0.815 wt (kg) + 21.09
Dore *et al.* (1982)	140/F/UK	NW, OW, OB	Variable	**REE** = 8.24 wt (kg) + 0.02 FFM (kg) − 3.25 age (years) + 712
Bernstein (1983)	202/M(154)/USA	OW, OB	40 ± 12 (X ± SD)	**RMR** = 7.48 wt (kg) − 0.42 ht (cm) − 3.0 age (years) + 844 **REE** = 22 FFM (kg) + 6.4 FM (kg) − 2.1 age (years) + 251
Garrow & Webster (1985)	104/F/UK	NW, OW, OB	Variable	**REE** = 24.2 FFM (kg) + 5.8 (% fat) + 310
Joint FAO/WHO/UN (1985)	11 000/M-F/Multi	NW, OW, OB	Variable	3–10 years F: **REE** = 22.5 wt (kg) + 499 3–10 years M: **REE** = 22.7 wt (kg) + 495 10–18 years F: **REE** = 17.5 wt (kg) + 651 10–18 years M: **REE** = 12.2 wt (kg) + 746 18–30 years F: **BMR** = 55.6 wt (kg) + 1397.4 ht (m) + 146 30–60 years F: **BMR** = 36.4 wt (kg) − 104.6 ht (m) + 3619 18–30 years M: **BMR** = 64.4 wt (kg) − 113.0 ht (m) + 3000 30–60 years M: **BMR** = 47.2 wt (kg) + 66.9 ht (m) + 3769
Schofield (1985)	7549/M-F/UK	NW, OW, OB	Range (<3 to >60)	Under 3 years F: **BMR** = 0.068 wt (kg) + 4.281 ht (m) − 1.730 Under 3 years M: **BMR** = 0.0007 wt (kg) + 6.349 ht (m) − 2.584 3–10 years F: **BMR** = 0.071 wt (kg) + 0.677 ht (m) + 1.553 3–10 years M: **BMR** = 0.082 wt (kg) + 0.545 ht (m) + 1.736 10–18 years F: **BMR** = 0.035 wt (kg) + 1.948 ht (m) + 0.837 10–18 years M: **BMR** = 0.068 wt (kg) + 0.574 ht (m) + 2.157 18–30 years F: **BMR** = 0.057 wt (kg) + 1.184 ht (m) + 0.411 18–30 years M: **BMR** = 0.063 wt (kg) − 0.042 ht (m) + 2.953 30–60 years F: **BMR** = 0.034 wt (kg) + 0.006 ht (m) + 3.530 30–60 years M: **BMR** = 0.048 wt (kg) − 0.011 ht (m) + 3.670 Over 60 years F: **BMR** = 0.033 wt (kg) + 1.917 ht (m) + 0.074 Over 60 years M: **BMR** = 0.038 wt (kg) + 4.068 ht (m) − 3.491
Owen (1986)	44/F/USA	NW, OW, OB	29 ± 14 (X ± SD)	F: **RMR** = 7.18 wt (kg) + 795
Owen (1987)	60/M/USA	NW, OW, OB	29 ± 14 (X ± SD)	M: **RMR** = 10.2 wt (kg) + 879
Owen (1988)	104/M-F/USA	NW, OW, OB	29 ± 14 (X ± SD)	**REE** = 23.6 FFM (kg) + 186
Ravussin & Bogardus(1989)	249/M-F/USA	NW, OW, OB	Variable	**REE** = 21.8 FFM (kg) + 392
Maffeis *et al.* (1990)	130/M-F/Italy	NW, OW, OB	Range (6–10)	F: **REE** = (35.8 wt (kg) + 15.6 ht (cm) − 36.3 age (years) + 1552)/4.18 M: **REE** = (28.6 wt (kg) + 23.6 ht (cm) − 69.1 age (years) + 1287)/4.18
Mifflin *et al.* (1990)	498/M-F/USA	NW, OW, OB	Range (19–78)	F: **RMR** = 9.99 wt (kg) + 6.25 ht (cm) − 4.92 age (years) − 161 M: **RMR** = 9.99 wt (kg) + 6.25 ht (cm) − 4.92 age (years) + 5 **REE** = 19.7 FFM (kg) + 413

Continued

Table 4 Continued

Cunningham (1991)	Meta-analysis	NW, OW, OB		**REE** = 21.6 FFM (kg) + 370
Hayter & Henry (1994)	2999/M/UK	NW, OW, OB	Range (18–30)	M: **RMR** = 51.0 wt (kg) + 3500
Piers *et al.* (1997)	39/M/Australia	NW, OW	Range (18–30)	M: **RMR** = 51.0 wt (kg) + 3415
van der Ploeg *et al.* (2001)	38/M/Australia	NW, OW	24.3 ± 3.3 (X ± SD)	18–30 years M: **RMR** = 48.2 wt (kg) + 25.8 ht (cm) − 49.6 age (years) + 113 18–30 years M: **RMR** = 21.0 wt (kg) − 56.2 age (years) + 76.1 FFM 4C (kg) + 2202
van der Ploeg *et al.* (2002)	41/M/Australia	NW, OW	44.8 ± 8.6 (X ± SD)	30–60 years M: **RMR** = 41.92 wt (kg) + 13.79 ht (cm) − 14.89 age (years) + 1939 30–60 years M: **RMR** = 91.85 FFM 4C (kg) + 1463
Siervo *et al.* (2003)	157/F/Italy	NW, OW, OB	23.8 ± 3.8 (X ± SD)	F: **RMR** = 11.5 wt (kg) + 542.2

M, male; F, female; NS, not specific; NW, normal weight; OW, overweight; OB, obesity; X, mean; SD, standard deviation; BSA, body surface area; wt, weight; ht, height; BMR, basal metabolic rate; RMR, resting metabolic rate; REE, resting energy expenditure; FFM, fat-free mass; FFM 4C, fat-free mass via the four-compartment body composition model; FM, fat mass.

Adapted from the following references: Altman P and Dittmer D (1968) *Metabolism*. Bethesda: Federation of American Societies for Experimental Biology; Bernstein RS, Thornton JC, Yang MU *et al.* (1983) Prediction of the resting metabolic rate in obese patients. *American Journal of Clinical Nutrition* **37**: 595–602; Cunningham JJ (1991) Body composition as a determinant of energy expenditure: a synthetic review and a proposed general prediction equation. *American Journal of Clinical Nutrition* **54**: 963–969; Dore C, Hesp R, Wilkins D, and Garrow JS (1982) Prediction of requirements of obese patients after massive weight loss. *Human Nutrition Clinical Nutrition* **36C**: 41–48; Garrow JS and Webster J (1985) Are pre-obese people energy thrift? *Lancet* **1**: 670–671; Harris JA and FG Benedict (1919) *A Biometric Study of Basal Metabolism in Man*, pp. 1–266. Washington DC: Carnegie Institution; Hayter JE and Henry CJK (1994) A re-examination of basal metabolic rate predictive equations: the importance of geographic origin of subjects in sample selection. *European Journal of Clinical Nutrition* **48**: 702–707; Joint FAO/WHO/UN Expert Consultation (1985) *Energy and Protein Requirements*. Technical Reports Series 724. Geneva: World Health Organization; Maffeis C, Schutz Y, Micciolo R Zoccante L, and Pinelli L (1993) Resting metabolic rate in six- to ten-year-old obese and non-obese children. *Journal of Pediatrics* **122**: 556–562; Mifflin MD, St Jeor TS, Hill LA *et al.* (1990) A new predictive equation for resting energy expenditure in healthy individuals. *American Journal of Clinical Nutrition* **51**: 241–247; Owen OE (1988) Resting metabolic requirements of men and women. *Mayo Clinic Proceedings* **63**: 503–510; Owen OE, Holup JL, D'Allessio DA *et al.* (1987) A reappraisal of the caloric requirements of men. *American Journal of Clinical Nutrition* **46**: 875–885; Owen OE, Kavle E, Owen RS *et al.* (1986) A reappraisal of caloric requirements in healthy women. *American Journal of Clinical Nutrition* **44**: 1–19; Piers LS, Diffey B, Soares MJ *et al.* (1997) The validity of predicting the basal metabolic rate of young Australian men and women. *European Journal of Clinical Nutrition* **51**: 333–337; Ravussin E and Bogardus C (1989) Relationship of genetics, age, and physical fitness to daily energy expenditure and fuel utilization. *American Journal of Clinical Nutrition* **49**: 968–975; Robertson JD and Reid DD (1952) Standards for the basal metabolism in normal people in Britain. *Lancet* **1**: 940–943; Schofield WN (1985) Predicting basal metabolic rate, new standards and review of previous work. *Human Nutrition Clinical Nutrition* **39C**: 5–41; Siervo M, Boschi V, and Falconi C (2003) Which REE prediction equation should we use in normal-weight, overweight and obese women? *Clinical Nutrition* **22**: 193–204; van der Ploeg and Withers RT (2002) Predicting the metabolic rate of 30–60-year-old Australian males. *European Journal of Clinical Nutrition* **56**: 701–708; van der Ploeg GE, Gunn SM, Withers RT, Modra AC, Keeves JP, and Chatterton BE (2001) Predicting the resting metabolic rate of young Australian males. *European Journal of Clinical Nutrition* **55**: 145–152.

Acknowledgments

This work was supported by Grants HL70298, DK42618 (Project 4), and HD42187 from the National Institutes of Health.

See also: **Bone**. **Obesity**: Definition, Etiology and Assessment; Childhood Obesity. **Older People**: Physiological Changes.

Further Reading

Baumgartner RN, Koehler K, Gallagher D *et al.* (1998) Epidemiology of sarcopenia among the elderly in New Mexico. *American Journal of Epidemiology* **147**: 755–763.

Ding M, Odgaard A, Linde F, and Hvid I (2002) Age-related variations in the microstructure of human tibial cancellous bone. *Journal of Orthopedic Research* **20**: 615–621.

Elia M (1992) Organ and tissue contribution to metabolic rate. In: Kinney JM and Tucker HN (eds.) *Energy Metabolism. Tissue Determinants and Cellular Corollaries*, pp. 61–77. New York: Raven Press.

Ellis KJ (2000) Human body composition: *in vivo* methods. *Physiological Reviews* **80**: 649–680.

Forbes GB (1987) *Human Body Composition: Growth, Aging, Nutrition, and Activity*. New York: Springer-Verlag.

Gallagher D, Belmonte D, Deurenberg P *et al.* (1998) Organ-tissue mass measurement allows modeling of REE and metabolically active tissue mass. *American Journal of Physiology* **275**: E249–E258.

Goran MI (2001) Metabolic precursors and effects of obesity in children: a decade of progress, 1990–1999. *American Journal of Clinical Nutrition* **73**: 158–171.

He Q, Horlick M, Fedun B, Wang J, Pierson RN Jr, Heshka S, and Gallagher D (2002) Trunk fat and blood pressure in children through puberty. *Circulation* **105**: 1093–1098.

Kleiber M (1961) *The Fire of Life, and Introduction to Animal Energetics*. New York: Wiley.

Roche AF, Heymsfield SB, and Lohman TG (eds.) (1996) *Human Body Composition*. Champaign: Human Kinetics.

Snyder WS, Cook MJ, Nasset ES *et al.* (1975) Reports of the task group on reference men. *International Commission on Radiological Protection* 23. Oxford: Pergamon.

Wang Z, Pierson RN Jr, and Heymsfield SB (1992) The five level model: a new approach to organizing body composition research. *American Journal of Clinical Nutrition* **56**: 19–28.

BONE

B M Thomson, Rowett Research Institute, Aberdeen, UK

Introduction

Bone serves as a framework for the body and as a metabolic reserve of calcium and phosphate at times of mineral deficiency. It consists of cells from two distinct lineages, bone-forming osteoblasts and bone-resorbing osteoclasts, and the calcified extracellular matrix that these cells secrete and remodel.

Bone formation begins in the embryo, either via a cartilaginous intermediate, as in the case of the long bones, or via a membranous intermediate, as in the case of the flat bones of the skull. Continued production of cartilage at specialized sites on the long bones, termed growth plates, and the subsequent conversion of this cartilage into bone results in longitudinal postnatal growth. Skeletal growth and development is regulated by genetic, mechanical and hormonal mechanisms. In general, genetic influences dictate the basic structure of the skeleton, while responses to mechanical loading adjust the strength of particular bones to their functional environment. Simultaneously, hormonal mechanisms coordinate the movement of calcium and phosphate to and from the skeleton, thereby enabling bone to act as a reservoir of these minerals at times of calcium stress (e.g., pregnancy and lactation). At the cellular level, bone growth is coordinated by an array of interacting cytokines and growth factors, which control bone cell division, maturation and activity.

Failure of the mechanisms controlling bone cell function, especially during bone turnover in adults, leads to bone loss, and this can produce clinical osteoporosis. Other skeletal disease states result from nutritional deficiency, e.g., rickets, or from genetic defects, e.g., osteopetrosis or osteogenesis imperfecta.

Bone Types, Composition and Structure

There are two types of bone in the skeleton – flat bones, e.g., the skull, and long bones, e.g., the femur. The principal anatomical features of a long bone are shown in **Figure 1**.

Bone matrix

Bone matrix is a composite material that derives its strength from a compression-resistant mineral phase and a tension-resistant network of collagen fibers. Bone's mineral phase – calcium hydroxyapatite,

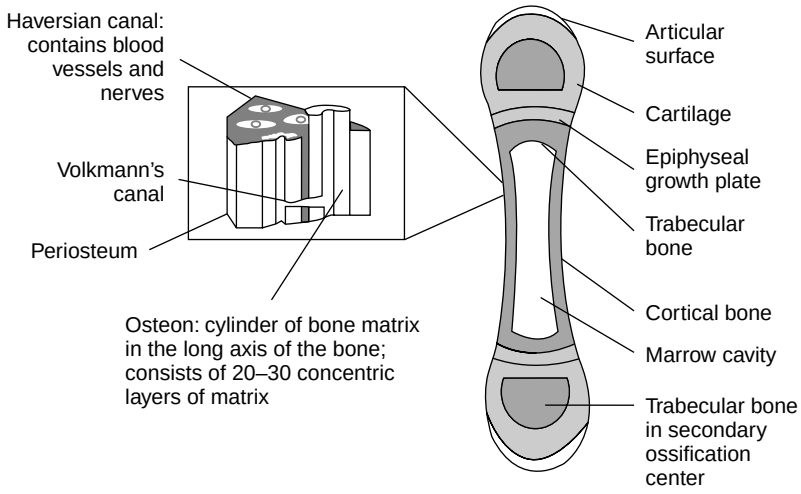


Figure 1 The anatomy of a long bone, e.g., the femur. Inset shows an enlarged section of cortical bone.

$Ca_{10}(PO_4)_6(OH)_2$ – is subdivided into a mosaic of tiny microcrystallites, thereby creating a large surface area for ion exchange and limiting the spread of cracks. Bone matrix also contains a number of specialized noncollagenous proteins, such as osteocalcin, osteonectin and osteopontin.

Macroscopic architecture

Two types of internal bone architecture are visible to the naked eye. Cortical bone, the stronger but heavier of the two forms, comprises the outer wall of all bones and fulfils a mainly mechanical function (see insert, **Figure 1**). It consists of parallel cylinders of matrix (osteons) arranged along the load-bearing axis of the bone.

Within each osteon the matrix is deposited in concentric layers, each 2–3 μm thick, with a predominant fiber direction (like multilayer plywood). The central canal of each osteon contains bone cells, blood vessels and nerves.

Trabecular bone, the second architectural form, is found at the ends of long bones and in the middle of the vertebrae. It consists of a latticework of bony struts, each 100–500 μm thick. Although weaker than cortical bone, it is more cellular and hence more metabolically active.

Bone Cells

Osteoblasts

Bone-forming cells are called osteoblasts. They are characterized by high levels of alkaline phosphatase, an enzyme required for matrix mineralization, and display structural features reflecting their intense secretory activity (e.g., prominent endoplasmic reticulum).

Osteoblasts are arranged as a closely packed layer of cells on growing bone surfaces, with each cell producing around three times its own volume of bone in about 3 days. Newly synthesized bone matrix is produced in an unmineralized form (termed osteoid) and consists of highly crosslinked collagen I fibers (which give the tissue its tensile strength) and a number of noncollagenous proteins such as osteocalcin. Osteoid is also rich in osteoblast-derived growth factors – insulin-like growth factor II (IGF-II) and transforming growth factor β – and these may regulate local bone turnover. Once formed, osteoid is mineralized. In cortical bone, crystal growth begins at sites along the collagen fibrils and is regulated by inhibitory molecules released by the osteoblasts.

Approximately 10–20% of osteoblasts become entombed in the matrix that they have produced and are termed osteocytes. They remain linked to the bone surface via long cell processes and appear to respond to mechanical loading. Osteocytes may therefore be responsible for coupling of mechanical stimulation to bone growth.

Osteoblasts contribute to the control of bone resorption, responding to bone-resorbing signals by producing degradative enzymes and by releasing molecules that increase osteoclast activity and recruitment. Osteoblasts may therefore coordinate bone turnover by switching from bone formation to the control of bone resorption.

Origin of the osteoblast Osteoblasts are derived from stem cells located near bone surfaces. These stem cells can give rise to cartilage, fat and fibrous tissues in experimental systems, suggesting that osteoblasts form part of a superfamily of connective-tissue cells.

The stem cells divide and mature into preosteoblasts, an intermediate cell type, which displays some osteoblast-like features, e.g., type I collagen production, alkaline phosphatase and osteonectin mRNA, but which lacks the intense alkaline phosphatase activity and highly developed endoplasmic reticulum of the mature cells.

Terminal differentiation into mature osteoblasts is associated with the cessation of cell division, the production of osteopontin and osteocalcin, and changes in both oncogene expression and nuclear protein–DNA interactions. Further changes in regulatory nuclear proteins accompany the onset of mineralization.

Osteoclasts

Osteoclasts are large ($3000–250000\,\mu m^3$) highly mobile multinucleate cells (10–20 nuclei), which contribute to bone remodeling and calcium homeostasis by resorbing bone. The osteoclast's resorptive apparatus consists of a central 'ruffled border', a highly folded region of cell membrane across which acid and degradative enzymes are extruded, and a peripheral 'clear zone', which seals the osteoclasts onto the bone.

Osteoclasts dissolve bone mineral by secreting acid across their ruffled borders using proton-pumping ATPases. Acid production also requires carbonic anhydrase – an enzyme used for acid production by the stomach, which is absent in some forms of the lethal bone disease osteopetrosis. The organic components of bone are degraded by lysosomal enzymes; one of these, acid phosphatase, is used as a marker for osteoclast activity.

Origin of the osteoclast Osteoclasts are descended from blood-cell-forming interleukin-3-dependent stem cells located in the bone marrow and are therefore part of the same superfamily of cells as macrophages, lymphocytes and red blood cells. Partially differentiated mononuclear preosteoclasts migrate via the circulation to resorption sites, where they proliferate and acquire differentiated features (e.g., acid phosphatase and calcitonin receptors), before fusing into multinucleate osteoclasts and beginning to resorb.

Regulation of Bone Cell Activity

Genetic and mechanical

While genetic influences dictate the basic structure of the skeleton, responses to mechanical loading adapt bone to its functional environment. In general, physical strain stimulates bone formation, while disuse (e.g., paralysis) results in bone loss. As a consequence, the cortical width of a professional tennis player's serving arm may be increased by 35% relative to the other arm. Such adaptation optimizes the balance between skeletal strength and weight.

Systemic regulation of serum calcium: the calciotropic hormones, parathyroid hormone, calcitonin, and dihydroxyvitamin D_3

Serum calcium is maintained within tight limits ($2.2–2.6\,mmol\,l^{-1}$), a process necessary for the maintenance of many cellular activities including neuromuscular function.

Parathyroid hormone (PTH) releases calcium into the blood by mobilizing bone mineral at times of calcium shortage. It binds to receptors on preosteoblastic cells and signals the release of osteoclast activating factors. It also promotes calcium retention by the kidney and uptake from the gut.

Calcitonin is produced in response to a rise in serum calcium. It acts directly on osteoclasts and inhibits bone resorption.

Vitamin D metabolites regulate serum calcium and are essential for skeletal growth and development. The parent compound, vitamin D, is essentially inactive and requires two hydroxylation steps before gaining biological activity. The second of these hydroxylation steps is tightly regulated by PTH. Once formed, the most potent vitamin D metabolite, 1,25-dihydroxyvitamin D_3 ($1,25(OH)_2D_3$), increases intestinal calcium absorption and promotes bone matrix mineralization. It also alters cell differentiation, influencing chondrocyte maturation within the growth plate and regulating gene expression in osteoblastic cells.

Systemic hormones that regulate skeletal growth or function: growth hormone, oestradiol, and vitamin A

Growth hormone stimulates bone growth by prompting cell division amongst the chondrocytes of the growth plate. It acts in part by stimulating IGF-I production.

While there is no doubt that the decreased levels of oestradiol that follow normal or artificially induced menopause lead to increased bone loss, oestradiol's role in skeletal biology remains obscure. The recent discovery of oestradiol receptors in osteoblasts suggests it may have a direct effect on osteoblastic cells.

Vitamin A is important for the maintenance of normal bone remodeling. It may also participate in the control of three-dimensional pattern formation during limb bud development.

Local regulation of bone cell proliferation, maturation, and function by polypeptide growth factors

A variety of growth factors (e.g., fibroblast growth factor and IGF-I) stimulate the division of preosteoblastic cells. Other signal molecules (e.g., transforming growth factor β) are associated with developmental events such as the formation of the vertebrae, jaws, and palate. Osteoblasts produce many of these growth factors themselves and deposit them in their extracellular matrix, suggesting that they coordinate small groups of cells at specific locations. Osteoclast recruitment is regulated by a variety of blood cell growth factors, termed 'colony stimulating factors'. Several of these are produced by osteoblastic cells. Osteoblasts have also been shown to produce factors that stimulate mature osteoclasts to resorb.

Mechanisms of Bone Growth

Flat bones and long bones arise by two distinct mechanisms.

Growth of long bones (e.g., femur): endochondral ossification

The principal stages of long-bone growth are shown in **Figure 2**. Long bones begin as cartilaginous regions in the early embryo. They grow as rapidly dividing peripheral cells add new chondrocytes to the outside of the structure and as older cells in the body of the cartilage divide, enlarge, and secrete matrix (**Figure 2A**).

The oldest chondrocytes (located in the middle) expand, calcify their matrix, and are termed 'hypertrophic' (**Figure 2B**). Osteoblasts then secrete a bony layer around the midshaft of the cartilage, forming the 'primary bone collar'. This structure is extended and thickened by successive generations of osteoblasts.

Once established, the primary bone collar is penetrated at several points by osteoclasts (**Figure 2C**). These rapidly erode the calcified cartilage of the interior to leave only a supportive framework inside the bone collar. As the peripheral bone gains in strength, osteoclasts remove this framework to leave the marrow cavity.

Continued growth of long bones: the growth plate Long-bone growth continues at specialized 'epiphyseal growth plates' (**Figure 3**). Proliferating chondrocytes at the top of the growth plate add new cells, while their more mature descendants secrete matrix and enlarge, thereby producing longitudinal growth. Ultimately, the chondrocytes hypertrophy and calcify their matrix. Osteoclasts present in the marrow cavity invade this calcified cartilage, destroying the horizontal septa separating the chondrocytes. This leaves vertical bars of calcified cartilage projecting into the marrow cavity, and these act as a framework for subsequent bone deposition.

Complete calcification of the growth plate at the end of puberty marks the end of longitudinal growth.

Intramembranous ossification (e.g., the bones of the cranium)

The flat bones of the skull begin as highly vascularized sheets of embryonic tissue. Undifferentiated cells within these sheets differentiate directly into osteoblasts and form a radiating network of bony spicules lying parallel to the surface of the brain. During growth, successive generations of osteoblasts add new bone to the outside and periphery of this structure, while osteoclasts resorb from the inner surface to maintain proportional thickness and shape.

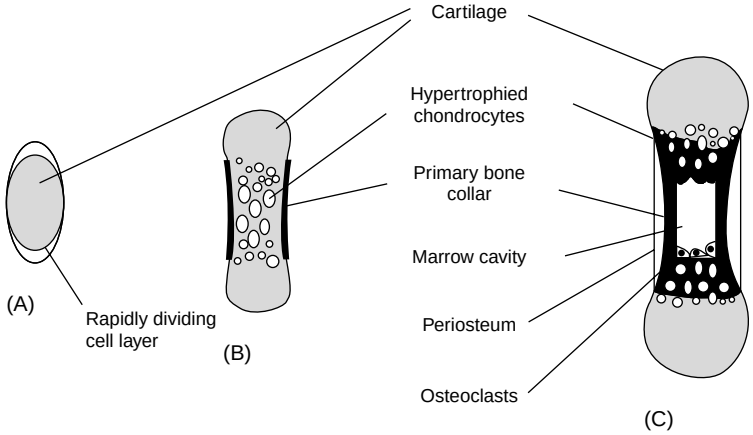


Figure 2 The principal stages of embryonic long-bone growth.

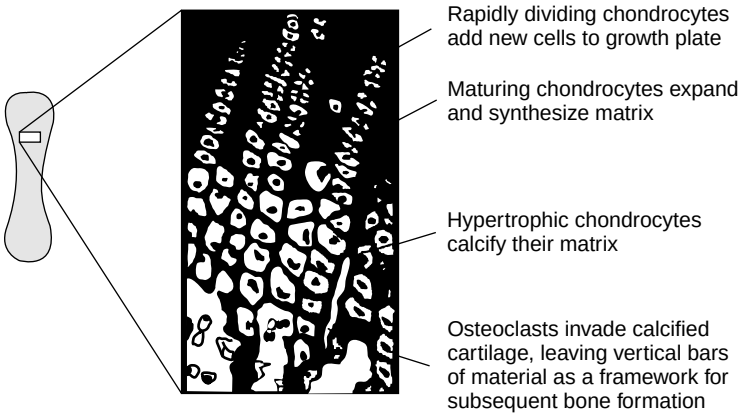


Figure 3 Continued longitudinal growth at the growth plate.

Bone Atrophy: Hormonal Influences and the Effect of Age

Bones grow in size and strength until peak bone mass is attained around the age of 35 years. Peak mass varies between demographic subgroups (25–30% higher in males than females 10% higher in blacks than whites), although there is a large variation within each group (1 woman in 40 has a peak mass less than the average for women aged 65 years). Bone mass subsequently declines with age and as a result of a variety of factors that predispose an individual to bone loss. The observed bone mass of an individual over 35 years therefore represents their peak bone mass minus both the cumulative age-related bone loss and any losses resulting from disease or specific risk factors.

Bone turnover and age-related bone loss

Bone is continuously broken down and replaced throughout life, thereby mobilizing calcium for systemic needs and preventing the accumulation of old fatigue-fractured material. Typically, 4% of cortical bone and 25% of trabecular bone are replaced each year.

Turnover begins with the recruitment of osteoclast precursors. These mature, fuse into multinucleate osteoclasts, and resorb a depression 60 μg deep in the bone surface over a period of about 10 days. Osteoblasts derived from undifferentiated cells located in the vicinity of the resorption site then replace the resorbed material. The entire process lasts approximately 4 months.

After the fourth decade of life, bone reformation fails to replace completely the resorbed bone, so that bone turnover produces a net bone loss. This so-called 'age-related bone loss' (0.3% and 1% of peak bone mass per year in males and females, respectively) occurs regardless of sex, physical activity, nutrition, or socioeconomic status. Its rate depends upon the frequency of remodeling cycles and the imbalance between the amount of bone resorbed and replaced at each remodeling event.

Age-related bone loss therefore occurs more rapidly in trabecular bone (which turns over more rapidly) and is increased by factors that promote bone turnover (transient calcium deficiency). Risk factors or disease states associated with either low peak bone mass or increased rates of loss include small body size, nulliparity, inactivity, early natural menopause, anorexia, thyrotoxicosis, and Cushing's syndrome.

Ultimately, the combination of reduced bone mass and disrupted trabecular architecture (as individual bony struts are severed or removed from the lattice) leads to reduced bone strength and increased fracture risk.

Other causes of bone loss: hyperparathyroidism and malignant disease

Hyperparathyroidism is an endocrine disorder characterized by raised serum calcium levels and increased cortical bone loss. It is usually caused by a benign adenoma, which secretes excessive PTH, thereby increasing both bone resorption and calcium retention by the kidney. Pathological bone loss is also a feature of many malignant disease states. Bone destruction may result from osteoclast activation in the immediate vicinity of invading tumors or from the systemic effects of circulating factors released by the tumor cells.

Oestrogen

Sex-hormone status and the pubertal growth spurt are important in determining bone mass at maturity. Conversely, bone loss accelerates in the first years after the menopause as the skeleton adapts to

declining oestrogen levels. In general, postmenopausal women who lose bone fastest have the lowest endogenous sex-hormone levels. Although the skeletal role of oestrogens at the cellular level remains unclear, recent findings suggest that oestrogens may have a direct effect on osteoblastic cells.

Osteopetrosis

Osteopetroses are a class of rare genetic disorders characterized by defective osteoclast function and a consequent failure of bone remodeling. Symptoms vary with severity but may include skeletal malformation with dense but fragile bones. Recurrent infection and spontaneous bleeding may result from the crowding out of the marrow cavity by unresorbed bone and mineralized cartilage.

See also: **Body Composition**. **Calcium**. **Carbohydrates**: Regulation of Metabolism. **Glucose**: Metabolism and Maintenance of Blood Glucose Level. **Growth and Development, Physiological Aspects**. **Osteoporosis**. **Vitamin A**: Physiology. **Vitamin D**: Physiology, Dietary Sources and Requirements; Rickets and Osteomalacia.

Further Reading

Favus MJ (1990) *Primer on the Metabolic Bone Diseases and Disorders of Mineral Metabolism* California: American Society for Bone and Mineral Research.

Goss RJ (1978) *The Physiology of Growth* New York: Academic Press.

Ham AW (1969) *Histology*, 6th edn. Philadelphia: Lippincott.

Riggs BL (1988) *Osteoporosis: Etiology, Diagnosis and Management* New York: Raven Press.

Thomson BM and Loveridge N (1992) Bone growth. In: *The Control of Fat and Lean Deposition*, 51st University of Nottingham Easter School in Agricultural Science. Oxford: Butterworth Heincmann.

Vaughan JM (1970) *The Physiology of Bone*. Oxford: Clarendon Press.

BRAIN AND NERVOUS SYSTEM

J D Fernstrom and M H Fernstrom, University of Pittsburgh, Pittsburgh, PA, USA

Design of the Nervous System

The nervous system has two principal cell types: neurons and glia. Neurons (like wires) conduct electrical signals and are organized into circuits to perform specific functions. They have a unique cellular architecture: small cellular extensions (dendrites) receive chemical and electrical signals from other neurons; a longer extension (the axon, which can be up to a meter in length) sends electrical signals down its length to one or more nerve terminals. Nerve terminals contain neurotransmitters, molecules released by arriving electrical signals that modify the electrical activity of adjacent neurons. Neurons have considerable energy needs; indeed the brain, which is 2% of body weight, consumes 15–20% of the body's daily energy intake. Glial cells, which constitute about 60% of the brain's cell mass, provide physical and metabolic support for neurons, and insulate axons and nerve terminals, to ensure privacy in electrical signaling. The glial cells found in peripheral nerves serve the same functions.

The nervous system is broadly divided into two parts: the central and peripheral nervous systems. The central nervous system (CNS) consists of the brain, retina, and spinal cord, and contains complex neuronal circuits that control body functions (e.g., blood pressure, breathing, hunger, movement). The peripheral nervous system consists of groups of neurons that mostly lie outside of the CNS, and either supply sensory information to the CNS, or send CNS commands to effector cells, such as muscle and gland cells.

The Blood–Brain Barrier

Each portion of the nervous system is separated from the blood (and thus the rest of the body) by a metabolic 'barrier,' which modulates the access of nutrients to and the removal of metabolites from the neurons and glia within it. For the brain and spinal cord, this barrier is termed the 'blood–brain barrier' (BBB; there is also a blood–cerebrospinal fluid (CSF) barrier; CSF is made from blood); for the retina, it is called the 'blood–retinal barrier,' and for peripheral neurons, the 'blood–nerve barrier.' The functions of

these barriers are very similar. The focus of the following discussion will be the BBB, because it has been studied the most.

The BBB is located in the endothelial cells that make up the brain's capillaries. Unlike capillaries elsewhere in the body, the endothelial cells of brain capillaries are tightly joined, such that nothing passes into (or out of) brain without passing through these cells. The BBB thus presents a continuous lipid barrier to molecules. One implication is that the ease with which molecules in blood gain access to brain should depend on their lipid solubility: the more lipid soluble, the greater the accessibility to brain by diffusion. However, most molecules of biologic importance to brain are not lipid soluble, and thus do not easily diffuse across lipid membranes into brain. Examples include glucose, amino acids, and water-soluble vitamins. Consequently, endothelial cell membranes must be more than just lipid barriers; indeed, embedded in them are specific transport carriers that mediate the brain uptake of most nutrients.

Energy Substrates

The brain uses glucose as its primary energy substrate. Glucose is not lipid-soluble, and thus requires a BBB transporter. The glucose transporter has a maximal transport capacity for glucose of 1.4 μmol per min per gram brain, or about $1200\,g\,day^{-1}$ for the entire brain (a human brain weighs 1400 g). The human brain consumes 15–20% of the body's oxygen; brain glucose utilization is therefore about $100\,g\,day^{-1}$. The BBB transporter thus has a maximal capacity for transporting glucose well in excess of that demanded daily by the brain.

Inside the brain, glucose is rapidly taken up into neurons by a cellular glucose transporter. Within the neuron, glucose enters the glycolytic pathway. The initial enzyme, hexokinase, has a very high affinity for glucose, and is fully saturated at normal brain glucose concentrations. Hence, overall, each step in the glucose pipeline from blood to brain neurons is designed to maximize glucose supply for neuronal energy production. It only fails when the blood glucose supply is abruptly curtailed, such as when a diabetic patient injects too much insulin and blood glucose levels rapidly fall (the transporter cannot compensate for such abrupt drops in blood glucose). The effect is dramatic: confusion, delirium, seizures, coma, and finally death occur as blood glucose drops to very low levels. Such effects are most rapidly reversed by the infusion of glucose, suggesting that no other compound in blood readily substitutes for glucose as the brain's primary energy substrate.

Normally, the body carefully maintains blood glucose concentrations. During starvation, however, blood glucose falls enough to cause the brain to recruit an additional energy source, i.e., ketone bodies. Ketone bodies are liver-produced by-products of the breakdown of stored fat (fatty acids), and provide an extended supply of energy when the input of food-derived energy is low. The brain uses ketone bodies whenever their blood levels rise; blood ketone body concentrations rise in starvation. The BBB ketone body transporter (ketone bodies are not lipid soluble) is induced in starvation, enhancing the flow of ketone bodies into brain. During prolonged starvation, more than half of the energy used by the brain is derived from ketone bodies. Continued use of some glucose appears obligatory, however, and is supplied via liver gluconeogenesis.

The chronic ingestion of high-fat diets also elevates blood ketone body concentrations, promoting their use by the brain for energy production. However, extremely high levels of fat must be consumed, and such diets are unpalatable. Hence, diet is not thought normally to influence cerebral energy production via dietary fat manipulation of ketone body supply to brain. Very high-fat diets are occasionally used clinically to treat intractable seizures. Though the beneficial effect is linked to levels of circulating ketone bodies, the mechanism is presently unknown.

Amino Acids and Protein

Neurons and glial cells in brain use amino acids to produce proteins. In addition, certain amino acids are used to produce small functional molecules such as neurotransmitters. Does diet influence amino acid flow into brain, and their use in generating proteins and transmitters? The path from diet to brain proceeds from amino acid absorption from the gastrointestinal tract, insertion into the circulation, and extraction by brain. This extraction process involves the BBB, which contains a number of transporters for amino acids. The properties of these transporters dictate how much of each amino acid enters (and exits) the brain. Currently, six carriers have been identified. Of special interest are two carriers: (1) the large neutral amino acid (LNAA) carrier, which is shared by several amino acids (some are precursors for neurotransmitters: phenylalanine, tyrosine, tryptophan, histidine). The carrier is competitive, allowing changes in the plasma concentration of any one LNAA to affect not only that amino acid's BBB transport, but also that of each of its transport competitors. Glutamine, an LNAA present in brain

in high concentrations, drives the brain uptake of the other LNAA, by serving as the principal amino acid counter-transported from brain to blood each time an LNAA is taken up into brain; and (2) the acidic amino acid carrier, which transports glutamic and aspartic acids. This carrier primarily transports glutamate and aspartate from the brain to the circulation. The other transporters include one selective for basic amino acids, two selective for subgroups of the small, neutral amino acids, and one selective for taurine.

The carriers that move amino acids into brain are those that transport primarily essential amino acids (the large, neutral, and basic amino acids), while those that move amino acids out of brain are those transporting nonessential amino acids (the acidic and small neutral amino acids). A small, net influx into brain of the essential amino acids no doubt reflects their consumption in brain by biosynthetic and metabolic pathways. The net efflux of the nonessential amino acids, notably aspartate, glutamate, glycine and cysteine may serve to remove from brain amino acids that act directly as excitatory transmitters or cotransmitters. The brain carefully compartmentalizes these amino acids metabolically, because they excite neurons, and a mechanism to remove them from brain may be a component of this compartmentalization design.

Changes in dietary protein intake have no effect on brain protein synthesis in adults. Indeed, the chronic ingestion of very low levels of dietary protein does not depress brain protein synthesis; brain cells may thus be efficient in retaining and reusing amino acids released during intracellular protein breakdown. In neonatal and infant animals, however, low levels of protein intake are associated with below normal rates of protein synthesis in brain. But, the presumed mechanism of this association, reduced uptake of essential amino acids into brain and abnormally low brain concentrations of these amino acids, has not been proven. Hence, at present, there is no convincing evidence linking dietary protein intake and brain protein synthesis via a limitation of amino acid availability to brain. For neurotransmitters, the evidence of this diet–brain link is more certain, and provides interesting examples of the fundamentally different manner in which the brain uses transport carriers to handle amino acids that are neurotransmitter precursors, and those that are neurotransmitters themselves. Good examples are tryptophan (an LNAA) and glutamate (an acidic amino acid), which have been most extensively studied.

Tryptophan (TRP) is the precursor for the neurotransmitter serotonin (5HT). The TRP concentration in brain rapidly influences the rate of 5HT synthesis: raising brain TRP concentrations increases synthesis, while lowering brain TRP decreases synthesis. Brain TRP uptake and concentrations are directly influenced by the plasma concentrations of TRP and its BBB LNAA transport competitors. The plasma concentrations of TRP and the other LNAA are readily modified by food intake, thereby linking diet to brain 5HT synthesis. Dietary proteins and carbohydrates are the food components that change brain TRP and 5HT: carbohydrate ingestion increases plasma TRP, while lowering the plasma concentrations of its LNAA competitors, causing BBB TRP uptake, brain TRP concentrations, and 5HT synthesis all to increase. The ingestion of a meal containing protein raises plasma concentrations of both TRP and its LNAA competitors. As a consequence, TRP experiences no change in competition for BBB transport (and sometimes a reduction, at high-protein intakes), and brain TRP concentrations and 5HT production do not change (or may decline). Hence, a key feature of the LNAA transporter, its competitive nature, explains the impact of meals containing or lacking protein on the production of a molecule important to normal brain function (5HT).

Chronic dietary effects are also observed. For example, chronic ingestion by rats of diets containing proteins with high ratios of one or more LNAA to TRP cause brain TRP and 5HT concentrations to decline. And, the chronic ingestion of diets low in protein causes the plasma concentrations of all LNAAs to decline (including TRP), and brain TRP and 5HT. In this case, brain TRP falls not because of a change in BBB competition, but simply because the BBB uptake of all LNAA declines with falling plasma concentrations (the transporter becomes unsaturated, eliminating competition).

Other LNAA are neurotransmitter precursors in substrate driven pathways in brain. Phenylalanine and tyrosine are substrates for catecholamine synthesis, and histidine is the precursor of histamine. Like TRP, the brain concentrations of these amino acids are influenced by their competitive BBB uptakes from the circulation, and thus the diet.

The nonessential amino acid glutamate (GLU) is an excitatory neurotransmitter, causing neurons that express GLU receptors to depolarize. Because GLU is excitatory, responsive neurons can become overexcited, when subjected to prolonged GLU exposure, and die. The term "excitotoxicity" was coined to describe this effect, and led to the concern that GLU ingested in food (as a constituent of dietary proteins; as a flavoring agent) might cause the brain to become flooded with GLU, causing

widespread neurotoxicity. The BBB acidic amino acid transporter prevents this from occurring: it primarily transports GLU out of, not into the brain. Consequently, the BBB functions as a 'barrier' to GLU penetration from the blood.

Another mechanism also protects brain neurons from excessive exposure to GLU. Glial cells rapidly remove GLU from brain extracellular fluid and convert it to an electrically inert amino acid, glutamine. While glial cells efficiently absorb neuronal GLU, they can just as readily clear any GLU that strays into the brain from the circulation.

Fatty Acids and Choline

Fatty Acids

The brain uses fatty acids to synthesize the complex fat molecules that form neuronal and glial cell membranes. This process is more active in growing animals than in adults. The brain synthesizes some fatty acids from smaller molecules, but their uptake from the circulation is also an important source, and is the only source for certain fatty acids (the essential fatty acids, which cannot be manufactured in the body). The details of the uptake process are not well understood.

From the nutritional perspective, diet influences essential fatty acid availability to brain, with potentially important functional consequences. In almost all mammals, there are two essential fatty acids: linoleic acid and α-linolenic acid (termed polyunsaturated fatty acids; PUFAs). In the nervous system (as elsewhere), linoleic and α-linolenic acids are incorporated into phospholipid molecules, and inserted into cellular membranes, where they influence membrane fluidity and membrane-associated functions (e.g., the functionality of receptors and transporters). In addition, the linoleic acid in membrane lipids can be released and converted into arachidonic acid, a key precursor in the synthesis of prostaglandins and leukotrienes, families of important signaling molecules. α-Linolenic acid can be converted into docosahexanoic acid, a molecule found in very large amounts in the rods and cones of the retina, and in nerve terminal membranes in brain. Docosahexanoic acid is thought to be a key component of phototransduction, and has been demonstrated to have important effects on vision. Dietary modifications in essential fatty acid intake might therefore be expected to influence membrane functions in brain, leading to alterations in brain function (as has been demonstrated for vision).

Choline

Choline occurs in the body as a constituent of lipid molecules in cell membranes, as a source of methyl groups, and as a precursor for the neurotransmitter acetylcholine. Choline is not an essential nutrient in humans, and deficiencies are rarely seen, since it is ubiquitous in the diet. However, in recent decades, dietary choline has been a focus of interest, because of the possibility that changes in choline intake could influence neuronal acetylcholine synthesis. Acetylcholine (ACh) is a neurotransmitter; its synthesis in and release by brain neurons is influenced by choline availability, which in turn can be altered by dietary choline intake, either in the form of free or fat-bound choline (phosphatidylcholine). In this context, oral choline and phosphatidylcholine have found some application in human diseases thought to involve ACh. For example, they have been used successfully to treat movement disorders such as tardive dyskinesia, a drug-induced muscular disorder in schizophrenic patients linked to low ACh function. However, they proved to be of little value in controlling abnormal muscle movements associated with Huntington's disease (also linked to low ACh function). Dietary choline and phosphatidylcholine supplements have also been studied as potential memory enhancers, since CNS ACh neurons play an important role in memory. Patients with Alzheimer's disease have been most studied but, in general, the disappointing outcome has been that neither choline nor phosphatidylcholine has afforded much improvement in memory.

Vitamins

Neurons and glia have the same functional demands for vitamins as do other cells in the body. Their access to brain is thus an important consideration, particularly given the existence of the BBB. Water-soluble vitamins are transported across the BBB and, in some cases, the blood-CSF barrier, most often by nonenergy requiring carriers. After they are taken up into neurons and glial cells, most are rapidly converted into their biologically active derivatives, namely cofactors in enzyme-mediated reactions. Since cofactors are recycled, dietary deficiencies in one or another vitamin do not immediately lead to brain dysfunction, inasmuch as cofactor pools may take extended periods of time to become depleted. Although fat-soluble vitamins are lipid soluble, their passage through the BBB most likely involves more than simply diffusion.

Water-Soluble Vitamins

Folic acid is transported into brain as methylenetetrahydrofolic acid, the major form of folic acid in the circulation. It is then transported rapidly into neurons and glia from the CSF/extracellular fluid. Once inside cells, folates are polyglutamated. Methylenetetrahydrofolate is used by neurons and glia in reactions involving single carbon groups, such as in the conversion of serine to glycine or homocysteine to methionine. Once methylenetetrahydrofolate is consumed in these reactions, folic acid is transported out of the brain into the circulation. Folate has become an issue of neurologic concern because of a link between folate deficiency and abnormal CNS development. The incidence of spina bifida, a serious spinal cord abnormality, rises above the population mean in the children of women who are folate-deficient during pregnancy. Moreover, the incidence of spina bifida can be reduced by folic acid supplementation during pregnancy, beginning prior to conception. Initiating supplementation before conception is essential, since the basic design of the CNS is laid down during the first trimester. At present, the mechanism(s) by which folic acid deficiency leads to the improper formation of the spinal cord is unknown. Folate deficiency may also be linked to depression in adults, and occasional studies suggest that folate supplementation can improve mood in depressed patients. The mechanism(s) by which folate modifies mood is presently unknown.

Ascorbic acid (vitamin C) is actively transported into the brain extracellular fluid through the blood–CSF barrier, from which it is actively transported into cells. Brain ascorbate pools show minimal fluctuations over a wide range of plasma ascorbate concentrations, which presumably explains the absence of CNS signs in ascorbate deficiency. To date, the only defined biochemical function of ascorbic acid in brain is as a cofactor for the enzyme that converts dopamine to noradrenaline (norepinephrine) (though ascorbate is thought by some to function as an antioxidant).

Thiamine (vitamin B_1) is taken up into brain by a BBB transporter; small amounts also gain entry via transport from blood into CSF. It is then transported into neurons and glia; conversion to thiamine pyrophosphate effectively traps the molecule within the cell. In nervous tissue, thiamine functions as a cofactor in important enzymes of energy metabolism. Severe thiamine deficiency in animals reduces thiamine pyrophosphate levels, and the activities of thiamine-dependent reactions. It causes loss of the coordinated control of muscle movement; the exact biochemical mechanism is not clear. The functional deficits are rapidly corrected with thiamine treatment, suggesting that neurons have not been damaged or destroyed. Thiamine deficiency in humans (beriberi; Wernicke's disease) produces similar deficits in the control of muscle movements, and also mental confusion. Korsakoff's syndrome, which occurs in almost all patients with Wernicke's disease, involves a loss of short-term memory and mental confusion. Severe thiamine deficiency in humans appears to produce neuronal degeneration in certain brain regions. The motor abnormalities can be corrected with thiamine treatment, but the memory dysfunction is not improved.

Riboflavin enters brain via a saturable BBB transport carrier. It is then transported into neurons and glia, and trapped intracellularly by phosphorylation and conversion to flavin adenine dinucleotide. Flavin adenine dinucleotide functions as a cofactor in carboxylation reactions. The brain contents of riboflavin and its derivatives are not notably altered in states of riboflavin deficiency or excess.

Pantothenic acid is transported into brain by a BBB transport carrier. Neurons and glial cells take up pantothenic acid slowly by a mechanism of facilitated diffusion. Inside the cell, the vitamin becomes a component of coenzyme A, the coenzyme of acyl group transfer reactions. Relative to other tissues, the brain contains a high concentration of pantothenate, mostly in the form of coenzyme A. Brain coenzyme A concentrations are not depleted in pantothenate deficiency states.

Niacin (vitamin B_3) is transported into brain as niacinamide, primarily via the BBB. Most niacin in brain is derived from the circulation, though brain may be able to synthesize small amounts. Niacin is taken up into neurons and glia and rapidly converted to nicotinamide adenine dinucleotide. The half-life of nicotinamide adenine dinucleotide in brain is considerably longer than in other tissues. Nicotinamide adenine dinucleotide and nicotinamide adenine dinucleotide phosphate are involved in numerous oxidation-reduction reactions. Dietary niacin deficiency in the presence of a low intake of TRP causes pellagra in humans, a deficiency disease that includes mental depression and dementia, loss of motor coordination, and tremor. The mechanism(s) for these effects have not been identified.

Pyridoxine (vitamin B_6) is taken up into brain via a transport carrier that has not been well described. The vitamin can be transported in any of its nonphosphorylated forms (pyridoxine, pyridoxal, pyridoxamine). Once within the brain extracellular fluid, the vitamin is readily transported into neurons and glia and phosphorylated (primarily to

pyridoxal phosphate or pyridoxine phosphate). Pyridoxal phosphate is a cofactor in a variety of neurotransmitter reactions, such as aromatic-L-amino acid decarboxylase (an enzyme of monoamine biosynthesis), glutamic acid decarboxylase (the enzyme of γ-amino butyric acid [GABA] synthesis), and GABA transaminase (the enzyme that catabolizes GABA). In humans, pyridoxine deficiency is rare, because of its widespread occurrence in foodstuffs. However, when identified, it has been associated with increased seizure activity, an effect dissipated by pyridoxine treatment. This effect may be linked to the production of GABA, an inhibitory neurotransmitter.

Biotin is transported into brain by a BBB carrier. It is a coenzyme for a variety of key carboxylation reactions in gluconeogenesis, fatty acid synthesis, and amino acid metabolism. Normally, biotin is recycled in cells during protein (enzyme) turnover, but not in brain; brain cells are thus more immediately dependent than other cells on circulating biotin availability. Biotin deficiency is rare; when it occurs, it can involve CNS symptoms (depression, sleepiness). The underlying basis for these effects is presently unknown.

Cobalamin (vitamin B_{12}) is thought to be transported into brain by a carrier-mediated mechanism. Little is known about this process, or about the function of vitamin B_{12} in the nervous system. Vitamin B_{12} deficiency is associated with neurologic abnormalities, which are presumed to derive from the demyelinization of CNS axons seen in advanced deficiency cases. These effects are reversed if vitamin B_{12} treatment is provided early enough; left untreated, axonal degeneration occurs. Vitamin B_{12} may be important in neuronal repair mechanisms, which may become compromised in deficiency states. Nervous system damage associated with vitamin B_{12} deficiency can occur at any age.

Fat-Soluble Vitamins

Of the fat-soluble vitamins, vitamin A (retinol) has been the most studied in relation to the CNS. The others have been much less well examined, though vitamin E is currently of some interest, because of its function as an antioxidant. The CNS is not thought to be a major focus of action for vitamins D and K, and thus little information is available regarding their roles in brain function.

The principal role of vitamin A in the CNS is as a component of the photoreceptive pigment of the eye, rhodopsin. In the blood, vitamin A circulates bound to retinol-binding protein and transthyretin (prealbumin). Its transport into retinal cells occurs at the blood–retinal barrier (the retinal pigmented epithelial (RPE) cells), after the retinol–protein complex binds to retinol-binding protein receptors. Once bound, retinol is released into the RPE cell. The retinol-binding protein and transthyretin molecules are released back into the circulation. Inside the RPE cell, retinol binds to a specific protein, and ultimately is esterified to a fatty acid. This molecule serves as the substrate for the conversion of retinol into the visually active form of the molecule, 11-*cis*-retinaldehyde, which then finds its way into the photoreceptor cell to be bound to opsin to form rhodopsin, the light-responsive pigment of the eye. When light strikes rhodopsin, phototransduction occurs and 11-*cis*-retinaldehyde is isomerized to all-*trans*-retinaldehyde, hydrolyzed from opsin, and released by the photoreceptor into the extracellular space (the opsin is retained and reused). The all-*trans*-retinaldehyde is shuttled into the RPE cell, where it is reconverted into 11-*cis*-retinaldehyde, and then recycled to the photoreceptor cells again to form rhodopsin.

From the nutritional perspective, retinal cells have an efficient system for managing and maintaining vitamin A pools. Hence, depletion of retinal vitamin A pools secondary to dietary deficiency only occurs over an extended time period. Deficiency appears functionally as 'night-blindness,' as rhodopsin levels decline. Extended vitamin A deficiency leads to a loss of photoreceptor elements, and eventually of the photoreceptor cells themselves. The cause of this cellular degeneration is not well understood.

Vitamin E is an antioxidant and free radical scavenger that protects fatty acids in cellular membranes. It is transported in blood associated with lipoproteins. The mechanism of its transfer into nervous tissue is unknown. Dietary vitamin E deficiency is extremely rare in humans. It occurs in association with certain abnormalities of vitamin E transport and fat absorption, and sometimes in individuals with protein-calorie malnutrition. The neurological manifestations are peripheral nerve degeneration, spinocerebellar ataxia, and retinopathy. Vitamin E has been proposed to play a role in a number of CNS diseases linked to oxidative damage. One example is Parkinson's disease, a movement disorder caused by the degeneration of certain groups of brain neurons. Evidence of oxidative damage is present in the brains of Parkinsonian patients, though controlled clinical trials of vitamin E supplementation have proved to be of no benefit. Such negative findings question the likelihood of a vitamin E link to the etiology of the degenerative changes. A second example is Alzheimer's dementia, which is associated with a progressive, ultimately catastrophic degeneration of the brain. Several

types of oxidative damage have been found in the brains of Alzheimer's patients, though it is presently unclear if this damage is cause or effect. Vitamin E supplementation can slow the progression of Alzheimer's disease. However, such findings do not indicate if vagaries in vitamin E intake over an extended period of time are a cause of the disease.

Minerals

All of the essential minerals are important for cellular functions in brain, as they are elsewhere in the body. These are sodium, potassium, calcium, magnesium, iron, copper, zinc, manganese, cobalt, and molybdenum. While most function as cofactors in enzymatic reactions, sodium and potassium are key ions in electrical conduction in neuronal membranes, calcium functions as a second messenger within neurons, and magnesium is an important component of certain neurotransmitter receptors. The diet normally provides more than adequate amounts of almost all minerals, except possibly for calcium, iron, magnesium, and zinc. The BBB permeability to most metals is quite low. For example, although the brain extracts 20–30% of the glucose in blood in a single capillary transit, it extracts $<0.3\%$ of any metal. The mechanisms of transport into brain for most metals are unknown. However, some details regarding the transport and/or functions of iron, calcium, and copper are available.

Iron circulates bound to the protein transferrin. Iron uptake into brain occurs primarily at the BBB, and involves a transferrin receptor-mediated endocytosis of the iron–transferrin complex by capillary endothelial cells. Iron dissociates from transferrin inside the cell, and is delivered into the brain interstitial fluid; the transferrin is returned to the circulation. Brain iron associates with ferretin, and is stored intracellularly. The bulk of the iron-ferretin stored in brain resides in glial cells and is laid down early in postnatal life. Marked regional differences in iron and ferretin concentrations occur in brain; levels in some areas are as high as those in liver. This distribution, however, does not correlate with the density of transferrin receptors in brain capillaries; it is presently unknown how or why the unequal distribution of iron develops. Numerous enzymes in brain are iron requiring, including several hydroxylases involved in neurotransmitter production, and a key metabolic enzyme, monoamine oxidase.

Iron deficiency can cause impairments in attention and cognition in children. Similar effects are seen in animals. In iron-deficient rats, brain iron concentrations decline, with newborn and infant animals showing more rapid declines than older animals. Iron repletion in brain occurs in infant and adult rats with iron supplementation, but not in animals depleted at birth. While outside of the brain, the activities of many iron-dependent enzymes are depressed by iron deficiency, their activities are unaffected inside the brain. However, a reduction in certain dopamine receptors occurs, along with aberrations in dopamine-dependent behaviors (dopamine is a CNS neurotransmitter). The inability of brain iron stores to recover in rats made iron deficient as newborns coincides with a persistence of dopamine-linked behavioral deficits, despite normal repletion of iron stores elsewhere in the body. Restoration of normal behavior with iron supplementation, along with brain iron stores, is seen in animals made iron deficient at other ages.

Iron deficiency also interferes with myelinization. Since marked glial proliferation and myelin formation occur early in infancy, iron deficiency during this period could prevent the optimal development of neuronal communications (glial cells provide insulation for axons and synapses). This effect could account for some of the behavioral deficits associated with neonatal iron deficiency.

Calcium is actively transported into the CNS, primarily via the blood–CSF barrier, and is not vitamin D sensitive. Since calcium concentrations in the circulation are regulated, under most circumstances, this process should also help to maintain brain calcium uptake and levels in the face of vagaries in calcium intake. Deficiencies in brain calcium should thus be a relatively rare occurrence.

Copper functions as a cofactor for numerous enzymes, including dopamine β-hydroxylase (DBH), which converts dopamine to noradrenaline. Dietary copper deficiency in humans is fairly rare. When produced in animals, it leads to reduced DBH activity in neurons and cells anywhere in the nervous system that synthesize noradrenaline. The mechanism of copper transport into the brain is presently unknown. Copper deficiency occurs as an X-linked genetic disease of copper transport in Menkes' syndrome, in which tissue and brain copper levels become extremely low, and produce neurodegeneration. Children with Menkes' syndrome die at a very young age.

See also: **Amino Acids**: Chemistry and Classification; Metabolism; Specific Functions. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Biotin**. **Calcium**. **Choline and Phosphatidylcholine**.

Cobalamins. **Copper**. **Fatty Acids**: Metabolism. **Folic Acid**. **Glucose**: Metabolism and Maintenance of Blood Glucose Level. **Niacin**. **Pantothenic Acid**. **Protein**: Synthesis and Turnover. **Riboflavin**. **Thiamin**: Physiology. **Vitamin A**: Physiology; Biochemistry and Physiological Role. **Vitamin E**: Metabolism and Requirements.

Further Reading

Davson H, Zlokovic B, Rakic L, and Segal MB (1993) *An Introduction to the Blood-Brain Barrier*. Boca Raton, CRC Press.

Fernstrom JD and Fernstrom MH (2001) Diet, monoamine neurotransmitters and appetite control. In: Fernstrom JD, Uauy R, and Arroyo P (eds.) *Nutrition and Brain, Nestlé Nutrition Workshop Series Clinical and Performance Program,* vol. 5, pp. 117–134. Basel: Karger.

Fernstrom JD and Garattini S (eds.) (2000) International Symposium on Glutamate. *Journal of Nutrition* **130**(4S): 891S–1079S.

Innis SM (1994) The 1993 Borden Award Lecture: Fatty acid requirements of the newborn. *Canadian Journal of Physiology and Pharmacology* **72**: 1483–1492.

Spector R (1989) Micronutrient homeostasis in mammalian brain and cerebrospinal fluid. *Journal of Neurochemistry* **53**: 1667–1774.

Zigmond MJ, Bloom FE, Landis SC, Roberts JL, and Squire LR (eds.) (1999) *Fundamental Neuroscience*. New York: Academic Press.

BREAST FEEDING

C K Lutter, Pan American Health Organization, Washington, DC, USA

Nutritional demands during the first year of life are greater than at any other time. During this period, a healthy newborn triples its birth weight and doubles its length and the size of its brain. The benefits of breast feeding during this critical period of development, even in the most privileged environments, are undisputable. Consequently, there are also measurable risks for those infants not breast fed, which include increases in diarrhea, acute respiratory infections, and otitis media and short- and long-term deficits in intellectual development. Not breast feeding may also be associated with increased risk of some chronic diseases and obesity, although the evidence for these relationships is still accumulating. Not breast feeding may also be associated with increased risk of some chronic diseases. In developing countries, the risks of not breast-feeding are magnified many times over and also include increased mortality. Because of these benefits, breast feeding should be promoted as a cultural and behavioral norm rather than as interchangeable with formula feeding.

Breast feeding also benefits women's health. Breast feeding reduces the risk of ovarian and premenopausal breast cancer and also promotes postpartum weight loss. Because it delays the return of menses, it reduces fertility and, in the absence of modern contraceptives, increases birth intervals. This article provides a broad overview of the physiological and nutritional aspects of breast milk and behavioral aspects of breast feeding, followed by a summary of global and national initiatives to improve breast feeding practices and data on breast feeding trends in Latin America.

Breast Feeding Recommendations

Both the World Health Organization (WHO) and the United Nations Children's Fund (UNICEF) recommend exclusive breast feeding for 6 months and continued breast feeding together with provision of safe, appropriate, and hygienically prepared complementary foods until 2 years of age or beyond. The American Academy of Pediatrics Section on Breastfeeding also recommends exclusive breast feeding for 6 months. Breast feeding is defined as exclusive if breast milk is the sole source of infant nutrition with no other liquids (including water) or food given, although medicinal and/or vitamin drops are permitted. Partial or mixed breast-feeding is used to describe infants who are not exclusively breast-fed. In a comprehensive review, WHO provided the scientific underpinnings of the recommended duration of exclusive breast feeding and noted that infants who were exclusively breast fed for 6 months experienced less morbidity from gastrointestinal infection than those who were exclusively breast fed for only 3 or 4 months. Also, exclusive breast feeding for 6 months, as opposed to only 3 or 4 months, resulted in no measurable deficits in growth among infants from either developing or developed countries.

The public health challenge is to support women to follow global breast feeding recommendations so as to ensure the healthiest start in life for all the world's children. Adherence to the recommended

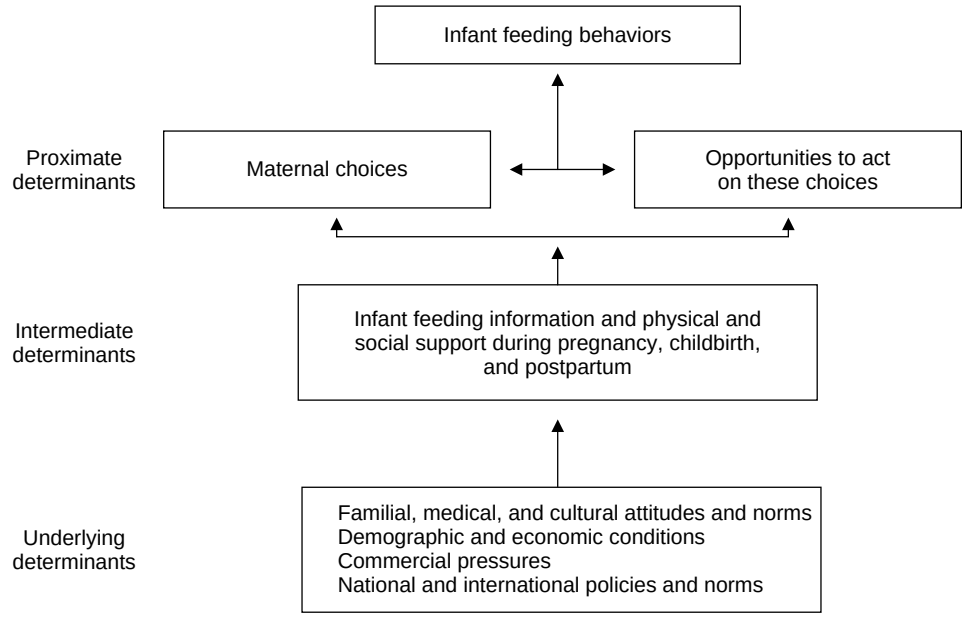


Figure 1 Determinants of infant feeding behaviors.

breast feeding behaviors or lack thereof results from a complex series of physiological and behavioral interactions between a mother and her infant—interactions that take place within a larger familial, community, and global setting (**Figure 1**). Although breast feeding occurs when a mother puts her child to the breast or allows her toddler to suckle, a woman's decision to breast feed and to act on this decision are dependent on a number of determinants, not all of which favor breast-feeding or are within her control. These determinants include infant feeding attitudes and norms among family members, the medical profession, peers, and employers; the availability of information and access to skilled assistance to prevent and/or address breast-feeding (BF) problems; and, during the period of BF initiation and exclusive breast feeding, nearly unrestricted access between mother and infant.

Breast Milk

Breast milk is a unique bioactive substance that changes composition, within and between feedings and over time, to suit the needs of the growing infant. More than 200 different constituents of breast milk have been identified, many of which have dual roles, and more continue to be discovered as analytic techniques improve. Breast milk includes true solutions, colloids, membranes, membrane-bound globules, and living cells. Its three distinct stages occur when colostrum, transitional, and mature milk are secreted; each aids newborns in their physiologic adaptation to extrauterine life. Compared to mature milk, colostrum, produced during the first week of life, has a higher protein and lower fat content and is rich in immunoglobulins, antibodies, and antioxidants. Compared to colostrum, transitional milk, produced from 7 days to between 10 days and 2 weeks postpartum, has lower immunoglobulins and total protein content and higher lactose, fat, and total caloric content.

Women experience the shift to transitional milk as a feeling of fullness in the breasts, which occurs between 40 and 72 h after birth. This occurs sooner for multiparas than primiparas. Milk volume increases dramatically after birth from less than 100 ml/day to approximately 500 ml/day by day 5 and approximately 650 ml/day by month 1 and 750 ml/day by month 3 (**Figure 2**). Exclusively breast fed infants consume 714, 784, and 776 ml/day between 0 and 2, 3 and 5, and 6 and 8 months, respectively. Mature milk has an energy density of approximately 75 kcal/ml, which translates into a

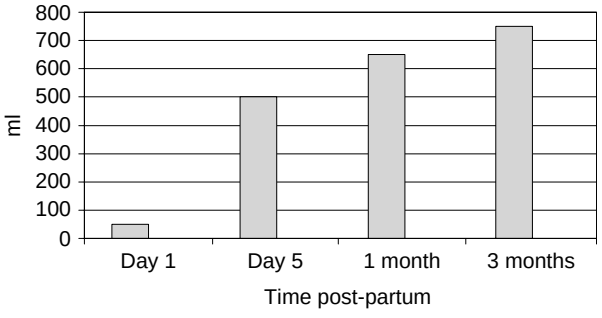


Figure 2 Average volume of breast milk production.

daily caloric intake of 450–550 kcal/day. This is entirely sufficient to satisfy current energy recommendations for infants for the first 6 months of life.

Once breast milk supply has been established, the volume produced depends on infant demand; frequent exclusive breast feeding is critical for stimulating optimal milk production. This is why it is so important that infants be breast fed on demand, day and night: The more often milk is removed from the breasts, the more milk a woman produces. The process of lactation requires both milk synthesis and its release into the alveoli and the lactiferous sinuses for removal by the suckling infant. Milk production is mediated by prolactin and its release by oxytocin. The decrease in plasma progesterone after birth initiates the process of lactogenesis. For several days following birth, lactogenesis does not depend on suckling; however, by day 3 or 4, milk secretion declines if milk is not removed from the breast. Therefore, any practices, such as the use of prelacteal feeds, supplemental bottles, or feeding on a schedule, that interfere with the infant's desire or ability to nurse effectively are likely to undermine the successful establishment of lactation. Also, because infant demand is the primary determinant of milk production, the early introduction of other liquids or complementary foods will displace the energy and nutrients provided by breast milk rather than provide an additional source of nutrition.

Lipids comprise the primary source of energy and are the most variable constituent of breast milk and vary in concentration during a feeding, between breasts, during the day, and over time. They also vary in concentration among women by as much as 50%. Research on the lipid component of breast milk has focused on its association with improved cognitive development and possible role in the prevention of obesity and other chronic degenerative diseases in breast fed children compared to formula-fed children. Particular attention has been paid to the dietary omega-3 polyunsaturated fatty acid, docosahexaenoic acid (DHA), which has been shown to enhance retinal development and visual acuity and may provide a physiologic explanation for the superior cognitive development that is documented in breast fed children. Following the lead of Japan and numerous European countries, the US Food and Drug Administration has permitted the addition of DHA to infant formula.

Proteins in breast milk include casein, serum, albumin, α-lactabumin, β-lactoglobulins, immunoglobulins, and other glycoproteins. Lactoferrin, an iron binding protein, inhibits the growth of certain iron-dependent bacteria in the gastrointestinal track and may protect against certain gastrointestinal infections. Lactose is the predominant carbohydrate of breast milk. It has been shown to enhance calcium absorption as well as provide a readily available source of galactose, which is essential to central nervous system development. Although concerns have been raised about the adequacy of breast milk to satisfy protein requirements for 6 months, a number of well-controlled studies have shown that protein needs can met through exclusive breast feeding.

The nutrients in breast milk most affected by maternal nutritional status are the water-soluble vitamins and the fat-soluble vitamins. With few exceptions, maternal stores and intake do not affect the mineral content of breast milk. Micronutrients affected by maternal intake and nutritional status include thiamin, riboflavin, vitamin B_6, vitamin B_{12}, vitamin A, iodine, and selenium. Those not affected include folate, vitamin D, calcium, iron, copper, and zinc. Particular attention has been paid to iron content of breast milk because of the relatively low breast milk content compared to theoretical needs. Breast milk iron, however, is highly bioavailable and exclusively breast fed term infants of normal birth weight are not at risk for iron deficiency anemia or depletion of iron stores. Iron supplements, beginning at 2 months, are recommended for preterm or low-birth-weight infants. Zinc, another mineral essential to human development, is also highly bioavailable in breast milk. Breast milk volume does not appear to be influenced by maternal nutritional status, except in situations of extreme food deprivation and famine.

Benefits of Breast Feeding

Breast feeding contributes to both maternal and infant nutrition and health through a number of important mechanisms. It provides a complete source of nutrition for the first 6 months of life for normal, full-term infants and provides one-half and one-third of energy needs for the second half of the first year and the second year of life, respectively. It also contributes significantly to protein and micronutrient requirements. Numerous studies have shown that during illness, whereas intake of complementary foods declines significantly, breast milk intake does not decrease. Because of the well-established superiority of breast milk over other infant feeding modes, women cannot ethically be randomized in infant feeding studies and as a result most data on the benefits of breast-feeding and the risks of not breast feeding are observational. However, the dose–response effect observed in such studies, even when donor breast milk is provided through a nasogastric tube to premature newborns, provides evidence of causality.

A large-scale study involving more than 17 000 infants in which breast feeding promotion was randomized and morbidity results analyzed on an 'intention to treat' basis, with breast feeding promotion as the treatment, also provides evidence of causality. Infants born in hospitals and provided care in clinics randomized to breast feeding promotion were 40% less likely to have more than one case of gastrointestinal infection and 50% less likely to have atopic eczema than infants not randomized to this intervention (**Figure 3**). The intervention significantly increased the duration of exclusive breast feeding at 3 months from 6 to 43% and the duration of partial breast-feeding at 1 year from 11 to 20%. Therefore, this study proved through a causal design that better breast feeding practices reduce risk of diarrhea and eczema, and that hospital and clinic-based interventions can result in large-scale shifts in behavior.

A pooled analysis of longitudinal data from Brazil, Pakistan, and the Philippines showed that during infancy breast feeding resulted in a 6-fold reduction in mortality during the first month of life, a 4-fold reduction in the second month, and a 2-fold reduction thereafter (**Figure 4**). In none of the studies was exclusive breast feeding sufficiently prevalent to examine the additional preventive effect of exclusive breast feeding over partial breast feeding. However, case–control studies that have examined breast feeding and mortality show that infants who are exclusively breast fed for the first 2 months of life have a 24-fold reduced risk of diarrhea compared to those exclusively bottle fed (**Figure 5**).

It is well established that breast-fed infants have a different pattern of growth. The fact that the nutrient composition of breast milk is qualitatively and quantitatively different than formula, coupled with the epidemic of obesity among children in developed countries, has led a number of investigators to examine its association with infant feeding mode. In one of the largest studies to date of more than 9000 German schoolchildren, a dose–response between the duration of breast feeding and risk of obesity at age 5 or 6 years was reported (**Figure 6**). However, not all studies have found a relationship. Because most of the research to date has been carried out in developed countries, where breast feeding is more common among the better educated and affluent who are also more likely to practice other healthful behaviors, the possibility of bias persists.

More than 11 studies have documented an association between breast feeding and cognitive development. In the longest follow-up to date, a study

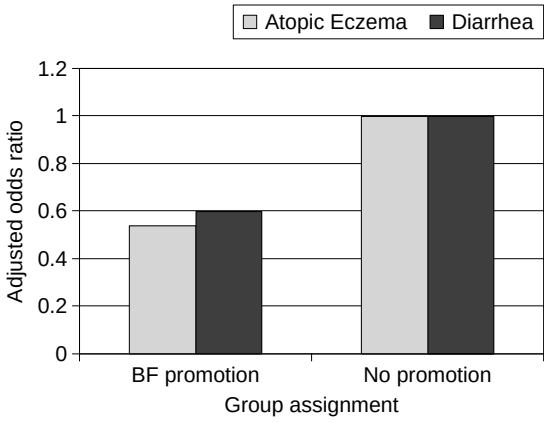


Figure 3 Breast feeding promotion reduces risk of atopic eczema and diarrhea. (Source: Kramer MS, Chalmers B, Hodnett E *et al.* (2001) Promotion of breastfeeding intervention trial (PROBIT): A randomized trial in the Republic of Belarus. *Journal of the American Medical Association* **285**(4):413–420.)

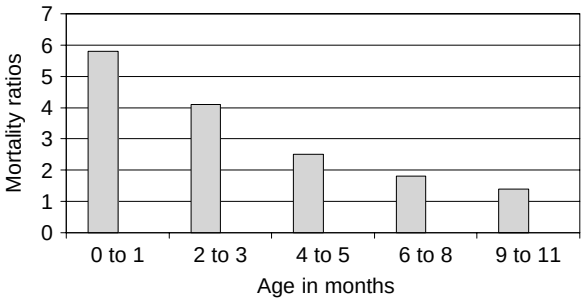


Figure 4 Increased risk of mortality associated with not breast feeding. (From World Health Organization (2000) WHO Collaborative Study Team on the role of breastfeeding on the prevention of infant mortality: Effect of breastfeeding on infant and child mortality due to infectious diseases in less developed countries: A pooled analysis. *Lancet* **355:** 451–455.)

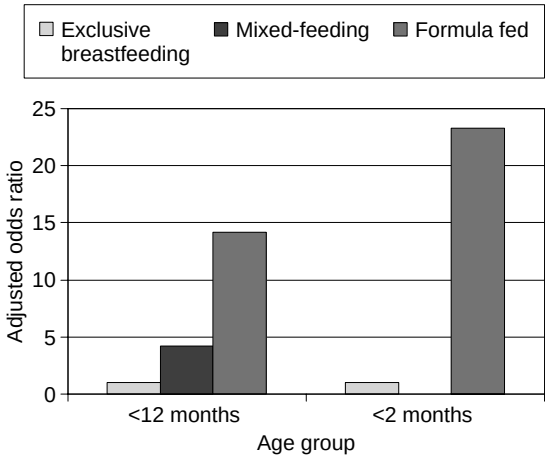


Figure 5 Exclusive breast feeding and risk of mortality from diarrhea in Brazil. (From Victora CG, Vaughan JP, Lombardi *et al.* (1987) Evidence for protection by breast-feeding against infant deaths from infectious diseases in Brazil. *Lancet* **2**(8554): 319–321.)

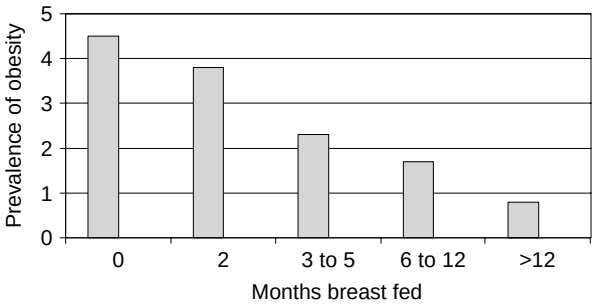


Figure 6 Months of breast feeding and prevalence of obesity in German schoolchildren. (From Von Kries R, Koletzko B, Sauerwald T *et al.* (1999) Breast feeding and obesity: Cross sectional study. *British Medical Journal* **319:** 147.)

also showed that adults who as infants had been breast fed for at least 9 months grew up to be significantly more intelligent than those breast-fed 1 month or less. After adjusting for a variety of factors, a dose–response relationship between the length of breast feeding and a number of different intelligence quotient measures was found in Danish adults who had been breast fed as infants. The total difference for those breast fed between 7 and 9 months versus less than 1 month was 6.6 points.

The longer women breast feed, the more they are protected against breast cancer. A reanalysis of data from 47 studies in 30 countries included more than 50 000 women with breast cancer and nearly 97 000 without it. Women who developed breast cancer were less likely to have breast fed, or if they had, the length of time they had breast fed was significantly shorter than that of women who were free of breast cancer. The effect of breast feeding on risk of breast cancer did not differ between women in developed and developing countries, and it did not vary significantly by age, menopausal status, ethnic origin, the number of births a woman had, her age when her first child was born, or any of a number of other personal characteristics examined. Based on the size of the effect, it is estimated that the cumulative incidence of breast cancer in developed countries would be reduced by more than half, from 6.3 to 2.7 per 100 women by age 70, if women had the average number of births and lifetime duration of breast feeding that has until recently been prevalent in developing countries.

Not surprisingly, not breast feeding results in significantly greater health care expenditures. Among middle-class children from the United States and Scotland, the excess cost of health care services for diarrhea, lower respiratory tract illness, and otitis media during the first year of life was estimated to be between $331 and $475 per never-breast fed infant in 1995. These costs were calculated on the basis of 2033 excess office visits, 212 excess days of hospitalization, and 609 excess prescriptions for each 1000 never-breast fed infants compared to 1000 infants breast fed exclusively for at least 3 months.

Breast Feeding Initiatives

In response to concerns about the use of infant formula in environments where lack of breast feeding resulted in large numbers of infant who became severely ill or died, a grassroots global initiative took hold in the 1970s to promote international and national efforts to protect, promote, and support breast feeding. These efforts culminated in 1981 with the nearly unanimous adoption by the World Health Assembly (WHA) of the International Code of Marketing of Breast-Milk Substitutes. This document and subsequent relevant WHA resolutions, collectively known as the Code, provide guidelines for the marketing of breast milk substitutes, bottles, and teats. To ensure infant feeding decisions free from the influence of marketing pressures, the Code provides guidelines on a number of issues associated with increases in formula feeding, including direct promotion to the public, donations to health care institutions, free supplies to mothers, and the use of baby images on labels that glorify bottle feeding. Its implementation is monitored by a 2-year reporting cycle by countries to the WHA and by the International Code Documentation Centre in Penang, Malaysia. Despite continued violations by infant food companies and the lack of enforcement in many developed countries, the Code has provided an important tool for regulating and monitoring the infant food industry to ensure that its marketing practices do not undermine breast feeding.

The 1990 Innocenti Declaration, which set four operational targets that all governments should achieve by 1995, was endorsed by the 45th WHA. These targets included appointment of a national breast feeding coordinator and establishment of a multisectoral national breast feeding committee; ensuring that all health facilities providing maternity services fully practice the 10 steps to successful breast feeding (**Table 1**) set out in the WHO/UNICEF statement; taking action to give effect to the Code; and enacting imaginative legislation to protect the breast feeding rights of working women. This declaration provided the basis for the WHO/UNICEF Baby Friendly Hospital Initiative (BFHI), which was developed in 1991, piloted in 12 countries, and inaugurated as a global initiative in 1992. BFHI promotes hospital practices consistent with early initiation, an environment conducive to BF, appropriate clinical management of BF, and compliance

Table 1 WHO/UNICEF 10 steps to successful breast feeding

Step 1. Have a written breast-feeding policy that is routinely communicated to all health care staff.
Step 2. Train all health care staff in skills necessary to implement this policy.
Step 3. Inform all pregnant women about the benefits and management of breast feeding.
Step 4. Help mothers initiate breast feeding within a half-hour of birth.
Step 5. Show mothers how to breast feed and how to maintain lactation even if they should be separated from their infants.
Step 6. Give newborn infants no food or drink other than breast milk, unless medically indicated.
Step 7. Practice rooming-in—allow mothers and infants to remain together—24 h a day.
Step 8. Encourage breast feeding on demand.
Step 9. Give no artificial teats or pacifiers (also called dummies or smoothers) to breast feeding infants.
Step 10. Foster the establishment of breast feeding support groups and refer mothers to them on discharge from the hospital or clinic.

with certain key provision of the Code, such as no donations of free or subsidized infant formula. Certification is awarded to hospitals that comply with a set of standardized Baby Friendly criteria developed by WHO and UNICEF. Worldwide, nearly 20 000 hospitals are certified as Baby Friendly.

At the same time that the Code and BHFI were being implemented, numerous projects worked to develop training materials in lactation management and counseling skills and provided funding for widespread dissemination of training courses. National governments also actively implemented campaigns to promote breast feeding. In Latin America, these efforts are likely responsible for a resurgence of breast feeding.

Global Breast Feeding Practices

The most comprehensive data on breast feeding come from the Demographic and Health Surveys conducted with support from the US Agency for International Development. These surveys are nationally representative and conducted throughout the developing world. In a number of countries, multiple surveys permit the analysis of trends. Overall, the data show that although the vast majority of women—more than 90% in all countries—initiate breast feeding, the duration of exclusive breast feeding is far less than the recommended 6 months (**Table 2**). In most countries, the duration of breast feeding is unchanged. Several countries are showing increases and in only one does there appear to be a decrease. However, concurrent with the time period during which the surveys took place, numerous demographic changes occurred that are negatively associated with breast-feeding, such as increased female employment and education and increased urbanization. When adjusted for these changes, highly significant improvements in breast feeding are seen in many countries, particularly those in which breast feeding promotion efforts have been most active.

Table 2 Trends in breast feeding practices

Country	*Year*	*Initiation (%)*	*Median duration BF (months)*	
			Exclusive BF	*Any BF*
Bolivia	1989	96.4	NA	16.9
	1994	96.3	3.3	17.5
	1998	96.6	3.9	18.0
Colombia	1986	NA	NA	11.6
	1990	93.4	0.6	12.7
	1995	94.5	0.5	11.3
	2000	95.5	0.7	13.1
Costa Rica	1981	NA	NA	7.2
	1986	NA	NA	9.3
	1993	NA	NA	9.1
Dominican Republic	1986	92.7	NA	8.1
	1991	92.0	0.4	9.0
	1996	93.2	0.6	7.6
Ecuador	1994	95.0	2.0	15.7
	1999	97.0	2.2	15.5
El Salvador	1988	93.1	NA	15.2
	1993	91.2	0.8	15.5
	1998	94.0	0.9	17.7
Guatemala	1987	NA	NA	19.9
	1995	95.6	1.7	19.8
	1998	96.5	0.9	19.9
Haiti	1977	NA	NA	15.6
	1994	96.3	0.4	17.5
	2000	97.4	0.4	18.5
Honduras	1986	NA	NA	17.3
	1991	NA	NA	17.2
	1996	96.0	2.1	17.3
Jamaica	1989	96.0	NA	12.4
	1993	94.0	NA	12.4
Nicaragua	1992–1993	91.9	0.6	12.3
	1998	92.4	0.7	12.2
Paraguay	1990	92.8	0.4	10.5
	1995/6	93.6	0.3	11.4
	1998	94.2	NA	11.5
Peru	1991–1992	96.0	0.8	17.3
	1996	96.8	2.7	19.5
	2000	97.8	4.2	21.6

BF, breast feeding; NA, not available.

See also: **Infants**: Nutritional Requirements; Feeding Problems. **Lactation**: Physiology; Dietary Requirements. **United Nations Children's Fund**. **World Health Organization**.

Further Reading

Aguayo VM, Ross JS, Kanon S *et al.* (2003) Monitoring and compliance with the Code of Marketing of Breast-Milk Substitutes in West Africa: Multisiet cross sectional survey in Togo and Burkina Faso. *British Medical Journal* **326**: 1–6.

Ball TM and Wright AL (1999) Health care costs of formula-feeding in the first year of life. *Pediatrics* **103**: 870–876.

Butte NF (2001) The role of breastfeeding in obesity. *Pediatric Clinics of North America* **48**(1): 189–198.

Collaborative Group on Hormonal Factors in Breast Cancer (2002) Breast cancer and breastfeeding: Collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease. *Lancet* **360**: 187–195.

Kramer MS, Chalmers B, Hodnett E *et al.* (2001) Promotion of breastfeeding intervention trial (PROBIT): A randomized trial in the Republic of Belarus. *Journal of the American Medical Association* **285**(4): 413–420.

Kramer MS and Kakuma R (2001) *The Optimal Duration of Exclusive Breastfeeding: A Systematic Review*, WHO/NHD/01.08, WHO/FCH/CAH/01.23. Geneva: World Health Organization.

Lawrence RA and Lawrence RM (1999) *Breastfeeding: A Guide for the Medical Profession*. St. Louis: Mosby.

Mortensen EL, Michaelsen KF, Sanders SA *et al.* (2002) The association between duration of breastfeeding and adult intelligence. *Journal of the American Medical Association* **287**: 2365–2371.

Victora CG, Vaughan JP, Lombardi *et al.* (1987) Evidence for protection by breast-feeding against infant deaths from infectious diseases in Brazil. *Lancet* **2**(8554): 319–321.

Von Kries R, Koletzko B, Sauerwald T *et al.* (1999) Breast feeding and obesity: Cross sectional study. *British Medical Journal* **319**: 147.

World Health Organization (1981) *The International Code of Marketing of Breast-Milk Substitutes*. Geneva: World Health Organization.

World Health Organization (1989) *Protecting, Promoting, and Supporting Breastfeeding: The Special Role of Maternity Services*. Geneva: World Health Organization.

World Health Organization (2000) WHO Collaborative Study Team on the role of breastfeeding on the prevention of infant mortality: Effect of breastfeeding on infant and child mortality due to infectious diseases in less developed countries: A pooled analysis. *Lancet* **355**: 451–455.

BURNS PATIENTS

S A Hill, Southampton General Hospital, Southampton, UK

Of all insults to the body, burns elicit the most profound stress response. This response encompasses hormonal, metabolic, and immunologic changes, which are complicated by the loss of the many protective functions of an intact skin. The initial hypermetabolic state induces intense protein catabolism, which must be checked by aggressive nutritional support in order to limit morbidity and mortality. Following this intense period of catabolism, which lasts 10–14 days, there is a gradual reduction in the metabolic rate, catabolic processes decrease, the wound heals, and anabolic processes predominate. Nutritional support must also continue throughout these latter stages, which may take many weeks, so that anabolism can be supported and fuel reserves replenished. A diet high in carbohydrate (CHO) and protein but low in fat will not only suppress catabolism and favor anabolism but also support an appropriate immunological response.

Hypermetabolism and Hypercatabolism

The burn wound is the focal point of all the circulatory, metabolic, and inflammatory responses associated with injury (**Figure 1**).

Metabolic Response

Increased glucose demand is initially met by glycogenolysis. When glycogen stores are exhausted, lipolysis and protein catabolism increase to supply gluconeogenic substrates. This hypermetabolic response is accompanied by increased cardiac output, increased oxygen consumption, and increased thermogenesis. The physical loss of skin cover has other major effects, including fluid loss, increased heat loss by evaporation, and loss of local immune function. Skin grafting will provide some cover for burned areas but the use of allografts increases the total area of damaged skin. Early excision and grafting is associated with increased survival in patients with more than 70% burns compared with conservative management. Donor split-skin graft sites heal within 7–14 days, unlike the burned area, which continues to make increased metabolic demands for weeks after the initial insult. Measurements of energy requirements take into account whole body metabolism and include any demands made by donor sites. Recent work suggests that early coverage with artificial 'skin' may reduce the hypercaloric requirements of the patient.

The sympathoadrenal axis is stimulated as a result of thermal injury, with increased plasma levels of epinephrine, norepinephrine, and cortisol. In addition, levels of both glucagon and insulin

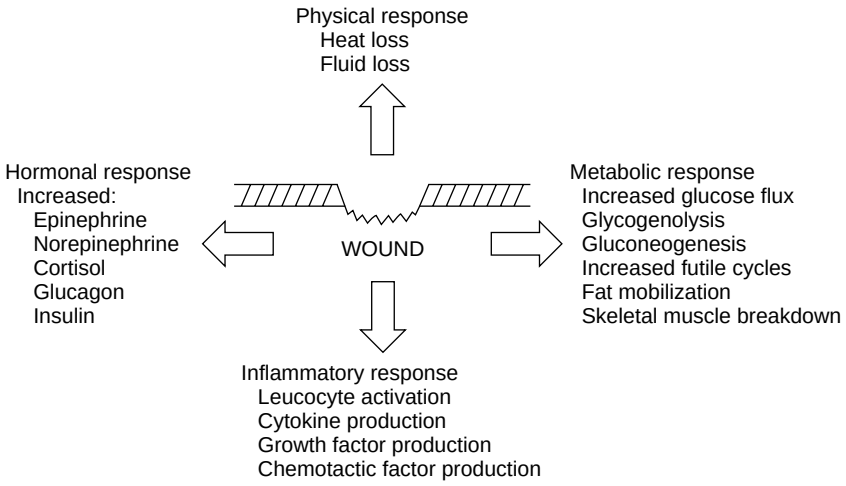


Figure 1 Schematic representation of the central nature of the burn wound in triggering physical, hormonal, metabolic, and inflammatory responses.

increase, although an apparent insulin resistance develops in peripheral tissues. There is an increase in core temperature, which appears to be mediated centrally by the hypothalamus in response to cytokine release, possibly interleukin-1 (IL-1). Plasma levels of glucose are maintained and may even increase, although glucose flux is greatly increased. Metabolic demands for glucose and amino acids increase and the body responds to meet these demands (**Table 1**). The degree of hypermetabolism and oxygen consumption are closely related to the extent and depth of burn injury. As a result, basal energy expenditure increases and is doubled for a 60% burn (**Figure 2**). Catecholamines augment cardiac and circulatory performance, which increases blood flow to the wound. Liver and kidney blood flow also increase, with increased delivery of gluconeogenic precursors, increased glucose release into the circulation, and increased nitrogen clearance. The release of fat from adipose tissue is stimulated by catecholamines. In the liver, fat metabolism to glycerol and free fatty acids produces energy as long as an adequate glucose supply replenishes oxaloacetic acid for oxidation of acetyl CoA, the product of triacylglycerol oxidation.

Table 1 Metabolic and circulatory responses to burn injury

Wound	*Whole body*
Damage	Increase in catecholamines, cortisol, glucagon, insulin
	Hepatic switch to synthesis of acute phase proteins
Increased blood flow to the wound	Increased cardiac output
Increased metabolism of glucose	Increased gluconeogenesis
	Increased free fatty acid flux
	Increased oxygen consumption
	Futile substrate cycling of carbohydrate intermediates and fatty acids
Increased heat loss	Increased core temperature: hypothalamic mediated
Attempted repair	Increased amino acid flux
	Release of arginine and glutamine from skeletal muscle
	Increased nitrogen loss
Inflammatory response	Inhibition of maximum inflammatory response by cortisol
	Cytokine and eicosanoids increase inflammation

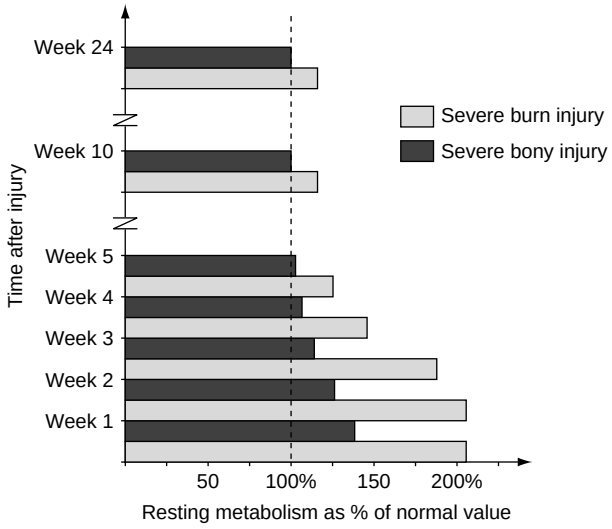


Figure 2 The stress response to thermal injury is greater than that to any other insult. The resting energy requirements in burns patients are greater for longer than for any other injury. (Adapted with permission from Long CL, Schaffel N, Geiger JW *et al.* (1979) Metabolic response to injury and illness: Estimation of energy and protein needs from indirect calorimetry and nitrogen balance. *Journal of Parenteral and Enteral Nutrition* **3**(6): 452–456.)

Thus, both continued oxygen and glucose must be supplied to prevent ketoacidosis. Heat production and energy wastage occur as a result of a 2- or 3-fold increase in futile cycling of substrates; glucose, pyruvate, and fructose-6-phosphate are all involved in these reactions.

Gluconeogenesis can occur only in the liver and is increased by catecholamines and glucagon. The plasma levels of gluconeogenic amino acids (alanine and glutamine) initially increase during the first 2 days, when glycogen is preferentially metabolized, but subsequently decrease. Days 4–7 are associated with a maximal decrease in plasma levels of gluconeogenic amino acids, whereas muscle production and hepatic consumption are both increased.

Catabolic Response

Release of gluconeogenic amino acids from skeletal muscle results in loss of muscle mass. Deamination of these amino acids, during the generation of carbon skeletons for glucose synthesis, increases nitrogen production with subsequent conversion to urea, which is excreted by the kidneys. Urinary nitrogen loss following thermal injury is largely from skeletal muscle breakdown, but a significant contribution of approximately 25–30% comes from the burn exudate. The rate of nitrogen loss is related to total burn area (TBA) and can be as much as 3 or 4 g/kg/day at its peak (**Figure 3**). These high rates of nitrogen loss persist for the first 7–10 days and then gradually decline until the burn area is healed and body stores of nitrogen are replenished; this may take many months. Nutritional support provides an exogenous source of calories and protein, which limits autocanabalism of skeletal muscle with a reduction in infective complications, increased survival rates, and reduced hospital stay.

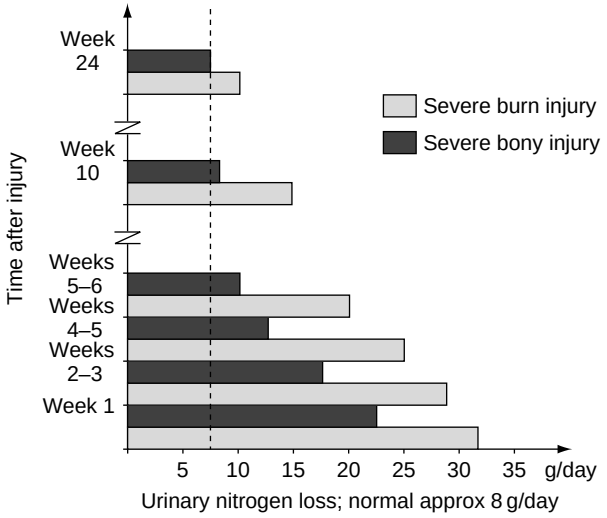


Figure 3 Nitrogen losses increase in thermally injured patients and remain high for a longer period than for any other injury. (Adapted with permission from Long CL, Schaffel N, Geiger JW *et al.* (1979) Metabolic response to injury and illness: Estimation of energy and protein needs from indirect calorimetry and nitrogen balance. *Journal of Parenteral and Enteral Nutrition* **3**(6): 452–456.)

Inflammatory Response

Local and systemic factors lead to an increase in inflammatory cell infiltration of the burn wound with removal of damaged tissue in preparation for epithelialization and wound healing. The cytokine cascade precipitated by stress hormones in response to injury involves tumor necrosis factor-α (TNF-α), which subsequently activates leucocyte production of IL-1, IL-6, platelet-derived growth factor, and eicosanoids (prostaglandins, thromboxanes, and leukotrienes). These in turn amplify the cellular and cytokine responses by release of chemotactic factors. TNF-α and the cytokines are also mediators of the increased metabolic and catabolic response seen in burns patients. TNF-α will increase acute-phase protein synthesis and increase loss of amino acids from skeletal muscle. However, it is also associated with the healing process in that wound healing is stimulated as a result of vascular proliferation and collagen synthesis. Excessive release of TNF-α can be harmful and is associated with muscle wasting, excessive weight loss, increased nitrogen loss, systemic inflammatory response syndrome, and debility.

These early immune reactions give way, in the second and third week postburn, to injury immunosuppression. This takes the form of reduced responsiveness of lymphocytes, impaired production of IL-2, and changes in immune cell phenotypes. Burn injury inhibits the T-helper 1 response but promotes a T-helper 2 response. As a result, IL-2 and interferon-γ (IFN-γ) production is reduced, which increases the risk of infection. Membrane lipid composition also influences lymphocyte and macrophage functions in terms of signaling and eicosaniod production. There is a reduction in *n*-6 (mainly arachidonic) fatty acids and an increase in PGE_2, which can lead to immunosuppression. Dietary replacement of *n*-6 by *n*-3 polyunsaturated fatty acids (PUFAs) reduces immunosuppression by altering membrane composition and eicosanoid series production. Nutritional studies have focused on the influence that enteral feed composition has on immune function—so-called immune-enhancing diets. The theoretical elements of interest are *n*-3 PUFAs and the amino acids arginine and glutamine. The evidence base for a clinically measurable

advantage of such immune-enhancing preparations is yet to be fully established. Small studies have shown a reduction in wound infection rate and a possible reduction in length of stay per percentage total body surface area burned.

Nutritional Requirements

In the absence of an exogenous nutrient supply, autocanabalism would result in major morbidity and mortality. Nutritional supply of energy and protein in excess of normal requirements prevents such complications.

Calories

Adults Much work has focused on developing an easy-to-apply formula for predicting the number of calories required to maintain weight in the severely burned patient. Many of these formulas are simply based on percentage burn area and body surface area, but others are complex—arrived at by regression analysis. Recent evaluations of these formulas, compared to indirect calorimetry, suggest that none perfectly predict a patient's true requirements. The most reliable are summarized in **Table 2**. Requirements vary considerably depending on multiple factors including time after burn injury, not all of which can be taken into consideration in the formulas used (**Table 3**). Because of major within- as well as between-patient variation, it is agreed that indirect calorimetry should be used to predict resting energy requirements (REEs) throughout the recovery period.

By measuring oxygen uptake and carbon dioxide production, the REE of the patient can be derived. Recent work has established that although body weight can be maintained on a regimen of a caloric intake of 1.3–1.5 times the REE, this reflects increasing fat accumulation despite persistent catabolism of skeletal protein. The catabolism appears to persist despite nutritional manipulation. It has been suggested that lean mass can only be maintained by pharmacological means with the use of insulin, insulin-like growth factor-1 (IGF-1), or anabolic steroids such as oxandrone.

Table 2 Formulas for predicting calorie requirements in adults that compare favorably with indirect calorimetric measurements (1.3 × REE)

Xie formula[a]
$1000 + (25 \times \text{BSAB})$ kcal/m^2/day *or* $4184 + (105 \times \text{BSAB})$ kJ/ m^2/day
Zawacki formula[b]
1440 kcal/m^2/day *or* 6025 kJ/ m^2/day
Milner formula[c]
$(\text{BMR} \times 24 \times \text{BSA}) + (0.274 + (0.0079 \times \text{BSAB}) - (0.004 \times \text{DPB}))$ kcal/m^2/day *or* $(\text{BMR} \times 24 \times \text{BSA}) + (1.146 + (0.0331 \times \text{BSAB}) - (0.017 \times \text{DPB}))$ kJ/m^2/day

[a]Xie WG *et al*. Estimation of the calorie requirements of burned Chinese adults. *Burns*. 1993, **19**: 146–9.
[b]Zawacki BE *et al*. Does increased evaporative water loss cause hypermetabolism in burned patients? *Annals of Surgery*. 1970, **171**: 236–40.
[c]Milner EA, Cioffi WG, Mason AD, McManus WF, and Pruitt BA Jr. A longitudinal study of resting energy expenditure in thermally injured patients. [Journal Article] *Journal of Trauma-Injury Infection & Critical Care*. **37**(2): 167–70.
BSAB, body surface area burned; BMR, basal metabolic rate; DPB, days postburn.

Table 3 Factors that may alter total energy requirements in burns patients

Number of days after burn injury
Changes in environmental temperature and humidity
Changes in core body temperature, including sepsis, infection
Inhalation injury
Activity level
Surgical interventions; grafting
Dressing changes
Pain and anxiety
Sedative drugs

Children Pediatric patients, who account for at least 35% of all burn injuries, are a challenge to nutritional support teams. Compared to adults, they have lower lean body mass and fat reserves and have a higher basal metabolic rate. Extra allowances are needed for growth and development, particularly during the infant and adolescent growth spurts. Many different pediatric formulas are used (**Table 4**). Indirect calorimetry indicates that those predicting lower calorie requirements may be more accurate.

For pediatric patients, for whom requirements have been particularly difficult to predict using formulas, indirect calorimetry has been of great importance in determining adequate calorie intake and measured energy expenditure should be multiplied by a factor of 1.5 to provide adequate calories for weight maintenance in children with burns.

Balance of Energy-Producing Substrates

Energy requirements may be met by glucose, fat, and protein. There has been much interest in the relative proportions of these three sources, and it now seems that 20–25% should be supplied as protein but only 15% as fat.

Table 4 Examples of formulas used to predict energy requirements in children

Wolfe[a]

BMR × 2 kcal/24 h

Males	*BMR equation*	*Females*	*BMR equation*
0–3	$(60.9 \times W) - 54$	0–3	$(61 \times W) - 51$
3–10	$(22.7 \times W) + 459$	3–10	$(22.5 \times W) + 499$
10–18	$(17.5 \times W) + 651$	10–18	$(12.2 \times W) + 746$

Modified Galveston formulas[b]

Less than 1 year old: 2100 kcal/m^2 BSA + 1000 kcal/m^2 BSA burned

Less than 12 years old: 1800 kcal/m^2 BSA + 1300 kcal/m^2 BSA burned

12–18 years old: 1500 kcal/m^2 BSA + 1500 kcal/m^2 BSA burned

Curreri junior formulas[c]

Daily calorie needs[d]

Birth–1 year: basal RDA in kcal + (15 kcal/% burn)

1–3 years: basal RDA in kcal + (25 kcal/% burn)

4–15 years: basal RDA in kcal + (40 kcal/% burn)

[a]O'Neil CE, Hutsler D, and Hildreth MA. Basic nutritional guidelines for pediatric burn patients. *Journal of Burn Care & Rehabilitation*. 1989, **10**(3): 278–84.

[b]Hildreth MA, Herndon DN, Desai MH, and Duke MA. Reassessing caloric requirements in pediatric burn patients. *Journal of Burn Care Rehabilitation* 1988, **9**(6): 616–8.

[c]Day T, Dean P, Adams MC, Luterman A, Ramenofsky ML, and Curreri PW. Nutritional requirements of the burned child: the Curreri junior formula. *Proceedings of the American Burn Association* 1986, **18**: 86.

[d]RDA (kcal) varies with age: 0–0.5 years, 320; 0.5–1 years, 500; 1–3 years, 740; 4–6 years, 950; 7–10 years, 1130; 11–14 years, 1440 (male) and 1310 (female); 15–18 years, 1760 (male) and 1370 (female).

BMR, basal metabolic rate; BSA, body surface area; RDA, recommended daily allowance; *W*, weight in kilograms.

Carbohydrate

Adults A high carbohydrate:fat ratio is associated with better maintenance of body weight. However, this may reflect increased fat accumulation rather than an increase in protein synthesis. Hyperglycemia alone can increase alanine efflux from skeletal muscle, without stimulating protein synthesis. Euglycemia, using exogenous insulin with high glucose delivery, can inhibit amino acid oxidation and favor amino acid synthesis. This may reflect an effect of IGF-1, which is released in response to insulin. In addition, hyperglycemia stimulates hepatic lipogenesis and increased CO_2 production, which may prevent weaning from ventilatory support. Hyperglycemia must therefore be prevented.

Children In children, carbohydrate is more effective than fat in promoting nitrogen retention by reducing the need for protein catabolism and subsequent gluconeogenesis. In infants, 5% dextrose in water parenterally can be used at 5 mg/kg/min initially and increased to a maximum of 15 mg/kg/min over the course of the first few days postinjury to provide 40–50% of calorie requirements. In older children, as in adults, glucose administration at a maximum rate of 5–7 mg/kg/min is recommended. These are parenteral recommendations; enteral feeding guidelines have not been established, although in general, carbohydrate should be limited to 50% of calorie intake.

Fat

Adults Feeding regimens that simply overfeed with a normal diet lead to problems in the recovery phase; muscle wasting persists together with central obesity. Reduction in fat administration largely prevents these problems if protein calories replace lipid calories (**Figure 4**). Fat cannot be excluded from the diet; a minimum fat content of 4% of total calories will ensure a supply of essential fatty acids. A diet containing 15% fat will meet such requirements as well as provide a delivery medium for fat-soluble vitamins. Dietary fat is largely composed of long-chain triacylglycerols (LCTs), and excess LCTs are associated with hepatic steatosis, reticuloendothelial system blockage, and immunosuppression. Varying the composition of the fats supplied to burns patients may alleviate some of these problems. Medium-chain fatty acids, particularly *n*-3 PUFA found in fish oil, appear beneficial in maximizing whole body protein synthesis in an animal model of burn injury. There is a decrease in plasma *n*-6 fatty acids after burn injury, so replacement with *n*-3 PUFA results in the production of prostanoid and eicosanoid series associated with less immunosuppression than those arising from *n*-6 PUFA metabolism. The use of low-fat feed, supplemented with *n*-3 PUFAs, reduces protein catabolism and increases IGF-1, particularly 2 or 3 weeks

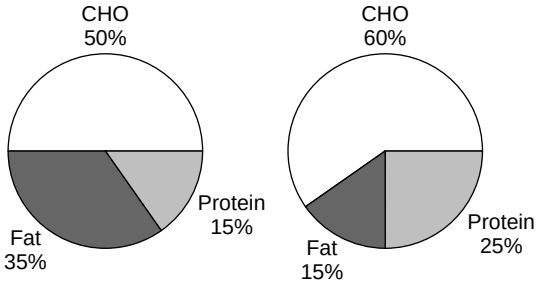


Figure 4 Pie charts demonstrating the difference in proportion of carbohydrate, fat, and protein calories required by burns patients (right) compared to normal patients (left).

postburn. Limiting fat to 15% total calorie intake reduces wound infection rate, improves healing rate, reduces the incidence of pneumonia, improves nutritional markers, and reduces hospital stay. Such benefits have been seen both with and without the addition of *n*-3 PUFAs.

Children A fat intake of 2 or 3% of total calories is the minimum recommended for the prevention of essential fatty acid deficiency in pediatric patients. For intravenous fat administration in infants, a maximum of 4 g/kg lean body weight is suggested.

Protein

Adults Protein calories comprise a significant proportion of the energy requirement of a severely burned patient. Intact protein, rather than amino acids, is associated with better weight maintenance and improved survival. Nitrogen loss must be estimated regularly in a burn patient in order to ensure adequate nitrogen replacement. Total nitrogen loss (TNL) is impossible to measure accurately since 20–30% of nitrogen loss occurs in the exudate from wounds. There is some doubt regarding the use of urinary urea nitrogen (UUN) to estimate total urinary nitrogen (TUN), from which TNL is usually calculated. In healthy, unstressed subjects, urea comprises 80% of the TUN, but ureagenesis is inconsistent after burn injury and varies widely depending on the extent and course of illness. If measurement of TUN is available this will reflect nitrogen loss more accurately:

$$\text{TNL} = \text{TUN} + 4\,\text{g/day or } (\text{UUN} \times 1.25) + 4\,\text{g/day}$$

Total nitrogen loss must then be compared with nitrogen supply (NS) to calculate the nitrogen balance (NB):

$$\text{NB} = \text{TNL} - \text{NS}$$

The aim is to keep a positive balance, and a suitable starting point would be 2 or 3 g protein/kg lean body weight/day; 6.25 g protein is equivalent to 1 g nitrogen. Urinary excretion of 3-methyl histidine has been used as a measure of skeletal protein catabolism. Nitrogen input from blood products is appreciable, accounting for 15% of total nitrogen intake, but is often ignored when calculating nitrogen balance, which therefore underestimates protein intake.

Amino acids play an important role in adaption to burn injury both as gluconeogenic substrates and as substrates for acute phase protein synthesis and wound repair. Arginine flux appears to be increased in burns patients, but plasma levels of arginine and glutamine appear to be greatly reduced following burn injury. These changes have prompted supplementary feeding with particular amino acids. Interest has focused on ornithine α-ketoglutarate (OAK) and its metabolites arginine and glutamine. OAK is also a precursor for proline, the incorporation of which into collagen is a rate-limiting step in collagen synthesis. Arginine also increases collagen deposition, and in animal models of burn injury increased arginine provision has been associated with increased wound healing and improved immune function. There is evidence of a clinically important reduction in healing time and infectious episodes following OAK-, arginine-, or glutamine-supplemented feeding in human studies. Glutamine is the most abundant amino acid in the body and is the major fuel source for the intestine. Its presence prevents villous atrophy and maintains mucosal integrity as well as stimulates blood flow to the gut. Glutamine supplementation of feed reduces the incidence of gram-negative bacteremia in patients with severe burns. The proposed mechanism is the reduction in bacterial translocation across the gut wall; glutamine has been shown to reduce bacterial translocation in a rat model. However, the evidence for a significant clinical effect of arginine- or glutamine-supplemented enteral feed is still equivocal.

Children Protein needs are often estimated by formulas based on lean body weight. To estimate preburn weight, the 50th centile weight for height should be used. For children younger than 1 year old, 3 or 4 g protein/kg lean body weight is suggested to provide adequate nutritional support for graft coverage and healing; this should be reduced to 2.5–3 g protein/kg lean body weight for children 1–3 years of age. In older children, protein requirements are further reduced to 1.5–2.5 g/kg lean body weight. When nitrogen balance is calculated for children, the following formulas have been suggested:

$$\text{Age } 0{-}4\,\text{years: } \text{TNL} = \text{UUN} + 2$$

$$\text{Age } 4{-}10\,\text{years: } \text{TNL} = \text{UUN} + 3$$

$$\text{Age } >10\,\text{years: } \text{TNL} = \text{UUN} + 4$$

Protein-enriched diets, containing 25% calories as protein compared to 16% in normal diets, have been associated with improved nitrogen balance, improved immune function, and fewer infective episodes in children with severe burns. Until recently, albumin was a mainstay of fluid requirements in children with severe burns, which contributes to protein provision. However, studies have shown increased morbidity and mortality in critically ill

patients given albumin. It seems that the outcome for children with severe burns is no worse if they receive albumin supplementation only when albumin levels decline below 10 g/l (or 15 g/l in the presence of intolerance of enteral feed). Use of albumin should be reviewed regularly.

Vitamins

Specific vitamin requirements in burns patients have not been established, but levels may decline in the hypermetabolic state. As a minimum, the recommended daily allowance should be given following injury. For minor burns of 10–20% body surface area, supplementation with a single multivitamin tablet orally should replace vitamin losses sustained during injury. For larger burns, additional supplementation is advised, especially vitamin C (ascorbic acid), which is of benefit in wound healing and has experimentally been shown to possess free radical scavenging properties that may help to limit tissue damage. A recommended dose is 1 g daily in two divided doses for all patients with major burns; children younger than age 3 years should receive half this dose daily. The essential, fat-soluble, vitamin A may also confer some advantages in wound healing and immunomodulation. A dose of 10 000 IU daily is recommended for all patients with major burns who are older than the age of 3 years; younger children should receive half this amount daily. Wherever possible, both vitamin and micronutrients should be administered by the enteral route. If such supplements are added to total parenteral nutrition (TPN), dosing schedules should take into account the increased bioavailablity via this route and a dose reduction may be advisable. Monitoring of levels of micronutrients should guide replacement by the parenteral route.

Trace Elements

Trace elements are present in the body in amounts less than one part per million by weight; many are essential components of metalloenzymes. Following burn injury, significant amounts of these trace elements may be lost. The acute phase reaction is characterized by a decrease in plasma levels of copper, iron, selenium, and zinc and an increase in the plasma levels of the carrier proteins ferritin and caeruloplasmin. Although iron levels decline following burn trauma, it has been shown that excessive administration of iron is harmful and that low plasma levels of iron appear to be of benefit in reducing microbial replication. In contrast, increased intravenous administration of copper, zinc, and selenium during the first week following burn injury resulted in fewer complications, an improved leucocyte response, a rapid return of the plasma levels of these trace minerals, as well as a shorter hospital stay. Zinc, copper, and manganese are essential for wound healing; serum zinc levels decrease following stress and burn injury largely due to increased urinary loss. Zinc supplementation, 220 mg daily, should be considered for all patients with major burns. Inclusion of trace elements in both enteral and parenteral nutrition is essential.

Adjunctive Treatments

There is great interest in a pharmacological role for growth hormone (GH) and/or IGF-1 in reversal of the catabolic state and stimulation of anabolic processes. GH stimulates production of IGF-1, which improves amino acid transport and enhances gluconeogenesis from exogenously supplied amino acids. Blood levels of IGF-1 are markedly reduced in burn patients following injury and remain so for the first week, after which levels increase; these changes correlate with IGF binding protein-3 levels. This binding protein prevents plasma proteolysis of IGF-1. GH and IGF-1 have both been used in experimental models of burn injury, and their effectiveness at limiting catabolism and enhancing mucosal proliferation is encouraging. In children GH treatment accelerates donor site healing and increases protein synthesis. GH has also been shown to exert immunomodulatory effects, which may contribute to a reduced incidence of infection. Other growth factors have also been used experimentally and in animal models improve the rate of healing and strength of burn wounds (**Table 5**). However, the role for such hormonal therapies has yet to be firmly established in clinical management plans.

Table 5 Growth factors identified as potential adjunctive therapy in wound healing

Growth hormone
Insulin-like growth factor-1
Epidermal growth factor
Transforming growth factors (α and β)
Platelet-derived growth factor
Fibroblast growth factors (1–7)
Erythropoeitin
Granulocyte macrophage colony-stimulating factor

Nutritional Management

Nutritional management of the burn patient is an important facet of overall management. It is important to involve a multidisciplinary feeding team to manage nutrient intake and organize nutritional assessment. A warm, ambient temperature is essential for reducing fluid and heat loss and keeping the patient comfortable. Metabolic rate increases with discomfort and dressing changes can be a continual source of stress. Thus, analgesic requirements must be adequate or anesthesia administered. There has been a move toward continuing enteral feed in the immediate perioperative period. The risk of aspiration seems to be very low, particularly if jejunal feeding is used. Even when nasogastric feeding is used, starvation times can be minimized without apparent increased risk. As a result, 60% rather than 6% of caloric requirements can be met on days of surgery/dressing changes.

Route of Feeding

Wherever possible, the enteral route should be used. The American Gastroenterological Association has strongly endorsed this view and stated that routine parenteral nutrition is contraindicated if the enteral route is available. Nasogastric, nasojejunal, and percutaneous enteral access tubes have all been used successfully when feeding is introduced as soon after burn injury as possible. Jejunal feeding is associated with a higher success rate than gastric feeding and may be continued even in the presence of gastric stasis. Increased mortality has been associated with

Table 6 Scheme for nutritional monitoring in a patient with a burn injury

Day 0	Record: a. Age b. Height c. Urine output d. Oral fluid/food intake	Estimate: a. Ideal body weight b. % TBA c. Need for dressings/ surgery/activity d. Fluid loss
	Investigate: a. Plasma prealbumin b. Electrolytes and urea c. Hemoglobin and hematocrit	
	Intervention: a. If burn area >20%, place nasoenteral feeding tube, nasojejunal if possible, start feeding according to calculated values b. Intravenous crystalloid, blood, and colloid according to center protocol	
Day 2/3	Investigate: a. Indirect calorimetry, energy requirement, calorie balance b. 24-h UUN c. Urinary myoglobin d. Hematocrit	
	Intervention: a. Adjust calorie intake to match calculated need b. If intolerant of enteral feeding, reduce to 10–30 ml/h and start TPN to supplement calorie and nitrogen delivery	
Day 4/5 and twice weekly	Investigate: a. Plasma prealbumin b. Hematocrit c. Indirect calorimetry, as above	
	Intervention: a. Calculate nitrogen balance; adjust nitrogen intake b. Adjust calorie intake	
Day 6/7 and weekly	Investigate: a. Weight b. Trace elements c. Weekly need for dressing changes and/or surgery d. Is enteral absorption improving? Stop/reduce TPN e. Is oral intake increasing? Is enteral supplementation still needed?	
	Intervention: a. Adjust calorie intake to account for (c) above b. Add trace elements if indicated c. Review route(s) of feeding	

the use of central venous catheters and TPN in patients with severe burn injury. This is related to both catheter-associated morbidity and depression of gut function. Glutamine is relatively unstable and has not been included in parenteral formulations. New preparations containing the dipeptide or acetylated form of glutamine will be available in the future that may be of benefit to patients who are dependent on TPN.

Patients with burn injuries greater than 10% are often unable or unwilling to increase their oral intake to meet calorie needs, which are higher by a factor of 1.3 compared to normal. For patients with <20% burns, calorie intake can often be met by supplemental nocturnal feeding through a fine-bore feeding tube. For burns >20%, nasoenteral supplementation is essential. Evidence suggests that the earlier nutrition is started, the greater the attenuation of the hypermetabolic and catabolic response. Early feeding, within 6 h of injury, is optimal. Enteral delivery of glucose and glutamine maintains mucosal integrity and reduces gut ischaemia, as shown by a reduction in arterial-to-intraluminal CO_2 gap. This latter measure can identify an imbalance between calories presented and oxygen delivered to the gut and may be used to adjust enteral feeding levels to prevent excessive delivery.

Some patients are intolerant of enteral feeding, especially those needing mechanical ventilation, who require high levels of opiate analgesia and exogenous norepinephrine support. In these patients, TPN is needed. Wherever possible, a slow, continuous presentation of enteral feed should also be provided to prevent intestinal mucosal atrophy and preserve immune function. Whichever routes are required, calorie provision should be guided by nutritional monitoring of energy expenditure and nitrogen balance.

Nutritional Monitoring

There are a variety of ways to monitor nutritional status in the burn patient; body weight, nutritional intake, nitrogen balance, and laboratory indices all play a role in nutritional assessment. A scheme for continuing nutritional assessment is given in **Table 6**.

Success of the chosen nutrition plan must be monitored and the plan adjusted accordingly. Energy expenditure and nitrogen losses should be measured once or twice weekly for calorie and nitrogen balance calculations. The choice of additional biochemical markers to assess overall nutritional state is difficult and no one marker adequately predicts actual nutritional state at all times during the course of injury. The constitutively produced prealbumin, with its short half-life and independence from exogenous albumin administration, is often used.

Pediatric patients often exhibit growth delays and special attention must be paid to younger children who are less than their ideal body weight at the time of injury. Nutritional support is often required for many weeks after discharge from hospital, and outpatient follow-up of growth in children is essential.

See also: **Amino Acids**: Metabolism. **Cytokines**. **Nutritional Support**: Adults, Enteral.

Further Reading

ASPEN board of directors (2002) Guidelines for the use of parenteral and enteral nutrition in adult and pediatric patients. *Journal of Parenteral and Enteral Nutrition* **26**(1, supplement): 88SA–90SA.

Bessey PQ and Wilmore DW (1988) The burned patient. In: Kinney JM, Jeejeebhoy KN, Hill GL *et al.* (eds.) *Nutrition and Metabolism in Patient Care*, pp. 672–700. Philadelphia: WB Saunders.

Cynober L (1989) Amino acid metabolism in thermal burns. *Journal of Parenteral and Enteral Nutrition* **13**(2): 196–205.

Dickerson RN, Gervasio JM, Riley ML *et al.* (2002) Accuracy of predictive methods to estimate resting energy expenditure of thermally-injured patients. *Journal of Parenteral and Enteral Nutrition* **26**(1): 17–29.

Gamliel Z, DeBiasse MA, and Demling RH (1996) Essential microminerals and their response to burn injury. *Journal of Burn Care and Rehabilitation* **17**: 264–272.

Herndon DN, Nguyen TT, and Gilpin DA (1993) Growth factors: Local and systemic. *Archives of Surgery* **128**: 1227–1234.

Martindale RG and Cresci GA (2001) Use of immune-enhancing diets in burns. *Journal of Parenteral and Enteral Nutrition* **25**(2): S24–S26.

Peck M (2001) American Burn Association Clinical Guidelines. Initial nutrition support of burn patients. *Journal of Burn Care and Rehabilitation* **22**: 595–665.

Pratt VC, Tredget EE, Clandin T *et al.* (2002) Alterations in lymphocyte function and relation to phospholipid composition after burn injury in humans. *Critical Care Medicine* **30**(8): 1753–1761.

Rettmer RL, Williamson JC, Labbé RF *et al.* (1992) Laboratory monitoring of nutritional status in burn patients. *Clinical Chemistry* **38**(3): 334–337.

Rodriguez DJ (1996) Nutrition in patients with severe burns: State of the art. *Journal of Burn Care and Rehabilitation* **17**: 62–70.

C

CAFFEINE

M J Arnaud, Nestle S.A., Vevey, Switzerland

During the period 1820–1827, three white crystalline substances called 'caffein' or 'coffein,' 'guaranin,' and 'thein' were isolated from green coffee beans, guarana, and tea, respectively. These substances were shown in 1838–1840 to be identical. Later, caffeine was also discovered in maté prepared from *Ilex paraguariensis* and kola nuts. Since then, caffeine has been shown to be a natural constituent of more than 60 plant species.

Two other related compounds, theophylline and theobromine, have also been isolated from tea and cocoa beans, respectively while a third, paraxanthine, was isolated from human urine (**Figure 1**). By the end of the nineteenth century, all of these methylated xanthines had been synthesized. Caffeine, both natural and synthetic, has been used as a flavoring agent in food and beverages and as an active component of a variety of over-the-counter pharmaceutical products and drugs. A regulation adopted by the European Commission requires the compulsory labeling 'high caffeine content' when soft drinks contain more than 150 mg caffeine per liter.

In addition to natural caffeine obtained by the industrial decaffeination process, caffeine can also be obtained by the methylation of theobromine and also by total chemical synthesis using dimethylcarbamide and malonic acid.

Chemistry

Caffeine (M_r 194.19) is also called, more systematically, 1,3,7-trimethylxanthine, 1,3,7-trimethyl-2,6-dioxopurine, or 3,7-dihydro-1,3,7-trimethyl-l*H*-purine-2,6-dione and has been referred to as a purine alkaloid.

Caffeine is odourless and has a characteristic bitter taste. It is a white powder (density ($d^{18/4}$) 1.23) moderately soluble in organic solvents and water. However, its solubility in water is considerably increased at higher temperatures (1% (w/v) at 15 °C and 10% at 60 °C). Its melting point is 234–239 °C and the temperature of sublimation at atmospheric pressure is 178 °C. Caffeine is a very weak base, reacting with acids to yield readily hydrolyzed salts, and relatively stable in dilute acids and alkali. Caffeine forms unstable salts with acids and is decomposed by strong solutions of caustic alkali.

In aqueous solution, caffeine is nonionized at physiological pH. Dimers as well as polymers have been described. The solubility of caffeine in water is increased by the formation of benzoate, cinnamate, citrate, and salicylate complexes. In plants, chlorogenic acid, coumarin, isoeugenol, indolacetic acid, and anthocyanidin have been shown to complex with caffeine.

Caffeine exhibits an ultraviolet absorption spectrum with a maximum at 274 nm and an absorption coefficient of 9700 in aqueous solution. Upon crystallization from water, silky needles are obtained containing 6.9% water (a 4/5 hydrate).

Determination

Caffeine has traditionally been identified in foods by ultraviolet spectrophotometry of an organic solvent extract after suitable cleanup by column chromatography. Such methods tend to be laborious and may be subject to interference from other ultraviolet-absorbing compounds. Recently, high-performance liquid chromatography has been more extensively used. This technique, often in conjunction with solid phase extraction, can provide accurate data for the determination of caffeine in foods and physiological samples.

Absorption, Distribution, and Elimination

Following oral ingestion, caffeine is rapidly and virtually completely absorbed from the gastrointestinal tract into the bloodstream. Mean plasma

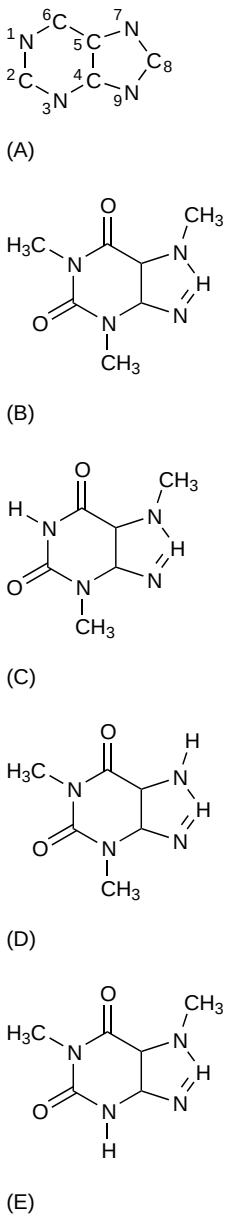


Figure 1 Chemical structures of caffeine and the dimethylxanthines. (A) Purine ring nomenclature according to E. Fischer; (B) caffeine; (C) theobromine; (D) theophylline; (E) paraxanthine. From Dews (1984).

concentrations of 8–10 mg l^{-1} are observed following oral or intravenous doses of 5–8 mg kg^{-1}. The plasma kinetics of caffeine can be influenced by a number of factors, including the total dose of caffeine, the presence of food in the stomach, and low pH values of drinks, which can modify gastric emptying. Caffeine enters the intracellular tissue water and is found in all body fluids: cerebrospinal fluid, saliva, bile, semen, breast milk, and umbilical cord blood. A higher fraction of the ingested dose of caffeine is recovered in sweat compared to urine. The fraction of caffeine bound to plasma protein varies from 10 to 30%.

There is no blood–brain barrier and no placental barrier limiting the passage of caffeine through tissues. Therefore, from mother to fetus and to the embryo, an equilibrium can be continuously maintained.

The elimination of caffeine is impaired in neonates because of their immature metabolizing hepatic enzyme systems. For example, plasma half-lives of 65–103 h in neonates have been reported compared to 3–6 h in adults and the elderly.

Gender, exercise, and thermal stress have no effect on caffeine pharmacokinetics in men and women. Cigarette smoking increases the elimination of caffeine, whereas decreases have been observed during late pregnancy or with the use of oral contraceptives and in patients with liver diseases. Drug interactions leading to impaired caffeine elimination are frequently reported.

There is no accumulation of caffeine or its metabolites in the body and less than 2% of caffeine is excreted unchanged in the urine. Some rate-limiting steps in caffeine metabolism, particularly demethylation into paraxanthine that is selectively catalyzed by *CYP1A2*, determine the rate of caffeine clearance and the dose-dependent pharmacokinetics in humans.

Important kinetic differences and variations in the quantitative as well as qualitative metabolic profiles have been shown between species, thus making extrapolation from one species to another very difficult. All of the metabolic transformations include multiple and separate pathways with demethylation to dimethyl- and monomethylxanthines, formation of dimethyl- and monomethylurates, and ring opening yielding substituted diaminouracils (**Figure 2**). The reverse biotransformation of theophylline to caffeine is demonstrated not only in infants but also in adults.

From metabolic studies, an isotopic caffeine breath test has been developed that detects impaired liver function using the quantitative formation of labeled carbon dioxide as an index. From the urinary excretion of an acetylated uracil metabolite, human acetylator phenotype can be easily identified and the analysis of the ratio of the urinary concentrations of other metabolites represents a sensitive test to determine the hepatic enzymatic activities of xanthine oxidase and microsomal 3-methyl demethylation, 7-methyl demethylation, and 8-hydroxylation. Quantitative analyses of paraxanthine urinary metabolites may be used as a biomarker of caffeine intake. Fecal excretion is a minor elimination route, with recovery of only 2–5% of the ingested dose.

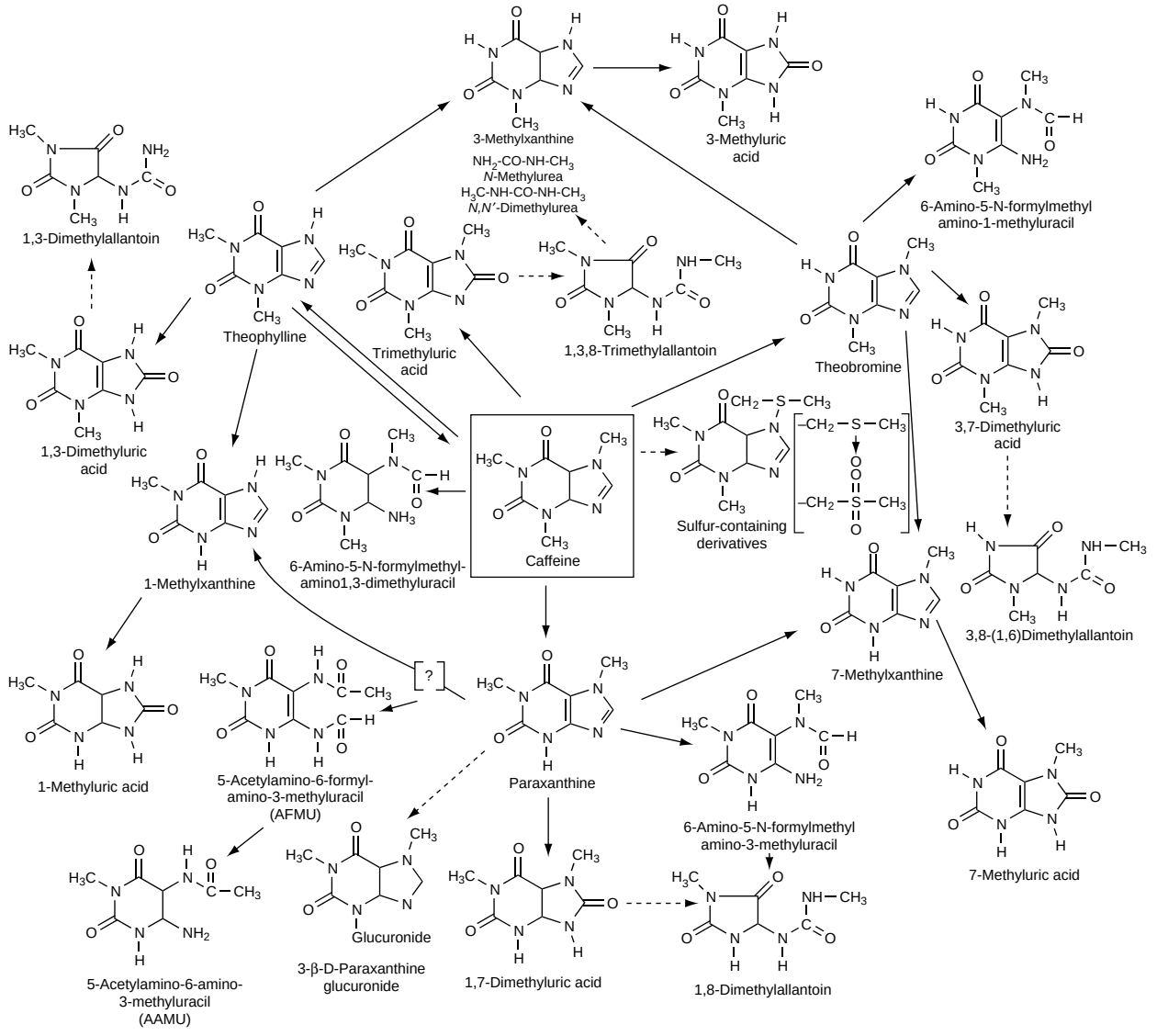


Figure 2 Metabolic pathways of caffeine in the human (⟶), the rat (⇢) and the mouse (⇢). From Garattini (1993).

Physiological and Pharmacological Properties

Because the physiological and pharmacological properties of caffeine represent the cumulative effects of not only the parent compound but also its metabolites, it is quite possible that effects attributed to caffeine *per se* are in fact mediated by one or more of its metabolites. It must also be noted that most of the knowledge about caffeine's effects has been derived from acute administration to fasted subjects submitted to a period of caffeine abstinence in order to ensure low plasma caffeine concentrations. It is thus difficult to extrapolate the results to the usual pattern of caffeine consumption in which most people consume it at different intervals throughout the day and over periods of years.

Effects on the Central Nervous System

Animal experiments have shown caffeine-mediated effects at the neuroendocrine level, such as increased serum corticosterone and β-endorphin and decreased serum growth hormone and thyrotropin, but it is expected that habitual human consumption has only marginal or inconsistent neuroendocrine effects. Caffeine is described as a central nervous system (CNS) stimulant, and the increased formation and release of neurotransmitters such as catecholamines, serotonin, γ-aminobutyric acid, norepinephrine, and acetyl-choline have been reported.

Behavioral effects can be observed in humans after acute and moderate doses of 1–5 mg kg^{-1} of caffeine. In these studies, the subjects felt more alert and active with improved cognitive function, including

vigilance, learning, memory, and mood state. It was claimed that they would be better able to cope with their jobs when bored or fatigued and after night work and sleep deprivation. Population-based studies of the effect of caffeine intake on cognition showed a positive trend, especially among elderly women. Comparative studies on regular and deprived caffeine consumers suggest that reversal of caffeine withdrawal is a major component of the effects of caffeine on mood and performance.

A dose-dependent delay in sleep onset is found as well as a decrease in total sleep time and an impairment of sleep quality characterized by an increased number of spontaneous awakenings and body movements. In premature infants, sleep organization appears to be unaffected by treatment with 5 mg/kg/day caffeine to prevent apnoea.

The observation that sensitive subjects are more likely to have trembling hands is considered to be a CNS effect and not a direct effect on muscle. Caffeine doses higher than 15 mg kg^{-1} induce headaches, jitteriness, nervousness, restlessness, irritability, tinnitus, muscle twitchings, and palpitations. These symptoms of chronic excessive caffeine intake are part of the criteria used to make the diagnosis of caffeinism. The same symptoms have been reported in adults on abrupt cessation of caffeine use.

With 100–200 mg kg^{-1} doses, mild delirium appears, followed by seizures and death. Although tolerance with low doses led to a pleasant stimulation, alertness, and performance benefits, on withdrawal, headache, drowsiness, fatigue, and anxiety were reported.

Epidemiology and laboratory studies suggest beneficial effects of caffeine consumption in the development of Parkinson's disease and the mechanisms involved may be mediated through adenosine A_{2A} receptors. The role of these receptors in neuronal injury and degeneration, as well as in other diseases such as Alzheimer's disease, has important therapeutic potential but needs further investigation.

Effects on the Cardiovascular System

Caffeine produces a direct stimulation of myocardial tissue leading to an increase in the rate and force of contraction. This direct cardiac effect can be inhibited by a depressant effect on the heart via medullary vagal stimulation. These opposing effects may explain why bradycardia, tachycardia, or no change can be observed in individuals receiving similar doses of caffeine. The traditional clinical view that caffeine induces arrhythmias in humans has not been confirmed by controlled experimental studies.

Caffeine decreases peripheral resistance by direct vasodilatation and increases blood flow to a small extent. This effect results from the relaxation of smooth muscle of blood vessels. For coronary arteries, vasodilatation is also observed *in vitro*, but the effects of caffeine in human coronary arteries *in vivo* are unknown. Different effects of caffeine on circulation can be observed in different vascular beds and, for example, the treatment of migraine headaches by caffeine is mediated through the vasoconstriction of cerebral arteries. It has also been shown that caffeine is capable of attenuating postprandial hypotension in patients with autonomic failure.

The observed cardiovascular effects consist of a 5–10% increase in both mean systolic and diastolic blood pressure for 1–3 h. A significant association was found between caffeine-related increase in systolic blood pressure and caffeine-related increase in pain tolerance. However, in contrast to the acute pressor effect reported, several epidemiological studies showed that habitual caffeine intake lowers blood pressure. Heart rate is decreased by 5–10% during the first hour, followed by an increase above baseline during the next 2 h. These effects are not detectable in regular coffee drinkers, suggesting that a complete tolerance can be developed. The tolerance to chronic caffeine intake can explain contradictory results reported in the literature. A few studies suggest that caffeine is partly responsible for the homocysteine-raising effect of coffee. This effect is associated with increased risk of cardiovascular disease, but it is uncertain whether this relation is causal.

Epidemiological studies designed to establish a relationship between caffeine intake and the incidences of myocardial infarction, mortality from ischaemic heart disease, or cerebrovascular accidents have provided conflicting results and have failed to establish a significant correlation.

Effects on Renal Functions

In humans, the administration of a single dose of 4 mg kg^{-1} caffeine increases the urinary excretion of sodium, calcium, magnesium, potassium, chloride, and urine volume. The mechanism of this mild diuresis has been attributed to an increase in renal blood flow, an increased glomerular filtration, and a decrease in tubular reabsorption of sodium ions and other ions. Although these effects appeared more pronounced for a higher acute dose of 10 mg kg^{-1}, a review concluded that caffeine consumption stimulates a mild diuresis similar to water. There was no evidence of a fluid–electrolyte

imbalance as well as disturbed thermoregulation, and caffeine was not detrimental to exercise performance or health.

Tolerance to the diuretic action of caffeine was demonstrated more than 50 years ago and was shown to develop on chronic caffeine intake so that the clinical significance of hypokalemia and calciuria is difficult to evaluate. Although controversial, some epidemiological studies have implicated caffeine in the increased risk for poor calcium retention. For calcium intakes lower than 750 mg per day, increased rate of bone loss and lower bone density were reported. However, it has been suggested that the effect on bone of high caffeine intake requires a genetic predisposition toward osteoporosis. In individuals who ingest calcium recommended daily allowances, there is no evidence of any effect of caffeine on bone status and calcium economy.

Effects on the Respiratory System

In caffeine-naive subjects, a dose of $4\,mg\,kg^{-1}$ increases the mean respiratory rate. This effect is not found in chronic caffeine ingestion. Several mechanisms have been suggested, such as an increase in pulmonary blood flow, an increased supply of air to the lungs due to the relaxation of bronchiolar and alveolar smooth muscle, an increase in sensitivity of the medullary respiratory center to carbon dioxide, stimulation of the central respiratory drive, an improved skeletal muscle contraction, and an increase in cardiac output.

At higher doses ($7\,mg\,kg^{-1}$), caffeine ingested by trained volunteers alters ventilatory and gas exchange kinetics during exercise, leading to a transient reduction in body carbon dioxide stores.

Effects on Muscles

Caffeine has been shown to have a bronchial and smooth muscle relaxant effect and to improve skeletal muscle contractility. Significant increases in hand tremor and forearm extensor electromyogram were observed in human subjects after the ingestion of $6\,mg\,kg^{-1}$ of caffeine. This effect is more likely due to a CNS stimulatory effect than to direct action on the muscle fibers. Skeletal muscle fatigue can be reversed by high concentrations of caffeine obtained only *in vitro* but not *in vivo*.

Effects on the Gastrointestinal System

Caffeine relaxes smooth muscle of the biliary and gastrointestinal tracts and has a weak effect on peristalsis. However, high doses can produce biphasic responses, with an initial contraction followed by relaxation. Caffeine seems to have no effect on the lower oesophageal sphincter. The increase in both gastric and pepsin secretions is linearly related to the plasma levels obtained after the administration of a dose of $4–8\,mg\,kg^{-1}$. In the small intestine, caffeine modifies the fluid exchange from a net absorption to a net excretion of water and sodium.

The role of caffeine in the pathogenesis of peptic ulcer and gastrointestinal complaints remains unclear, and no association has been found in clinical and epidemiological studies.

Effects on Energy Metabolism

Acute administration of caffeine produces a 5–25% increase in the basal metabolic rate. Inactive subjects exhibit a greater increase in resting metabolic rate than do exercise-trained subjects. It is concluded that endurance training seems to result in a reduced thermogenic response to a caffeine challenge.

These modifications of energy metabolism were associated with significant increases in serum free fatty acids, glycerol, and lactate concentrations, whereas inconsistent findings were reported for blood glucose levels. Acute administration of caffeine was shown to decrease insulin sensitivity and to impair glucose tolerance, possibly as a result of elevated plasma epinephrine. However, it is not understood why a large and long-term epidemiological study associated significant lower risks for type 2 diabetes in both men and women with total caffeine intake. The lipolytic effect is generally explained by the inhibition of phosphodiesterase, the release of catecholamine, or adenosine receptor antagonism. The increased availability of free fatty acids and their oxidation may have a glycogen-sparing effect. However, increasingly more results do not support the hypothesis that caffeine improves endurance performance by stimulating lipolysis, and some of the ergogenic effects in endurance exercise performance may occur directly at the skeletal muscle and CNS levels. In addition, this effect may be suppressed by the simultaneous ingestion of a high-carbohydrate meal, which is a common practice prior to competition.

Despite the controversy among scientists concerning the ergogenic potential of caffeine on sport performance, it is accepted that caffeine will not improve performance during short-term, high-intensity work, whereas an increase in both work output and endurance in long-term exercise is expected. Most studies also show that the duration and the magnitude of the ergogenic effect of caffeine are greater in nonusers than in users.

Based on the assumption that caffeine may enhance athletic performance, the International

Olympic Committee defined an upper concentration limit of 12 μg/ml in urine samples, above which an athlete was disqualified. However, in the World Anti-Doping Agency Executive Committee Meeting (September 2003), it was observed that the stimulant effect of caffeine is obtained at levels lower than 12. As a consequence, caffeine was removed from the 2004 list of prohibited substances because athletes must be allowed to behave like other people in society and may thus be allowed to drink coffee.

Safety and Toxicology

The acute oral LD_{50} (dose sufficient to kill one-half of the population of tested subjects) of caffeine is more than 200 mg kg^{-1} in rats, 230 mg kg^{-1} in hamsters and guinea pigs, 246 mg kg^{-1} in rabbits, and 127 mg kg^{-1} in mice. The sensitivity of rats to the lethal effects of caffeine increases with age, and higher toxicity is observed in male than in female rats.

Vomiting, abdominal pain, photophobia, palpitations, muscle twitching, convulsions, miosis, and unconsciousness were described in several reports of nonfatal caffeine poisonings in children who ingested 80 mg kg^{-1} caffeine. In several fatal accidental caffeine poisonings, cold chills, stomach cramps, tetanic spasms, and cyanosis were reported. The likely lethal dose in adult humans has been estimated to be approximately 10 g, which corresponds roughly to 150–200 mg kg^{-1}. With daily doses of 110 mg kg^{-1} given via intragastric cannula to female rats over 100 days, hypertrophy of organs such as the salivary gland, liver, heart, and kidneys was reported. Caffeine also induced thymic and testicular atrophy. Developmental and reproductive toxicity was associated with high, single daily doses of caffeine. The no-effect level for teratogenicity is 40 mg kg^{-1} caffeine per day in the rat, although delayed sternebral ossification can be observed at lower doses. This effect has been shown to be reversed in the postnatal period. Available epidemiological evidence suggests that maternal caffeine consumption does not cause morphological malformation in the fetus. Caffeine intake has been linked with reduced fetal size in some studies, particularly when intake was more than 600 mg per day, whereas others have not shown an impact on growth. High daily levels given as divided doses in rats were less toxic than when given as a single dose, in which case reduced fetal body weight was the only effect observed.

Caffeine at high concentration levels has mutagenic effects in bacteria and fungi and causes chromosomal damage in vitro. However, there is consensus that caffeine is not mutagenic in higher animals.

An epidemiological study showed no chromosomal aberrations in lymphocytes of normal, caffeine-exposed people, and other studies reported an increased frequency of micronucleated blood cells and the absence of mutagenic compounds in urine. In long-term studies, caffeine was shown to have no carcinogenic potential in rodents. Caffeine has not been classified as carcinogenic in animals or humans by the International Agency for Research on Cancer.

Therapeutic Uses

The most extensively investigated and most firmly established clinical application of caffeine is the control of neonatal apnoea in premature infants. The respirogenic properties of theophylline were first reported, and caffeine is increasingly being used as a substitute for theophylline because of its wider therapeutic index. For infants with a body weight of 2.5 kg, the therapeutic loading doses varied from 5 to 30 mg kg^{-1}, followed by a maintenance dose of 3 mg kg^{-1} per day. Plasma caffeine levels must be controlled carefully to reach 10–20 mg l^{-1}.

Because of the bronchial muscle relaxant effect, caffeine is used in chronic obstructive pulmonary disease and for the treatment of asthma. The use of caffeine in the treatment of children with minimal brain dysfunction, to increase the duration of electroconvulsive therapy-induced seizure, for allergic rhinitis, as well as for atopic dermatitis has also been described. Recently, caffeine has been used as a diagnostic test for malignant hyperthermia and in the diagnosis of neuroleptic malignant syndrome, a complication of neuroleptic therapy.

Caffeine is found in many drug preparations, both prescription and over-the-counter. Caffeine is present in drugs used as stimulants, pain relievers, diuretics, and cold remedies. When used as an analgesic adjuvant, the potency of the analgesic drug is significantly enhanced by the addition of caffeine.

Although caffeine has been shown to promote thermogenesis in humans, it is no longer allowed as an ingredient in weight-control products in the US market because long-term clinical studies demonstrate that it does not help those wishing to lose weight.

Biochemical Mechanisms of Action

The physiological and pharmacological properties of caffeine cannot be explained by a single biochemical mechanism. Three principal hypotheses have been investigated to explain the diverse actions of caffeine.

The first biochemical effect described was the inhibition of phosphodiesterase, the enzyme that catalyzes the breakdown of cyclic adenosine 3′,5′-phosphate (cAMP). Caffeine was shown to increase cAMP concentrations in various tissues. This inhibition occurs at large concentrations (millimolar range) and is of

limited importance with regard to the physiological effects of caffeine at levels at which it is normally consumed.

Calcium translocation is the second mechanism frequently suggested from experiments using skeletal muscles. However, high concentrations of caffeine are also necessary to modify intracellular calcium ion storage.

In the plasma, increased levels of β-endorphin, epinephrine, norepinephrine, corticosterone, ACTH, renin, and angiotensin I and decreased levels of growth hormone, thyroxine, triiodothyronine, and thyrotropin were reported with high caffeine doses. The mechanisms responsible for these various effects are largely unknown, and the mediation of adenosine receptors is suggested. The antagonism of benzodiazepine at the receptor level is observed at lower caffeine concentrations (0.5–0.7 mM) than those required for phosphodiesterase inhibition.

The third mechanism, antagonism of the endogenous adenosine, is the most plausible mode of action because caffeine exerts its antagonism at micromolar levels. Its main metabolite, paraxanthine, is as potent as caffeine in blocking adenosine receptors. Caffeine is more potent at A_{2A} receptors and less potent at A_3 receptors compared to A_1 and A_{2B} receptors. An upregulation of adenosine receptor is the postulated biochemical mechanism of caffeine tolerance.

Adenosine receptor antagonism appears to be the mechanism that explains most of the effects of caffeine on CNS activity, intestinal peristalsis, respiration, blood pressure, lipolysis, catecholamine release, and renin release. However, some effects, such as opiate antagonism or effects that are similar to those of adenosine, must be mediated by other mechanisms, such as the potentiation by caffeine of inhibitors of prostaglandin synthesis.

See also: **Brain and Nervous System**. **Diabetes Mellitus**: Etiology and Epidemiology. **Energy**: Balance; Requirements. **Exercise**: Beneficial Effects; Diet and Exercise. **Sports Nutrition**. **Tea**.

Further Reading

Armstrong LE (2002) Caffeine, body fluid–electrolyte balance and exercise performance. *International Journal of Sport Nutrition and Exercise Metabolism* **12**: 205–222.

Arnaud MJ (1987) The pharmacology of caffeine. *Progress in Drug Research* **31**: 273–313.

Clarke RJ and Macrae R (eds.) (1988) Physiology. In *Coffee*, vol. 3. London: Elsevier.

Clarke RJ and Vitzhum OG (eds.) (2001) *Coffee. Recent Developments*. London: Blackwell Science.

Debry G (1994) *Coffee and Health*. Paris: John Libbey Eurotext.

Dews PB (ed.) (1984) *Caffeine*. Berlin: Springer-Verlag.

Dews PB, O'Brien CP, and Bergman J (2002) Caffeine: Behavioural effects of withdrawal and related issues. *Food and Chemical Toxicology* **40**: 1257–1261.

Garattini S (ed.) (1993) *Caffeine, Coffee and Health*. New York: Raven Press.

Graham TE (2001) Caffeine and exercise, metabolism, endurance and performance. *Sports Medicine* **31**: 785–807.

James JE (1991) *Caffeine and Health*. London: Academic Press.

Lorist MM and Tops M (2003) Caffeine, fatigue, and cognition. *Brain and Cognition* **53**: 82–94.

Nawrot P, Jordan S, Eastwood J *et al.* (2003) Effects of caffeine on human health. *Food Additives and Contaminants* **20**: 1–30.

Schmitt JAJ (2001) *Serotonin, Caffeine and Cognition. Psychopharmacological Studies in Human Cognitive Functioning*. Maastricht, The Netherlands: Neuropsych.

Snel J and Lorist MM (eds.) (1998) *Nicotine, Caffeine and Social Drink, Behaviour and Brain Function*. Amsterdam: Harwood Academic.

World Health Organization–International Agency for Research on Cancer (1991) *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Coffee, Tea, Mate, Methylxanthines and Methylglyoxal*, vol. 51. Lyon, France: WHO–IARC.

CALCIUM

L H Allen, University of California at Davis, Davis, CA, USA
J E Kerstetter, University of Connecticut, Storrs, CT, USA

Calcium is an essential nutrient. Although most of the calcium in the body is found in bones and teeth, the other 1% has critical, life-sustaining functions. Most people in the world, including those in industrialized countries, fail to consume the recommended amounts of calcium, which will ultimately result in poor bone health and increase the risk of osteoporosis. Adequate calcium intake is critical to the achievement of peak bone mass in the first several decades of life, the retention of bone during middle adulthood, and the minimization of bone loss during the last several decades. Without adequate intake, the intestine, bone, and renal systems

have intricate ways of retaining more calcium and normalizing serum calcium levels. These three primary tissues of calcium homeostasis (intestine, bone, and kidneys) are dynamic in their handling of calcium, reacting to dietary intake, physiological need, or disease processes. This article discusses calcium absorption, regulation, function, metabolism, and excretion as well as the changes in calcium physiology during the lifespan.

Absorption and Transport

Intake and Distribution

The dietary intake of calcium in the United States is approximately 20 mmol (600–1200 mg) per day unless supplements are consumed. Approximately 73% of dietary calcium is supplied from milk products, 9% from fruits and vegetables, 5% from grains, and the remaining 12% from all other sources. Approximately 25% of women take a nutritional supplement that contains calcium, but supplement use by men and children is much lower.

Approximately 25–50% of dietary calcium is absorbed and delivered to the exchangeable calcium pool. Of the 25–30 mol (1000–1200 g) of calcium in the body, 99% is found in the skeleton and teeth. The remaining 1% is in the blood, extracellular fluid, muscle, and other tissues. The extracellular pool of calcium turns over 20–30 times per day in adults, whereas bone calcium turns over every 5 or 6 years. A remarkably large amount is filtered through the kidneys, approximately 250 mmol (10 000 mg) per day, of which approximately 98% is reabsorbed, so that urinary excretion of the mineral is only 2.5–5 mmol (100–200 mg) per day (**Figure 1**).

Intestinal Calcium Absorption

The efficiency of dietary calcium absorption depends on two major factors: its interaction with other dietary constituents and physiological/pathological factors. Dietary factors that reduce the total amount of calcium absorbed by the intestine include phosphate, oxalate, phytate, fiber, and very low calcium intakes, whereas those that increase absorption include protein (or specific amino acids, lysine and arginine) and lactose in infants. The physiological/pathological factors that decrease intestinal calcium absorption include low serum $1,25(OH)_2$ vitamin D (the form of the vitamin that effects calcium absorption), chronic renal insufficiency, hypoparathyroidism, aging, and vitamin D deficiency, whereas increased absorption is observed during growth, pregnancy, primary hyperparathyroidism, sarcoidosis, and estrogen and growth hormone administration. The interindividual variability in intestinal calcium absorption is very high for reasons that are not entirely clear. On tightly controlled diets, a homogenous group of subjects can have intestinal calcium absorptions ranging from 10 to 50%.

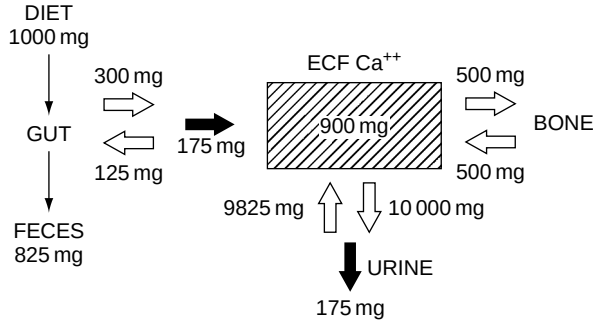


Figure 1 Daily calcium turnover.

Dietary calcium is complexed to food constituents such as proteins, phosphate, and oxalate, from which it needs to be released prior to absorption. The role of gastric acid (or the lack thereof induced by commonly used proton pump inhibiting drugs) in intestinal calcium absorption is not well established, although achlorhydria can impair absorption in the fasted state.

Calcium crosses the intestinal mucosa by both active and passive transport. The active process is saturable, transcellular, and occurs throughout the small intestine. The transcellular pathway is a multistep process, starting with the entry of luminal calcium into the enterocyte (possibly via a calcium channel) and translocation of calcium from the microvillus border of the apical plasma membrane to the basolateral membrane followed by extrusion out of the enterocyte. Calbindin, a calcium binding protein that is regulated by the hormonal form of vitamin D, $1,25(OH)_2D_3$, affects every step in the movement of calcium through the enterocyte, including entry into the cell, movement in the cell interior, and transfer into the lamina propria. Although details of the movement of calcium through intestinal cells are still under investigation, it appears that the vitamin D-dependent calcium binding protein calbindin-D_{9k} and the plasma membrane calcium-pumping ATPase 1b (PMCA1b) are critical transport molecules in the cytoplasm and basolateral membrane, respectively. The active transport pathway is predominant at lower levels of calcium intake, and it becomes more efficient in calcium deficiency or when intakes are low and also when calcium requirements are high during infancy, adolescence, and pregnancy. It becomes less efficient in vitamin D-deficient individuals and in elderly women after menopause.

The passive transport pathway is nonsaturable and paracellular. It occurs throughout the small intestine

and is unaffected by calcium status or parathyroid hormone (PTH). It is relatively independent of $1,25(OH)_2D_3$, although this metabolite has been found by some investigators to increase the permeability of the paracellular pathway. A substantial amount of calcium is absorbed by passive transport in the ileum due to the relatively slow passage of food through this section of the intestine. The amount of calcium absorbed by passive transport will be proportional to the intake and bioavailability of calcium consumed.

Fractional calcium absorption increases in response to low intake but varies throughout life. It is highest during infancy (60%) and puberty (25–35%), stable at approximately 25% in adults, and then declines with age (by approximately 2% per decade after menopause). There is little difference in calcium absorption efficiency between Caucasians and African Americans. The lower urinary calcium and better calcium conservation in African Americans probably contributes to their higher bone mineral density.

Storage

The skeleton acts as the storage site for calcium. Bone calcium exists primarily in the form of hydroxyapatite ($Ca_{10}(PO_4)_6(OH)_2$), and this mineral comprises 40% of bone weight. In the short term, the release of calcium from bone serves to maintain serum calcium concentrations. In the longer term, however, persistent use of skeletal calcium for this purpose without adequate replenishment will result in loss of bone density. The storage of very small amounts of calcium in intracellular organelles and its subsequent release into cytosol acts as an intracellular signal for a variety of functions.

Between 60 and 80% of the variance in peak bone mass is explained by genetics, including polymorphisms in the vitamin D–receptor gene and in genes responsible for insulin-like growth factor-1 (IGF-1) and collagen production.

Metabolism and Excretion

Regulation by Hormones

The concentration of ionized calcium in serum is closely regulated because it has profound effects on the function of nerves and muscles, blood clotting, and hormone secretion. The principal regulators of calcium homeostasis in humans and most terrestrial vertebrates are PTH and the active form of vitamin D, $1,25(OH)_2$ vitamin D_3 (**Figure 2**).

PTH is a single-chain polypeptide that is released from the parathyroid when there is a decrease in the calcium concentration in extracellular fluid. The calcium-sensing

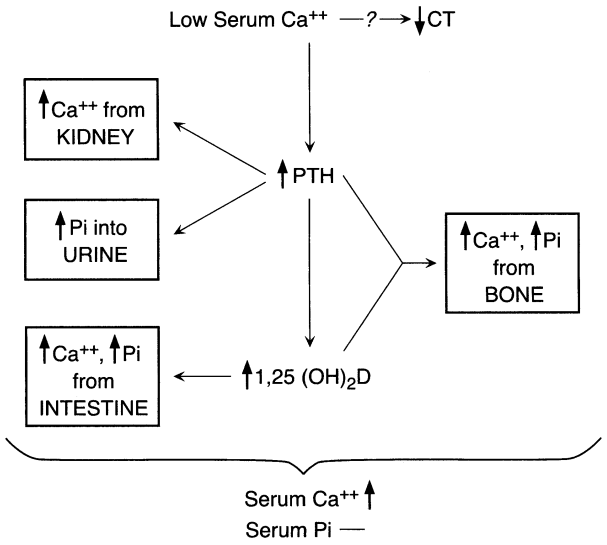


Figure 2 Hormonal regulation of calcium metabolism.

receptor (acting as the thermostat for calcium) is found on the parathyroid gland, where it detects small perturbations in serum ionized calcium. The decline in serum calcium induces an increase in PTH secretion. PTH has the effect of restoring extracellular calcium concentrations by stimulating the resorption of bone to release calcium, by increasing the renal reabsorption of calcium, and by enhancing the renal conversion of $25(OH)D_3$ to the active, hormonal form of the vitamin, $1,25(OH)_2$ vitamin D_3. The active form of vitamin D then increases the synthesis of intestinal calcium binding protein (CaBP), leading to more efficient intestinal calcium absorption. PTH release is inhibited when serum calcium and $1,25(OH)_2$ vitamin D_3 increase or when serum phosphate is decreased. The highly regulated interactions among PTH, calcium, $1,25(OH)_2$ vitamin D_3, and phosphate maintain blood calcium levels remarkably constant despite significant changes in calcium intake or absorption, bone metabolism, or renal functions. The extracellular calcium concentration is the most important regulator of PTH secretion and occurs on a minute-by-minute basis. Acute PTH administration leads to release of the rapidly turning over pool of calcium near the bone surface. Chronic administration of PTH increases osteoclast cell number and activity. Interestingly and paradoxically, intermittent PTH administration is anabolic, increasing formation of trabecular bone. In the kidney, PTH has three major functions: It increases calcium reabsorption, inhibits phosphate reabsorption, and enhances the synthesis of the active form of vitamin D. All of these actions are designed to defend against hypocalcemia.

There are two sources of vitamin D: the diet (where it is found as the fortificant vitamin D_2 or

natural D_3) or synthesis in skin during exposure to ultraviolet radiation (sunlight). The vitamin enters the circulation and is transported on a vitamin D binding protein to the liver, where it is hydroxylated to 25(OH) cholecalciferol, which leaves the liver, is bound again to the binding protein in the circulation, and enters the kidney where it is hydroxylated again to $1,25(OH)_2D_3$, the most active metabolite of the vitamin. The primary biological effect of $1,25(OH)_2D_3$ is to defend against hypocalcemia by increasing the efficiency of intestinal calcium absorption and by stimulating bone resorption. $1,25(OH)_2D_3$ interacts with the vitamin D receptor on the osteoblasts, and via RANKL/RANK it stimulates the maturation of osteoclasts that function to dissolve bone, releasing calcium into the extracellular space. The recently discovered RANK ligand, a member of the tumor necrosis factor superfamily, and its two receptors (RANK and osteoprotegerin) are pivotal regulators of osteoclastic bone resorption, both *in vivo* and *in vitro*. More of the active metabolite is produced during calcium deficiency or after a low calcium intake in order to restore serum calcium by increasing intestinal calcium absorption, renal calcium reabsorption, and bone turnover.

Serum vitamin D concentrations decline in winter and are generally related to vitamin D intake and sunlight exposure. When serum $25(OH)D_3$ concentrations decline below 110 nmol/l, PTH levels increase, contributing to the bone loss that occurs in vitamin D deficiency and that is evident in northern Europe, the United States, Japan, and Canada during winter months. Rickets is becoming a more recognized health problem, particularly in infants of African American mothers who are not taking vitamin D supplements or consuming adequate amounts of vitamin D-fortified milk and who are exclusively breast-feeding their infants. In the US national nutrition survey conducted in the 1990s, 42% of African American women had low 25(OH) vitamin D concentrations in plasma. Thus, vitamin D deficiency is more prevalent than once believed, and it is particularly a risk for the elderly due to their reduced capacity for synthesizing vitamin D precursors in their skin, in those who are infirm and/or in nursing homes or living at more northern or southern latitudes, or in other situations in which the skin is not exposed to sunlight. The result of vitamin D deficiency is normal serum calcium and elevated PTH and alkaline phosphatase. The secondary hyperparathyroidism causes increased osteoclastic activity, calcium loss from bone, and ultimately bone loss.

Several other hormones also affect calcium metabolism. Notably, estrogens are necessary for the maintenance of balance between bone resorption and accretion. The decrease in serum estrogen concentrations at approximately the time of menopause is the primary factor contributing to the elevated rate of bone resorption that occurs at this stage of life and that is the primary contributory factor to osteoporosis. Estrogen treatment will reduce bone resorption within a few weeks and subsequently lead to higher serum concentrations of PTH and $1,25(OH)_2D_3$ and improved intestinal absorption and renal reabsorption of calcium. Testosterone also inhibits bone resorption, and lack of this hormone can cause osteoporosis in men. Glucocorticoids, sometimes used to treat conditions such as osteoid arthritis, inflammatory bowel disease, and asthma, inhibit both osteoclastic and osteoblastic activity, impair collagen and cartilage synthesis, and reduce calcium absorption. Consequently, excessive bone loss often results from glucocorticoid treatment or occurs when excessive amounts of the hormone are secreted, such as in Cushing's disease. Oral calcium supplements should be considered for patients receiving exogenous glucocorticoids. Thyroid hormones stimulate bone resorption so that bone abnormalities occur in both hyper- and hypothyroidism. Growth hormone stimulates cartilage formation, the formation of $1,25(OH)_2D$, and intestinal calcium absorption. Insulin stimulates collagen production by osteoblasts and impairs the renal reabsorption of calcium.

Excretion

Ionized calcium in the plasma is freely filtered by the kidney so that very large amounts enter the kidney each day, but 99.8% of this calcium is reabsorbed throughout the nephron. Regulated active transport occurs in the distal convoluted tubule and involves vitamin D, CaBP, and PTH. Typically, approximately 2.5–6 mmol (100–240 mg) of calcium is excreted in urine daily.

Dietary calcium has a relatively small impact on urinary calcium (e.g., only 6–8% of an increase in dietary calcium intake will appear in the urine). The major food components that affect urinary calcium are protein, phosphorus, caffeine, and sodium. For each 50-g increment in dietary protein, approximately 1.5 mmol (60 mg) of additional calcium is lost in urine. The higher amounts of phosphorus consumed concurrently with a high-protein diet can blunt, but not eliminate, this phenomenon. Dietary phosphorus (as well as intravenously administered phosphorus) increases PTH synthesis and subsequently stimulates renal calcium reabsorption and reduces the urinary excretion of calcium. Caffeine causes a reduction in renal reabsorption of calcium and a subsequently increased loss of urinary calcium

soon after it is consumed. It has been shown repeatedly in animals and humans that dietary sodium, in the form of salt (NaCl), increases urinary calcium excretion. On average, for every 100 mmol (2300 mg) of sodium excreted in urine, there is an approximately 0.6–1 mmol (24–40 mg) loss of calcium in free-living healthy populations of various ages. Because most of the urinary calcium is of bone origin, it is commonly hypothesized that those nutrients or food components that are hypercalciuretic are also detrimental to the skeleton. On the other hand, thiazide medications are hypocalciuric and, as such, may have modest positive effects on bone.

Metabolic Functions

The most obvious role of calcium is to provide structural integrity and strength in bones and teeth. Approximately 99% of the total body calcium content is used for this purpose. Bone also serves as a reservoir of calcium that can be drawn upon when serum calcium concentrations decline. The remaining 1% of the body's calcium is contained in blood, extracellular space, muscle, and other tissues, in which calcium concentrations are kept relatively constant. The maintenance of constant serum calcium concentrations at approximately 2.5 mmol/l is critical for a number of cellular functions.

The extracellular concentration of calcium is in the 10^{-3} M range, whereas in the cytosol it is approximately 10^{-6} M. Almost all of the intracellular calcium is bound within organelles such as the nucleus, endoplasmic reticulum, and vesicles. Cytosolic calcium concentrations are very low and influenced greatly by release of some calcium from cellular organelles. Therefore, a small change in the release of calcium from intracellular sites or transport across the cell membrane results in a relatively large change in cytosolic calcium concentration. Binding of a hormone or a growth factor to a plasma membrane receptor increases inositol triphosphate release, which in turn increases the free intracellular calcium concentration. The ionized calcium then binds to calmodulin, followed by a conformation change in the protein to trigger cellular events such as muscle contraction, nerve conduction, cell movement and differentiation, cell division, cell-to-cell communication, and secretion of hormones such as insulin. In these roles, calcium acts as an intracellular messenger.

Calcium may play a role in energy regulation and risk of obesity. Dietary calcium regulation of circulating $1,25(OH)_2D_3$ in turn regulates the concentration of calcium in adipocytes. When adipocyte intracellular calcium concentration increases, this promotes the expression of lipogenic genes, and fat breakdown is reduced leading to accumulation of lipid in adipocytes. Through this pathway, low-calcium diets appear to promote fat deposition, whereas high-calcium intakes afford some protection from obesity.

Changes in Calcium Metabolism during the Life Span

The total body calcium content of the newborn infant is approximately 0.75 mol (30 g), which increases during growth to approximately 1000 g in adult women and 1200 g in adult men. This represents an average daily accumulation of approximately 2.5–3.7 mmol (100–150 mg) from infancy to adulthood.

The efficiency of calcium absorption is highest during infancy (approximately 60%), and the amount absorbed from breast milk does not appear to be affected by calcium consumed in solid foods. During the growth spurt of adolescence, calcium retention and accretion increase to peak at approximately 200–300 mg per day in girls and boys, respectively. It involves the action of growth hormone, IGF-1, and sex steroids. The onset of menstruation in girls is associated with a rapid decline in bone formation and resorption. Intestinal calcium absorption is predictably more efficient during the growth spurt and also decreases subsequently. Importantly, it is thought that calcium intakes during the period of growth can affect the peak bone mass achieved and therefore influence the amount of bone mineral remaining when osteoporosis begins in later life. Bone mass may continue to accumulate up to approximately age 30 years, although the amount gained is relatively small after age 18 years.

During pregnancy, a relatively small amount of calcium, approximately 625–750 mmol, is transported to the fetus. Most of this calcium is thought to be obtained through greater efficiency of maternal intestinal calcium absorption, possibly induced by increases in $1,25(OH)_2D_3$ production. For this reason, a higher calcium intake during pregnancy is probably not required.

Most studies have reported that there is no increase in intestinal calcium absorption during lactation even when dietary intake of the mineral is relatively low. Changes in biochemical markers and kinetic studies using isotopes indicate that the source of much of the calcium secreted in breast milk is the maternal skeleton, as well as more efficient renal reabsorption and subsequently lower urinary excretion of the mineral. Bone calcium is restored at the

end of lactation as the infant is weaned, when ovarian function returns and menstruation resumes. At this time, intestinal calcium absorption increases, urinary calcium remains low, and bone turnover rates decline to normal levels. There is no strong evidence that lactation per se or maternal calcium intake during lactation affect later risk of osteoporosis in women. Thus, there is no strong rationale for increasing maternal calcium intake during lactation. Breast milk calcium concentration is relatively unaffected by maternal intake, and it remains stable throughout lactation.

Menopause begins a period of bone loss that extends until the end of life. It is the major contributor to higher rates of osteoporotic fractures in older women. The decrease in serum estrogen concentrations at menopause is associated with accelerated bone loss, especially from the spine, for the next 5 years, during which approximately 15% of skeletal calcium is lost. The calcium loss by women in early menopause cannot be prevented unless estrogen therapy is provided. Calcium supplements alone are not very helpful in preventing postmenopausal bone loss. Upon estrogen treatment, bone resorption is reduced and the intestinal calcium absorption and renal reabsorption of calcium are both increased. Similarly, amenorrheic women have reduced intestinal calcium absorption, high urinary calcium excretion, and lower rates of bone formation (compared to eumenorrheic women). In both men and women, there is a substantial decline in intestinal absorption of calcium in later life.

Calcium Deficiency

When calcium absorption is chronically low, because of low intakes, poor bioavailability, or conditions that impair intestinal absorption, there is a decrease in the serum ionized calcium concentration. This in turn stimulates the release of PTH, which returns serum calcium to normal by increasing renal calcium reabsorption, stimulating the renal production of $1,25(OH)_2D_3$, and inducing bone reabsorption. The result of long-term calcium deficiency is accelerated bone loss in older individuals or the inability to fully achieve peak bone mass in younger individuals.

Dietary Sources

Food Sources

The majority of dietary calcium in industrialized countries comes from milk products; one serving (i.e., 250 ml milk or yogurt or 40 g cheese) contains approximately 7.5 mmol (300 mg). Nondairy sources (fruits, vegetables, and grain products) supply approximately 25% of total calcium. When substantial amounts of grains are consumed, for example, in breads or as maize products, these can be important sources, although the calcium in cereals tends to be less bioavailable than that in dairy products. Other foods high in calcium include tofu set with a calcium salt, kale, broccoli, and, increasingly, calcium-fortified juices and cereals. No matter what the source, a high percentage of people in both industrialized and less wealthy countries fail to meet recommended guidelines for optimal calcium intake.

Bioavailability

Several dietary constituents decrease the bioavailability of calcium in food. Increasing fiber intake by, for example, replacing white flour by whole wheat flour in a typical Western diet has long been associated with negative calcium balance even when calcium intakes meet recommended levels. Likewise, the fiber in fruits and vegetables can cause negative calcium balance. In cereals, phytic acid is the main constituent of fiber that binds calcium, making it unavailable for absorption. The fermentation of bread during leavening reduces phytate content substantially, making calcium more bioavailable. In fruits and vegetables, the uronic acids in hemicellulose are strong calcium binders, as is the oxalic acid present in high concentrations in foods such as spinach. Calcium bioavailability from beans is approximately half and that from spinach approximately one-tenth of the bioavailability from milk. In contrast, calcium absorption from low-oxalate vegetables, such as kale, broccoli, and collard greens, is as good as that from milk. The difference in calcium absorption between the various forms of supplements is not large.

Dietary fat does not affect calcium absorption except in individuals with diseases that impair fat malabsorption (e.g., short bowel syndrome, celiac disease, and pancreatitis). In these conditions, the calcium forms an insoluble and unabsorbable 'soap' with the unabsorbed fat in the alkaline lumen of the small intestine, potentially resulting in impaired bone mineralization. In addition, the luminal calcium is not available to precipitate the oxalates, meaning that the free oxalates will be hyperabsorbed leading to increased risk for renal oxalate stones. Neither dietary phosphorus nor a wide range of phosphorus-to-calcium ratios affect intestinal calcium absorption in very low-birth-weight infants and adults.

Lactose improves calcium absorption in young infants, in whom absorption of calcium is predominantly by passive transport. In adults, the presence

of lactose in the diet has little effect on the efficiency of calcium absorption.

Effects of High Calcium Intakes

Calcium can inhibit the absorption of both heme iron (found in meat, fish, and poultry) and non-heme iron. The mechanism by which this occurs remains controversial, but the inhibition probably occurs within the mucosal cells rather than in the intestinal lumen. This interaction is of concern because calcium supplements are taken by many women who may have difficulty maintaining adequate iron stores. Approximately 300–600 mg of calcium, as a supplement or in foods, reduces the absorption of both heme and non-heme iron by approximately 30–50% when consumed in the same meal. The inhibitory effect on iron absorption is inversely related to iron status so that it is relatively unimportant above a serum ferritin concentration of approximately 50–60 μg/l. Thus, consideration should be given to monitoring the iron status of menstruating women with low iron stores who take calcium supplements. There is no inhibitory effect when calcium and iron supplements are consumed together in the absence of food, and inhibition may be less with calcium citrate.

In the past, it was common to restrict dietary calcium in patients with a history of calcium oxalate stones. However, recent data suggest that a severe calcium restriction in patients with oxalate stones is not only ineffective but also can lead to bone demineralization. For the prevention of recurrent stone formation, a diet restricted in oxalate, sodium, and animal protein is probably most effective. Only if absorptive hypercalciuria is present should a moderate calcium restriction be imposed.

Long-term consumption of approximately 1500–2000 mg calcium per day is safe for most individuals, although there will be some reduction in the efficiency of iron absorption. However, higher intakes from supplements (62.5 mmol or 2.5 g per day) can result in milk–alkali syndrome (MAS), with symptoms of hypercalcemia, renal insufficiency, metabolic alkalosis, and severe alterations in metabolism. Based on risk of developing MAS, the upper limit for calcium intake is 2500 mg per day for adults and children.

See also: **Bioavailability**. **Bone**. **Dairy Products**. **Lactation**: Dietary Requirements. **Pregnancy**: Nutrient Requirements; Safe Diet for Pregnancy. **Vitamin D**: Physiology, Dietary Sources and Requirements; Rickets and Osteomalacia.

Further Reading

Bronner F (1992) Symposium: Current concepts of calcium absorption. *Journal of Nutrition* **122**: 641–686.

Heaney RP (2003) How does bone support calcium homeostasis? *Bone* **33**: 264–268.

Institute of Medicine (1997) *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D and Fluoride*. Washington, DC: National Academy Press.

Prentice A (2000) Calcium in pregnancy and lactation. *Annual Reviews in Nutrition* **20**: 249–272.

Prentice A (2004) Diet, nutrition and the prevention of osteoporosis. *Public Health Nutrition* **7**: 227–243.

Specker BL (2004) Nutrition influences bone development from infancy through toddler years. *Journal of Nutrition* **134**: 691S–695S.

Wasserman RH and Fullmer CS (1995) Vitamin D and intestinal calcium transport: Facts, speculations and hypotheses. *Journal of Nutrition* **125**: 1971S–1979S.

Weaver CM and Heaney RP (1999) Calcium. In: Shils ME, Olson JA, Shike M *et al.* (eds.) *Modern Nutrition in Health and Disease*. Philadelphia: Lea & Febiger.

Zemel MB and Miller SL (2003) Dietary calcium and dairy modulation of adiposity and obesity risk. In *Primer on the Metabolic Bone Diseases and Disorders of Mineral Metabolism*, 5th edn. Washington, DC: American Society for Bone and Mineral Research.

Calories *see* **Energy**: Balance; Requirements. **Energy Expenditure**: Indirect Calorimetry; Doubly Labeled Water

CANCER

Contents

Epidemiology and Associations Between Diet and Cancer

G A Colditz, Harvard Medical School, Boston, MA, USA

Sources of Evidence Linking Diet and Cancer

Laboratory scientists have known since the early twentieth century that various nutritional manipulations can influence the occurrence of tumors in animals. Despite this discovery of the relationship between diet and cancer in animals, widespread interest in the study of diet and cancer in humans did not develop until more recently when the large international differences in cancer rates were correlated with variations in dietary factors. In fact, investigators have found strong correlations between estimated per capita fat consumption and breast cancer rates internationally, raising the possibility that dietary fat may have an important role in the etiology of breast cancer. Other observations such as those demonstrating that migrating populations adopted, sooner or later, the cancer rates of their new host population strengthened the evidence that international differences were the result not of genes, but of noninherited factors, including diet. The study designs used to investigate diet and cancer in humans are discussed below.

Descriptive Studies

Rates of cancer show large differences between countries for many malignancies. International correlations compare disease rates with lifestyle factors such as per capita consumption of specific dietary factors.

Age-adjusted rates of colon and breast cancer are up to five times higher in some countries than others. Dakar in Senegal (0.6) and Poona (3.1) and Bombay (3.5) in India, have the lowest incidence rates of colon cancer per 100 000 males; in contrast, the USA has the highest recorded rates of 32.2 in Connecticut and 31.4 in New York.

Strong nutritional correlates exist for specific cancers. These studies, also known as ecological studies use the country or other geographic area as the unit of measure rather than the individual. For example, Armstrong and Doll in 1975 compared per capita total fat intake and national breast cancer mortality rates among women and found a correlation of 0.89: Countries with higher fat intake had higher breast cancer mortality. They also compared per capita fat intake and mortality from colon cancer and observed a correlation of 0.85 for men and 0.81 for women.

The most important of the existing strengths of correlation studies is that the contrasts in dietary intake are very large. For example, the range of fat intake within a population tends to be small compared with the range of far intake among different populations.

Although correlation studies have opened the door to new leads in the study of diet and cancer, certain limitations have prevented these investigations from advancing past the level of hypothesis generation. First and foremost, there are many factors other than dietary differences that distinguish countries with a high incidence from those with a low incidence. This makes it difficult to identify dietary factors as the primary explanation for the differences in the etiology of cancers. For example, besides consuming a diet with a higher proportion of energy from fat, populations of countries that are more industrialized will also have shifted from an agrarian to an urbanized, sedentary society with lower total energy expenditure. Therefore, with increasing industrialization exposure to many aspects of life will decrease exercise and increase fat intake. Consider the example of colon cancer. The international correlation between fat and colon

cancer mortality in men is 0.85, and for meat it is 0.85. There is also a correlation between gross national product and colon cancer mortality (0.77 for men); more industrialized countries have higher economic production and higher rates of cancer. Owing to the many factors that are associated with industrialization it is not possible to separate our which factor is important in the etiology of colon cancer, lack of physical activity or increased consumption of fat or meat. Studies with data on lifestyle factors at the individual level are needed to clarify which of these variables is important (see below).

Special-Exposure Groups

Within populations there are groups that have atypical dietary patterns which may provide valuable information in the probe for further information on the relationship of diet and cancer. These groups are called special-exposure groups and are often defined by ethnic or religious characteristics. In addition to offering many of the advantages of correlation studies, the number of alternative explanations for any observations may be reduced if the special-exposure group lives in the same area as the comparison group.

As a largely vegetarian group, the Seventh-Day Adventists have been used in studies of meat earing and cancer. Studies of these groups, however, are limited in the same ways that other ecological studies are limited. For example, although lower rates of colon cancer have been observed among Seventh-Day Adventists—supporting the hypothesis that meat is related to colon cancer—there are other lifestyle choices that characterize the group, such as low rates of tobacco use and alcohol intake, which could also modify their rates of colon cancer.

Evidence from Descriptive Studies

In 1981, Doll and Peto made an estimate based largely on descriptive studies that 35% of cancers in the USA may be attributable to dietary factors; but reflecting uncertainty in the sources of data used for this estimate, they noted that the range of possible dietary contribution was from as low as 10% to as high as 70%. The marked variation in the rates of most cancers among countries is evidence that dietary factors may influence the development of cancer. The potential range of dietary factors that may influence cancer risk are presented in **Table 1**. Despite the fact that descriptive studies provide an excellent source of hypotheses, it is necessary to conduct analytical studies to collect data that will provide more definitive evidence.

Table 1 Examples of suspected dietary factors influencing cancer risk

Dietary factor	*Site of cancer*
Increased risk	
Overnutrition/Obesity	Endometrium, gallbladder, breast, colon
Alcohol	Liver, oesophagus, larynx, pharynx, breast, colon
Beer	Rectum
Fat (especially saturated)	Colorectum, breast, prostate
Red meat	Colorectum
Salt	Stomach, nasopharynx
Heterocyclic amines (from cooked meat)	Colorectum
Decreased risk	
Fiber	Colorectum, breast
Vitamins A, C, E	Many sites
Protease inhibitors	Colorectum
Calcium, vitamin D	Colorectum
Folate	Colorectum
Lycopene	Prostate
Carotenoids	Lung
Phytooestrogens	Breast

Time Trends within Countries

The analysis of cancer trends over time can lead to useful findings in the study of diet and cancer. By looking at the change in cancer rates in a specific population over time and comparing these rates with changes in specific factors over the same period (e.g., changes in dietary habits), investigators can uncover possible associations supporting the dietary factors hypotheses. For example, researchers have examined vital statistics for Japanese natives and US whites to unveil changes in cancer mortality and related antecedent patterns of lifestyle in the two populations. These investigations have uncovered that animal fat consumption in Japan steadily increased from a daily level of 6.5 g per person in 1955 to 27.6 g in 1987; at the same time the Japanese rate of colon cancer in men rose at a rapid pace; in fact, the mortality rates owing to colon cancer in men almost trebled over this time. This evidence lends more support to the hypothesis that mortality from colon cancer in men is influenced by high dietary fat consumption.

Similar data were collected in Singapore to determine trends in the incidence of breast cancer: in 1996 an average annual increase in breast cancer incidence of 3.6% over a 25-year period for all women was reported. The most convincing evidence that the observed trend was real was that it was clearly cohort-related rather than period-related. The risk was observed to increase in successive birth cohorts from the 1890s to the 1960s. Changes

in dietary consumption patterns (e.g., the adoption of a more 'Western' diet) fall among other factors cited as having a possible effect on the continuing increase in rates of breast cancer among women in Singapore. Like descriptive studies, time-trend studies are a valuable source for hypotheses generation, but more definitive evidence is required from analytical epidemiology to uncover any real associations between dietary factors and cancer rates.

Migrant Studies

Migrant studies examine the rates of specific diseases in migrating populations. These studies are important in addressing the possibility that observed correlations in ecological studies are owing to genetic factors. Generally, results from migrant studies have so far found that the migrating group takes on the rate of cancer of the new country. Hence genetic factors are excluded as the dominant cause for varying rates of cancer between countries. A good example of this is seen in the Japanese migrant population to the USA. Japan has low rates of cancers of the breast, colon, and prostate, while the rates of these cancers among Japanese migrants to the USA move toward the higher US rates. The increased risk of breast cancer among migrants occurs primarily in later generations, leading investigators to believe that the causal factors operate early in life. Investigators also consider major changes in the rate of disease that occur within a population over time as evidence that nongenetic factors play an integral role in the etiology of cancer. The limitations of migrant studies are similar to those of ecological studies.

Analytical Studies

Cohort studies Cohort studies involve the collection of information from healthy participants who are followed over time and observed for the occurrence of new cases of disease (incident cases). During or at the end of follow-up, the disease frequency within a cohort may be measured as either a cumulative incidence rate (the number of cases divided by the entire base population) or an incidence density rate (the number of cases divided by the total follow-up time accumulated by all members of the population, or 'person-time' follow-up). The relative risk is the rate of disease (cumulative incidence rate or incidence density rate) in the exposed (e.g., those with a high intake of dietary fat) divided by the rate of disease in the unexposed (e.g., those on a low-fat diet). A relative risk of 2 implies that the exposed group has twice the rate of disease compared with the unexposed group.

For illustration, in a study of 121 700 women, a group of participants who completed dietary questionnaires and had no previous diagnosis of cancer in 1980, were followed through 1988 to address the hypothesis that dietary fat increases and fiber intake decreases the risk of breast cancer. This outcome was defined by histologically confirmed cases of breast cancer. In one analysis, the primary exposure of interest was energy-adjusted intake of total dietary fiber. Among the women in the highest quintile of energy-adjusted dietary fiber intake there were 299 cases of breast cancer compared with 305 cases among the women in the lowest quintile. This gave a relative risk (with adjustment for established breast cancer risk factors as well as alcohol intake) of 1.02 for those in the highest quintile of energy-adjusted dietary fiber intake compared with those in the lowest quintile.

There is also a growing body of evidence available from cohort studies for the assessment of dietary fat intake and breast cancer in developed countries. The average relative risk was 1.03. This observation was based on the results from nine prospective studies with at least 50 incident breast cancer cases each ($n = 2742$) and a large comparison series (i.e., noncases). At the same time, as these results suggest no overall association for total fat intake, emerging evidence suggests that monounsaturated fat may be protective against breast cancer.

The use of cohort studies can be advantageous in many ways when studying the relationship between diet and cancer. A cohort study allows the assessment of multiple effects of a given dietary exposure. Dietary data can be updated during follow-up and the temporal relation between diet and cancer can be addressed. For example, the beneficial effects of alcohol in reducing the risk of gallstone formation and coronary heart disease, and the potentially deleterious effects of alcohol on cancer and hemorrhagic stroke, can be weighed against each other in a cohort study. It is also possible to measure the absolute rates of disease according to the level of food or nutrient intake.

Among the limitations of cohort studies is the concern that current practice, usage, or exposure may change over the duration of the follow-up, limiting the ability to come to any relevant conclusions in studies of diet and cancer that have measured exposure just once at the beginning of the study. Controlling for extraneous variables such as smoking, which are related both to risk of cancer and to dietary intake, and separating the effects of specific dietary factors from those that exist together, also limit the range of knowledge that can be extracted from cohort studies.

Some investigators believe that the large number of subjects required to study rare disease and the high expense of management and maintenance also limit the usefulness of cohort studies. Others believe that the larger overall monetary investment most cohort studies require can be advantageous: More variables can be studied and in the long run further hypotheses can be generated and more conclusions produced than in a single case–control study that relies on recall of past habits.

Case–control studies In case–control studies information is obtained from diseased participants and compared with information provided by disease-free controls with respect to a possible risk factor (e.g., level of a dietary factor). Data collected from these studies can be used to evaluate the hypothesis that the risk factor is a cause of the disease. The cases are selected from a defined population, such as a country population. The population represents those at risk of developing the disease under study. Each time someone in the defined population is diagnosed with the disease during the duration of the study, this individual joins the case series. As each case arises from the population, one or more controls should be sampled to estimate the prevalence of the exposures among those remaining free from disease. The controls may be chosen from any population of individuals that provides valid information about those at risk for the disease. It is important to choose controls so that their probability of selection is unrelated to the exposure being studied.

In the study of the relationship of diet and cancer, case–control studies may be used to evaluate the hypotheses that individual or multiple dietary factors are the cause of the cancer under investigation. For example, a study in 1977 identified all cases of lung cancer diagnosed during an 18-month period from 1972 in three Singapore hospitals. Controls were chosen from other hospital patients free of any smoking-related diseases. There were a total of 233 cases and 300 controls interviewed regarding their frequency of consumption of dark-green leafy vegetables and food preparation habits. The investigation found a substantially increased risk of lung cancer among those reporting a low consumption of dark-green leafy vegetables.

Case–control studies are better suited to the study of rare diseases because in cohort studies tens of thousands of individuals must be followed in order to study the most common cancers. It is also thought that case–control studies may be quicker and less expensive to conduct because they require fewer subjects, and they are therefore often employed as an alternate mode of investigation to cohort studies.

Among the limitations of case–control studies is the comparability of information between the cases and the controls. While in a cohort study the exposure of interest is measured before the onset of disease, in case–control studies the exposure is assessed in individuals who (in most cases) already know their own disease status. Often the person collecting the data will also know the disease status of the patient. This may influence the accuracy of the data collected, either through differential recall by cases and controls, or by an interviewer being more persistent in questioning cases than controls. In cohort studies neither the participant nor the investigator knows whether or not the subject will be a case or noncase by the end of the follow-up period.

Intervention studies In principle, the most powerful means of determining the effects of dietary factors on cancer risk is an intervention study (i.e., a randomized trial). In randomized trials bias is removed because of the equal distribution of risk factors in each group. For example, it has been proposed that a randomized trial of fat reduction could help uncover the mystery of the relationship between dietary fat intake and breast cancer. The Women's Health Initiative was started by the US National Institutes of Health with the goal of enrolling and randomizing several tens of thousands of women, half of whom will be trained to follow a diet deriving less than 20% of energy from fat. Unfortunately, such a trial would not be able to address the most promising modification of the dietary fat hypothesis—that dietary fat reduction at an early age may reduce breast cancer risk several decades later. Other problems with such a randomized trial include the difficulty of maintaining compliance with a diet incompatible with prevailing food consumption habits, and the gradual secular decline in total fat consumption already under way which may reduce the size of the comparison of fat intake between the intervention group and the control groups. The Women's Health Initiative Trial will also counsel the women in the intervention group to adopt a diet that is high in fruits, vegetables, and grain products as well as low in total and saturated fat, therefore making it more difficult to distinguish between the effect of the fat reduction and that of increasing intake of fruits, vegetables, and grain products. All in all, intervention studies may in principle have a great chance of determining effects of dietary factors on cancer risk, but trials of sufficient duration and size may not be feasible because of long-term compliance and cost.

Epidemiological Issues in the Study of Diet and Cancer

Resolved and Unresolved Issues

Some of the issues that researchers have encountered in their attempt to uncover the mystery of the dietary factors linked to cancer include the difficulty of distinguishing the importance of parts of dietary factors from the overall effect of each dietary factor (e.g., total dietary fat intake compared with type of dietary fat intake). In a meta-analysis in 1990 of 12 case–control studies of dietary fat intake and cancer, 4 studies observed a significant positive association, 6 uncovered nonsignificant positive associations, and 2 saw inverse associations. When the data were analyzed together there was a positive association observed for both total fat intake and saturated fat intake. Investigators must ask themselves which factor has larger implications in the study of diet and cancer, as not all studies have included analyses of the individual types of fats along with their data on overall fat consumption. In the study of the influence of dietary fiber intake (which includes crude fiber and many soluble fiber fractions) on cancer rates, there is debate about the most appropriate method of biochemical analysis for determining fiber content of individual foods. This same issue arises with the study of most dietary factors and could affect any important advances in the study of diet and cancer.

Biochemical indicators of food and nutrient intake have two fundamental uses in epidemiological studies. Most often they serve as a 'surrogate' for actual dietary intake in studies of disease occurrence. For nutrients that vary widely in concentration within foods and for which food composition tables are inaccurate, biochemical indicators may be the most feasible way of measuring intake. Within-food variation may occur owing to differences in food storage, processing, or preparation, or may be owing to geographical differences in soil nutrient content. For example, it has been found that selenium content in US soil can vary by as much as 100-fold, which in turn causes the selenium content of swine muscle to vary more than 15-fold. Another example is that of fat. When the composition of fats in commercial food products is not known to study participants, it is possible to assess the fat components of the diet by subcutaneous fat aspirates which reflect long-term dietary patterns.

Like most exposures in chronic disease, nutrient exposures relevant to disease are usually long-term. As the promotion period for cancers may be years or decades, it is usually desirable that a biomarker indicates the cumulative effect of diet over an extended period of time. There are a couple of methods to surpass the barrier of an indicator that is only sensitive to short-term intake, and to overcome the day-to-day intake fluctuations that occur with most nutrients: (1) experimental studies, in which nutrient levels are manipulated; and (2) sampling levels in individuals longitudinally. Biomarkers of nutrient levels in blood or other tissues can provide a useful assessment of intake of certain nutrients, although the above considerations must be acknowledged, and careful attention must be given to specimen collection, storage, and analysis in order to avoid misclassification or bias. With an expanding array of biochemical indicators that have been validated as measures of dietary intake, their use in nutritional epidemiology will continue to grow.

The limited range of diet within most populations adds its own set of complexities to the epidemiological study of nutrition and cancer. For example, in the majority of populations where foods high in fat are readily available, very few individuals consume less than 30% of their energy from fat. This makes it difficult to study the impact of reducing fat intake to less than 30% of total energy intake. At the same time, some individuals of a relatively homogencous population may have very different dietary patterns: For example, a range of dietary fat intake from 25% to 40% of total energy was seen within a cohort of 52 000 male health professionals in the USA.

Given that most neoplasms have a long induction period (the time from an exposure to a carcinogen to the development of cancer), often spanning several decades, accurate measure of long-term dietary intake is of utmost importance in the study of the implications of diet on cancer. Therefore, short-term methods of dietary assessment such as 24-h recalls are usually insufficient. In the context of case–control studies these short-term methods are inappropriate because they measure current diet, and it has been found that individuals alter their diet after the diagnosis of cancer. The most feasible method of measuring long-term intakes in large numbers of individuals is the food frequency questionnaire: These questionnaires measure the usual frequency of a selected list of foods.

Food frequency questionnaires to assess dietary intake need to be carefully designed. First of all, the food items on the questionnaire must represent the major source of nutrients of interest within the study population. Depending on the consistency of the concentration of a nutrient in a given food, the precision of dietary questionnaires varies. Food frequency questionnaires may provide rankings by level of intake but they do not quantify actual intake. A dietary questionnaire may efficiently distinguish between participants with low-fiber and high-fiber

intakes in a given population, but it will not necessarily provide a precise assessment of the absolute fiber intake. In the case of larger studies, it is possible for a random sample of participants to provide a more comprehensive assessment of intake by keeping several weeks of dietary records. This additional information will provide a more precise quantification of dietary intake by helping estimate true dose–response relationships between a nutrient and diet expressed in absolute intake.

Summary of Known Relations between Diet and Cancer

A wealth of studies since the 1970s have clearly documented the relations between diet and a growing number of cancers (**Table 2**). Convincing evidence based on consistent findings from epidemiological studies conducted in diverse populations now shows that diet is an established cause of prostate, breast, digestive tract, airway, and urinary tract cancers. With these rich epidemiological data we can more confidently conclude that some 30% of cancer is attributable to diet. Public health officials have taken the accumulated evidence and developed strategies for minimizing cancer risk. Among these recommendations is a diet high in vegetables, fruits, and legumes and low in red meat, saturated fat, salt, and sugar. They suggest that carbohydrates be consumed as whole grains such as whole meal bread and brown rice rather than as white bread and rice. Any added fats should come from plant sources and should be unhydrogenated, an example being olive oil, which may potentially be beneficial. Given the evergrowing knowledge of the association between diet and cancer, and the subsequent recommended prevention strategies, it is time that researchers and public health officials combined their efforts

Table 2 Levels of evidence for major forms of cancer by foods, energy-generating nutrients, dietary exposure to selected nonnutrients, and nutrition-related indicators

	Suggestive evidence	*Strong evidence*	*Convincing evidence*
Increased risk			
Major food groups			
– Cereals	Stomach		
– Meat	Pancreas	Large bowel	
– Eggs and egg products	Stomach		
– Sugars	Large bowel		
Macronutrients			
– Proteins (animal)	Large bowel, pancreas, endometrium		
– Carbohydrates (total)	Stomach, pancreas		
– Saturated fat (animal)	Large bowel, lung, endometrium		Prostate
Nonnutrients			
– Alcohol		Oesophagus, pancreas	Oral cavity, large bowel, cervix uteri, breast
– Salt (NaCl)		Stomach	Nasopharynx
Nutritional covariates			
– Height	Ovary, prostate	Breast, large bowel	Large bowel, premenopausal breast = inverse association postmenopausal breast = positive association, endometrium, kidney, oesophagus
– Obesity			
– Hot drinks		Oesophagus	
Decreased risk			
Major food groups			
– Vegetables	Liver, pancreas, breast, endometrium, cervix uteri, ovary, prostate	Oesophagus, stomach, larynx, lung, urinary bladder	Oral cavity, large bowel, kidney
– Fruits	Liver, pancreas, breast, endometrium, cervix uteri, ovary, kidney	Oral cavity, oesophagus, stomach, larynx	Large bowel, lung, urinary bladder
Macronutrients			
– Fiber	Large bowel, pancreas		
– Monounsaturated fat	Breast		
Nutritional covariates			
– Physical activity	Endometrium, prostate	Breast	Large bowel, breast
– Folate			Colon, breast

not only to uncover the mysteries of diet and cancer but also to balance the 'war on cancer' treatment with more extensive efforts in prevention.

See also: **Alcohol**: Disease Risk and Beneficial Effects. **Cancer**: Epidemiology of Gastrointestinal Cancers Other Than Colorectal Cancers; Epidemiology of Lung Cancer; Effects on Nutritional Status. **Dietary Fiber**: Potential Role in Etiology of Disease. **Dietary Surveys**. **Vegetarian Diets**.

Further Reading

Ames BN, Gold LS, and Willett WC (1995) The causes and prevention of cancer. *Proceedings of the National Academy of-Sciences USA* **92**: 5258–5265.

Armstrong B and Doll R (1975) Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. *International Journal of Cancer* **15**: 617–631.

Colditz GA and Willett W (1991) Epidemiologic approaches to the study of diet and cancer. In: Alfin-Slater RB and Kritchevsky D (eds.) *Cancer and Nutrition*. New York: Plenum.

Doll R and Peto R (1981) The causes of cancer: Quantitative estimates of avoidable risks in the United States today. *Journal of the National Cancer Institute* **66**: 1191–1308.

Giovannucci E, Stampher MJ, Colditz GA *et al.* (1993) A comparison of prospective and retrospective assessments of diet in the study of breast cancer. *American Journal of Epidemiology* **137**: 502–511.

Seow A, Duffy SW, McGee MA *et al.* (1996) Breast cancer in Singapore: Trends in incidence 1968–1992. *International Journal of Epidemiology* **25**: 40–45.

Trichopoulos D, Li FP, and Hunter DJ (1996) What causes cancer? *Scientific American* **275**: 80–87.

Trichopoulos D and Willett WC (eds.) (1996) Nutrition and cancer. *Cancer Causes and Control* **7**: 3–180.

Willett WC (1989) *Nutritional Epidemiology*. Oxford: Oxford University Press.

Willett WC, Colditz GA, and Mueller NE (1996) Strategies for minimizing cancer risk. *Scientific American* **275**: 88–95.

Wydner EL, Yasuyuki F, Harris RE *et al.* (1991) Comparative epidemiology of cancer between the United States and Japan. *Cancer* **67**: 746–763.

Epidemiology of Gastrointestinal Cancers Other Than Colorectal Cancers

H-Y Huang, Johns Hopkins University, Baltimore, MD, USA

This article addresses the epidemiology of esophageal cancer, stomach cancer, pancreatic cancer, and small intestine cancer. People with any of these cancers are often diagnosed at 60–80 years of age. The incidences are higher among men than among women and vary widely with geographic location and population, suggesting that environmental factors are important in the development of these cancers.

Esophageal Cancer

The esophagus is a hollow tube, approximately 10 in long in adults. It conveys food from the pharynx to the stomach. Mucous glands in the lining of the esophagus secret mucus to aid in lubrication. Absorption in the esophagus is nil.

Descriptive Epidemiology

Worldwide, esophageal cancer is the eighth most common cancer and the sixth most common cause of cancer death, accounting for approximately 450 000 new cases and a similar number of deaths in 2002. More than 70% of esophageal cancer is squamous cell carcinoma, and approximately 20% is adenocarcinoma. Squamous cell carcinoma arises from dysplasia in the middle and lower third of the esophagus epithelial lining, whereas adenocarcinoma usually develops in the glandular tissue in the distal esophagus. The incidence of esophageal cancer varies tremendously with geographic location and populations throughout the world, with a maximum ratio of 500 to 1. In central and Southeast Asia, the Far East, and the Middle East, squamous cell carcinoma is the predominant form of esophageal cancer, whereas in the United States and Europe, adenocarcinoma of the esophagus has been rapidly increasing since 1970s, particularly in Caucasian men, to approach or surpass the rate of squamous cell carcinoma. In the United States, African Americans have an approximately 2-fold increased risk for esophageal cancer compared to Caucasians, possibly because of an unhealthy lifestyle.

Disease Process

There may be no symptoms of esophagus cancer during the early stages. As the cancer develops, nonspecific symptoms occur, including dysphagia, weight loss, chronic cough, and pain in the retrosternal, back, or right upper abdomen. In more than 50% of esophagus cancer cases, the cancer is either unresectable or has metastasized at the time of diagnosis. The prognosis of esophagus cancer depends on disease stages and tumor sizes. For resectable esophagus cancer, the 5-year survival rate ranges from 15 to 24%. For metastasized esophagus cancer, the 5-year survival rate is less than 5%.

Although both squamous cell carcinoma and adenocarcinoma of the esophagus are responsive to chemotherapy, the treatment effect rarely lasts more than 1 year. Radiotherapy may reduce the chance of perioperative morbidity and mortality, but it may increase the risk for local and regional complications such as esophagotracheal fistulas. Research is under way to determine whether an improved treatment efficacy can be achieved by combined chemotherapy, radiotherapy, and surgery.

Risk Factors

Squamous cell carcinoma Factors that cause chronic irritation and esophageal mucosa inflammation may increase the risk for esophageal squamous cell carcinoma. These factors include moderate to heavy alcohol drinking, smoking, achalasia, diverticuli, and consumption of extremely hot beverages, coarse grains or seeds, lye, and caustic spices.

The importance of alcohol consumption in the carcinogenesis of esophageal squamous cell carcinoma is well recognized. However, the mechanisms by which alcohol increases cancer risk have not been elucidated. Alcohol may cause chronic irritation to the esophagus, and it may increase cell proliferation and enhance the permeability of carcinogens to cells. An alcohol metabolite, acetaldehyde, is known to be a carcinogen. Risk for esophageal squamous cell carcinoma is higher for spirits drinkers, followed by wine and beer drinkers.

Cigarette smoke is a rich source of carcinogens, such as benzo(*a*)pyrene and volatile nitrosamines. It also contains free radicals, reactive oxygen species, and reactive nitrogen species that are capable of initiating and propagating oxidative damage to lipids, proteins, and DNA, leading to several degenerative diseases including cancer. Alcohol drinking may account for approximately 80% of squamous cell esophageal cancer cases, whereas tobacco use may account for approximately 60%. Simultaneous use of alcohol and tobacco further increases esophageal cancer risk.

Achalasia is a swallowing disorder caused by degeneration of the intrinsic autonomic nerves in the esophagus wall and lower esophageal sphincter, leading to decreased or absent peristalsis in the esophageal smooth muscle or impaired relaxation of the lower esophageal sphincter. Approximately 20–29% of achalasia patients may develop esophageal cancer within 15–20 years, predominantly squamous cell carcinoma, possibly because of increased inflammation, bacterial growth, and chemical irritation caused by prolonged contact of food ingredients with esophageal mucosa. In contrast, the likelihood of malignant transformation from diverticuli is less than 1%, although the mechanisms of carcinogenesis are speculated to be the same as those for achalasia.

Low income is associated with squamous cell carcinoma of the esophagus, independent of alcohol and tobacco use, suggesting that other factors associated with poverty may play a role. In Africa and Far East countries, incidences of esophageal cancer are high in regions where starchy food is the predominant food in the diet, and this may have been an indication of poor nutritional status. Several studies have reported that very low intake of fresh fruits and vegetables is associated with higher risk of esophagus cancer. Conversely, high intake of fruits and vegetables, particularly citrus fruits, may confer preventive benefits. Frequent consumption of highly salted meat, pickled vegetables, cured meat, and smoked meat was found to be associated with esophageal cancer risk; these foods contain carcinogenic compounds such as heterocyclic amines and *N*-nitroso compounds.

Familial aggregation of esophageal squamous cell carcinoma has been reported, but it may reflect genetic predisposition as well as common environmental exposures. Hereditary squamous cell carcinoma of the esophagus develops in approximately 95% of people with a genetic abnormality at chromosome 17q25 that causes a rare autosomal dominant disorder, nonepidermolytic palmoplantar keratoderma.

Adenocarcinoma The risk factor profile of esophageal adenocarcinoma is quite different from that of squamous cell carcinoma. Tobacco use is associated with adenocarcinoma of the esophagus, but the association is less strong than that with squamous cell carcinoma. High intakes of fiber, vitamin C, vitamin B_6, folate, and β-carotene were found to be associated with a lower risk. However, unlike squamous cell carcinoma, esophageal adenocarcinoma does not consistently develop more often in people with frequent alcohol consumption or low income.

Gastroesophageal reflux disease (GERD) is strongly associated with adenocarcinoma of the esophagus. In the process of gastroesophageal reflux, acid fluid regurgitates into the gastroesophageal junction and causes a sensation of heartburn. GERD can be caused by hiatal hernia, esophageal ulcer, and use of drugs that relax the lower gastroesophageal sphincter and increase reflux. Alcohol, tobacco, obesity, and pregnancy may also contribute to GERD.

Barrett's esophagus represents intestinal metaplasia of the squamous epithelium in the distal esophagus. Barrett's esophagus develops in approximately

5–10% of people with GERD and is associated with a 30- to 125-fold increased risk for esophageal adenocarcinoma.

In the United States, the incidence of adenocarcinoma of the esophagus has increased more than 350% since the 1970s. Obesity has been hypothesized to be one of the factors responsible for this increase by augmenting abdominal pressure and gastroesophageal reflux frequency. However, evidence has not been consistent to support this hypothesis.

Prevention

For primary prevention, smoking cessation and avoidance of heavy alcohol intake may significantly reduce the risk for squamous cell carcinoma. A healthful diet with fresh fruits and vegetables but no highly salted, preserved, or smoked food should lead to a reduction in the risk for both of the major forms esophageal cancer. For secondary prevention, routine screenings by endoscopes may confer benefits to individuals with Barrette's esophagus. Treatment with endoscopic ablation combined with proton pump inhibitors may retard Barrett's esophagus to normal squamous mucosa.

Research has been under way to determine the chemopreventive effects of 13-*cis*-retinoic acid, nonsteroidal antiinflammatory drugs (e.g., aspirin and sulindac), selenium, and ornithine decarboxylase inhibitor, α-difluoromethylornithine, in patents with Barrett's esophagus. These agents also hold promise for preventing squamous cell carcinoma. Other chemopreventive agents that may be useful for reducing both types of esophageal cancer include ascorbic acid, polyphenols (e.g., ellagic acid and epigallocatechin-3-gallate), and sulfhydryl compounds. These agents have been shown in animal models to inhibit nitrosamine formation and enhance the activities of detoxifying enzymes such as glutathione-*S*-transferase and glutathione peroxidase, but evidence from humans is sparse.

Stomach Cancer

The stomach is located between the esophagus and the duodenum on the left side of the abdominal cavity. It serves as a short-term reservoir of foodstuff and provides digestive functions. Gastric epithelium secrets mucus, hydrochloric acid, hormones (e.g., gastrin), protease, lipase, gelatinase, and other enzymes. The movement of the stomach is controlled by the autonomic nervous system and several hormones in the digestive system.

Descriptive Epidemiology

Worldwide, stomach cancer is the fourth most common cancer and the second most common cause of cancer deaths, accounting for approximately 989 000 new cases and 850 000 deaths in 2001 and 2002, respectively. Stomach cancer can be classified as diffuse or intestinal. The former has an earlier onset with similar occurrences by sex and by geographic areas, whereas the latter has a later onset and develops more often in men than women. There is a wide variation (more than 10-fold) in the incidence of the intestinal type, suggesting that environmental factors are important determinants. Japan, Korea, China, Eastern Europe, Central America, and South America have higher incidences, whereas southern Asia, India, North America, and Africa have lower incidences. A wide range of incidence also occurs within countries.

The incidence and mortality rates of stomach cancer have been declining for several decades because of a reduction in childhood *Helicobacter pylori* infection, improved nutritional status, and reductions in exposures to carcinogens in preserved food. However, because of increases in life expectancy, the absolute number of stomach cancer cases has been increasing.

Disease Process

Approximately 90% of stomach cancer is adenocarcinoma. Other forms of stomach cancer include lymphomas and sarcomas. Symptoms such as excessive belching, heartburn, stomachache, and back pain may occur. Internal bleeding may appear as blood in the vomit or as black, tar-like feces, or it may be so slight that it is undetected. The prognosis of stomach cancer is poor and dependents on disease stages; 5-year survival rate is approximately 20% in the United States.

Risk Factors

***Helicobacter pylori* infection** *Helicobacter pylori* infection can cause inflammatory responses that induce atrophic gastritis and intestinal metaplasia of gastric mucosa, resulting in reduced gastric acidity, which in turn facilitates in vivo formation of carcinogenic *N*-nitroso compounds and leads to the intestinal type of stomach cancer. In addition, *H. pylori* infection can trigger a cascade of inflammatory responses and oxidative damage to induce cell proliferation and malignant transformation, leading to the diffuse type of stomach cancer. It was estimated that *H. pylori* infection accounted for approximately 50–60% of stomach cancer cases and was associated with an approximately sixfold increased risk at least 10 years prior to

diagnosis. These may have been underestimated because of the possibility of loss of the infection or antibody due to extensive replacement of gastric mucosa with intestinal metaplasia in people with stomach cancer.

Infection of *H. pylori* is common—50% worldwide and 90% in developing countries. However, only a small percentage develops into stomach cancer, suggesting that factors such as diet and genetic susceptibility modify risk.

Dietary factors Pickled vegetables and smoked, cured, salted, or dried fish or meat contain nitrite or *N*-nitroso compounds. These preserved foods, as well as grilled or charcoal flame-broiled food that contains polycyclic aromatic hydrocarbons, have been shown to be associated with increased risk of stomach cancer in most studies. Despite the fact that vegetables are a major source of nitrate, evidence suggests an inverse association between fresh fruits and vegetables and stomach cancer risk; the associations for yellow- or green-colored vegetables and citrus fruits are particularly strong. A few studies have reported a lower risk for stomach cancer among people consuming more allium vegetables, onions, and garlic. Some, but not all, studies have found a positive association between starchy food consumption and stomach cancer risk.

Vitamin C intake is consistently found to be inversely associated with stomach cancer risk in observational studies. Vitamin C can act as a powerful water-soluble antioxidant as well as an effective scavenger of nitrite. Protective roles of α-tocopherol and β-carotene are suggestive but less strong. These micronutrients may also be surrogate markers of healthy dietary pattern or lifestyle.

Evidence is inconsistent regarding the role of alcohol, coffee, or black tea consumption in the development of stomach cancer. However, green tea consumption was associated with a lower risk in several studies, presumably because of its polyphenol content.

High intake of salt is associated with a higher risk of stomach cancer. Animal studies have demonstrated that salt per se can damage gastric mucosa and induce gastritis. However, in humans, high salt intake correlates positively with intake of processed meat or fish that contains nitrosamines. Hence, it is unclear whether salt evokes stomach cancer or is merely a marker of other exposures.

Cigarette smoking Tobacco use is associated with a 1.5- to 2.0-fold increased risk for stomach cancer, and it has been estimated to account for 10–17% of stomach cancer cases. These estimates may have been confounded by other factors such as poor diet.

Familial factors Familial aggregation of stomach cancer derives mostly from common environmental exposures and lifestyle factors. Hereditary stomach cancer is rare. Germline mutations in the gene coding for cell adhesion protein E-cadherin (CDH1) were found to be associated with stomach cancer of the diffuse type. Germline mutation of *p53* has also been reported. People with hereditary nonpolyposis colorectal cancer are also at higher risk for stomach cancer.

Prevention

Evidence points to the importance of improving diet and eradicating *H. pylori* infection. A diet rich in fresh fruits and vegetables without highly salted, preserved, or smoked food will theoretically offer benefits in primary prevention. In countries where the incidence of stomach cancer is high, screening for *H. pylori* may be effective for secondary prevention. To this end, programs of mass screening and eradication of *H. pylori* by antibiotics are being performed in Japan. However, because only a small proportion of individuals with *H. pylori* colonization develop stomach cancer, concerns have been raised regarding the possibility of antibiotic resistance by a mass *H. pylori* eradiation program. Use of vaccines against *H. pylori* may be an alternative approach.

Pancreatic Cancer

The pancreas is an elongated organ locating in close proximity to the duodenum. It consists of three parts—head, body, and tail—and is partitioned into lobules by connective tissue. Approximately 85% of the pancreas is composed of exocrine cells called acini that secret digestive enzymes such as proteases, lipase and amylase, ribonuclease, gelatinase, deoxyribonuclease, and elastase. These digestive enzymes, together with bicarbonate secreted from the epithelial cells lining small pancreatic ducts, enter into pancreatic ducts and subsequently to the lumen of the duodenum. Embedded in the exocrine tissue are endocrine tissues called Islets of Langerhans that secret endocrine enzymes, such as insulin and glucagon.

Descriptive Epidemiology

Worldwide, there were approximately 230 000 pancreatic cancer cases and a similar number of deaths due to pancreatic cancer in 2002. Pancreatic cancer is the fifth and sixth most common cause of cancer death in men and women, respectively, in most Western countries. The incidence of pancreatic

cancer has declined slightly, with an average annual change of −0.04%, from 1975 to 1998 in the United States, presumably as a result of smoking cessation. In contrast, the incidences in Japan and European countries are increasing.

Disease Process

Adenocarcinoma in the head of the pancreas accounts for 80–90% of pancreatic cancer. Pancreatic cancer is a devastating disease because it is rapidly fatal; the case fatality ratio is 0.99, median survival is 6 months or less, 1-year survival is approximately 20–30%, and 5-year survival is less than 5%. There is no effective screening modality for pancreatic cancer. The disease is difficult to diagnose and detect because the disease process is either silent or present with nonspecific symptoms, such as unexplained weight loss, back pain, nausea, jaundice, and altered intestine habits. In approximately 80–90% of cases, the cancer is diagnosed at a nonresectable stage when even small tumors have metastasized to other organs, most commonly the liver. Patients undergo cachexia, a complex metabolic syndrome clinically presenting with progressive weight loss and depletion of reserves of adipose tissue and skeletal muscle. Pancreatic cancer cells are particularly resistant to radiotherapy and chemotherapy, rendering the treatment unsuccessful. The lack of a useful screening tool and the poor prognosis of this disease highlight the importance of primary prevention.

Risk Factors

The etiology of pancreatic cancer is largely unknown. Prospective follow-up epidemiologic studies are the better study designs for determining a temporal relationship between exposures and disease outcomes. However, the rarity of pancreatic cancer makes it difficult to examine an association with sufficient statistical power. Most studies are case–control designs in which information of lifestyle and environmental exposures is collected from pancreatic cancer cases or proxies after cancer diagnoses and from selected controls with no pancreatic cancer. Such study design is prone to recall biases and information biases, and a temporal relationship cannot be determined. Once nonspecific symptoms occur, the aggressive disease processes make it difficult to complete data collection before a patient dies of the disease.

To date, the only risk factors of pancreatic cancer that have been well accepted are oldage and cigarette smoking. Pancreatic cancer is more common in men than women, possibly because of differences in lifestyle factors and environmental exposures. African Americans, New Zealand Maoris, native Hawaiians, and Jews have higher incidences, whereas individuals in India or Nigeria and Seventh-Day Adventists have lower incidences. Hereditary pancreatitis and germline mutations may account for 10–15% of pancreatic cancer cases. Purported but unproven risk factors include diet, obesity, diabetes mellitus, chronic pancreatitis, *H. pylori* colonization, gastric or duodenal acidity, and occupational exposures to carcinogens. Socioeconomic status is not associated with pancreatic cancer risk.

Cigarette smoking Cigarette smoking has been consistently shown to be associated with a 2- or 3-fold increased risk and accounts for 25–30% of pancreatic cancer cases. Higher risk has been associated with increased numbers of cigarettes smoked. Cigarette smoking may interact with hereditary factors to increase pancreatic cancer risk. It was estimated that smokers who had a family history of pancreatic cancer had a sixfold increased risk compared to nonsmokers who did not have a family history, whereas a three- or fourfold increased risk was found in nonsmokers who had a family history or smokers who did not have a family history compared to nonsmokers with no family history.

Inherited gene mutations Inherited mutations of genes account for approximately 10% of pancreatic cancer cases. *BRCA2* mutations are the most common, accounting for approximately 7%. Pancreatic cancer caused by these mutations often presents as 'apparently sporadic' because of the low penetrance of *BRCA2* mutations. High-penetrance germline mutations in the *CDKN2A* (*p16*) gene that cause the familial atypical multiple mole and melanoma syndrome are also associated with higher risk for pancreatic cancer. Inherited mutations of *LKB/STK11* gene cause the Peutz–Jeghers syndrome, characterized by hamartomatous gastrointestinal polyps, mucocutaneous melanotic spots, and, in 30% of patients, pancreatic cancer. Inherited defects in a DNA mismatched repair gene causing hereditary nonpolyposis colorectal cancer and inherited mutations of the cationic trypsinogen gene causing acute pancreatitis at young age may also cause pancreatic cancer.

Dietary factors Higher intakes of fat, carbohydrate, animal protein, fried food, cured meat, or smoked meat have been associated with a higher risk for pancreatic cancer. In contrast, higher intakes of vitamin C, fiber, or, more generally, fresh fruits and vegetables and higher serum concentrations of

folate and pyridoxine are associated with lower risk for pancreatic cancer. Alcohol, tea, or coffee consumption is not associated with pancreatic cancer.

Diabetes mellitus The temporal relationship between diabetes and pancreatic cancer is uncertain. A twofold increased risk of pancreatic cancer has been reported for people diagnosed with diabetes at least 5 years prior to pancreatic cancer diagnosis. The latency period of pancreatic cancer is unknown, but an estimate of at least 10 years has been reported. Hence, diabetes may be a consequence rather than a cause of pancreatic cancer. Interestingly, familial pancreatic cancer was not found to be associated with diabetes. In addition, approximately 50% of individuals who have non-insulin-dependent diabetes mellitus are not aware of the disease, and many pancreatic cancer patients are diagnosed with diabetes at the time of the cancer diagnosis.

Chronic pancreatitis Both hereditary and sporadic forms of chronic pancreatitis have been found to increase pancreatic cancer risk. In the inflammatory processes of chronic pancreatitis, cytokines, reactive oxygen species, and mediators of the inflammatory pathway (e.g., NF-κB and COX-2) may increase cell turnover, cause loss of tumor suppressor genes, stimulate oncogene expression, and lead to pancreatic malignancy. Heavy alcohol consumption may increase the risk of chronic pancreatitis.

Prevention

Smoking cessation may be the first choice for the primary prevention of pancreatic cancer. However, evidence to support this rationale is lacking. An effective screening modality for pancreatic cancer has not been developed.

Small Intestine Cancer

The small intestine is approximately 20 feet long and consists of three sections: the duodenum, jejunum, and ileum. The small intestine performs extensive digestion and absorption functions. It also secrets secretin, which stimulates the pancreas to produce digestive enzymes.

Descriptive Epidemiology

Cancer of the small intestine is very rare; the age-adjusted incidence is approximately 1.4 per 100 000—less than 2% of all gastrointestinal malignancies. The incidence of small intestine cancer is higher in Maori of New Zealand and Hawaiians, and it is lower in India, Romania, and other areas of Eastern Europe. In the United States, the incidences of adenocarcinoma, lymphoma, and carcinoid have only slightly increased since 1980s; even for lymphoma, which has had the largest increase, the annual rate of increase has been no more than 1 per 1 million.

Disease Process

There are four types of small intestine cancer, each with unique characteristics: adenocarcinoma, carcinoid, lymphoma, and sarcoma. In Western developed countries, approximately 30–40% of small intestine cancer is adenocarcinoma, predominantly in the duodenum, and carcinoid and lymphoma occur more often in the jejunum or ileum, whereas sarcoma may develop anywhere in the small intestine. In developed countries, lymphoma is very rare and occurs more often in older people with relatively good survival. In contrast, in developing countries, lymphoma is the main type of small intestine cancer, and it occurs more often in younger individuals, anywhere in the small intestine, with poor survival. Hence, prognosis of small intestine cancer depends on the type, geographic location (which may be an indication of etiology and/or the advancement of treatment), and disease stages. Clinical presentation may include abdominal pain, weight loss, abdominal mass, anemia, nausea/vomiting, bleeding, obstruction, jaundice, and anorexia before diagnosis. Overall, the 5-year survival rate is approximately 80% for carcinoid, 60% for lymphoma, 45% for sarcoma, and 20% for adenocarcinoma.

Risk Factors

Due to the rarity of small intestine cancer, etiologic investigation has relied on only a few small case–control studies. A lack of histology data has further undermined the strength of the evidence.

Tobacco use, alcohol consumption, and dietary factors such as high animal protein, high animal fat, sugar, and salted, cured, or smoked food were associated with small intestine cancer in some but not all studies. Small intestine adenoma, familial adenomatous polyposis, hereditary nonpolyposis colorectal cancer, peptic ulcer, celiac sprue, and cholescystectomy have been found to be associated with increased risk for small intestine adenocarcinoma. In people with Crohn's disease, a 16- to more than 100-fold increased risk for small intestine adenocarcinoma has been reported, but unlike most adenocarcinomas that occur in the duodenum, these patients tend to have adenocarcinomas in the elium. The reasons for the increased risk are uncertain, but

it has been hypothesized to be due to the medication for treating Crohn's disease.

Prevention

Because very little is known about the etiology of small intestine cancer, no preventive strategy has been proposed.

Conclusion

The wide variation in the incidences of cancers of the esophagus, stomach, pancreas, and small intestine by geographic location and by population suggest that environmental factors play an important role in the etiology. Indeed, several risk factors are commonly shared by these cancer sites, including tobacco use, a diet low in fresh fruits and vegetables, and a diet high in salted, cured, or smoked food. Strategies for gastrointestinal cancer prevention should aim to counteract these risk factors. In addition, avoidance of heavy alcohol consumption and eradication of *H. pylori* may significantly reduce the incidence of esophageal cancer and stomach cancer, respectively. Studies are under way to test the efficacy of chemoprevention agents in the prevention of esophageal cancer and stomach cancer in high-risk populations. The development of noninvasive screening tests, such as molecular or imaging technology, is needed for early detection and better prognosis.

See also: **Alcohol**: Disease Risk and Beneficial Effects. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Diabetes Mellitus**: Etiology and Epidemiology. **Fruits and Vegetables**. **Small Intestine**: Disorders. **Stomach**: Disorders.

Further Reading

Enzinger PC and Mayer RJ (2003) Esophageal cancer. *New England Journal of Medicine* **349**(23): 2241–2252.

Ghadirian P, Ekoe JM, and Thouez JP (1992) Food habits and esophageal cancer: An overview. *Cancer Detection and Prevention* **16**(3): 163–168.

Heath EI, Limburg PJ, Hawk ET, and Forastiere AA (2000) Adenocarcinoma of the esophagus: Risk factors and prevention. *Oncology (Huntington)* **14**(4): 507–514.

Lange J and Siewert JR (eds.) (2000) *Esophageal Carcinoma: State of the Art*. New York: Springer.

Levin B (1999) An overview of preventive strategies for pancreatic cancer. *Annals of Oncology* **10**(supplement 4): 193–196.

Lowenfels AB and Maisonneuve P (1999) Pancreatic cancer: Development of a unifying etiologic concept. *Annals of the New York Academy of Sciences* **880**: 191–200.

Neugut AI, Jacobson JS, Suh S, Mukherjee R, and Arber N (1998) The epidemiology of cancer of the small bowel. *Cancer Epidemiology Biomarkers & Prevention* **7**(3): 243–251.

Palli D (2000) Epidemiology of gastric cancer: An evaluation of available evidence. *Journal of Gastroenterology* **35**(supplement 12): 84–89.

Plummer M, Franceschi S, and Munoz N (2004) Epidemiology of gastric cancer. *IARC Scientific Publications* **2004**(157): 311–326.

National Cancer Institute (2001, February) Report of the Pancreatic Cancer Progress Review Group. Available at http://prg.nci.nih.gov/pancreatic/default.html.

National Cancer Institute (2002, December) Report of the Stomach/Esophageal Cancers Progress Review Group. Available at http://prg.nci.nih.gov/stomach/finalreport.html.

Sharma P and Sampliner RE (eds.) (2001) *Barrett's Esophagus and Esophageal Adenocarcinoma*. Malden, MA: Blackwell Science.

Epidemiology of Lung Cancer

A J Alberg and J M Samet, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

At the start of the twentieth century, lung cancer was a rare disease, whereas by its end it had become a leading cause of death and the most common cause of cancer death in the United States. The occurrence of most cases can be explained by environmental agents; cigarette smoking was identified as its predominant cause in the 1950s. With the role of smoking well established, research on lung cancer has focused on environmental and genetic factors that may determine lung cancer risk in smokers. Genetic factors have been the subject of increased scrutiny because the risk of cancer may be determined in part by interindividual variation in the metabolism and detoxification of environmental agents, such as cigarette smoke, as well as variation in susceptibility to DNA damage and in DNA repair capability.

The role of diet as an environmental factor that may determine lung cancer risk in smokers remains a topic of considerable interest. Studies on diet and lung cancer in humans began in the 1970s as part of a broader search for factors determining susceptibility to the carcinogenic effects of tobacco smoke. Early animal studies showed that vitamin A depletion caused loss of differentiation of the respiratory epithelium, a histopathological state analogous to the dysplasia found in cigarette smokers at risk for lung cancer. In animal studies, these changes reverted with nutrient repletion, raising the prospect that dietary interventions could reduce lung cancer risk in cigarette smokers. Early epidemiological studies indicated that indices of vitamin A consumption were associated, in the expected protective direction,

with lung cancer risk. An influential 1981 publication by Buckley and colleagues shifted emphasis to β-carotene rather than retinol, and by the early 1980s clinical trials of β-carotene as a chemopreventive agent had been initiated. Observational evidence continued to show that measures of vitamin A and carotene intake were inversely associated with lung cancer risk; the weight of evidence favored fruits and vegetables as the carriers of the chemopreventive agents. In the mid-1990s, the hypothesis that β-carotene and retinol were protective was disproved in large clinical trials that unexpectedly showed increased risk in smokers randomized to the active agents. Hypotheses concerning specific carotenoids and other dietary components have since been advanced along with more general theories involving dietary antioxidants and oxidant stress from smoking and other factors.

Lung Cancer

Respiratory Carcinogenesis

The term 'lung cancer' refers to a histologically and clinically diverse group of malignancies arising in the respiratory tract, primarily but not exclusively from cells lining the airways of the lung. Beginning with the trachea, the airways branch dichotomously through 20 or more generations. Most cancers arise in the larger airways of the lung, typically at the fourth through the eighth generations. There, the airways are lined by a ciliated epithelium that includes secretory cells and glands and also neuroepithelial cells. The specific cells of origin of lung cancer are unknown; candidates include the secretory cells, pluripotential basal cells, and neuroepithelial cells. Only a small proportion of lung cancers in smokers have been considered as originating in the lung's periphery, but with the current trend of increasing adenocarcinoma this proportion may be increasing.

Lung cancer is thought to arise from a sequence of genetic changes that move a cell from a normal to a malignant state. Diverse genetic changes in oncogenes and tumor suppressor genes have been found in lung cancers, although the specific longitudinal sequence of these changes has not been characterized. Nonetheless, our evolving understanding of respiratory carcinogenesis, as a sequential progression from normal cell to clinical cancer, implies that there may be multiple points for interrupting the sequence and thereby preventing cancer.

Risk Factors for Lung Cancer

The increase in the incidence of lung cancer during the first half of the twentieth century prompted intensive epidemiological investigation of the disease, with the identification of a number of causal agents. Cigarette smoking is by far the most prominent cause of lung cancer, and the worldwide epidemic of lung cancer is largely attributable to smoking. However, occupational exposures have placed a number of worker groups at high risk and there is evidence that indoor and outdoor air pollution also increases lung cancer risk generally. The observed familial aggregation of lung cancer suggests that genetic factors may also determine risk. Extensive research is in progress on the specific genes that may determine risk in smokers; experience to date has supplied some leads but the evidence has been mixed for most of the genes studied.

In smokers, the risk of lung cancer depends largely on the duration of smoking and the amount smoked; risk increases exponentially with both, but more steeply with duration than amount. A safe level of smoking has not been shown, and even the secondhand tobacco smoke involuntarily inhaled by nonsmokers increases lung cancer risk. Fortunately, lung cancer risk declines in those who stop smoking, although not to the level of those who have never smoked; risk is present even after 30 years of abstention and currently in the United States approximately half the lung cancer cases occur in former smokers.

A number of occupational exposures increase lung cancer risk; the substances involved include radon (found in underground mines), arsenic, asbestos, chromium, chloromethyl ethers, nickel, and polycyclic aromatic hydrocarbons. For these agents, risk increases with the level of exposure, and synergism with smoking has been shown for several, such as asbestos and radon. Many other agents are suspected occupational carcinogens.

Indoor and outdoor air also contains respiratory carcinogens. Combustion sources contaminate outdoor air with polycyclic aromatic hydrocarbons and radionuclides, and outdoor air pollution is thought to contribute to a few percent of lung cancers in general. Carcinogens in indoor air vary with the setting but may include radon, tobacco smoke, smoke from wood or coal burning, and cooking fumes. In the United States, radon is estimated to cause approximately 14 000 lung cancer deaths annually.

Lung Cancer Histopathology

As assessed by the clinical approach of light microscopy, primary cancer of the lung occurs as multiple histological types, the most common being squamous cell carcinoma (epidermoid carcinoma), adenocarcinoma, large cell carcinoma, and small cell

carcinoma. The other malignancies include adenosquamous carcinomas, carcinoid tumors, and bronchial gland carcinomas. The pathogenetic bases of the four principal histological types are uncertain, and various cells of origin and pathways of differentiation have been hypothesized. In addition, a careful examination of multiple sections from the same case shows tumors to be frequently heterogeneous, with elements of several histological types. Observer variation in classifying histological types of lung cancer is well documented and should be considered when interpreting research findings on a histological basis.

Few links have been made between specific histological types and particular etiological agents. Cigarette smoking increases risk for squamous cell carcinoma, adenocarcinoma, large cell carcinoma, and small cell carcinoma, although the risks tend to increase less steeply with extent of smoking for adenocarcinoma, the most common type in never-smokers. Most occupational carcinomas are not associated with risk for a particular histological type of lung cancer. The evidence for specific links is strongest for chloromethyl ether and radon progeny. Some studies of diet and lung cancer risk have provided analyses stratified by histological type, but these histology-specific analyses provide no specific biological insights.

Although knowledge of the etiological and pathological bases of the different types of lung cancer remains limited, trends of lung cancer histology have been monitored in the general population. A trend of increasing proportion of adenocarcinoma has been documented in many regions throughout the world. For example, in the United States during the past three decades adenocarcinoma remained the most common type of lung cancer in women and increased in men so that it is now also the most common histologic type of lung cancer in men. The hypothesis has been advanced that this shift reflects temporal changes in the carcinogens delivered by smoking as well as a changing topography of smoking.

Diet

Dietary Hypotheses and Mechanisms

Epidemiological research on diet and lung cancer has been both hypothesis driven and descriptive, exploring associations between foods or nutrient indexes and lung cancer risk. Interest in macronutrients has emphasized indices of dietary fat, which was long ago noted to have the capacity to act as a tumor promoter. Micronutrients have been extensively studied, spurred initially by the pioneering epidemiological work of Bjelke and the original vitamin A and β-carotene hypotheses. Bjelke and subsequent researchers originally focused on vitamin A because of its role in cellular differentiation and the promise of the initial observational findings, but this line of inquiry was subsequently expanded to include antioxidant micronutrients, with an emphasis on β-carotene. The more general hypothesis has been advanced that antioxidant micronutrients may protect against oxidative damage to DNA and thereby protect against cancer. Hypotheses concerning specific beverages have also been proposed; for example, animal studies have shown alcohol consumption to be associated with changes in lung lipids, including surfactant, and in levels of enzymes that can activate procarcinogens and mutagens. Another epidemiologic approach, empirical rather than hypothesis driven, has been to explore the intakes of several specific foods or food groups for associations with lung cancer risk. The identification of protective associations between fruit and vegetable consumption and lung cancer resulted from the use of this more empirical approach.

Certain methodological issues are relevant to a discussion of diet and lung cancer. When investigating a potential link between diet and lung cancer, the potent role of cigarette smoking in the etiology of lung cancer, along with the current differences in the diets of smokers compared with nonsmokers, makes the potential confounding effects of cigarette smoking an acute concern. Even when there is an attempt to control for smoking, residual confounding of diet–lung cancer associations may still occur. Cigarette smoke can directly affect circulating concentrations of dietary factors (**Figure 1**); for example, smokers tend to have lower levels of circulating antioxidant micronutrients even after accounting for differences in dietary intake.

Aspects of the design and conduct of epidemiological studies in general further limit interpretation of

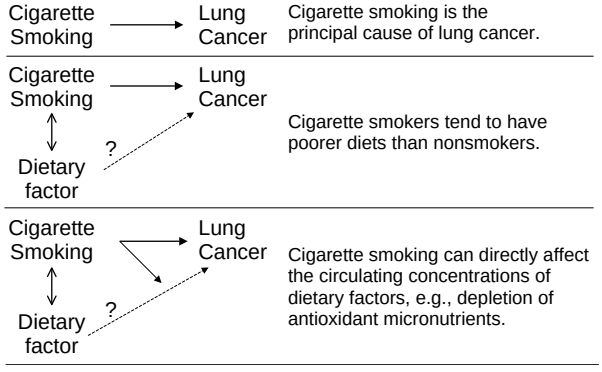


Figure 1 Cigarette smoking complicates the study of diet and lung cancer.

specific dietary studies. Approaches to dietary assessment are not fully standardized, and there may be differences between studies in the number of foods queried, the measurement of serving sizes, and the interview approach employed. There is also uncertainty regarding the biologically relevant exposure window or windows for lung cancer, and dietary agents may plausibly act in early and/or later stages of carcinogenesis. Clinically diagnosed lung cancer reflects a series of complex molecular genetic events that occur over a long period, and the relevant windows for dietary exposures are uncertain. Case–control studies usually measure past diet during some reference period, whereas cohort studies tend to focus on current diet. Most of the epidemiologic research has taken the form of case–control studies, many of which focus on diet during the 5 years preceding diagnosis. These studies provide direct information concerning dietary factors in the later stages of carcinogenesis. To the extent that such measures reflect usual adult (or lifetime) diet, the results of these studies may also be relevant to the role of diet in earlier stages of carcinogenesis. However, because lung cancer tends to be rapidly fatal, many case–control studies include data collected from deceased subjects' next-of-kin. Data from surrogate respondents are probably less accurate than self-reported data, and using such information is certain to introduce substantial misclassification.

Evidence concerning relationships between lung cancer and fruits, vegetables, micronutrients, phytochemicals, fat, body mass index, beverages, and meat intake is reviewed here. For these agents, there is no basis for anticipating that some threshold of intake is relevant to protection (or risk). Rather, a dietary factor that protects against lung cancer would theoretically be expected to confer greater protection when present in greater amounts, and vice versa. Consequently, this review emphasizes dose–response trends, with a monotonic dose–response relationship considered to provide the strongest evidence favoring an association. To focus on the most relevant evidence, the summary tables include only studies that controlled for age and cigarette smoking and case–control studies with more than 200 lung cancer cases. Smaller studies would not be expected to have sufficiently precise estimates to be informative.

Dietary Associations with Lung Cancer

Fruit In total, the evidence favors a protective association between greater fruit consumption and lower lung cancer risk, with associations in the protective direction in 18 of 31 studies (**Table 1**). When stratified by gender, the overall protective association holds more often than not for both males and females. No clear pattern emerges when studies have examined specific fruits or classes of fruits. For example, apples and citrus fruits are associated with reduced risk of lung cancer in some studies but not in others.

Vegetables Evidence for a protective association for vegetable consumption parallels the evidence for fruit consumption, with 18 of 33 studies showing associations at least weakly in the protective direction (**Table 1**). The overall evidence thus points strongly toward a protective association, which has been observed in both males and females. In addition to vegetable intake as a whole, the results for a number of specific vegetables, such as carrots and cruciferous vegetables, have been consistently associated with a reduced risk of lung cancer, at least for the highest versus the lowest categories of consumption.

Micronutrients Two different strategies have been used to evaluate the relationship of micronutrients to lung cancer. One approach has been to use data summarized from food-frequency questionnaires to estimate micronutrient intake. A second approach has been to draw blood samples from study participants and assay the concentrations of micronutrients in circulation. The former approach provides a better average measure of micronutrient 'exposure,' whereas the latter approach has the advantage of measuring micronutrient concentrations closer to the level of cells, where the biologic effect is postulated to occur. However, a single assay of circulating micronutrient concentrations may not reflect the biologically appropriate window of exposure. The evidence is most abundant for vitamins A, C, and E and for total carotenoids and β-carotene. A body of evidence is also accumulating for α-carotene, β-cryptoxanthin, lutein, and lycopene.

The majority of studies on dietary retinol show no association with lung cancer risk (**Table 2**). On the other hand, studies of dietary intake of β-carotene, total carotenoids, and vitamin C point more consistently toward an inverse association, with results at least equally divided between showing some evidence of protective association and showing no association (**Table 2**). As with β-carotene, studies of additional provitamin A carotenoids α-carotene and β-cryptoxanthin have also been evenly divided between showing no association and showing a protective association with lung cancer. Conversely, only a minority of studies of the non-provitamin A carotenoids lycopene and lutein have suggested that higher intakes are associated with decreased risk of lung cancer.

Studies based on micronutrients measured from blood samples drawn after an individual is diagnosed with lung cancer showed that lung cancer

Table 1 Estimated relative risk of lung cancer according to frequency of fruit and vegetable consumption[a]

First author (year)	*Sex*	*Fruits*					*Vegetables*					*Also adjusted for*
		1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*	
Case–control studies												
Alavanja (1993)	F	1.0	1.1	0.8	1.1	—	1.0	0.8	0.9	0.8	1.0	Previous lung disease, daily energy intake
Axelsson (1996)	M	1.0	0.8	0.7	—	—	1.0	0.7	0.4	—	—	Marital status, occupation
Brennan (2000)	M/F	1.0	0.9	1.0	—	—	1.0	0.9	0.7	—	—	Sex, center (all nonsmokers)
Darby (2000)	M/F	—	—	—	—	—	1.0	0.8	0.9	0.9	—	Sex
De Stefani (1999)	M	1.0	0.6	0.5	—	—	1.0	0.7	0.5	—	—	Residence, education, family history, body mass index, total energy and fat intake
Dorgan (1993)	M/F	1.0	1.0	0.9	—	—	1.0	0.9	0.7	—	—	Passive smoking, education, occupation
Fontham (1988)	M/F	1.0	0.8	0.7	—	—	1.0	0.8	0.7	—	—	Race, sex
Gao (1993)	M	1.0	0.8	0.5	—	—	1.0	0.7	0.6	—	—	
Hu (1997)	M/F	—	—	—	—	—	1.0	1.1	0.8	—	—	Sex, area of residence, income
Jain (1990)	M/F	1.0	0.9	1.0	1.1	—	1.0	0.7	0.7	0.6	—	
Kreuzer (2002)	F	1.0	0.6	0.7	—	—	1.0	0.6	0.5	—	—	Region (all nonsmokers)
Kubik (2002)	F	1.0	0.7	—	—	—	1.0	0.8	—	—	—	Residence, education
Le Marchand (1989)	M	—	—	—	—	—	1.0	0.9	0.7	0.4	—	Ethnicity, cholesterol intake
Mayne (1994)	M	1.0	0.4	0.7	0.7	—	1.0	1.1	1.3	0.6	—	
	F	1.0	0.6	0.5	0.6	—	1.0	1.0	0.8	0.5	—	
Mohr (1999)	M/F	1.0	1.0	1.2	0.8	—	1.0	1.0	1.0	1.0	—	Sex, race, education
Swanson (1992)	M	1.0	0.7	1.5	0.9	—	—	—	—	—	—	Income, education
Swanson (1997)	F	1.0	1.0	1.0	0.9	0.8	1.0	0.9	0.7	0.8	0.6	
Takezaki (2001)	M	1.0	1.2	1.0	1.0	— (AC)	1.0	1.1	1.1	1.0	— (AC)	Season and year of visit, prior lung disease, green vegetable and meat intake
		1.0	0.9	0.8	0.6	— (SCC)	1.0	1.3	0.7	0.8	— (SCC)	
	F	1.0	0.7	0.8	0.7	— (AC)	1.0	0.7	0.9	0.8	— (AC)	
Wu-Williams (1990)	F	1.0	1.0	1.4	1.5	—	1.0	1.1	1.0	0.9	—	Education
Prospective studies												
Breslow (2000)	M/F	1.0	1.2	0.8	0.9	—	1.0	1.0	1.3	0.9	—	Sex
Chow (1992)	M	1.0	0.8	0.8	0.7	—	1.0	1.1	1.1	1.2	—	Occupation
Feskanich (2000)	M/F	1.0	1.1	1.0	1.1	0.9	1.0	0.9	0.8	0.9	0.8	Energy intake
Fraser (1991)	M/F	1.0	0.3	0.3	—	—	1.0	1.4	1.1	—	—	Sex
Hirvonen (2001)	M	1.0	0.9	0.9	0.7	—	1.0	1.0	0.9	0.7	—	Group
Jansen (2001)	M	1.0	0.6	0.7	—	—	1.0	0.7	0.9	—	—	Country, energy and fruit/vegetable intake
Knekt (1999)	M	1.0	0.9	0.6	—	—	1.0	0.9	0.8	—	—	
Kromhout (1987)	M	1.0	1.1	1.0	—	—	—	—	—	—	—	
Neuhouser (2003)	M/F	1.0	0.8	0.7	0.7	0.6	1.0	0.6	0.8	0.6	0.8	Sex, asbestos exposure, ethnicity, enrollment center
Ocke (1997)	M	1.0	0.5	0.5	—	—	1.0	0.8	0.9	—	—	Energy intake
Shibata (1992)	M	1.0	1.1	1.0	—	—	1.0	1.1	1.0	—	—	
	F	1.0	0.8	0.7	—	—	1.0	0.7	0.6	—	—	
Steinmetz (1983)	F	1.0	0.8	0.5	0.8	—	1.0	0.7	0.5	0.5	—	
Voorips (2000)	M/F	1.0	0.7	0.6	0.6	0.8	1.0	1.1	1.0	1.0	0.7	Sex, family history, education

[a]1 = lowest consumption category. The exposure categories of low to high are for summary purposes only and do not correspond to identical categories across studies.

AC, adenocarcinoma; F, female; M, male; SCC, squamous cell carcinoma.

cases had circulating concentrations of retinol, β-carotene, total carotenoids, vitamin E, and vitamin C that were 20% or more lower than those of noncases. However, the possibility that preclinical and clinically diagnosed lung cancer and concomitant changes in diet can lead to decreases in circulating micronutrient levels limits the inferences that may be drawn from studies based on blood samples taken after the diagnosis of lung cancer.

Data from prospective cohort studies are not subject to the previous limitation. In these studies, blood is collected from a population that is initially cancer-free and the population is then followed for the occurrence of lung cancer. The results of such

Table 2 Estimated relative risk of lung cancer according to intake of selected micronutrients[a]

First author (year)	*Sex*	*Retinol*					*β-Carotene*					*Total carotenoids*					*Vitamin C*				
		1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*
Case-control studies																					
Alavanja (1993)	F	1.0	1.4	0.8	0.9	1.3	1.0	0.9	0.9	1.0	1.0	1.0	0.9	0.9	0.9	1.3	1.0	1.5	1.1	1.3	1.5
Bond (1987)	M	1.0	0.9	0.9	—	—	—	—	—	—	—	1.0	1.0	0.9	—	—	—	—	—	—	—
		1.0	0.8	0.5	—	—	—	—	—	—	—	1.0	0.7	0.4	—	—	—	—	—	—	—
Byers (1984)	M	1.0	0.8	0.7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
		(SCC)					—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
		1.0	0.9	0.6	—	— (SM)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
		1.0	1.0	1.2	—	— (AD)															
Byers (1987)	M	1.0	1.1	1.1	0.7	—	—	—	—	—	—	1.0	1.0	0.9	0.8	—	1.0	1.0	0.9	0.8	—
	F	1.0	0.8	1.4	0.8	—	—	—	—	—	—	1.0	0.7	1.1	0.9	—	1.0	0.9	0.7	0.9	—
De Stefani (1999)	M/F	—	—	—	—	—	1.0	0.8	0.6	0.4	—	1.0	0.8	0.6	0.4	—	1.0	0.9	1.0	1.0	—
Fontham (1988)	M/F	1.0	0.9	0.9	—	—	—	—	—	—	—	1.0	1.0	0.9	—	—	1.0	0.9	0.7	—	—
Hinds (1984)	M/F	1.0	0.8	0.9	0.7	—	—	—	—	—	—	1.0	0.8	0.7	0.6	—	1.0	0.8	0.8	0.8	—
Jain (1990)	M/F	1.0	1.0	0.1	1.2	—	1.0	0.9	0.8	0.9	—	1.0	0.9	0.8	1.1	—	—	—	—	—	—
Le Marchand (1989)	M	1.0	1.0	1.1	1.1	—	1.0	1.3	0.8	0.5	—	1.0	1.0	0.5	0.4	—	1.0	1.0	0.5	0.4	—
Mayne (1994)	M	1.0	0.9	—	—	—	1.0	0.6	—	—	—	—	—	—	—	—	—	—	—	—	—
	F	1.0	1.0	—	—	—	1.0	0.8	—	—	—	—	—	—	—	—	—	—	—	—	—
Mettlin (1979)	M	1.0	0.9	0.6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Mettlin (1989)	M/F	—	—	—	—	—	1.0	0.9	0.8	0.6	0.5	—	—	—	—	—	—	—	—	—	—
Samet (1985)	M/F	1.0	0.8	0.7	—	—	—	—	—	—	—	1.0	0.9	0.8	—	—	—	—	—	—	—
Ziegler (1986)	M	1.0	1.2	1.2	—	—	—	—	—	—	—	1.0	0.9	0.8	—	—	—	—	—	—	—
		—	—	—	—	—	—	—	—	—	—	1.0	0.2	1.5	—	—	—	—	—	—	—
Prospective studies																					
Chow (1992)	M	1.0	1.1	1.2	0.8	0.9	1.0	0.8	1.0	1.0	0.8	1.0	1.0	0.8	1.1	0.8	1.0	0.7	0.9	0.7	0.8
Holick (2002)	M	1.0	1.0	1.0	1.0	1.0	1.0	0.9	0.9	0.8	0.9	1.0	1.0	0.9	0.8	0.8	—	—	—	—	—
Knekt (1999)	M	—	—	—	—	—	1.0	1.1	0.8	—	—	1.0	1.2	0.9	—	—	—	—	—	—	—
Kromhout (1987)	M	—	—	—	—	—	1.0	0.5	0.7	—	—	1.0	0.4	0.4	—	—	—	—	—	—	—
Michaud (2000)	M/F	—	—	—	—	—	1.0	1.1	1.0	1.0	0.8	1.0	1.0	1.0	0.8	0.7	—	—	—	—	—
Neuhouser (2003)	M/F	—	—	—	—	—	1.0	0.9	0.9	1.0	1.0	1.0	1.0	0.9	1.0	0.9	1.0	0.7	0.8	0.8	0.7
Ocke (1997)	M	—	—	—	—	—	1.0	0.7	0.7	—	—	—	—	—	—	—	1.0	0.8	0.4	—	—
Paganini-Hill (1987)	M	1.0	1.2	1.0	—	—	1.0	1.3	0.7	—	—	—	—	—	—	—	—	—	—	—	—
	F	1.0	0.7	0.9	—	—	1.0	0.3	0.7	—	—	—	—	—	—	—	—	—	—	—	—
Rohan (2002)	F	—	—	—	—	—	1.0	1.8	1.8	1.4	—	—	—	—	—	—	—	—	—	—	—
Shekelle (1981)	M	1.0	2.2	1.4	2.0	—	—	—	—	—	—	1.0	0.8	0.4	0.1	—	—	—	—	—	—
Shibata (1992)	M	—	—	—	—	—	1.0	0.9	1.1	—	—	1.0	0.9	1.1	—	—	—	—	—	—	—
	F	—	—	—	—	—	1.0	0.7	0.6	—	—	1.0	0.8	0.6	—	—	—	—	—	—	—
Speizer (1999)	F	—	—	—	—	—	1.0	1.1	0.8	1.0	0.8	1.0	1.2	1.0	1.0	0.9	1.0	0.9	1.2	0.6	1.4
Steinmetz (1993)	F	—	—	—	—	—	1.0	0.8	0.7	0.8	—	1.0	1.1	0.7	1.4	—	—	—	—	—	—
Voorrips (2000)	M	—	—	—	—	—	1.0	0.8	0.9	1.0	1.0	—	—	—	—	—	1.0	0.8	0.8	—	—
Yong (1997)	M/F	—	—	—	—	—	—	—	—	—	—	1.0	0.7	0.8	0.7	—	1.0	0.9	0.7	0.7	—

[a]1 = lowest consumption category. The exposure categories of low to high are for summary purposes only and do not correspond to identical categories across studies.
AD, adenocarcinoma; F, female; M, male: SCC, squamous cell carcinoma; SM, small cell carcinoma.

prospective studies bolster the evidence supporting the premise that in general, the higher the circulating concentrations of carotenoids (α-carotene, β-carotene, β-cryptoxanthin, lutein, lycopene, and total carotenoids), the lower the risk of lung cancer. Circulating concentrations of retinol, tocopherol, and selenium have not been associated with a reduced risk of lung cancer in most studies.

Studies of both dietary intake and prediagnostic blood concentrations favor a protective association between provitamin A carotenoids (specifically β-carotene, α-carotene, and β-cryptoxanthin) and lung cancer. It is not known, however, if a generally protective association is specific to these carotenoids or whether carotenoid intake merely serves as a marker of the intake of other protective substances or healthier dietary habits in general. The evidence for vitamin C is scant but suggestive of a protective association, whereas the data on vitamin A, vitamin E, and selenium have yielded null findings.

Phytochemicals Phytochemicals are low-molecular-weight molecules produced by plants. Of the many classes of phytochemicals, those studied in relation to lung cancer include phytoestrogens, flavonoids, and glucosinoids.

The tumor-promoting effects of steroid hormones can be blocked by phytoestrogens. Soybeans are a primary source of a specific class of phytoestrogens known as isoflavonoids. The relatively few studies on isoflavonoids in relation to lung cancer have not provided evidence of a link.

Flavonoids exhibit potent antioxidant activity. Flavonoid intake has been at least weakly associated with reduced risk of lung cancer in three out of four studies to date.

Isothiocyanates are metabolites of the class of phytochemcials known as glucosinolates. Isothiocyanates could exert anticancer effects by blocking carcinogens via induction of phase II detoxification enzymes, such as glutathione *S*-tranferase. Cruciferous vegetables contain high concentrations of glucosinolates, and hence consumption leads to higher endogenous isothiocyanate levels. As with cruciferous vegetables, lung cancer risk is also consistently lower with higher intakes or urinary levels of isothiocyanates.

A postulated link between isothiocyanates and a common polymorphism in the *GSTM1* gene provides an example of a potential gene–diet interaction relevant to lung carcinogenesis. A growing focus in cancer epidemiology is to characterize interindividual susceptibility to cancer by studying polymorphisms in genes involved in DNA repair and in the metabolism and detoxification of potential carcinogens. Of further interest is how such genetic traits interact with environmental exposures to contribute to cancer risk. The role of glutathione *S*-transferase as a phase II detoxification enzyme has made a common polymorphism in the glutathione *S*-transferase M1 (*GSTM1*) gene of interest in relation to lung cancer. Results combined across studies show that compared to people with the *GSTM1* present genotype, those with the *GSTM1* null genotype had an increased risk of lung cancer.

When isothiocyanates have been studied in combination with *GSTM1*, the decreased risk of lung cancer associated with isothiocyanates has been especially pronounced in people with the *GSTM1* null genotype. This association may represent either the cancer-blocking activity of isothiocyanates playing an enhanced role in *GSTM1* null individuals or more efficient metabolism of isothiocyanates in those with the *GSTM1* present genotype. Regardless, this is one example of the potential interactions between genetic and dietary factors, an approach that may eventually advance our understanding of the nutritional epidemiology of lung cancer.

Fat and cholesterol Evidence that dietary fat may facilitate tumor growth was reported as early as 1940. Correlation exists between international or regional dietary fat consumption and lung cancer mortality. In case–control studies, total fat intake is consistently associated with lung cancer risk among men and women, but saturated fat, unsaturated fat, and cholesterol intake tend to be associated with lung cancer risk only among men (**Table 3**). The prospective evidence shows a slightly different picture, with both total fat and saturated fat intake strongly associated with lung cancer in men but not women, and unsaturated fat and cholesterol not consistently associated with lung cancer risk in men or women (**Table 3**). The equivocal nature of the evidence is reflected in the lack of consistent findings between the sexes and the results of a large, pooled cohort study of both sexes that found lung cancer risk was not strongly associated with fat (total, saturated, or unsaturated) or cholesterol intake.

Regarding cholesterol, there is inconsistency between the dietary data presented previously and serologic data. A review of 33 prospective cohort studies indicated that lower circulating cholesterol levels were predictive of greater lung cancer risk. Similar results were obtained after accounting for the possible preclinical effects of cancer on cholesterol levels by limiting analyses to cases of lung cancer that were diagnosed 5 or more years after the initial cholesterol measurement. This association may be due to a direct effect of cigarette smoking on lipid profiles or to differences in dietary patterns between smokers and nonsmokers. The lack of consistency between the serologic

Table 3 Estimated relative risk of lung cancer according to fat intake or cholesterol intake[a]

First author (Year)	*Sex*	*Total fat*					*Unsaturated fat*					*Saturated fat*					*Cholesterol*				
		1	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*	*1*	*2*	*3*	*4*	*5*
Case–control studies																					
Alavanja (1993)	F	1.0	1.4	1.4	2.2	2.8	—	—	—	—	—	1.0	1.5	1.6	2.3	4.9	1.0	0.6	0.7	1.1	1.1
Byers (1984)	M	1.0	0.9	1.1	—	— (SCC)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
		1.0	0.8	0.8	—	— (SM)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
		1.0	1.2	1.0	—	— (AD)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Byers (1987)	M	1.0	1.6	1.8	2.0	—	—	—	—	—	—	—	—	—	—	—	1.0	1.3	1.7	1.4	—
	F	1.0	1.9	1.0	1.4	—	—	—	—	—	—	—	—	—	—	—	1.0	1.5	1.1	0.9	—
De Stefani (1997)	M	1.0	1.2	1.4	2.9	—	1.0	1.3	1.7	1.8	—	1.0	0.9	2.3	2.1	—	1.0	1.4	1.3	2.3	—
							(monounsaturated)														
							1.0	1.2	2.2	1.8	—										
							(polyunsaturated)														
Goodman (1988)	M	1.0	2.3	2.0	2.2	—	1.0	3.5	2.1	2.5	—	1.0	1.8	2.3	2.1	—	1.0	2.3	1.8	2.2	—
	F	1.0	0.4	0.8	0.9	—	1.0	0.5	0.6	0.9	—	1.0	0.6	1.0	1.4	—	1.0	0.6	1.5	0.9	—
Hinds (1983)	M/F	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.0	1.3	1.4	2.0	—
	M	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.0	1.2	1.4	2.3	—
	F	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.0	1.7	1.3	1.2	—
Jain (1990)	M/F	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.0	0.9	1.0	1.6	—
Swanson (1997)	F	1.0	1.4	1.3	2.3	—	—	—	—	—	—	—	—	—	—	—	1.0	1.2	0.9	1.0	1.2
Prospective studies																					
Bandera (1997)	M	1.0	1.3	1.4	—	—	1.0	1.2	1.4	—	—	1.0	1.3	1.4	—	—	1.0	1.1	1.1	—	—
	F	1.0	0.9	1.1	—	—	(monounsaturated)					1.0	0.7	0.8	—	—	1.0	0.8	1.0	—	—
		—	—	—	—	—	1.0	1.1	1.0	—	—	—	—	—	—	—	—	—	—	—	—
		—	—	—	—	—	(polyunsaturated)					—	—	—	—	—	—	—	—	—	—
							1.0	0.9	1.0	—	—										
							(monounsaturated)														
							1.0	0.9	0.9	—	—										
							(polyunsaturated)														
Heilbrun (1994)	M	—	—	—	—	—	—	—	—	—	—						1.0	0.7	1.0	1.0	—
Knekt (1991)	M	1.0	1.3	1.6	—	—	1.0	0.9	1.1	—	—	1.0	1.4	1.6	—	—	1.0	0.8	1.1	—	—
							(monounsaturated)					—	—	—	—	—	—	—	—	—	—
							1.0	0.6	0.9	—	—										
							(polyunsaturated)														
Shekelle (1991)	M	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.0	1.3	1.9	—	—
Smith-Warner (2002)	M/F	1.0	1.0	1.0	1.0	—	1.0	1.0	1.0	1.0	—	1.0	1.0	1.0	1.0	—	1.0	1.0	1.0	1.0	—
							(monounsaturated)					—	—	—	—	—	—	—	—	—	—
							1.0	1.0	1.0	1.0	—										
							(polyunsaturated)														
Speizer (1999)	F	1.0	1.0	0.9	0.9	1.1	—	—	—	—	—	1.0	1.0	1.1	1.1	1.1	1.0	1.0	1.3	1.3	1.3
Wu (1994)	F	1.0	0.9	0.9	0.8	—	—	—	—	—	—	—	—	—	—	—	1.0	0.6	0.9	0.9	—
		1.0	0.7	0.6	0.9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
		(animal)					—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
		1.0	0.6	0.7	0.7	— (plant)															

[a]1 = lowest consumption category. The exposure categories of low to high are for summary purposes only and do not correspond to identical categories across studies.
AD, adenocarcinoma; F, female; M, male; SCC, squamous cell carcinoma; SM, small cell carcinoma.

and dietary cholesterol data is not unreasonable given that dietary cholesterol intake is not strongly associated with serum cholesterol levels. The results of a trial of 846 men living at a veteran's home in Los Angeles showed that compared to men randomized to receive a conventional US diet, those randomized to receive a diet that reduced cholesterol intake by half and reduced serum cholesterol by 13% had a 20% increased risk of lung cancer after 8 years of dietary intervention and 2 additional years of follow-up.

Body mass index Prospective studies consistently show low body mass index (BMI) and relative weight to be associated with an increased risk of lung cancer (**Table 4**). This association is observed in some case–control studies, which rely on retrospective ascertainment of BMI, but not in others. Confounding by cigarette smoking should be considered as an explanation for these findings because cigarette smoking is strongly associated both with the risk of lung cancer and with leanness. The need to further test the hypothesis that leanness is a susceptibility factor for lung cancer is indicated by the results of studies in which this association is still observed even after potential confounding by cigarette smoking has been carefully addressed by stratifying by smoking status and by the number of cigarettes smoked per day for smokers.

Beverages Confounding by cigarette smoking is ubiquitous to the study of diet and lung cancer, but perhaps no topic better epitomizes the challenges of controlling confounding by smoking than does beverage consumption. Several beverages, including alcohol, coffee, tea, and milk, have been studied for a possible link to lung cancer. The majority of studies that have adjusted for age and cigarette smoking have observed either null or weak associations between alcohol drinking and the risk of lung cancer (**Table 5**).

Three prospective cohort studies have shown heavy coffee consumption to be associated with an elevated risk of lung cancer after adjustment for cigarette smoking, whereas seven case–control studies have yielded findings that tend to fluctuate around the null (**Table 6**). The issue of confounding between coffee

Table 4 Relative risk of lung cancer according to body mass index[a]

First author (year)	*Sex*	*Average follow-up (years)*	*Body mass index*							*Subgroup*	*Adjusted for*
			1	*2*	*3*	*4*	*5*	*6*	*7*		
Chyou (1994)	M	25	1.0	0.9	0.9	0.7	—	—	—		Age, smoking
Drinkard (1995)	F	6	1.0	0.5	0.5	0.5	—	—	—	Total	Age, smoking, physical activity
			1.0	0.6	0.7	—	—	—	—	Never smokers	
			1.0	1.2	0.6	—	—	—	—	Former smokers	
			1.0	0.8	1.0	—	—	—	—	Current smokers	
Henley (2002)	M	14	0.9	1.0	—	—	—	—	—	Nonsmokers	Age, race, former smoker, marital status, education, asbestos exposure, socioeconomic status, intake of alcohol, fat, fruits, and vegetables
	F		1.2	1.0	—	—	—	—	—	Nonsmokers	
Kark (1995)	M	23	2.3	1.3	1.1	1.1	1.0	—	—	Total	Age, smoking, area of residence
			3.7	2.1	1.7	2.0	1.0	—	—	Smokers	
Knekt (1991)	M	15	1.8	1.5	1.4	1.0	—	—	—	Total	Age, smoking, social class, health status, stress
Lee (1992)	M	22/26	1.8	1.5	1.0	—	—	—	—	Follow-up 11–15 years	Age, cigarettes per day, physical activity
			1.0	1.1	1.0	—	—	—	—	Follow-up >15 years	
Olson (2002)	F	12	1.0	0.9	0.7	0.5	0.4	—	—	Total	Age, pack years smoking, smoking status, physical activity, education, beer consumption, height, BMI at age 18, waist circumference

[a]1 = lowest BMI. The exposure categories of low to high are for summary purposes only and do not correspond to identical categories across studies.
BMI, body mass index; F, female; M, male.

Table 5 Estimated relative risk of lung cancer according to alcohol intake[a]

First author (year)	*Sex*	*Total alcohol*						*Beer*						*Wine*						*Hard liquor*					
		1	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*	*5*	*6*
Case–control studies																									
Mettlin (1989)	M/F	—	—	—	—	—	—	1.0	0.5	0.7	1.3	—	—	1.0	0.5	0.8	0.9	—	—	1.0	0.7	0.6	0.6	—	—
Wu-Williams (1990)	F	1.0	1.3	1.0	1.3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Bandera (1992)	M	1.0	—	—	—	1.6	—	1.0	—	—	—	1.9	—	1.0	—	—	—	0.7	—	1.0	—	—	—	1.1	—
De Stefani (1993)	M	1.0	1.4	1.6	2.2	—	—	1.0	0.7	1.4	3.4	—	—	1.0	1.2	1.3	1.5	—	—	1.0	0.9	1.3	1.1	—	—
Mayne (1994)	M/F	—	—	—	—	—	—	1.0	1.1	0.9	1.2	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Carpenter (1998)	M/F	1.0	0.5	0.9	1.1	—	—	1.0	0.4	0.9	—	—	—	1.0	0.7	0.8	—	—	—	1.0	1.2	1.9	—	—	—
Korte (2002)	M/F	1.0	0.6	1.3	1.1	1.9	1.4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Dosemeci (1997)	M	1.0	1.6	1.7	1.7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.0	1.7	1.9	1.6	—	—
Swanson (1997)		1.0	0.6	1.1	1.0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Murata (1996)	M	1.0	1.0	2.4	1.8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Zang (2001)	M	1.0	1.1	1.2	1.1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Prospective studies																									
Pollack (1984)	M	1.0	0.7	1.3	1.7	1.9		—	—	—	—	—	—	1.0	2.2	—	—	—	—	1.0	2.6	—	—	—	—
Chow (1992)	M	—	—	—	—	—		1.0	1.2	1.4	1.7	1.1	1.8	—	—	—	—	—	—	1.0	1.3	1.3	1.3	1.0	1.9
Potter (1992)	F	—	—	—	—	—		1.0	0.6	1.9	—	—	—	—	—	—	—	—	—	1.0	1.1	—	—	—	—
Kvale (1983)	M	1.0	—	1.3	—	—		—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Gordon (1984)	M	1.0 (continuous variable)						—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	F	0.7 (continuous variable)						—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Kono (1986/87)	M	1.0	0.6	0.4	0.8	0.9		—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Stemmermann (1990)	M	1.0	0.7	0.9	1.4	1.1		—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Breslow (2000)	M/F	1.0	0.9	1.2	2.1	—		—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
		1.0 (nonsmokers)	—	—	2.3	—		—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Bandera (1997)	M	1.0	0.8	1.1	—	—		—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	F	1.0	1.2	1.0	—	—		—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Prescott (1999)	M	1.0	0.8	1.0	0.9	1.2	1.6	1.0	1.1	1.4	—	—	—	1.0	0.8	0.4	—	—	—	1.0	1.2	1.5	—	—	—
	F	1.0	0.9	1.0	1.0	1.0	0.8	1.0	0.9	1.5	—	—	—	1.0	0.9	0.2	—	—	—	1.0	0.8	0.7	—	—	—
Woodson (1999)	M	1.0	1.0	1.0	0.9	1.0	—	1.0	1.0	0.8	0.9	0.9	—	1.1	1.0	0.8	—	—	—	1.1	1.0	1.0	1.1	1.1	—
Hirvonen (2001)	M	—	—	—	—	—	—	—	—	—	—	—	—	1.0	0.7	—	—	—	—	—	—	—	—	—	—
Djousse (2002)	M/F	1.0	1.2	1.1	1.3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Korte (2002)	M/F	1.0	1.0	0.9	1.0	1.5	1.2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
CPS I and II, from Korte (2002)	M/F	1.0	1.0	1.0	1.2	1.4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Omenn (1996)		1.2	2.0	0.8	2.0	2.0	1.7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

[a]1 = lowest consumption category. The exposure categories of low to high are for summary purposes only and do not correspond to identical categories across studies.
F, female; M, male.

Table 6 Estimated relative risk of lung cancer according to frequency of tea, coffee, or milk consumption[a]

First author (year)	*Sex*	*Tea*						*Coffee*						*Milk*					
		1	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*	*5*	*6*	*1*	*2*	*3*	*4*	*5*	*6*
Case–control studies																			
Brennan (2000)	M/F	—	—	—	—	—	—	—	—	—	—	—	—	1.0	1.0	0.8	—	—	—
Darby (2001)	M/F	—	—	—	—	—	—	—	—	—	—	—	—	1.0	0.9	1.1	2.1	—	—
Mendilaharsu (1998)	M smokers	1.0	0.7	0.7	0.9	0.5	0.3	1.0	1.1	1.3	0.8	1.2	1.2	—	—	—	—	—	—
Kreuzer (2002)	F nonsmokers	—	—	—	—	—	—	—	—	—	—	—	—	1.0	1.0	0.7	—	—	—
Kubik (2002)	F	1.0	1.0	—	—	—	— (black)	1.0	0.6	—	—	—	—	1.0	0.8	—	—	—	—
		1.0	0.9	—	—	—	— (green)	—	—	—	—	—	—	—	—	—	—	—	—
		1.0	1.3	—	—	—	— (herbal)	—	—	—	—	—	—	—	—	—	—	—	—
Zhong (2001)	F nonsmokers	1.0	0.8	0.6	0.5	—	— (green)	—	—	—	—	—	—	—	—	—	—	—	—
	F smokers	1.0	1.4	0.6	—	—	— (green)	—	—	—	—	—	—	—	—	—	—	—	—
Mohr (1999)	M/F	1.0	1.0	1.4	1.6	—	—	1.0	1.0	1.2	1.3	—	—	1.0	1.6	1.8	—	—	— (whole)
		—	—	—	—	—	—	—	—	—	—	—	—	1.0	0.8	—	—	—	— (2%)
		—	—	—	—	—	—	—	—	—	—	—	—	1.0	0.6	0.7	—	—	— (skim)
Swanson (1997)	F	—	—	—	—	—	—	—	—	—	—	—	—	1.0	1.3	1.2	1.0	1.2	—
Takezaki (2001)	M, AC	1.0	1.1	1.1	1.3	—	—	1.0	0.9	1.2	—	—	—	1.0	1.0	0.9	0.8	—	—
	F, AC	1.0	1.0	1.1	1.1	—	—	1.0	0.8	0.8	1.3	—	—	1.0	0.8	1.0	0.7	—	—
	M, SCC	1.0	1.0	1.2	1.1	—	—	1.0	1.0	1.2	1.6	—	—	1.0	0.9	0.8	0.7	—	—
Tewes (1990)	F	1.0	1.4	—	—	—	— (black)	—	—	—	—	—	—	—	—	—	—	—	—
		1.0	2.7	—	—	—	— (green)	—	—	—	—	—	—	—	—	—	—	—	—
Yoshiyuki (1995)	M	1.0	0.7	0.9	0.9	0.6	— (Okinawan)	—	—	—	—	—	—	—	—	—	—	—	—
	F	1.0	0.7	0.8	0.8	0.4	— (Okinawan)	—	—	—	—	—	—	—	—	—	—	—	—
Mayne (1994)	M/F	—	—	—	—	—	—	—	—	—	—	—	—	1.0	0.9	3.3	1.6	—	— (whole)
		—	—	—	—	—	—	—	—	—	—	—	—	1.0	1.3	0.8	0.8	—	— (skim)
Mettlin (1989)	M/F	1.0	0.9	0.9	1.1	—	—	1.0	1.0	0.9	1.3	—	—	1.0	1.6	1.6	2.1	—	— (whole)
		—	—	—	—	—	—	—	—	—	—	—	—	1.0	0.5	0.7	0.5	—	— (2%)
		—	—	—	—	—	—	—	—	—	—	—	—	1.0	0.8	0.5	0.7	—	— (skim)
Axelsson (1996)	M	1.0	0.9	1.2	0.7	—	—	1.0	0.9	1.2	1.6	—	—	1.0	0.9	1.4	1.7	—	—
Prospective studies																			
Goldbohm (1996)	M/F	1.0	0.9	1.0	0.8	0.9	1.1 (black)	—	—	—	—	—	—	—	—	—	—	—	—
Zheng (1996)	F	1.0	0.9	1.2	1.1	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Breslow (2000)	M/F	—	—	—	—	—	—	—	—	—	—	—	—	1.0	1.4	1.0	—	—	— (whole)
		—	—	—	—	—	—	—	—	—	—	—	—	1.0	0.9	0.5	—	—	— (2%)
		—	—	—	—	—	—	—	—	—	—	—	—	1.0	1.0	0.6	—	—	— (skim)
Hirvonen (2001)	M smokers	1.0	0.7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Fraser (1991)	M/F	—	—	—	—	—	—	—	—	—	—	—	—	1.0	1.0	0.9	—	—	—

[a]1 = lowest consumption category. The exposure categories of low to high are for summary purposes only and do not correspond to identical categories across studies.
AC, adenocarcinoma; F, female; M, male; SCC, squamous cell carcinoma.

drinking and other health behaviors, particularly cigarette smoking, has not been addressed adequately, indicating that much stronger evidence is needed for coffee drinking to be considered a risk factor for lung cancer. Despite numerous in vitro and in vivo studies that have observed potential tumor-inhibitory effects of tea, the epidemiologic evidence does not provide support for a link between tea drinking and the risk of lung cancer (**Table 6**).

The associations observed between milk drinking and lung cancer depend on milk fat content. Milk drinking is not strongly associated with lung cancer risk when milk fat content is ignored. The associations between whole milk and lung cancer tend to be either null or in the direction of increased risk, whereas the associations for reduced fat or nonfat milk tend to be either null or in the protective direction (**Table 6**). Perhaps milk consumption, including the type of milk, is merely serving as a marker of fat intake.

Meat and fish Associations have been observed between red meat intake and increased lung cancer risk, but this evidence is counterbalanced by an equal number of null studies. The cooking method may play a role because heterocyclic amines from cooked meat may contribute to an increased lung cancer risk. The evidence does not support a strong link between fish consumption and lung cancer.

Diet and Prevention

Chemoprevention trials Three randomized, double-blind, placebo-controlled trials were undertaken in the 1980s and 1990s to test whether β-carotene supplementation protects against lung cancer. All three studies indicated that β-carotene supplementation in later adulthood does not protect against lung cancer (**Table 7**). To the contrary, β-carotene supplementation was associated with an increased risk of lung cancer among the high-risk populations of heavy smokers in the ATBC Cancer Prevention Study and smokers and asbestos-exposed workers in the CARET Study. No beneficial effect was observed for α-tocopherol supplementation in the ATBC Cancer Prevention Study.

These experimental results fail to corroborate the evidence from observational studies that favors a protective association between β-carotene and lung cancer. In fact, it is possible that β-carotene may exhibit prooxidant properties. Consistent with the results of observational studies, a protective association was noted when the data from the placebo controls in the ATBC Cancer Prevention Study were analyzed according to baseline serum and dietary β-carotene.

In interpreting the results of the ATBC and CARET studies, it is important to recognize that the studies enrolled older, high-risk individuals who had high cumulative exposure to tobacco smoke and/or asbestos. The results therefore presumably apply mainly to the latter stages of carcinogenesis. The doses administered were far higher than the normal dietary range, and the dose–response relationship for preventive effects, anticipated from the observational evidence, may not be applicable. Because antioxidant nutrients may exert their protective effect in the earlier stages of carcinogenesis, β-carotene may have been administered too late to halt the evolution of cellular changes that lead to lung cancer. Alternatively, compounds present in fruits and vegetables other than the micronutrients studied in the trials may protect against lung cancer. The protective associations for fruit and vegetable consumption were allied to the micronutrient

Table 7 Summary of randomized chemoprevention trials of micronutrients and lung cancer

Study (year)	*Location*	*N*	*Number of cases*	*Study population*	*Years of follow-up*	*Regimen*	*Relative risk*	
							Incidence	*Mortality*
ATBC (1994)	Finland	29 133	876	Male smokers, age 50–69 years	6 (median)	1. Placebo 2. AT (50 mg per day) 3. BC (20 mg per day) 4. AT + BC	0.98 (AT) 1.19 (BC)	1.02 (AT) 1.16 (BC)
CARET (1996)	United States	18 314	388	Asbestos-exposed smokers and heavy smokers Males and females, age 45–69 years	4 (average)	1. Placebo 2. BC (15 mg per day) + retinol (25 000 IU per day)	1.28	1.17 (all causes)
PHS (1996)	United States	22 071	170	Male physicians, age 40–84 years	12 (average)	1. Placebo 2. BC (50 mg on alternate days)	0.93	Not reported

AT, α-tocopherol; BC, β-carotene.

hypothesis, but the results of the chemoprevention trials raise questions about the potential payoff from large trials designed to test single micronutrients, unless there is a strong mechanistic basis or substantial observational evidence pointing to an individual micronutrient as the primary protective agent. Indeed, fruits and vegetables contain an abundance of antioxidants and phytochemicals with diverse anticarcinogenic activities. However, perhaps fruit and vegetable intake is acting as a marker of a healthier lifestyle that is associated with a low risk of cancer.

Conclusions

The past three decades have witnessed a tremendous increase in the information on diet and lung cancer. People who eat more fruits and vegetables in general have a lower risk of lung cancer than people who consume less of these foods. In observational studies, the same holds true for specific micronutrients, such as provitamin A carotenoids and vitamin C. The specific constituents of fruits and vegetables that confer protection are not known, and the results of the chemoprevention trials suggest a more complex role than researchers had previously thought. An important unanswered question is whether fruits and vegetables directly confer protection against cancer or whether estimates of fruit and vegetable consumption are indicators of differences between individuals who eat healthy and unhealthy diets that are leading to uncontrolled confounding. Nevertheless, the protective association noted for fruit and vegetable consumption has the potential to contribute to prevention. A diet adequate in fruit and vegetables is prudent for preventing chronic diseases in general.

Even for a dietary factor such as fruit or vegetable consumption that is generally associated with a reduced risk of lung cancer, the highest exposure category is usually associated with at most a halving of the risk of lung cancer. An association of this magnitude may result from residual confounding by cigarette smoking. Studies that control for cigarette smoking in the design are best suited to address the persistent concern about residual confounding by cigarette smoking. Examples are case–control studies in which cases and controls are closely matched on cigarette smoking history and studies limited to never-smokers.

Advances in the understanding of the role of diet in the etiology of lung cancer should not obscure the irrefutable fact that cigarette smoking is the predominant cause of lung cancer. Many important questions about the complex relationship between dietary habits and the development of lung cancer remain, but the path to permanently ending the epidemic of lung cancer is known: Prevent children from starting cigarette smoking and effectively assist addicted smokers to stop smoking cigarettes.

Acknowledgments

The authors thank Brant W. Hager for assistance in the preparation of the manuscript. This work was supported by funding from grants from the US National Cancer Institute (CA73790 and 5U01CA086308), the National Institute of Environmental Health Sciences (P30 ES03819), and the National Institute of Aging (5UO1AG018033).

See also: **Alcohol**: Disease Risk and Beneficial Effects. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Body Composition**. **Cancer**: Carcinogenic Substances in Food. **Carotenoids**: Chemistry, Sources and Physiology. **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Fruits and Vegetables**. **Phytochemicals**: Classification and Occurrence; Epidemiological Factors. **Vitamin A**: Physiology.

Further Reading

Alberg AJ and Samet JM (2003) Epidemiology of lung cancer. *Chest* **123**: 21S–49S.

Samet JM (ed.) (1995) *Epidemiology of Lung Cancer*. New York: Marcel Dekker.

Schottenfeld D and Fraumeni JF (eds.) (1996) *Cancer Epidemiology and Prevention*, 2nd edn. Oxford: Oxford University Press.

Willett WC (1998) *Nutritional Epidemiology*, 2nd edn. Oxford: Oxford University Press.

World Cancer Research Fund (1997) *Food, Nutrition, and the Prevention of Cancer: A Global Perspective*. Washington, DC: American Institute for Cancer Research.

Dietary Management

C Shaw, Royal Marsden NHS Foundation Trust, London, UK

Patients with cancer suffer from numerous eating difficulties due to the disease or treatments. However, if attention is paid to these problems and dietary intervention is early, patients can be relived from some of the symptoms and need not lose a great deal of weight. Many dietary problems may be anticipated if there is a known diagnosis and treatment

plan, and a prompt referral to a dietician is beneficial. All patients should undergo a nutritional screening and assessment, and high-risk patients or those experiencing problems should be referred for advice and appropriate support.

Weight loss and poor nutritional status may lead to poor wound healing, increased risk of local and systemic infection, reduced tolerance to treatment, poor postoperative recovery, and reduced quality of life. Maintenance of good nutritional status is important to enable patients to complete their course of anticancer treatment.

Nutritional Support

Many cancer patients will require some form of nutritional support during the course of their illness (**Table 1**). When patients have an eating difficulty, the first course of action is to assess their oral intake. If patients are able to eat, then they should be given appropriate advice to maximize their oral intake. If patients are unable to swallow enough nourishment to maintain their weight, an enteral tube feed should be considered. The type of tube placed will depend on the following:

1. The anticipated length of time the feed will be required.
2. The physical state of the patient; for example, a nasogastric tube or percutaneous endoscopically placed gastrostomy tube may not be suitable for patients with complete oesophageal obstruction. A jejunostomy tube may be preferred following upper gastrointestinal tract surgery.
3. The wishes of the patient concerning the physical appearance of different tubes and the invasiveness of the procedure required to place them.

Table 1 Methods of nutritional support

Method	*Route*
Oral feeding	Oral feeding can be facilitated by
	Altering the consistency or timing of food or drink
	Fortifying food and drinks with protein and energy
	Altering the flavoring added to food
	Using sip feeds and dietary supplements
Enteral tube feeding	Nasogastric or nasojejunal tube Gastrostomy
	Percutaneous endoscopically guided Gastrostomy
	Radiologically inserted gastrostomy
	Percutaneous gastrostomy with a jejunal extension
	Jejunostomy
Parenteral nutrition	Central line
	Peripheral line

Numerous brands of enteral feeds are available. Most cancer patients will require complete, whole protein feeds providing 4–6 kJ ml^{-1} (1–1.5 kcal ml^{-1}). Only in cases of severe malabsorption, gastrointestinal fistula, or pancreatic insufficiency may elemental, peptide, or low-fat feeds be necessary.

The choice of feeding regimen will depend on the patient's mobility and activity during the day and on the volume of feed tolerated. It may be administered in the following ways: pump feeding overnight and/or during the day; gravity feeding, which is usually provides a faster rate of feeding that does not require the precision of a pump; and bolus feeding.

Parenteral nutrition is required where the gastrointestinal tract cannot be used, such as in patients with complete bowel obstruction or severe malabsorption.

Practical Management of Eating Difficulties

Anorexia

Anorexia (loss of appetite) is often associated with other eating difficulties, such as nausea, taste changes, and constipation, and addressing these problems may improve the patient's appetite. Pain may also contribute to anorexia, and regular analgesia for pain may in turn help improve appetite, as may dietary alterations (**Table 2**). For patients who have severe anorexia, an appetite stimulant should be considered, such as dexamethasone, medroxyproesterone acetate, or megestrol acetate.

Taste Changes

Cancer patients may suffer from lack of taste or 'taste blindness,' they may find that foods taste metallic or excessively salty or sweet, or they may find that foods taste abnormal. Depending on the taste change experienced, it is often worth excluding certain foods from the diet or using certain flavorings to try to stimulate the taste buds (**Table 3**).

Table 2 Dietary management of anorexia

Give small, frequent meals and snacks in preference to three meals daily.
Serve food on a small plate.
Ensure food looks appetizing.
Encourage any food the person prefers, even if it is all of one type (e.g., puddings).
Distract from eating (e.g., by conversation, watching television, or listening to music).
Give an alcoholic drink to be sipped before meals or with food.

Table 3 Suggestions for overcoming taste changes

Taste change	*Suggestions*
Excessively sweet	Reduce sugar content of food and drink. Add a pinch of salt to drinks and puddings.
Excessively salty	Avoid packet soups, gravy, and sauces. Avoid salted snacks (e.g., crisps and nuts) or try unsalted varieties. Avoid bacon and other cured or tinned meat. Add a pinch of sugar to sauces or soups.
Metallic taste	Soak red meat in acidic marinate (e.g., vinegar and wine). Eat white meat, fish, eggs, and cheese in preference to red meat. Avoid tea, coffee, and chocolate.
Taste blindness	Use extra flavorings: salt, pepper, pickles, mustard, herbs, and spices. Eat highly flavored food (e.g., curry).

Nausea and Vomiting

Nausea and vomiting must be controlled with antiemetic drugs. Some dietary suggestions may help patients with food choice when they are feeling nauseous (**Table 4**).

Dysphagia

Dysphagia (difficulty swallowing) may occur with solid food, semisolid foods such as porridge, or liquids. For the person who cannot manage solid food but is able to eat semisolids, altering the consistency of the food may be the only dietary change needed, encouraging food with extra sauce, soft puddings, and nourishing drinks.

For the patient who is only able to swallow fluids, close attention must be paid to their intake and dietary supplements are likely to be necessary. Some people who can only manage liquids choose to liquidize their food; this dilutes the nutrients, so meals should be fortified with butter, cream, glucose, cheese, etc. to add protein and energy.

If there is complete dysphagia to both solids and liquids, feeding by an enteral tube should be considered.

Table 4 Suggestions for food and fluids when person has nausea

Have cold food and drink in preference to hot because these have less odor.
Sip fizzy drinks.
Drink through a straw.
Try ginger flavors (e.g., ginger ale and ginger biscuits).
Eat small, frequent snacks to avoid the stomach from becoming completely empty.

In some instances, people can swallow solid food but aspirate liquids. Patients should undergo a complete assessment from a speech and language therapist to ascertain which textures are safe to swallow. It may be that thickened liquids such as milk shakes or those thickened with a commercial thickener are suitable, whereas thin liquids, such as tea and water, are aspirated. If thick fluids are also aspirated, it is usually safer to give nothing by mouth and to maintain hydration and nutrition through an enteral tube.

Mucositis and Stomatitis

If the mouth or throat is sore, eating can become very difficult. An analgesic taken before meals can help ease the pain and enable the person to eat a little more. Modifying the diet is also helpful (**Table 5**).

Xerostomia

Xerostomia (dry mouth) may be a long-term side effect of cancer treatment, and patients may need to use extra sauce with their foods or have soft food, and they usually need to sip a drink while eating. Chewing gum, preferably sugar-free, can stimulate saliva, although it should be avoided by those with no saliva because it will stick to their teeth. Pineapple can also stimulate saliva and eating it between meals may make the mouth more comfortable.

Good dental hygiene is particularly important because saliva protects the mouth against infection. If people with xerostomia also get mouth infections, the resulting mucositis makes it increasingly difficult for them to eat.

Trismus and Difficulty Chewing

Trismus (difficulty opening the mouth) and difficulty chewing may be overcome with soft food or, failing that, with nourishing drinks and dietary supplements. If the person loses weight and can manage very little orally, an enteral tube feed should be considered.

Table 5 Suggestions to relieve mucositis and stomatitis

Avoid citrus fruits and drinks.
Avoid salty, spicy food, vinegar, pickles, and other strong flavors.
Avoid carbonated drinks.
Have tepid food and drinks.
Iced drinks may be soothing (or may increase the pain).
Avoid dry foods that need extra chewing (e.g., toast).
Eat soft food and use extra sauce.

Gastrointestinal Fistulas

A fistula may develop anywhere in the gastrointestinal tract. The site of the fistula will determine the dietary management (**Table 6**).

Constipation

The cause of constipation must be considered initially. If it is due to a tumor pressing on the bowel (e.g., cancer of the ovary or colon), a low-fiber diet may be helpful. Low-fiber food is less bulky and may pass through the bowel more easily, particularly if accompanied by appropriate laxatives (e.g., stool softener).

If constipation is due to lack of fiber in the diet, then an increased fiber and fluid intake may be helpful. If constipation is due to analgesia, then appropriate laxatives need to be used in conjunction with any changes in the diet. In addition to fiber, a good fluid intake must be maintained to avoid constipation; approximately 2 litres per day is recommended.

Diarrhea

Diarrhea may be due to overflow from constipation, in which case the advice for constipation should be followed. Diarrhea due to intestinal hurry caused by bowel disease or drugs may be controlled with drugs and by avoiding excessive intake of high-fiber foods, which naturally pass through the bowel quickly. When malabsorption is suspected, a low-fat, elemental enteral tube feed should be considered. When diarrhoea is severe, it is important to replace the fluid lost to prevent dehydration. Oral rehydration solution is useful to replace fluid loses. Diarrhea caused by radiotherapy needs to be controlled with drugs, and a low-fiber diet is not thought to be helpful in this instance.

Intestinal Failure

A long-term side effect of pelvic radiotherapy may be enteritis resulting in intestinal failure. Extensive gastrointestinal surgery leaving less than 100 cm of small bowel, or a fistula in the small bowel causing high stoma losses, may also cause intestinal failure. Previous chemotherapy that may affect the function of the bowel can contribute to this condition. Intestinal failure is more likely to occur when the patient does not have a functioning colon (e.g., in the case of ileostomists or when the ileo-caecal valve is absent).

Dietary manipulation can greatly alleviate the symptoms of intestinal failure, such as thirst, dehydration, and high stoma losses or large volumes of diarrhea (**Table 7**).

An oral rehydration solution consisting of 20 g glucose, 3.5 g sodium chloride, 2.5 g sodium bicarbonate, and 1000 ml water provides 90 mmol of sodium per liter. It may be used chilled and to dilute weak fruit squashes. If the patient remains dehydrated despite following the advice detailed in **Table 7**, intravenous fluid replacement is necessary.

Drugs may be given to increase gut transit time or reduce fluid losses. If medication is in the form of capsules, these should be opened and the drugs given 60 min before meals. Suitable drugs include codeine phosphate, loperamide, rantidine, and octreotide. In the longer term, the following should be monitored:

Plasma electrolytes, ferritin, and vitamin D levels
Serum albumin, magnesium, zinc, calcium, phosphate, and alkaline phosphate
Folate and vitamin B_{12} concentrations
Prothrombin time
Body weight
Urinary sodium concentration

Bowel Obstruction

Bowel obstruction may be subacute or complete. In cases of complete bowel obstruction, the clinical condition of the patient must be considered. If it is

Table 6 Sites of fistulas and their management

Site	*Management*
Neck, salivary fistula	'Nil by mouth' and enteral tube feed until healed
Chyle leak (e.g., in neck)	Low-fat diet initially; if unsuccessful, a low-fat, medium-chain triglyceride enteral tube feed If unsuccessful, consider parenteral nutrition
Large bowel	Low-residue diet or elemental enteral tube feed
Small bowel	See **Table 7**

Table 7 Dietary management to reduce gut losses in intestinal failure

Restrict fluids to 500–1000 ml daily, increasing to 1500 ml.
Avoid drinks for 30 min before and 45 min after meals.
Avoid foods that are particularly high in fiber.
Sprinkle salt liberally on food.
Consider fat restriction if patient has a colon and there is evidence of steatorrhea.
Take salt and carbohydrate foods together to help sodium absorption.
If gut losses are 1000 ml or more, part or all of fluid intake should consist of an oral rehydration solution.

anticipated that the obstruction will resolve, or if aggressive treatment such as surgery is planned, parenteral nutritional support should be considered. Total parenteral nutrition may be inappropriate and is unlikely to be useful in cases in which the prognosis is poor and no treatment is possible.

Depending on the degree of obstruction, in cases of subacute obstruction the following action may be taken under medical supervision:

First day: sips of clear fluid, approximately $10\,ml\,h^{-1}$
Second day: $30\,ml\,h^{-1}$ clear fluid
Third day: $60\,ml\,h^{-1}$ clear fluid
Fourth day: free clear fluids
Fifth day: free fluids, including milk, low-fiber soup, custard, and jelly
Sixth day: low-fiber diet, avoiding all fruit and vegetables, nuts, pulses, and whole grain cereals, whole meal bread, etc.

A patient who starts to vomit should return to the diet prescribed for the preceding day. If symptoms of bowel obstruction, such as abdominal pain and indigestion, remain controlled, fruit and vegetables may be introduced as tolerated, starting with small amounts.

Weight loss

Weight loss is often the consequence of the dietary problems described previously. The measures in **Table 8** should be considered to help prevent weight loss or encourage weight gain. It must be remembered that energy requirements may be elevated due to the physiological effects of malignancy. Much interest has focused on attempts to influence the metabolic alterations in cachexia via nutrients. Research has examined the possible role of eicosapentaenoic acid (EPA), an n-3 fatty acid, in reducing the inflammatory response in cachexia. A randomized trial in pancreatic cancer patients compared a high-energy drink fortified with EPA to a standard high-energy drink to examine whether this was more effective at promoting weight gain. The study failed to show any additional benefit of EPA in terms of weight gain.

Table 8 Dietary advice to help prevent weight loss

Fortify food with cream, butter, cheese, oil, sugar, honey, glucose, jam, etc.
Have small, frequent snacks.
Use full-fat and full-sugar products.
Avoid large amounts of lower energy foods (e.g., fruit and vegetables).
Try dietary supplements, such as milky drinks and glucose polymer power.
Consider an overnight enteral tube feed to supplement the diet if weight loss continues despite following the previous advice.

Palliative Care

In some people, cancer will not be cured. Palliative care focuses on the relief of symptoms rather than aggressive curative treatment. The majority of people receiving palliative care will suffer from at least one eating difficulty. Much of the advice detailed previously for overcoming dietary problems is relevant, but it is often upsetting for these patients to have to pay close attention to their dietary intake. If patients are unconcerned about their poor dietary intake, it may be appropriate not to offer any advice; conversely, for those who are very concerned, the problem should be addressed seriously.

Alternative and Complementary Diets

The alternative and complementary diets considered here are modifications of a normal diet that are claimed to cure or treat cancer. Such diets are often followed for their anticipated antitumor effect. Often, they have not been tested or demonstrated to be effective in scientifically acceptable clinical trials. Patients may use other complementary therapies, such as healing, relaxation, visualization, homeopathy, and herbalism, in addition to making dietary changes. Dietary regimens may share common features:

Mainly vegetarian or vegan—alternatively, diets may limit red meat and allow limited free-range chicken and deep-sea fish
No manufactured or processed foods
Low in salt
Low in sugar
Low in fat
High in fiber, including raw fruit and vegetables and whole grains (these may be organic)
May include fruit and vegetable juices
High-dose vitamins and minerals

Nutritional inadequacies may arise in the patient who has a poor appetite. The diets may cause weight loss and are restrictive and time-consuming to prepare. Some ingredients may be difficult to obtain and are often costly. Studies appear to show no difference in survival rates between patients following complementary therapies and patients receiving conventional treatment alone. Patients who use complementary therapies, however, do report psychological benefits, such as feelings of hope and optimism. Patients should have enough information about the possible advantages and disadvantages

before embarking on strict complementary or alternative diets.

The Potential Therapeutic Role of Vitamins

Much interest has been expressed in the therapeutic role of vitamins in cancer patients. This has led a number of alternative and complementary practitioners to advocate the use of high-dose vitamins for cancer patients. It has been know for some time that some vitamin-deficiency states may predispose some individuals to develop cancer. In a study of 29 000 vegetarian Chinese with a high frequency of oesophageal cancer, subjects were given supplements of β-carotene and vitamin E. Raising their daily intake above the minimum requirement reduced the incidence of deficiencies and reduced the number of oesophageal cancers. This type of study on vitamins and the etiology of cancer has led many practitioners and laypeople to extrapolate he role of vitamins into cancer treatment.

Although vitamins in food, especially vegetables and fruits, have been shown to be beneficial in reducing the incidence of particular types of cancer when included in the diet, the beneficial effects have not always been shown with vitamin and mineral supplements. Some supplements may promote tumor growth, as was seen in a study using β-carotene supplementation in patients with lung cancer. Supplementation increased the rate of tumor recurrence in such patients.

The potential therapeutic role of vitamins, such as vitamins D, K, B_6, B_{12}, and folate, has been investigated. However, additional studies are required to determine the role, if any, of such vitamins. It may be that some vitamins help protect against the side effects of tumor therapy, whereas some may modify tumor growth. Excessive dietary supplementation in cancer patients should be avoided until further evidence is available on the effects of vitamins on tumor growth.

See also: **Cancer**: Epidemiology and Associations Between Diet and Cancer; Epidemiology of Gastrointestinal Cancers Other Than Colorectal Cancers; Epidemiology of Lung Cancer; Effects on Nutritional Status. **Cobalamins**. **Colon**: Nutritional Management of Disorders. **Diarrheal Diseases**. **Eating Disorders**: Anorexia Nervosa. **Folic Acid**. **Nutritional Support**: Adults, Enteral; Adults, Parenteral; Infants and Children, Parenteral. **Supplementation**: Dietary Supplements. **Vitamin B_6**. **Vitamin D**: Physiology, Dietary Sources and Requirements. **Vitamin E**: Metabolism and Requirements; Physiology and Health Effects. **Vitamin K**.

Further Reading

Bozzetti F (2001) Nutrition support in patients with cancer. In: Payne-James J, Grimble G, and Silk D (eds.) *Artificial Nutrition Support in Clinical Practice*, pp. 639–680. London: Greenwich Medical Media.

Cerhan JR, Potter JD, Gilmore JME *et al.* (2004) Adherence to the AICR cancer prevention recommendations and subsequent morbidity and mortality in the Iowa Women's Health Study Cohort. *Cancer Epidemiology, Biomarkers and Prevention* **13**(7): 1114–1120.

Downder SM, Cody MM, McCluskey P *et al.* (1994) Pursuit and practice of complementary therapies by cancer patients receiving conventional treatment. *British Medical Journal* **309**: 86–89.

Fearon KCH, von Meyenfeldt MF, Moses AGW *et al.* (2003) Effect of a protein and energy dense n-3 fatty acid enriched oral supplement on loss of weight and lean tissue in cancer cachexia: A randomised double blind trial. *Gut* **52**: 1479–1486.

Food Standards Agency Expert Group on Vitamins and Minerals (2003) *Safe Upper Limits of Vitamins and Minerals*. London: Food Standards Agency.

Gianotti L, Braga M, Nespoli L *et al.* (2002) A randomised controlled trail of preoperative oral supplementation with specialised diet in patients with gastrointestinal cancer. *Gastroenterology* **122**: 1763–1770.

Effects on Nutritional Status

C Shaw, Royal Marsden NHS Foundation Trust, London, UK

Many patients with cancer experience nutritional problems during their treatment. The physiological effects of malignancy can cause increased nutritional requirements and a reduced nutritional intake. Anticancer treatment can produce side effects, including anorexia, mucositis, nausea, and vomiting, that can further reduce nutritional intake.

Weight loss and nutritional depletion of the cancer patient may interfere with anticancer treatment. Patients with cancer who lose weight may have a reduced tolerance to treatment due to poor wound healing and an increased susceptibility to infection. Weight loss may also contribute to a poor quality of life. Nutritional support of patients with cancer should be an integral part of treatment. Modification of oral intake may be sufficient to maintain nutritional status. There is also interest in nutrients

that may modify cachexia in patients for whom altered metabolism is also contributing to weight loss. If the patient is unable to take sufficient nutrition orally, then food may be given through an enteral tube. Parenteral nutrition may be required when the gastrointestinal tract cannot be used.

Physiological Effects of Malignancy

Cancer cachexia is a syndrome suffered by many, but not all, patients with cancer. It is most prevalent in patients with tumors of the lung, head and neck, or gastrointestinal tract. Features of cachexia include weight loss, muscle wasting, lethargy, anorexia, early satiety, anemia of a nonspecific type, and altered host metabolism.

Weight loss is the most obvious feature of cachexia. In patients with cancer, an unintentional weight loss of 10% is deemed to be a significant change. Cachexia may be masked in people who were previously overweight or who have oedema. These people may be especially at risk of suffering from the consequences of cachexia, such as poor wound healing and increased risk of infection because their nutritional problems may not be addressed as promptly as those of patients who are obviously underweight. The causes of cancer cachexia are multifactorial (**Figure 1**).

Physiological Causes of Cancer Cachexia

Weight-losing patients with cancer may have a reduced energy intake, increased energy expenditure, or a combination of both. It has been estimated that 50–60% of cancer patients in hospital have abnormal resting energy expenditures, although this may be restricted to certain diagnoses, such as pancreatic, lung, and gastrointestinal cancers. Cachexia differs from uncomplicated starvation in that the usual adaptation to a reduced food intake, such as a reduction in energy expenditure and conservation of body protein stores, does not appear to take place. In the starving state, metabolic rate is decreased, weight loss occurs mostly from fat stores, and nitrogen losses are reduced; however, the reverse occurs in many patients with cancer.

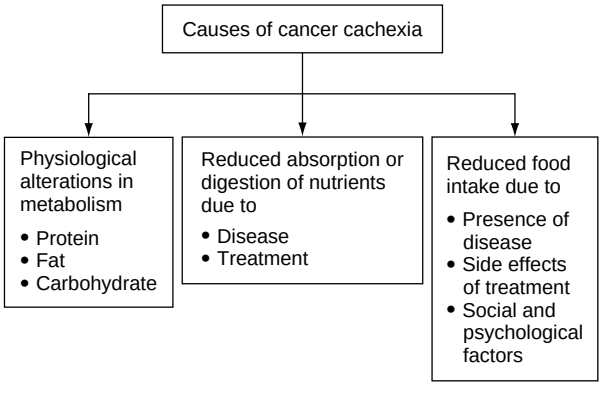


Figure 1 Causes of cancer cachexia.

Carbohydrate Metabolism

Some patients with cancer have been demonstrated to have a high rate of glucose turnover in both the fasting and the fed states and one cause may be increased Cori cycle activity. In the Cori cycle, glucose is metabolized anaerobically by the tumor, producing lactate (**Figure 2**). The tumor cannot utilize lactate and so it is returned to the liver where it is converted back into glucose. This utilizes energy, and the process is thus an energy-wasting cycle: Only 2 mol of ATP is produced by anaerobic glycolysis, and gluconeogenesis requires 6 mol of ATP. Had the lactate been metabolized aerobically via the Krebs cycle, 30 mol of ATP would have been synthesized. It has been estimated that Cori cycle activity may account for increased energy requirements of 1260 kJ (300 kcal) per day. Lactate has been shown to cause nausea, so it is possible that another side effect of the Cori cycle is reduced food intake.

Insulin resistance has been established as a common hormonal alteration, and raised levels of growth hormone seen in some patients with cancer may contribute to insulin resistance. Glucose intolerance is associated with sepsis, bed rest, starvation, and malnutrition, all of which may occur in cancer patients, making it difficult to establish how much the tumor contributes to glucose intolerance.

Protein Metabolism

Muscle wasting or loss is common in patients with cancer. It has been estimated that protein–energy malnutrition may be present in 50–80% of patients with cancer. Data from human studies indicate increased whole-body protein turnover, synthesis, and catabolism, increased hepatic protein synthesis, and reduced rates of albumin and skeletal muscle protein synthesis. It has been estimated that whole-body protein turnover in cancer patients is 32%

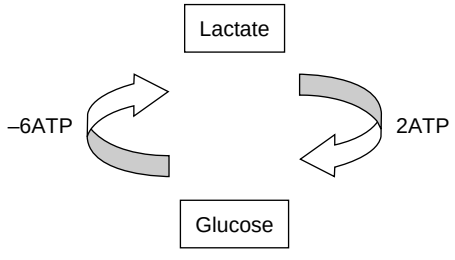


Figure 2 The Cori cycle.

higher than in noncancer patients and 35% higher than in starved normal subjects.

Glucose required by the tumor may be supplied from dietary glucose or by the conversion of amino acids into glucose. Weight-losing patients with cancer may have reduced plasma levels of amino acids alanine, glycine, and glutamine, possibly because these compounds are used for gluconeogenesis, and this may be a cause of increased protein catabolism.

Fat Metabolism

Loss of fat or adipose tissue is common in patients with cancer and is one of the most obvious signs of cachexia. Adipose tissue consists of triglycerides, which can be metabolized to yield free fatty acids and glycerol. In patients without cancer, glycerol concentrations are decreased postprandially because the body does not need to break down adipose tissue; however, patients with cancer have been shown to have raised glycerol concentrations in the fed state, which suggests they are in a hypermetabolic condition.

Alterations in carbohydrate and protein metabolism are greater than those for fat, and it is likely that loss of body fat is mostly due to raised energy expenditure as a consequence of glucose intolerance and cycling.

Differences between Malnourished Patients with and without Cancer

During starvation, the metabolism of people without cancer adapts in order to conserve body tissue. Patients with cancer, however, do not exhibit these mechanisms and therefore lose weight (**Table 1**).

The Role of Cytokines in Patients with Cancer

Cytokines are a range of polypeptides produced by cells of the immune system in response to an inflammatory action. Cytokines, including tumor necrosis factor, interleukins-1 and -6, and interferon-gamma, have been shown to induce some of the features of cancer cachexia when administered to humans. The tumors of patients with cancer may induce an inflammatory reaction resulting in raised levels of cytokines, or it may be that the cytokines produce proinflammatory cytokines. There may be amplification of the effect of cytokines by interaction between two or more.

Table 1 Metabolic differences between cancer patients and noncancer patients during starvation

Noncancer patients	*Cancer patients*
Reduced glucose tolerance	Increased glucose turnover
Reduced total body protein turnover, including reduced hepatic protein synthesis	Total body protein turnover maintained or increased
Slow weight loss, preferentially of stored fat	Rapid weight loss of fat and protein

There is much interest in the possible modulation of cytokine activity by fish oils that contain eicosapentenoic acid. Fish oil supplements rich in n-3 fatty acids may reduce production of cytokines and have been shown to inhibit fat and protein breakdown in animal models of cancer cachexia. Small studies of humans have demonstrated a reduction of proinflammatory cytokines in pancreatic cancer patients when given a supplement equivalent to 2 g of eicosapentaenoic acid (EPA) per day. However, in a clinical trial EPA-supplemented high-energy drinks and standard high-energy drinks slowed weight loss to the same extent.

Mechanisms of Cancer-Related Anorexia

Anorexia is the term given to loss of appetite, and it is thought that varying levels of neurotransmitters within the hypothalamus influence appetite. The neurotransmitter serotonin reduces appetite and neuropeptide Y stimulates appetite. Levels of neurotransmitters in the brain may be altered by plasma nutrients, hormones, and nerve impulses arriving from the gastrointestinal tract.

Although not all mechanisms controlling appetite are known, several theories have been suggested to account for anorexia in cancer patients. The amino acid tryptophan is usually bound to albumin. In cancer patients, because albumin synthesis may be reduced, there may be more free tryptophan circulating in the plasma. Tryptophan is a precursor of serotonin, which is known to inhibit appetite. This model has been used to examine appetite in animal models of cachexia but there is still debate as to whether this is the main cause of the anorexia of cachexia. Cytokines are also known to cause anorexia and, as discussed previously, cancer patients have raised levels of cytokines.

Effects of Treatment on Nutritional Status

Various treatments may be used for cancer with the goal of curing it or palliating symptoms (**Table 2**). All may potentially affect nutritional status.

The treatment chosen depends on the position of the tumor, its extent, and its sensitivity to radiotherapy or chemotherapy. Often, different treatments will be used in succession (e.g., surgery followed by chemotherapy, or radiotherapy followed by surgery).

Table 2 Treatment for cancer

Treatment	*Examples*
Surgery	Removal of tumor
	Removal of organ (e.g., gastrectomy, nephrectomy, colectomy)
	Palliation of symptoms (e.g., intestinal bypass, colostomy formation)
Chemotherapy	Single agent
	Combination of agents
	Drugs given as single dose or continuous infusion (e.g., methotrexate, epirubicin, mytoycin, fluorouracil)
Radiotherapy	External bean
	Single fractions daily
	Hyperfractionated
	Intensity modulated radiation therapy
	Brachytherapy—interstitial (e.g., iridium wires)
	Radioisotopes (e.g., iodine 131)
Biological therapies	Interferon
	Interleukin
Endocrine therapies	Tamoxifen
	Aminoglutethimide
	Goserelin
	Medroxyprogesterone acetate
	Megestrol acetate

Alternatively, treatments may be used concurrently (e.g., chemoradiation). The effect of treatment on nutritional status depends on the site of the tumor and the treatment given.

Surgery is used to remove all or part of the tumor or to bypass the tumor, thereby allowing organs to continue to function. There are often periods of starvation before and after surgery that may contribute to malnutrition in the cancer patient.

The trauma of surgery causes an increase in the production of catecholamines such as adrenaline, which results in the obligatory loss of nitrogen from the body. Repeated or extensive surgery contributes to an increase in metabolic rate and therefore contributes to nutritional depletion.

Chemotherapy is based on the use of drugs that interrupt the cell cycle and prevent cell multiplication. The drugs act on rapidly proliferating cells and also damage healthy cells, particularly those in the gastrointestinal tract and hair follicles. They may also cause nausea and vomiting, altered taste, nerve damage, and infertility. Drugs may be given orally, intravenously (bolus or continuous infusion), or intrathecally. High-dose chemotherapy may be used with a stem cell rescue. **Table 3** lists commonly used chemotherapeutic agents and their side effects that affect nutritional status.

Radiotherapy is the use of ionizing radiation to destroy malignant cells. External beam radiation is most commonly used. It may be used in combination with chemotherapy to enhance the effects of both treatments.

Table 3 Side effects of chemotherapy that may affect nutritional status

Agent	*Side effects*
Methotrexate	Severe mucositis
	Nausea (dose dependent)
	Vomiting
Vincristine	Paralytic ileus
Fluorouracil	Diarrhea
	Occasional nausea
Cisplatin	Anorexia
	Severe prolonged nausea
	Nausea and vomiting
	Taste changes (particularly a metallic taste)
	Diarrhea (high doses)
Doxorubicin	Nausea
	Some vomiting
	Mucositis throughout gastrointestinal tract
Docetaxol	Diarrhea

Biological therapies are based on the use of cytokines derived from cells in the immune system. Cytokines are administered to stimulate the body's immune response to reduce or prevent tumour growth and often cause anorexia.

Endocrine therapies are primarily used to control the growth of hormone-dependent cancers, such as cancer of the breast or prostate. Drugs may be used to block production of hormones or to block hormone receptors.

Head and Neck Cancer

Surgical resection is often used for cancers of the oropharynx. Removal of the tumor and reconstruction may lead to periods when the patient is not allowed to eat or drink in order to allow healing to take place. Resection of the mandible, tongue, maxilla, or pharynx may lead to difficulties with chewing or swallowing. There may be an increased risk of aspiration because of a poor ability to control food or fluids in the mouth or because of an alteration in the anatomy or cranial nerves required for swallowing.

Radiotherapy to the head and neck can have a significant impact on nutritional status with both early side effects, during or immediately after treatment, and late side effects, which may occur years after treatment (**Table 4**). Reduced food intake and weight loss due to soreness and dysphagia are particularly common and occur in up to approximately 90% of patients. A dry mouth (xerostomia) occurs when the salivary glands are irradiated and leave the teeth increasingly prone to tooth decay.

Table 4 Side effects of radiotherapy to the head and neck

Early	*Late*
Mucositis (inflammation of the mucosa)	Mucosal ulceration
Stomatitis (inflammation of the mouth)	Xerostomia
Xerostomia (dry mouth)	Increased viscosity of saliva
Increased viscosity of saliva	Dysphagia
Dysphagia (difficulty swallowing)	Altered taste
Altered taste	Mouth blindness
Mouth blindness	Dental caries
	Trismus (inability to open mouth)
Nausea	Fibrosis (formation of excessive fibrous tissue)
Anorexia	Stenosis (narrowing)
Loss of smell	Fistula
	Poor wound healing
	Osteoradionecrosis (degeneration of bone)

Chemotherapy may be used in combination with surgery and occasionally given concurrently with radiotherapy. Reduced food intake results from nausea, vomiting, learned food aversions, anorexia, and mucosititis. Artificial nutritional support may be required during treatment if oral intake is significantly reduced.

Gastrointestinal Cancer

Surgery, chemotherapy, and radiotherapy treatments may be used for cancers of the gastrointestinal tract, depending on the site and type of disease. The impact of both disease and treatment on nutritional status is often great, particularly in upper gastrointestinal cancers.

Surgical resection of the upper gastrointestinal tract often affects the capacity to ingest food and fluids. Surgery that changes the length or motility of the small intestine affects the ability to digest and absorb food and fluids (**Table 5**). Tumors of the small intestine are rare, but small intestinal resection may be necessary because of strictures caused by previous abdominal or pelvic radiotherapy or because of adhesions from previous surgery.

Radiotherapy is often used to treat tumors of the gastrointestinal tract. It can severely affect the ingestion, digestion, and absorption of food and fluids (**Table 6**). Chronic malabsorption of bile salts contributes to fat malabsorption. Bile salts entering the colon inhibit water absorption and stimulate colonic peristalsis, causing fluid and electrolyte deficiencies.

Chemotherapy for gastrointestinal tumors is often associated with side effects such as anorexia, nausea, vomiting, and mucositis (inflammation of the gastrointestinal mucosa). Learned food aversions, which arise through the association of particular foods with vomiting induced by chemotherapy, may also have an impact on nutritional intake.

Table 5 Nutritional consequences of surgery to the gastrointestinal tract

Area of gastrointestinal tract resected	*Impact on nutritional status*
Oesophagus	Intestinal hurry (due to vagotomy and pyloroplasty)
	Reduced gastric capacity (due to stomach pull up)
	Stricture at anastomosis (surgical junction)
	Dumping syndrome
Stomach	Reduced capacity for food and fluids
	Early satiety
	Dumping syndrome
	Intestinal hurry (due to vagotomy and pyloroplasty)
	Fat malabsorption
	Anemia (lack of intrinsic factor)
Small intestine	General malabsorption
	Diarrhea
Terminal ileum	Malabsorption of vitamin B_{12}
	Fat malabsorption
	Decreased absorption of bile salts, fat-soluble vitamins, and minerals
	Short bowel syndrome

Leukemias and Lymphomas

Systemic diseases such as leukemias and lymphomas are usually treated with chemotherapy. Chemotherapy is generally given intermittently to allow the bone marrow to recover. Patients experience anorexia, severe mucositis, nausea, vomiting, taste changes, food aversions, tiredness, and lethargy. Some drugs, such as doxorubicin, may reduce gut motility and increase the risk of gastrointestinal obstruction.

Table 6 Effects of radiotherapy on the gastrointestinal tract

Area irradiated	*Side effects*
Oesophagus	Dysphagia
	Mucositis
	Fibrosis
Stomach	Anorexia
	Nausea
	Vomiting
Abdomen and pelvis	Anorexia
	Nausea
	Vomiting
	Diarrhea
	Malabsorption
	Early enteritis
	Chronic enteritis

Radiotherapy may be used to treat isolated lymph nodes, as in the case of lymphoma, or it may be used to abolate the immune system prior to a bone marrow transplant or peripheral stem cell rescue. Profound neutropenia may results in bouts of sepsis and increased nutritional requirements due to pyrexia.

Bone Marrow Transplant

Bone marrow transplantation (BMT) carries a high risk of patients developing malnutrition due to the severe side effects of high-dose chemotherapy and whole-body irradiation. BMT may be used to treat acute myeloid leukemia, acute lymphoblastic leukemia, chronic myeloid leukemia, and lymphoma. Side effects that alter nutritional status are more likely to occur in allografts, in which the bone marrow from a matched or mismatched donor is used rather than the patient's own bone marrow. Graft versus host disease may occur in allogenic BMT. It is characterized by inflammation of the skin, which may cause increased nutritional requirements. Other symptoms include gastrointestinal involvement with severe diarrhoea; loss of blood, mucus, and tissue via the gastrointestinal tract; and increased fluid losses. There may also be liver involvement and altered liver function (**Table 7**).

Gynecological Cancer

Treatment for cancer of the ovary, uterus, or cervix can have a major impact on nutritional status (**Table 8**). Malignant disease, surgery, or radiotherapy in the pelvis can lead to adhesions, strictures of the bowel, malabsorption, and bowel obstruction. Irreversible damage may necessitate a gastrointestinal resection or the formation of an ileostomy or colostomy. Malignant ascites can impinge on the gastrointestinal tract due to pressure exerted by fluid within the abdomen, causing anorexia and early satiety. Chemotherapeutic drugs (e.g., cisplatin) used to treat gynecological cancers cause severe nausea and vomiting and therefore greatly reduce food intake.

Table 7 Side effects of bone marrow transplantation affecting nutritional status

Early	*Late*
Anorexia	Anorexia
Weight loss	Weight loss
Vomiting	Xerostomia
Stomatitis	Taste changes
Xerostomia	Chronic graft versus host disease
Diarrhea	Somnolence
Malabsorption	
Taste changes	
Acute graft versus host disease	

Table 8 Nutritional problems in gynecological cancer

Anorexia
Early satiety
Nausea
Vomiting
Taste changes
Radiation enteritis, early or late
Short bowel syndrome
Subacute bowel obstruction
Complete bowel obstruction

Brain Tumors

Brain tumors are generally treated with surgery followed by radiotherapy or chemotherapy. A small percentage of patients may develop dysphagia as a result of the tumor, the treatment, or both, particularly in tumors of the brain stem. Patients may also experience taste changes, which can be severe and permanent, and somnolence following brain radiotherapy. High-dose steroids, such as prednisolone or dexamethasone, may cause increased appetite, steroid-induced diabetes mellitus, and weight gain.

Breast Cancer

Breast cancer patients are often treated with surgery, radiotherapy, and chemotherapy. Because surgery and radiotherapy are performed away from the gastrointestinal tract, these treatment modalities rarely have a major impact on nutritional status. Chemotherapy with drugs such as epirubicin may cause nausea, vomiting, and stomatitis (inflammation of the mouth), and docetaxol may cause diarrhea. In advanced disease, mediastinal nodes may cause dysphagia, and secondary liver disease may cause anorexia and nausea. Chemotherapy may cause general anorexia, lethargy, and nausea in some patients, although there is evidence that many breast cancer patients may gain weight during chemotherapy or when taking hormone treatments.

Cancer in Children

Malnutrition is common in children undergoing treatment for cancer because treatment is often aggressive and multimodal. Malnurition may occur in approximately one-third of all pediatric patients but is more common in certain diagnostic groups. High-risk diagnoses include Ewing's sarcoma, Wilms' tumor, head and neck tumors, advanced lymphomas, and neuroblastoma. Malnutrition and anticancer treatment in children may affect their future growth. The possible causes of malnutrition are listed in **Table 9**.

Table 9 Causes of malnutrition in children with cancer

Increased metabolic rate
Mechanical gastrointestinal problems (e.g., tumor pressing on stomach or gastrointestinal tract)
Malabsorption
Nausea
Vomiting
Taste abnormalities
Mucositis
Stomatitis
Dysphagia
Diarrhea
Somnolence
Behavioral and environmental factors
Poor eating habits
Learned food aversions
Noncompliance with dietary regimens

See also: **Ascorbic Acid**: Physiology, Dietary Sources and Requirements; Deficiency States. **Cancer**: Effects on Nutritional Status; Epidemiology and Associations Between Diet and Cancer. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology. **Small Intestine**: Disorders; Structure and Function. **Stomach**: Disorders; Structure and Function.

Further Reading

Bozzetti F (2001) Nutrition support in patients with cancer. In: Payne-James J, Grimble G, and Silk D (eds.) *Artificial Nutrition Support in Clinical Practice*, pp. 639–680. London: Greenwich Medical Media.

Gibney E, Elia M, Jebb SA, Murgatroyd P, and Jennings G (1997) Total energy expenditure in patients with small-cell lung cancer: Results of a validated study using the bicarbonate–urea method. *Metabolism* **46**(12): 1412–1417.

Tisdale MJ (2003) The 'cancer cachectic factor.' *Supportive Care in Cancer* **11**(2): 73–78

Carcinogenic Substances in Food

D Anderson, University of Bradford, Bradford, UK
J C Phillips, BIBRA International Ltd, Carshalton, UK

Introduction

Chemicals that are known or suspected to be carcinogenic to experimental animals and man are widespread throughout the environment. They occur naturally in the physical environment and are found in a very large number of higher plants, fungi and micro-organisms, many of which are part of the human diet. Some carcinogens have also been introduced into the human diet as a result of traditional cooking and preserving practices. Although carcinogens act through a wide variety of mechanisms, a substantial number have a common mechanism of action in that they react with the genetic material of the body, DNA. These so-called genotoxic carcinogens generally require metabolic activation to express their carcinogenicity. Although substantial efforts are being made to develop short-term, non-animal tests to predict the carcinogenicity of chemicals, animal bioassays remain the only reliable method for establishing the potential of a chemical to be a carcinogen, and form the basis of current approaches for the control of potentially carcinogenic chemicals in the human diet.

Naturally Occurring Carcinogens

It has been estimated that the total number of known chemicals exceeds 7 million, and that the great majority are naturally occurring. Although only a very small proportion (perhaps less than 0.01%) of these chemicals have been tested for carcinogenic potential in laboratory studies, a high proportion (as high as 50% in some evaluations) have been found to be positive. Therefore, even allowing for the imperfect selection and testing process, it is likely that there are a very large number of naturally occurring carcinogenic chemicals in the universe of chemicals, and therefore in the food we eat.

Naturally occurring substances identified as carcinogens in animals and humans by the range of approaches available for this purpose include inorganic compounds, organometallic compounds, and both simple and complex organic chemicals (see **Table 1**). These materials are present in the environment either as naturally occurring minerals or as a result of natural processes acting in the environment such as combustion, radioactive decay, or biodegradation of plant materials to oils. They are also widespread throughout the plant kingdom in both edible and nonedible plants and in many fungi and in unicellular organisms.

Inorganic Chemical Carcinogens

Many metallic elements are present as contaminants in food, being derived from a range of sources including the water used in food processing, soil residues, packaging, and cooking equipment. A number of metals and some of their salts have been shown to be carcinogenic in animals and humans, particularly to the lungs. These include arsenic,

Table 1 Examples of naturally occurring carcinogens

Inorganic chemicals
Arsenic; beryllium; chromium; cobalt; cadmium; lead; manganese; nickel
Polonium; radium; uranium; radon (gas)
Asbestos; silica (glassfiber); talc
Organic chemicals – complex mixtures
Mineral oils; shale oil; soot; wood shaving/dust
Organic chemicals in higher plants
Cycasin (betel nuts); saffrole (sassafras); pyrrolizidine alkaloids (Borginaceaea, Compositae); ptaquilosides (bracken); nitrosoalkaloids (tobacco)
Organic chemicals in lower order plants and microorganisms
Agaratine (mushrooms); aflatoxin, ochratoxin, sterigmatocystin (*Aspergillus* spp. and others); mitomycin, streptozotocin, daunomycin, actinomycin (*Streptomyces* spp.)

beryllium, cadmium, chromium, and nickel. Little is known about the mechanism by which metals cause cancer, although evidence is emerging that some metal ions affect the fidelity of an enzyme involved in the biosynthesis of DNA resulting in abnormal DNA being produced. A number of naturally occurring radioactive elements are also carcinogenic, particularly to the lungs. These include uranium, radium, and radon gas and may act by damaging DNA directly or by increasing oxidative damage as a result of an increase in reactive radical species. In addition, some naturally occurring minerals such as asbestos, silica, and talc are known to be carcinogenic to animals and humans under some circumstances.

Organic Chemicals – Complex Natural Mixtures

The earliest association made between the development of cancer in humans and exposure to an essentially natural rather than man-made chemical was that between scrotal (skin) cancer and soot by Percival Pott in 1775. However, the specific chemical(s) responsible (polycyclic aromatic hydrocarbons such as benzo(a)pyrene and 7,12-dimethylbenzanthracene) were not identified until more than a century later. Since then, a number of other naturally occurring materials have been shown to be carcinogenic. These have included mineral oils, shale oils, and wood dust/shavings, the oils being carcinogenic to the skin and wood dust to the nasal cavity. Inadvertent ingestion of small amounts of such materials with food may be difficult to avoid.

Organic Chemicals in Higher Plants

Although the acute toxicity of many plant species has been known since written records first appeared, only comparatively recently has the carcinogenicity of plant-derived products been recognized. The list of confirmed animal carcinogens present in plants is still relatively small, and few, if any, are confirmed or suspected human carcinogens. However, developments in analytical chemistry will allow an increasingly detailed inventory to be made of chemicals in plants, which will undoubtedly result in the discovery of many more carcinogens in our foodstuffs. The recent identification of over 1000 chemicals in coffee beans, and the observation that whereas only 3% of the chemicals had been tested for carcinogenicity, nearly 70% of these tested positive, is a clear pointer to future directions. Although it can be argued that the majority of these compounds are present at very low levels in plants, and so the hazard to man from any individual compound may be small, reliable methods for assessing both hazard and risk of low-level exposure are not well developed. In addition, methods for assessing the hazard from complex mixtures of chemicals are also poorly developed, resulting in additional uncertainty in evaluating the hazard posed by natural materials. The identified chemical carcinogens in plants tend to be secondary metabolites, often present as part of the plants natural defense mechanism against predation (i.e., natural pesticides), and as such are widespread in fruit, vegetables, herbs, and spices (see **Table 2**).

Table 2 Some naturally occurring carcinogenic plant pesticides (a) and their sources (b)

(a)

Chemical class	*Examples*
Aldehyde	Crotonaldehyde; benzaldehyde; hexanal
Hydrazine/hydrazone	*N*-methyl-*N*-formylhydrazine; methylhydrazine; pentanal methylformylhydrazone
Alcohol	Methylbenzyl alcohol; catechol
Ester	Ethyl acrylate; benzyl acetate
Simple heterocycles	Coumarin; hydroquinone; saffrole; sesamol; 8-methoxypsoralen
Polyphenols	Quercetin

(b)

Generic source	*Examples*
Fruit	Apple; apricot; cherry; grapefruit; lemon; melon; peach pear; pineapple
Root vegetables	Carrot; onion; parsnip; radish; turnip
Brassica	Broccoli; Brussel sprout; cabbage
Herbs	Coriander; dill; fennel; mint; sage; tarragon
Spices	Allspice; carroway; cardamom; nutmeg; paprika; turmeric

One of the first classes of toxic compounds in plants to be identified were the pyrrolizidine alkaloids from the genus *Senecio*. Subsequently, more than 200 related compounds have been isolated from numerous families and species, many of which are potent liver toxins and liver carcinogens. Other classes of alkaloids found in the plant kingdom include derivatives of the nicotine alkaloids, such as *N*-nitrosonornicotine, which are present in tobacco leaves and are known to be carcinogenic to animals. Tobacco leaves also contain a range of compounds that have been shown to potentiate the carcinogenic effect of the alkaloids present.

Many other classes of carcinogenic plant products have been identified. These include glycosides of azoxy alcohols such as cycasin from betel nuts, a colon carcinogen; isoprene glycosides such as ptaquiloside found in bracken, a liver carcinogen; and phenolic alkylbenzenes such as safrole present in many herbs and vegetables, which are also principally liver carcinogens. Other phenolic compounds including flavanoids such as quercetin, rutin and kaemferol, and tannins such as trapain and brevifolin are potent mutagens but evidence for their carcinogenicity is lacking. In fact many of these compounds have been shown to exert anticarcinogenic effects.

Organic Chemical Carcinogens in other Edible Plants and in Microorganisms

Chemical carcinogens are also found in a wide range of lower plants, such as fungi, and in microorganisms. Simple and complex hydrazines are found in many species of mushroom and have been shown to produce tumors in many tissues of experimental animals. Mycotoxins such as aflatoxin B_1 and the related polynuclear compound produced by *Aspergillus* species are some of the most potent carcinogens known, being active at dose levels in the nanogram per kilogram range. Human exposure to such compounds occurs when cereal crops and nuts are stored in humid conditions, as they are in many parts of equatorial Africa and China. Aflatoxin B_1 is one of the few established human carcinogens found in the plant kingdom. Other carcinogenic compounds produced as natural products include the antibiotics adriamycin and daunomycin and the antineoplastic agent streptozotocin isolated from microorganisms of the genus *Streptomyces*.

Carcinogens Produced by Food Processing

Despite the widespread occurrence of potentially carcinogenic chemicals in the plant kingdom, most foodstuffs contain only low levels of these chemicals. However, it has now been recognized that a number of processes used in food preparation/processing can introduce significant amounts of carcinogens into the food or the local environment. The most widely studied of these processes are preservation of meats and fish by salting or smoking; grilling or broiling of meats, and cooking in vegetable oils.

Traditional methods for preserving meat and fish involve either salting or smoking. Epidemiological evidence has been found for an association between an increased incidence of cancer of the mouth and pharynx and intake of salted meat. It seems likely that a reaction between sodium nitrate and/or nitrite used for preserving the meat and alkylamides present in the meat results in the formation of *N*-nitrosamines and nitrosamides. These compounds have been shown to be potent carcinogens in animal experiments to the mouth, pharynx and other sites. Levels of nitrosamines in cured meats and fish can be as high as 100–200 ppb (parts per billion) for the simple alkylnitrosamines and between 10 and 100 ppb for volatile heterocyclic nitrosamines. Although dose levels required to induce tumor formation in animal studies are substantially higher than those likely to be ingested by man, there is a concern that the presence of nitrosamines in food presents a significant hazard to man.

Preservation of meats and fish by smoking has also been shown to introduce chemicals known to be carcinogenic to animals, particularly polycyclic aromatic hydrocarbons (PAHs), although direct evidence for an association between an increased incidence of human cancers and consumption of smoked meat and fish is lacking.

The frying or grilling of meats and fish has been found to generate significant quantities of heterocyclic nitrogenous compounds derived from amino acids present in foods. These so-called cooked food mutagens include 2-amino-3-methylimidazo[4,5-f]quinoline (IQ), 2-amino-3,8-dimethylimidazo[4,5-f]quinoxaline (methyl-IQ_x), 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine (PhIP), 3-amino-1,4-dimethyl-5H-pyrido[4,3-b]indole (Trp-P-1), and 2-aminodipyrido[1,2-a:3′,2-d]imidazole (Glu-P-2). They are some of the most potent bacterial mutagens known and have been shown to induce a wide range of tumors in animals. Levels as high as 500 ppb have been found in grilled chicken and it has been suggested that they may be implicated in the induction of colon and breast cancer in humans. PAHs can also be generated by the grilling of meat and fish and both carcinogenic and noncarcinogenic compounds have been

identified. Levels of one particular PAH in foods, the carcinogen benzo(a)pyrene, have been reported to vary from <1 ppb in grain to more than 30 ppb in singed meat.

Cooking foods in hot oils has also been found to generate a range of carcinogenic chemicals. Many of these are volatile and may therefore represent more of a hazard to the cook than to the food consumer. Thus, cooking with unrefined rapeseed or soya bean oil, which contain significant levels of the polyunsaturated fatty acid linolenic acid, has been shown to result in the release of aldehydes including formaldehyde, acetaldehyde and acrolein, hydrocarbons including 1,3-butadiene and benzene, and other chemicals. Many of these compounds are mutagenic to bacteria and carcinogenic in animals, and in areas of the world where such cooking practices are common (e.g., China), the incidence of lung cancer in the exposed population is high.

Mechanisms of Carcinogenicity

Chemical carcinogens induce neoplasia by a wide range of mechanisms involving either interaction with the hereditary material of the organism or interference with one of the many cellular control systems. The former compounds, known as genotoxic carcinogens, interact directly with DNA, resulting in a permanent heritable change to a cell following replication (i.e., an altered genotype). In contrast, nongenotoxic (epigenetic) carcinogens do not interact directly with DNA but cause cancer by other mechanisms.

Chemicals that react with DNA are invariably electrophiles (i.e., they possess one or more electron-deficient centers in the molecule) that target the nucleophilic (electron-rich) sites in the DNA. The electrophilic center may be present in the molecule itself (activation independent) as in β-propiolactone, dimethyl sulfate, and a,β-unsaturated aldehydes or be generated following metabolism (activation dependent) in the target species.

Examples of classes of compounds that are converted to reactive electrophiles by oxidative metabolism include nitrosamines, chlorinated alkanes, hydrazines, and polycyclic aromatic hydrocarbons. Because of the inherent reactivity of these species, they react not only with DNA, but also with other cellular macromolecules such as RNA and proteins. These reactions protect the cell against the carcinogenicity of the chemical by reducing the amount of electrophile available to react with DNA, but may lead to other forms of damage and ultimately cell death.

The enzyme system considered to be mainly involved in the activation of chemicals to carcinogenic species is the so-called mixed function oxidase system. This enzyme complex is centered on cytochrome P450 and is present in most, if not all, of the organs of the body. The enzyme system consists of a very large family of related isoenzymes of differing substrate specificity and has a widespread distribution in the animal kingdom. Early work with this enzyme system suggested that only certain isoenzymes were responsible for the activation of carcinogens, although it is now clear that different isoenzymes may activate the same compound in different species.

Most chemical carcinogens appear to be substrates of one particular isoenzyme called CYP1A1. Molecular modeling has shown that only relatively flat (planar) molecules are oxygenated by this cytochrome. Common carcinogens activated by this isoenzyme include PAHs, aflatoxin, and 9-hydroxyellipticine, whereas the related isoenzyme CYP1A2 activates arylamines and amides such as 2-acetylaminofluorene and the cooked food mutagens. Other subfamilies of cytochromes involved in activating carcinogens include CYP2E1, which is known to act on a wide range of small molecules such as dialkylnitrosamines, urethane, vinyl monomers and haloalkanes, and CYP3A, which also activates PAHs, aflatoxins, and cooked food mutagens.

The chemistry of the activation process varies with the type of carcinogen. The oxidation of aflatoxin B_1, for example, results in the formation of the 8,9-epoxide in a single step whereas the activation of PAHs, such as benzo(a)pyrene, is a multistep process involving an epoxide that is converted to a diol by epoxide hydrolase, which is then converted to the proximate carcinogenic species, a diol-epoxide. Activation of arylamines and amides to DNA reactive species, in contrast, frequently involves an initial oxidation step to an *N*-hydroxy derivative, which is then further metabolized to a highly reactive N-O-ester. This latter reaction is catalyzed by a transferase enzyme, usually sulfotransferase or acetyltransferase for arylamines and glucuronotransferase for arylamides. Other oxidative reactions result in the formation of unstable compounds that decompose spontaneously to the ultimate carcinogenic species. Thus, simple nitrosamines are oxidized by CYP2E1 to an α-hydroxy intermediate, which breaks down to the electrophilic alkyldiazonium ion.

Enzyme systems other than the mixed function oxidase system may also be involved in the metabolic activation of carcinogens. Thus, for aflatoxins,

there is evidence that prostaglandin H synthetase can activate this group of compounds and for arylamines, oxidation may be carried out by prostaglandin peroxidase, myeloperoxidase, or by flavin-containing monooxygenases.

The direct metabolic activation of compounds to carcinogenic species by phase II metabolism, a process normally associated with detoxification, can also occur. Thus safrole and related compounds are converted to their sulfate esters, the ultimate carcinogenic species by the phase II enzyme, sulfotransferase.

Metabolic Activation of Epigenetic Carcinogens

Since there is no common mechanism describing the action of epigenetic carcinogens, generalizations concerning the effect of metabolism on the activity of chemicals acting by a nongenotoxic mechanism are not possible. The activity of a number of epigenetic carcinogens is reduced as a result of metabolic activation, although in the case of one group of epigenetic carcinogens that produce renal tumors in the rat by binding to and preventing the degradation of a specific kidney protein, alpha-2-microglobulin, metabolic activation is required for carcinogenic activity. Compounds acting by this mechanism include isophorone and D-limonene, which are present naturally in many fruits.

Similarly, a wide range of structurally diverse chemicals induce liver tumors in rodents due to their ability to induce the proliferation of hepatic peroxisomes. Food contaminants such as phthalate diesters, which leach out of packaging materials, fall into this category, although no naturally occurring food chemical has yet been found to be a peroxisome proliferator. Some examples of nongenotoxic mechanisms of carcinogenesis are shown in **Table 3**.

Table 3 Some examples of nongenotoxic mechanisms of carcinogenesis

Mechanism	*Examples of chemical classes*
Promotion	Phorbol esters; barbiturates; chlorinated hydrocarbons
Peroxisome proliferation	Phthalate diesters; hypolipidemic drugs; chlorinated herbicides
Endocrine modulation	Androgens and estrogens; antithyroid agents
Cytotoxicity	Metal chelators; branched chain hydrocarbons

Carcinogenicity Tests

Animal Bioassays

As the mechanism of carcinogenesis in both humans and animals is not well understood, the only acceptable procedure for determining whether a chemical is likely to be a carcinogen is the examination of experimental animals exposed to the suspect material under carefully controlled conditions. This procedure relies on the assumption that animals will behave in essentially the same way as humans to carcinogen exposure, i.e., the mechanism of tumor induction will be similar in both animals and humans. Mechanistically based, short-term tests for carcinogenicity prediction not involving experimental animals are still a distant and elusive goal.

The basic approach for carcinogenicity testing involves administering the test material to two suitable animal species for a considerable proportion of their natural lifespan. Because of their small size and relatively short life expectancy, the rat and mouse are the species of choice, although the hamster is occasionally used. In the US, inbred strains of animals are widely used (the F344 rat and the $B6C3F_1$ hybrid mouse), although out-bred strains are commonly used in Europe. To examine the carcinogenic potential of food components, the test substance is usually given in the diet, although in some circumstances administration may be in the drinking water or by gavage. The study continues until a certain proportion in one or other of the treatment groups has died or has been killed in a moribund state. As a minimum, 50 animals are allocated at random to each of the experimental groups, allowing a statistically significant carcinogenic effect to be detected if five animals in a test group develop tumors and no animals in the control group do.

During the study, the animal's clinical state is regularly monitored and at the end of the study a complete necropsy is performed on all surviving animals. Any tumors found are classified as either neoplastic or non-neoplastic and some attempt is made to determine whether any tumors seen were the cause of the (early) death of the animal (fatal tumors) or were unrelated to the death (incidental tumors). The procedures of these bioassays are conducted under rigorous conditions defined by the Code of Good Laboratory Practice (GLP).

Tests are essentially of two types: the first, used widely under the National Toxicology Program (NTP) in the US, is designed to examine the ability of the test material to induce cancer in the species used; the second, is aimed at determining the cancer incidence in respect of dose – a classical dose–response study. The former requires a few treatment groups,

including a relatively high-dose group in order to maximize the chance of detecting a carcinogenic effect, whereas the latter requires a wide range of dose groups to define accurately the dose–response relationship.

The analysis of a carcinogenicity bioassay is aimed at determining whether the administration of the test chemical has resulted in an increase in the incidence of tumors at one or more sites. In order to accomplish this analysis, two major confounding factors may have to be taken into consideration. The first is the effect of differences in mortality rates between the control and treated groups and the second is the effect of differences in food intake and its consequence on body weight. Both factors can substantially alter the tumor pattern observed in different groups. Early deaths may prevent the animals reaching tumor-bearing age, and reduced food intake and the associated reduction in body weight may result in a considerable reduction in tumor incidence.

The interpretation of the results of a bioassay are complex but most authorities work to the 'weight of evidence' principle. This evidence is taken in the light of the 'adequacy' of the bioassay, which is dependent on some of the factors previously discussed. Strong evidence for the compound being a genotoxic carcinogen would be increased malignant tumor incidence in two species, with tumors at multiple sites showing a clear dose–response relationship. Rare or unusual tumors at a site would be given added weight. Equivocal evidence may result from a statistically marginal result or only an increase in commonly occurring benign tumors. Tumor development in only one species and in association with species-specific toxicity is characteristic of nongenotoxic (or epigenetic) carcinogens. Sometimes, problems associated with such findings may be clarified by further mechanistic studies or by reference to historical data. When the data from bioassays are considered in human risk assessment, other factors must clearly also be taken into consideration. These may include evidence of genotoxicity in short-term tests and data on metabolism and potential human exposure. Furthermore, a measure of risk at doses substantially below the bioassay dose may be needed. This may require an extrapolation using mathematical models. As yet no general agreement has been reached as to the most appropriate method, and so the calculated risk given by different methods may vary considerably. Thus, the final assessment may be made on quite pragmatic grounds, in which the experience and expertise of a number of individuals are drawn on to reach a consensus opinion.

Short-Term Predictive Tests

A large number of short-term tests have been developed in an attempt to predict carcinogenic potential and thereby reduce the reliance on animal tests. These include assays for detecting gene mutation, damage to chromosomes, or damage to the whole genome.

Gene mutation can be assessed in bacteria, yeasts, or mammalian cells in culture (see **Table 4**). Since many of the cell systems used are unable to activate metabolically the majority of test chemicals, an exogenous mammalian metabolizing system, the so-called S-9 mix, is incorporated into the assay. Chromosome damage can be measured in cell lines *in vitro* or by using animals exposed for a short time to the chemical. Structural damage produced can include chromosome and chromatid gaps and breaks, rings, fragments, dicentrics, translocations, and inversions. A short-term *in vivo* assay measuring

Table 4 Short-term test systems for predicting carcinogenic potential

Test system	*Cell used*	*End point*
Bacterial mutation	*Salmonella typhimurium* TA strains *Escherichia coli* WP2	Reversion to histidine independence
Mammalian gene mutation	Chinese hamster lung (V79) Chinese hamster ovary (CHO) Mouse lymphoma (L5178Y) Human transformed lymphoblastoid (TK6)	Loss of HPRT, TK, or Na^+/K^+ ATPase expression
Chromosome aberration *in vitro*	Chinese hamster fibroblast (CHL) Chinese hamster ovary (CHO) Human peripheral blood lymphocytes(PBL)	Chromosome/chromatid aberration (gaps, breaks, deletions)
Chromosome damage *in vivo*	Bone marrow erythrocytes (mouse)	Micronuclei induction
Heritable damage *in vivo*	Rodent germ cells	Dominant/lethal mutations; heritable translocations, etc.

HPRT, hypoxanthine phosphoribosyl transferase; TK, thymidine kinase.

unscheduled DNA synthesis (UDS) in rat liver or gut is recommended by most regulatory authorities if there is a positive response in any *in vitro* assay and a negative response in an *in vivo* cytogenetics assay. Other test methods and end points are under consideration by regulatory authorities as indicators of genotoxic potential including the COMET assay for assessing DNA damage, and aneuploidy, the change in chromosome number resulting from damage to the cellular architecture (spindle) controlling chromosome replication.

The last two decades have seen extensive efforts to determine whether short-term tests are suitable for predicting carcinogenic potential. The early validation studies suggested good predictability, with correct identification of over 90% of carcinogens (high sensitivity) and over 90% of noncarcinogens (high specificity). In later evaluations, a much lower figure (60%) was obtained. However, when carcinogens known to react by nongenotoxic mechanisms (e.g., hormones or peroxisome proliferators) were excluded, the predictability was improved suggesting that short-term tests are suitable for detecting those carcinogens that act by a genotoxic mechanism.

Although many regulatory authorities have guidelines for carcinogenicity evaluation, which include short-term tests, they all still require animal studies as the ultimate test for carcinogenicity. However, the use made of short-term tests varies. In the US, the Food and Drugs Administration (FDA) recommends a battery of short-term tests for all 'additives' for which cumulative dietary intake is expected to exceed 1.5 μg per person per day in order to assist in the interpretation of animal feeding studies. Some expert bodies, such as The International Agency for Research in Cancer, use short-term tests as an adjunct to animal carcinogenicity studies in their evaluation process, giving added weighting in their assessment of likely human hazard to an animal carcinogen that is also positive in short-term tests.

However, until a consensus can be reached as to what a positive or negative result in an animal feeding study means in terms of whether the compound may or may not be a human carcinogen, the further development of better (faster/cheaper) short-term tests may be a futile exercise.

Monitoring and Control of Hazards

The complex mixture of chemicals that constitute food, together with the uncertainty of the specific role of the various components in the diet, has made the control of potential carcinogens in food difficult. In particular, the realization that animal carcinogens, as identified by standard animal bioassays, are widely distributed in the general environment, including food, has made control by total elimination impossible.

Control of toxic agents in food particularly contaminants and additives has been achieved by examining their hazard in animal studies. Thus, the establishment of a no-observable-adverse-effect-levels (NOAEL) is followed by the setting of an allowable daily intake (ADI) through extrapolation based on the relative sensitivity of animals and humans to toxic events. This extrapolation may also take into consideration other properties of the chemical concerned, such as genotoxic potential. For genotoxic carcinogens, however, it is generally considered that there is no no-effect-level and therefore acceptable intake is based on estimation of likely risk. A maximum risk of 10^{-6} cancers in a lifetime is considered as an acceptable risk by some authorities, particularly those in the US, and acceptable exposure estimates extrapolated from animal data. For nongenotoxic carcinogens (and for some genotoxic carcinogens, particularly those that act as aneugens), no-effect levels are accepted, since the carcinogenic response is the result of a prior toxic event for which a no-effect-level can be determined. For additional safety, an arbitrary factor of 100 was applied to the NOAEL, to allow for interspecies variation ($\times 10$) and interindividual variability ($\times 10$). More recently, the two factors have been subdivided into variable factors (pharmacokinetic and pharmacodynamic) to reflect increased understanding of the mechanisms underlying the development of cancer and allow for factors associated with special groups such as infants and children. It must be said that the scientific basis to support either of these approaches (acceptable risk or no-effect-levels) is quite limited as even for the best-documented cases, the mechanism of the carcinogenic effect is poorly understood.

The unequivocal identification of human carcinogens is difficult since direct experimental approaches are precluded. Thus, epidemiological data involving both prospective and retrospective studies, and using case controls in certain investigations, has to be employed. These techniques have limited applications to diet-associated carcinogenesis and have proved most useful in identifying specific carcinogens in the work place or those used as therapeutic agents. The specific problem in identifying dietary carcinogens relate to the complexity of diet, the difficulty in identifying specific components, and the sensitivity of the epidemiological methods themselves. It would seem likely that epidemiological data will only be able to link specific chemical

carcinogens in food with a carcinogenic effect in a few favorable circumstances, since such chemicals are likely to be present at low levels and induce only a small increase in tumor incidence over background levels. One such example was the identification of a carcinogenic hydrazone in the mushroom *Gyromitra esculenta,* as a result of an epidemiological study in Finland. Such methods have also indicated the relative importance of 'life style' factors in carcinogenesis: in particular, associations have been made between lack of dietary fiber and colon cancer, between a low intake of fresh fruit and vegetables and stomach cancer, and between excess dietary fat and colon and breast cancer, although the specific chemicals responsible have not been identified with any certainty.

Most of the activity aimed at controlling carcinogens in food has been directed at preventing addition of potentially carcinogenic substances to the existing background level of natural carcinogens. This has been tackled through the application of laws governing the adulteration of food, the first of which were enacted in the mid-nineteenth century in the UK. The current UK legislation is the 1990 Food Safety Act, governing the nature and quality of food and its nutritive value. This Act, like its forerunner, the 1955 Food and Drug Act, requires that the constituents of food should not be injurious to health. Thus, while there is no specific requirement for carcinogenicity testing in the current act, consideration is given to all available data, including the result of mutagenicity tests and long-term tests in animals.

The position in the US up to 1958 was similar to that in the UK. Food was considered adulterated if injury could arise from its use. Legislation was based on traditional food, added substances, and unavoidable added substances (contaminants). For added substances, listed in an inventory of over 3000 chemicals and often referred to as 'Everything Added to Food in the United States' (EAFUS), the food was considered adulterated if the added substance could render the food injurious to health; for unavoidably added substances, a balance was applied between the essential nature of the food material and the degree of contamination. These strictures applied to both carcinogenic and noncarcinogenic toxicants. In 1958 a change in emphasis was introduced through the Food Additives Amendment. This established a licensing scheme for substances deliberately added to foods or for substances that could migrate into food, but excluded materials that were generally, through usage, regarded as safe (GRAS). For licensing purposes, the material has to be shown to be 'safe' for its intended use, although in theory at least the GRAS substance could be a carcinogen.

In 1958 the Delaney Clause was enacted; this required that if there was evidence of carcinogenicity in any test system, the material should be prohibited from food usage. Improved analytical techniques have shown that many foods contain both unintentionally added and natural carcinogens, such as polynuclear aromatic hydrocarbons, nitrosamines, mycotoxins and arylamines, and no form of regulation could control these materials. Furthermore, bulk components of food may themselves play an important role in the development of carcinogenesis. The recognition that the exclusion of all potentially carcinogenic additives (under the Delaney Clause) is a practical impossibility has given way to the concept of 'safe' tolerance, and that 'safe levels' may be set by appropriate, conservative risk assessment in which an 'insignificant life time risk' of developing tumors of, for example, 10^{-6} is considered acceptable.

See also: **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Fish**. **Food Safety**: Mycotoxins. **Meat, Poultry and Meat Products**. **Sodium**: Salt Intake and Health.

Further Reading

Arcos JC, Woo YT, Argus MF, and Lai D (1968–1985) *Chemical Induction of Cancer*, vols 1, 2A, 2B, 3A, 3B. New York: Academic Press.

Anderson D and Conning DM (1993) *Experimental Toxicology, The Basic Issues*, 2nd edn. London: Royal Society of Chemistry.

Ashby J and Tennant RW (1991) Definitive relationships among chemical structures, carcinogenicity and mutagenicity for 301 chemicals tested by the US NCI/NTP. *Mutation Research* **257**: 229–306.

Committee on Mutagenicity of Chemicals in Food, Consumer Products and the Environment (COM) (2000) *Guidance on a Strategy for Testing Chemicals for Mutagenicity*. London, UK: Department of Health.

Conning DM and Lansdown ABG (1983) *Toxic Hazards in Food*. London, UK: Croom Helm Ltd.

Hirono I (1987) *Naturally Occurring Carcinogens of Plant Origin. Toxicology, Pathology and Biochemistry* B.V. Amsterdam: Elsevier Science Publishers.

International Agency for Research on Cancer (1994) In: Hemminki K, Dipple A, Shuker DEG, Kadlubar FF, Segerback D, and Bartsch H (eds.) *DNA Adducts; Identification and Biological Significance*, IARC Scientific Publication No. 125. Lyon: IARC.

International Agency for Research on Cancer (1972–2002) *Evaluation of the Carcinogenic Risk of Chemicals to Humans*, Monograph Series No. 1–82. Lyon: IARC.

Lewis DFW, Bird MG, and Jacobs MN (2002) Human carcinogens: an evaluation study via the COMPACT and HazardExpert procedures. *Human and Experimental Toxicology* **21**: 115–122.

Nagao M and Sugimura T (1999) *Food Borne Carcinogens: Heterocyclic Amines*. Chichester, UK: Wiley.

Renwick AG (2000) The use of safety or uncertainty factors in the setting of acute reference doses. *Food Additives and Contaminants* **17**: 627–635.

Scheuplein RJ (1990) Perspectives on toxicological risk – an example; food borne carcinogenic risk. In: Clayson DB, Munroe IC, Shubik P, and Swenberg JA (eds.) Progress in Predictive Toxicology, pp. 351–371. Amsterdam: Elsevier Science Publishers.

Williams GM and Weisburger JH (1991) Chemical carcinogenesis. In: Amdur MO, Doull J, and Klassen CD (eds.) *Casarett and Doull's Toxicology. The Basic Science of Poisons*, 4th edn., pp. 127–200. New York: Pergamon Press.

World Health Organisation (1999) Principles for the assessment of risks to human health of exposure to chemicals. *Environmental Health Perspectives*, No. 210.

CARBOHYDRATES

Contents

Chemistry and Classification

C L Stylianopoulos, Johns Hopkins University, Baltimore, MD, USA

Introduction

Carbohydrates are the most abundant constituents of cereals, fruits, vegetables, and legumes. They are the major energy source in human nutrition and contribute to the texture and flavor of processed foods. They comprise a group of substances with different structures and varying physical, chemical, and physiological properties. Carbohydrates are polyhydroxy aldehyde or ketone molecules and their derivatives with the general formula $(CH_2O)_n$. Dietary carbohydrates are important in maintaining glycaemic homeostasis and gastrointestinal health. Furthermore, they contain necessary micronutrients, phytochemicals, and antioxidants.

Classification and Chemical Structure

Carbohydrates are classified into four categories according to their chemical structure and degree of polymerization: monosaccharides, disaccharides, oligosaccharides, and polysaccharides.

Monosaccharides

Monosaccharides are the simplest form of carbohydrate and cannot be further hydrolyzed into smaller subunits. According to their chain length, the monosaccharides can be divided into several categories, the more nutritionally important being the pentoses, with skeletons containing five carbon atoms (e.g., ribose), and the hexoses, with skeletons containing six carbon atoms (e.g., glucose).

The presence of asymmetrical carbon atoms in monosaccharides with different functional groups attached gives rise to optical activity, meaning that if polarized light is passed through a solution of these compounds, the plane of light will be rotated to the left (levorotatory or L form) or to the right (dextrorotatory or D form). As a result, mirror-imaged structures of the same compound exist and are called stereoisomers. Monosaccharides of the D form are more nutritionally important because the metabolic and digestive enzymes are specific for the D stereoisomers.

Monosaccharides demonstrate another type of stereoisomerism due to their formation of cyclic structures. The pentoses form furanose (five-membered rings), and the hexoses form pyranose (six-membered rings). Cyclization can produce two stereoisomers (the α and β configurations), and generally monosaccharide solutions contain an equilibrium mixture of these two forms. **Figure 1** illustrates D-glucose in its pyranose form in the α and β configurations. The isomerization produces compounds with different properties and has major metabolic importance, because of enzyme specificity for particular stereoisomers.

Glucose is the most abundant monosaccharide, and is a major cell fuel in the human body, and can be found unbound in body tissues and fluids. Glucose is the building block of several polysaccharides. Galactose and fructose are also used as cell

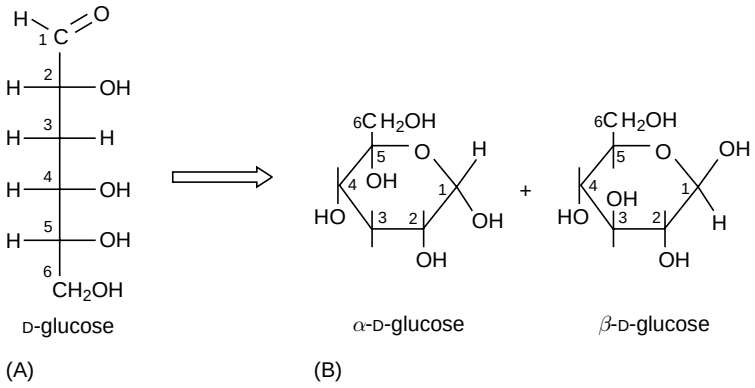


Figure 1 D-glucose molecule shown as (A) open chain and (B) a cyclic pyranose ring in the α and β configurations.

fuel. The most important monosaccharides and their significance are outlined in **Table 1**.

Several monosaccharide derivatives are constituents of polysaccharides, as well as food ingredients. Some nutritionally important monosaccharide derivatives and their significance are outlined in **Table 1**.

Table 1 Some nutritionally important monosaccharides

Class	*Species*	*Significance*
Hexoses	D-glucose	Major cell fuel, unbound in body fluids and tissues, building block of several polysaccharides
	D-fructose	Cell fuel, constituent of sucrose
	D-galactose	Cell fuel, constituent of lactose
	D-mannose	Constituent of plant cell wall polysaccharides and gums
Pentoses	L-arabinose, D-xylose	Constituent of plant cell wall polysaccharides
	D-ribulose, D-xylulose	Metabolite in pentose pathway
	D-ribose	RNA constituent
Uronic acids	D-glucuronic, D-galacturonic	Constituent of plant cell wall polysaccharides
	D-mannuronic, D-guluronic	Constituent of algal polysaccharides
Sugar alcohols	D-glucitol, D-xylitol	Food ingredient
	D-galactitol	Metabolite of galactose
Deoxysugars	D-deoxyribose	DNA constituent
	D-deoxygalactose	Constituent of algal polysaccharides
	L-fucose	Constituent of bacterial polysaccharides
	L-rhamnose	Constituent of pectic plant polysaccharides
Aminosugars	D-glucosamine, D-galactosamine	Constituent of aminosaminoglycans, cartilage

Disaccharides

Disaccharides consist of two monosaccharide units, linked by glycosidic bonds in the α or β orientation. The most important disaccharides are sucrose, lactose, and maltose. Sucrose consists of a molecule of α-glucose and a molecule of β-fructose linked together (**Figure 2A**). Lactose is found in milk and dairy products and consists of a molecule of galactose linked to a glucose molecule by a β-1,4glycosidic bond

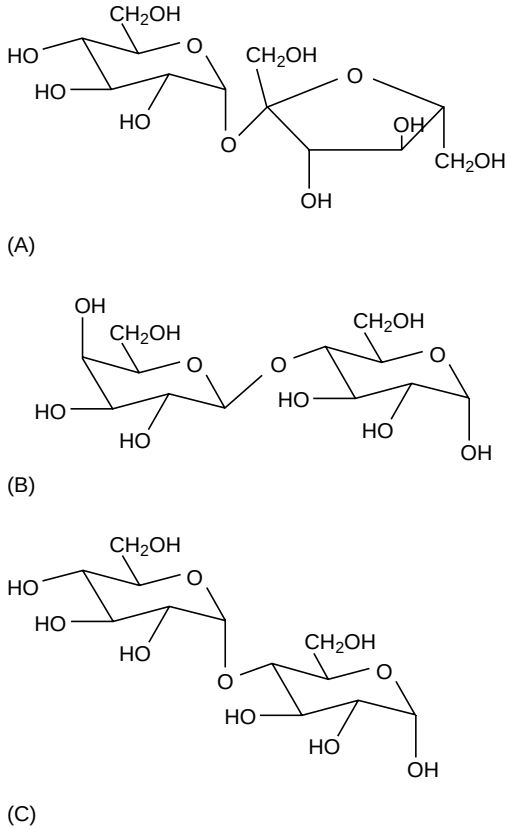


Figure 2 The molecular structures of (A) sucrose, (B) lactose, and (C) maltose.

Table 2 Some nutritionally important disaccharides

Class	*Species*	*Significance*
Disaccharide	Sucrose	Constituent of fruits, vegetables, and sweetener
	Lactose	Constituent of milk and dairy products
	Maltose, Isomaltose	Constituent of starch
	Trehalose	Food additive, constituent of mushrooms
	Lactulose	Lactose derivative, laxative
Disaccharide alcohols	Maltitol	Constituent of starch, sweetner
	Lactitol	Constituent of lactose, sweetener

(**Figure 2B**). Maltose is mainly produced by partial hydrolysis of starch and consists of two glucose units linked through an α-1,4glycosidic bond (**Figure 2C**). Some nutritionally important disaccharides and their significance are outlined in **Table 2**.

Oligosaccharides

Oligosaccharides consist of a chain of between three and nine monosaccharide units covalently linked to form large units and are named trioses, tetroses, etc, depending on the number of carbon atoms in their molecules. Oligosaccharides are distributed widely in plants and when digested yield their constituent monosaccharides. The major oligosaccharides are the raffinose series, formed by the linkage of galactose, sucrose, and glucose units, and the maltose series, formed by the linkage of glucose units. Some nutritionally important oligosaccharides and their significance are outlined in **Table 3**.

Polysaccharides

Polysaccharides consist of long chains of monosaccharide residues (more than nine) linked with glycosidic bonds. Polysaccharides found in nature usually have high molecular weights and are named after the component monosaccharide. These compounds consist of several hundred or even thousands of monosaccharide units. The properties of polysaccharides are determined by the species of monosaccharides in the polymer backbone, the types of linkages between residues, and the extent and type of chain branching.

Table 3 Some nutritionally important oligosaccharides

Class	*Species*	*Significance*
Maltoses	Maltotriose, Maltotetraose	Constituent of starch
Raffinoses	Raffinose, Stachyose, Verbascose	Constituent of vegetables and legumes
Fructoses	Fructotriose	Constituent of cereals and tubers
Lactoses	Fucosyl lactoses	Constituent of human breast milk

Glucans are polymers of glucose and the major polysaccharides in the diet. The most important glucans are starch, glycogen, and cellulose. Glycogen is the short-term storage form of glucose in animal tissues. Starch is the most common digestible storage polysaccharide in plants, and cellulose is a major structural component of plant cell walls (**Figure 3**). Some nutritionally important polysaccharides and their significance are outlined in **Table 4**.

Polysaccharides with α linkages have a helical shape (e.g., the amylose starch molecule), while those with β linkages generally have a linear or flat ribbon-like molecule (e.g., the cellulose molecule) (**Figure 3**).

Polysaccharide molecules can be linear or branched. Branches can be formed through any unlinked hydroxyl group and vary from alternating and consecutive single-unit branches to multiple-unit branches (ramified structure). Polysaccharides that are highly branched tend to be soluble in water, because the chain structure prevents hydrogen bonding. Linear polysaccharides tend to be insoluble in water, unless they possess structural irregularities.

Nutritional Importance

Some carbohydrate types of specific importance in human nutrition are sugars and sugar alcohols, starch, and dietary fiber.

Sugars

The Food and Agriculture Organization (FAO) and the World Health Organization (WHO) expert consultation on carbohydrates use the term 'sugar' to describe monosaccharides and disaccharides. Sugars can be separated analytically from the food matrix by gas–liquid chromatography (GLC), high performance liquid chromatography, and enzymatic methods. Sugars are widely used in the food industry as sweeteners and preservatives. They improve the texture, body, palatability, and viscosity of foods and beverages.

The UK Department of Health distinguishes between 'intrinsic' and 'extrinsic' sugars. Intrinsic sugars are defined as those that occur naturally as part of the plant cell walls. Extrinsic sugars were defined as added sugars or those present when the

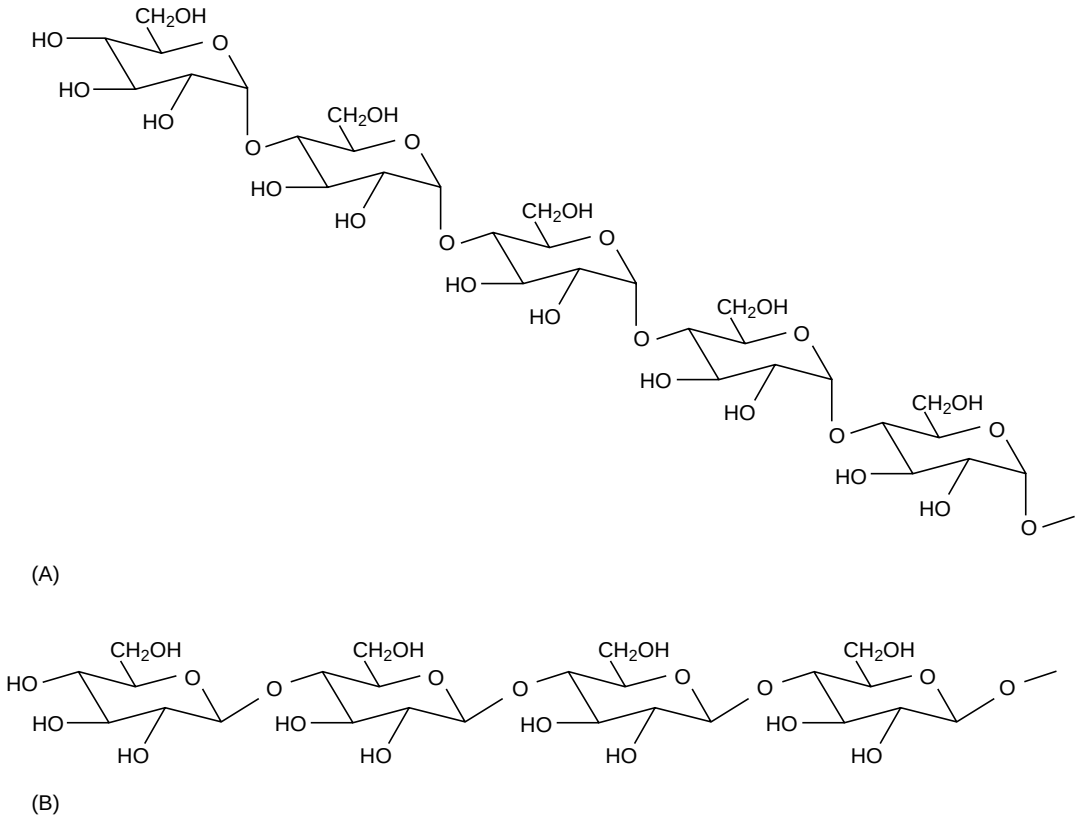


Figure 3 (A) Five units of an α-1,4-D-glucopyranose chain from a starch molecule (amylose). (B) Four units of a β-1,4-D-glucopyranose chain from a cellulose molecule.

food matrix has been disrupted. An additional term, 'non-milk' extrinsic sugar, is used to distinguish between milk and other extrinsic sugars. However, these terms have not been widely accepted.

Sugar Alcohols

Sugar alcohols are monosaccharide and disaccharide derivatives, such as sorbitol and xylitol, which are extensively used as sweeteners in the food industry. They have received increased attention because of their desirable properties of relative sweetness and limited digestion and absorption.

Starch

The most important, abundant, and digestible polysaccharide in human nutrition is starch. Starch comprises large chains of α-linked glucose residues, in the form of amylose or amylopectin. Amylose is a linear unbranched form of starch, which consists of α-1,4-linked glucose units (**Figure 3A**). Amylopectin is a branched-chain polymer, which consists of α-1,6-linked glucose units. Both forms of starch can be found in cereals, potatoes, legumes, and other vegetables, with amylopectin comprising 80–85% and amylose 15–20% of total starch.

Dietary Fiber

There are several definitions of fiber, and no consensus exists among international organizations. The Association of Official Analytical Chemists International defines fiber as nondigestible animal and plant carbohydrates, based on the analytical methods for fiber separation using an enzymatic–gravimetric method (**Table 5**).

According to the new definition of the expert panel on macronutrients appointed by the Institute of Medicine of the National Academies of Science (**Table 5**), dietary fiber comprises intact nondigestible carbohydrates and lignin derived from plant sources. Functional fiber consists of nondigestible carbohydrates, derived from either plant or animal sources, that have shown favorable health outcomes for humans. Total fiber consists of both dietary and functional fiber. With this new definition of dietary and functional

Table 4 Some nutritionally important polysaccharides

Class	*Species*	*Significance*
Glucans	Starch	Storage polysaccharide in plants
	Glycogen	Short-term storage form of glucose in animal tissues
	Cellulose	Major structural component of plant cell walls
Galactans		Major constituents of noncellulosic matrix of plant cell wall
Xylans		Constituents of mature plant tissues
Mannans		Storage forms in several plants
Uronans	Galacturonans	Major components of water-soluble pectic fraction of plants
	Mannuronans	Components of algal polysaccharides
	Guluronans	Components of algal polysaccharides
Starch	Amylose, amylopectin	Most common digestible plant polysaccharides
Nonstarch	Cellulose	Major component of plant cell wall
	Pectin	Constituent of plant cell wall, food additive
	Hemicellulose	Constituent of plant cell wall
	Gums, mucilages	Plant hydrocolloids, food additives
	Algal polysaccharides	Constituents of algae and seaweed, food additives

fiber, new analytical methods should be developed and implemented to quantify accurately the total fiber component of foods.

Dietary and functional fiber cannot be digested by mammalian enzymes, and therefore they pass almost intact through the small intestine. Fiber consumption has potential health benefits, including the promotion of general gastrointestinal health and the prevention of several noncommunicable diseases.

Table 5 Current definitions of dietary fiber

Source	*Definitions*
AOAC[a]	Fiber: nondigestible animal and plant carbohydrates
IOM[b]	Dietary fiber: intact nondigestible carbohydrates and lignin, derived from plant sources
	Functional fiber: nondigestible carbohydrates derived from either plant or animal sources that have shown favourable health outcomes for humans
	Total fiber: dietary and functional fiber

[a]AOAC: Association of Official Analytical Chemists
[b]IOM: Institute of Medicine

Chemistry

Monosaccharides share the same functional groups, but their isomeric forms often exhibit differences in chemical reactions. Disaccharides exhibit a similar range of reactions to monosaccharides owing to the presence of similar functional groups. Oligosaccharides generally exhibit properties similar to those of monosaccharides and disaccharides with similar functional groups, but some oligosaccharides with nine monosaccharide units may exhibit similar properties to polysaccharides. In general, polysaccharides show slower reaction rates because of steric effects.

Solubility

Monosaccharides, disaccharides, and oligosaccharides have similar solubilities. Overall, they are very soluble in water. Sucrose is extremely soluble in water, while lactose is soluble to a lesser extent. Furthermore, they are insoluble in nonpolar organic solvents. They exhibit limited solubility in pure alcohols but are very soluble in aqueous alcohol solutions (70–80% v/v), and therefore these solutions are widely used for extraction and analysis. Oligosaccharides are less soluble than monosaccharides in aqueous alcohol solutions, and their solubility decreases as the number of monosaccharide units increases.

In general, polysaccharides form colloidal solutions in water, while some other polymers are extremely insoluble in water and require prior treatment with acid, alkali, or organic solvents to get them to dissolve. For example, β-1,4-mannans and glucans (e.g., cellulose) are very insoluble owing to hydrogen bonding between parallel chains. On the other hand, arabinoxylans are readily soluble in water, because the arabinosyl chains inhibit hydrogen bonding. Galactomannans are also readily soluble in water, producing viscous solutions, and are used as food additive gums. The α-linked glucans (e.g., amylose and amylopectin) have completely different solubilities. The glucan α-1,4-amylose is very soluble in warm water and forms colloidal solutions. When the amylose chains cool down, they form an amylose gel, which subsequently forms an insoluble crystalline material. Amylopectins are also very soluble in hot water but do not form an insoluble crystalline material to the same degree as amylose.

Reducing Properties

Monosaccharides are powerful reducing agents to a range of metals in alkaline solution, owing to the presence of aldo and keto groups. The extent of

reduction varies among different monosaccharides. Disaccharides and oligosaccharides have the same reducing properties, except for sucrose, in which both hemi-acetal groups are combined. Polysaccharides usually contain one reducing group at the terminal end of the polymer chain and, as a result, have lower reducing properties.

Reactions in Acidic Solutions

When heated in strong acidic solutions, monosaccharides dehydrate and condense into a range of furans. The resulting furans condense with several reagents to generate colored products; hence the presence of monosaccharides and their derivatives can be verified. Under weaker acidic conditions, fructose is labile.

Reactions in Alkaline Solutions

In weak alkaline solutions, monosaccharides undergo isomerization of the aldose-keto group (enolization). In stronger alkaline solutions, they produce a series of degradation compounds, namely saccharinic acids. In the presence of ammonia, amino acids, and proteins, they condense repeatedly to generate a series of highly colored products (Maillard reaction); this reaction is used in the food industry to produce caramel colors.

Hydrolysis

Acid Disaccharides and oligosaccharides in mild acidic conditions are hydrolyzed to their constituent monosaccharides. The fructofuranosyl linkages of the fructooligosaccharides are quite susceptible to acid hydrolysis. Polysaccharides are also hydrolyzed to their constituent monosaccharides by acid hydrolysis, but the conditions necessary for complete hydrolysis depend on the solubilities of the polymers. The majority of polysaccharides (e.g., starch) are completely hydrolyzed under weak acidic conditions. However, cellulose requires treatment with strong acid for several hours prior to hydrolysis and subsequent heating under weak acidic conditions for the completion of the reaction. The uronans are very resistant to complete acid hydrolysis, and disaccharides of aldobiuronic acids are generally produced. Acid hydrolysis of polysaccharides results in extensive losses of their monosaccharide constituents.

Enzymatic Disaccharides are hydrolyzed in specific enzymatic solutions, and, therefore, this is a useful method for the analysis of sugar mixtures. Oligosaccharides are also susceptible to enzymatic hydrolysis. The maltooligosaccharides can be rapidly hydrolyzed by glucosidase enzymes.

Polysaccharides are more efficiently hydrolyzed to their monosaccharide constituents using specific enzymes. Fungal enzymes act specifically to hydrolyze different polysaccharides. The α-1,4 glycosidic linkages in starch can be hydrolyzed by various α amylases (e.g., salivary and pancreatic), producing maltose and isomaltose. The β-1,6 glycosidic linkages in amylopectin are not as easily hydrolyzed and require the presence of pullulanase – a fungal enzyme – to complete the hydrolysis.

Ester Formation

Monosaccharides contain hydroxyl groups and react with acids to form a variety of esters. The phosphate esters play a main role in carbohydrate metabolism. For example, the first step of glycolysis involves the production of the glucose-6-phosphate ester in a reaction catalyzed by the enzyme glucokinase in the presence of adenosine triphosphate. The uronic acids react with alcohols to form esters. The methyl esters of uronic acids are the most important in determining the physical properties of the uronans.

The presence of additional hydroxyl groups in disaccharides and oligosaccharides increases the number of sites for esterification reactions. Sucrose reacts with fatty acids to produce nondigestible esters, which have similar properties to the triacylglycerols.

The polysaccharide galacturonans, which are composed of an α-1,4 galacturonic acid chain with integrated rhamnose units, form salts with cations and may be esterified with methoxyl groups.

Substitution

Monosaccharides undergo substitution reactions with methyl iodide to produce methyl ether derivatives. These compounds have been used to identify the structure of polymers, because the sites of nonmethyl substituted groups are indicative of the branch points after hydrolysis. Monosaccharides undergo acetylation, which occurs on the free or the reduced molecule to produce acetylated alditols. These volatile compounds have been used to identify sugar mixtures by GLC. The presence of additional hydroxyl groups in disaccharides and oligosaccharides increases the number of sites for substitution reactions.

Abbreviations

D	dextrorotatory
FAO	Food and Agriculture Organization
GLC	gas–liquid chromatography
L	levorotatory
v/v	volume/volume
WHO	World Health Organization

See also: **Alcohol**: Absorption, Metabolism and Physiological Effects. **Carbohydrates**: Regulation of Metabolism; Requirements and Dietary Importance; Resistant Starch and Oligosaccharides. **Colon**: Structure and Function. **Dietary Fiber**: Physiological Effects and Effects on Absorption. **Fructose**. **Galactose**. **Glucose**: Chemistry and Dietary Sources. **Glycemic Index**. **Small Intestine**: Structure and Function. **Stomach**: Structure and Function. **Sucrose**: Nutritional Role, Absorption and Metabolism.

Further Reading

Brody T (1999) *Nutritional Biochemistry*, 2nd edn, pp. 1–56. San Diego: Academic Press.

Eastwood M (2003) *Principles of Human Nutrition*, 2nd edn, pp. 195–122, 418–426, 486–509. Oxford: Blackwell.

FAO/WHO (1998) Carbohydrates in human nutrition. Report of a Joint FAO/WHO Expert Consultation. FAO. *Food and Nutrition Paper* **66**: 1–140.

Gray GM (2000) Digestion and absorption of carbohydrate. In: In: *Biochemical and Physiological Aspects of Human Nutrition*, 1st edn, pp. 91–106. Philadelphia: WB Saunders Company.

Groff JL and Gropper SS (2000) *Advanced Nutrition and Human Metabolism*, 3rd edn, pp. 70–105. Belmont: Wadsworth.

Institute of Medicine of the National Academies (2002) *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, Amino Acids*. Washington, DC: The National Academies Press.

Lewis BA (2000) Structure and properties of carbohydrates. In: In: *Biochemical and Physiological Aspects of Human Nutrition*, 1st edn, pp. 3–22. Philadelphia: WB Saunders Company.

Nantel G (2003) Glycemic carbohydrate: an international approach. *Nutrition Reviews* **61**: S34–S39.

Sanchez-Castillo CP, Hudson GJ, Englyst HN, Dewey P, and James WPT (2002) The importance of dietary carbohydrates. *Archivos Latinoamericanos de Nutricion* **52**: 321–335.

Sullivan DM and Carpenter DM (eds.) (1993) *Methods of Analysis for Nutritional Labeling*. Arlington: Association of Official Analytical Chemists.

Williams SR and Schlenker ED (2003) *Essentials of Nutrition and Diet Therapy*, 8th edn, pp. 47–65. St Louis: Mosby.

Regulation of Metabolism

C L Stylianopoulos, Johns Hopkins University, Baltimore, MD, USA

Introduction

The three basic monosaccharides important in human nutrition are glucose, fructose, and galactose. Glucose is the product of the digestion of starch. In human metabolism, all simple sugars are converted into glucose. Glucose is the circulating form of carbohydrate in the bloodstream. Fructose is the sweetest of the simple sugars, and it is found in fruits and naturally occurring substances such as honey. Fructose consumption has increased greatly in the USA since the 1970s, when high-fructose corn syrup started to be widely used in food processing. High-fructose corn syrup is the major sweetening agent used by the food industry. Galactose is produced by the digestion of lactose, the major carbohydrate in milk.

Digestion

Most carbohydrates have to be converted to glucose in order to be used for energy production. The digestion of carbohydrates starts in the mouth, with mastication and the enzymatic action of salivary amylase, which converts starch to dextrins and maltose. Successive contractions of the stomach (peristalsis) move the food to the lower part of the stomach, while 20–30% of the carbohydrate is already converted to maltose. Peristalsis facilitates digestion in the small intestine, while the chemical digestion of carbohydrates is completed by pancreatic amylase (which continues the breakdown of starch to maltose) and intestinal disaccharidases (sucrase, lactase, and maltase for the breakdown of fructose, lactose, and maltose, respectively). The monosaccharide products of carbohydrate digestion are then absorbed into the portal circulation.

Absorption

Glucose accounts for the largest quantity of absorbed carbohydrate (80%), and galactose and fructose account for only a small amount (20%). The body quickly absorbs and transports the simple sugars, which enter the portal circulation via the capillaries of the intestinal villi and are transported to the liver. In the liver, fructose and galactose are converted to glucose, which is either used immediately for energy or stored in the form of glycogen. The liver can store approximately 5% of its mass in the form of glycogen, which can be readily converted to glucose for the production of energy.

Transport

Monosaccharides traverse the epithelial lining of the intestine by simple or facilitated diffusion or by active transport. The transport system for the passage of glucose and galactose through the apical membrane of the intestinal villi is called the Na^+-dependent glucose transporter (Na^+-dependent GLUT). Fructose uses a different transporter, called GLUT5, for the same passage. All monosaccharides are then transported from the enterocyte to the bloodstream by

another sugar transporter known as GLUT2. The passage of glucose and galactose across both membranes of the intestine requires the presence of Na^+, while the passage of fructose is dependent on fructose concentration not Na^+ concentration.

Carbohydrates and Energy Metabolism

Glucose

The breakdown of glucose can be divided into two major parts: the anaerobic conversion of glucose to pyruvate, known as glycolysis, and the aerobic breakdown of pyruvate to carbon dioxide and water, which involves the tricarboxylic acid cycle and the electron transport chain.

Glycolysis is the series of enzymatic steps leading to the breakdown of one molecule of glucose to produce two molecules of pyruvate (**Figure 1**). Glycolysis occurs in the cytosol of different cells, and all human cells are capable of carrying out this process. However, most glycolysis occurs in the liver, muscle, and adipose tissue.

The fate of pyruvate is determined by the cell type and the availability of oxygen. In the absence of oxygen, pyruvate is reduced to lactate in the cytosol. This occurs in the muscles during strenuous exercise, when the demands for energy are high. In cells that do not contain mitochondria, such as the erythrocytes, the glycolysis pathway is the only mechanism of energy production.

In the presence of oxygen, pyruvate is converted to acetyl coenzyme A (acetyl CoA) in the mitochondria and thus enters the tricarboxylic acid cycle and subsequently the electron transport chain. As a result, pyruvate is fully oxidized to carbon dioxide and water, and large amounts of energy are produced.

Fructose and Galactose

Fructose and galactose enter the glycolytic pathway through their conversion to intermediate compounds

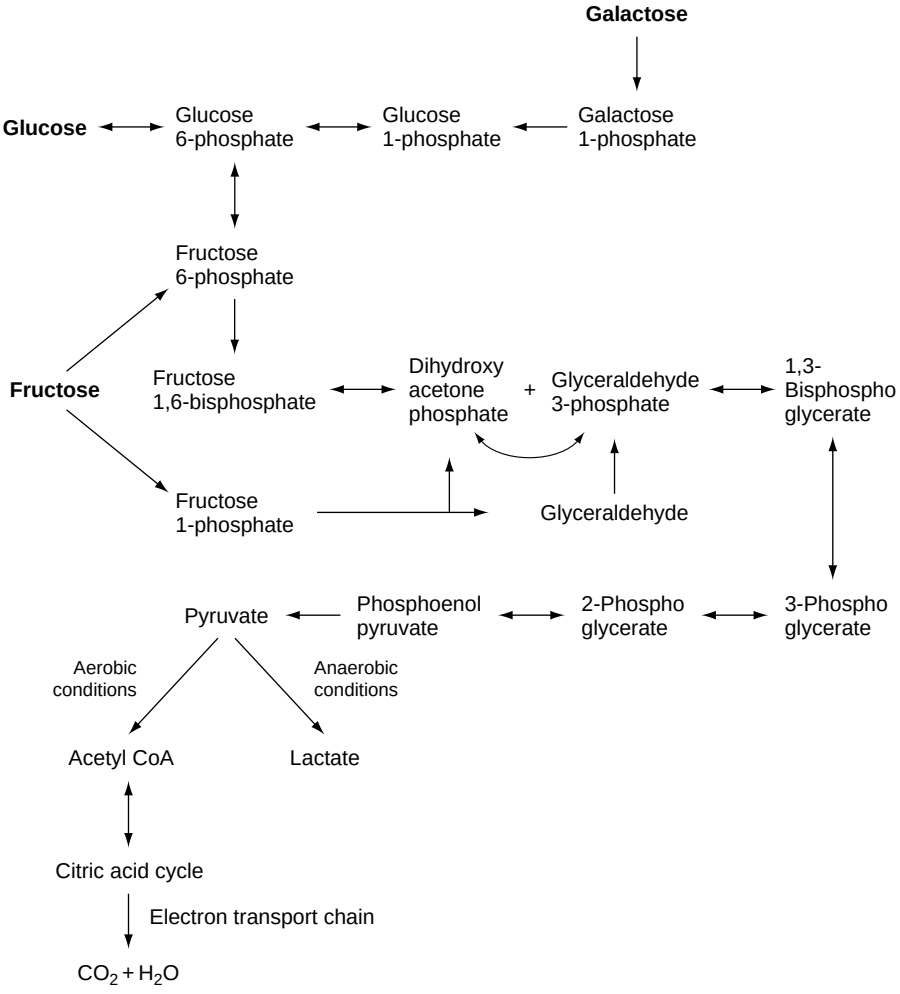


Figure 1 Outline of carbohydrate metabolism, including the points of entry of glucose, fructose, and galactose.

(**Figure 1**). This occurs primarily in the liver, and, as a result, these two monosaccharides are not generally available for uptake by other tissues. The end products of the catalysis of these monosaccharides are similar to glucose; however, when they are absorbed, they do not elicit the same hormonal response as glucose.

In the liver, breakdown of fructose, known as fructolysis, is initiated by the conversion of fructose to fructose 1-phosphate and subsequent hydrolysis to glyceraldehyde and dihydroxyacetone phosphate, a reaction catalyzed by fructose 1-phosphate aldolase. These products of hydrolysis can be used for further glycolytic conversion. Fructolysis in the liver bypasses the highly regulated step of phosphofructokinase and can produce a large amount of glycolytic metabolites. In the muscle and kidney cells, fructose can enter the glycolysis pathway through its conversion to fructose 6-phosphate, prior to the highly regulated phosphofructokinase step.

In the liver, galactose enters the glycolytic pathway through its phosphorylation to galactose 1-phosphate and subsequent epimerization to glucose 1-phosphate. This metabolic intermediate can either enter glycolysis by its conversion to glucose 6-phosphate or be used in glycogen synthesis, depending on the nutritional state of the organism.

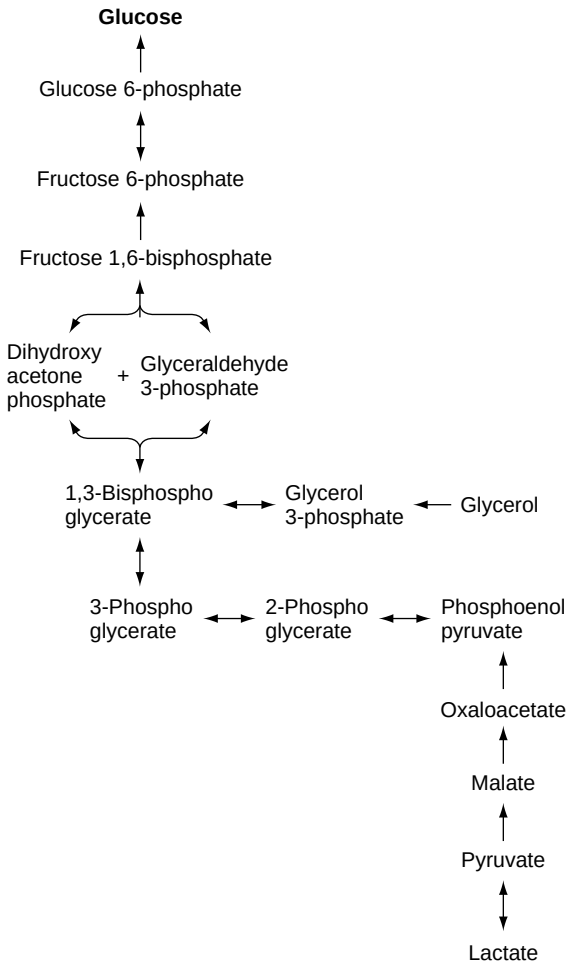


Figure 2 Outline of gluconeogenesis.

Glucose Production by the Liver and Kidneys

Gluconeogenesis

The biosynthesis of glucose from pyruvate, lactate, or other precursors is known as gluconeogenesis. It is not a direct reversal of glycolysis, since several steps of glycolysis are irreversible. Gluconeogenesis occurs mainly in the liver and less so in the kidney. These tissues contain all the necessary enzymes for gluconeogenesis and, furthermore, for the enzymatic activity of glycerol kinase, which allows glycerol to enter the gluconeogenic pathway at the level of glyceraldehyde 3-phosphate (**Figure 2**).

It is vital that the organism synthesizes glucose for those tissues that are unable to synthesize glucose. In humans, liver glycogen stores can sustain the body for 18 h without the ingestion of dietary carbohydrates. After this period, the liver must produce glucose for transport to other organs. The liver is the main gluconeogenic contributor (90%), while the kidney contributes gluconeogenically produced glucose to a lesser extent (10%).

Glycogenolysis

Glycogen is a branched polymer of glucose, which contains as many as 100 000 glucose units. The breakdown of glycogen for the production of glucose is known as glycogenolysis. Glycogen breakdown is initiated at the nonreducing ends of its branches. It consists of phosphorolysis of single glucose units by the cooperating enzymatic action of glycogen phosphorylase and the debranching enzyme. The product of phosphorolysis, glucose 1-phosphate, needs the additional action of phosphoglucomutase to convert it to glucose 6-phosphate. The liver contains the enzyme glucose 6-phosphatase for the hydrolysis of glucose 6-phosphate to free glucose, which can then be exported to the target tissues. However, the muscle and brain do not contain this enzyme, and the glucose 6-phosphate they produce enters the glycolytic pathway for energy production. Glycogen is a very efficient storage from of glucose, having an overall efficiency of storage of approximately 97%.

Control of Carbohydrate Metabolism

Hormonal Regulation

Hormones regulate (activate or inhibit) specific enzymes that catalyze the reactions of metabolic pathways. This is achieved mainly by covalent regulation or by conversion of the enzymes into their active or inactive form. Furthermore, hormones can control enzymes by induction or regulation of their transcription. Regulation of the expression of specific genes controls the concentrations of the enzymes and transport proteins necessary for carbohydrate metabolism.

Insulin When a meal is ingested, glucose is liberated as a result of the hydrolysis of dietary carbohydrate in the small intestine, and it is then absorbed into the blood. Increased glucose concentrations stimulate the production and secretion of insulin by the β cells of the pancreas. Insulin promotes the transfer of glucose into the target cells (i.e., skeletal muscle, liver, and adipose tissue) for use as energy and for storage in the form of glycogen, primarily in the liver.

Insulin also stimulates glycolysis by increasing the activity of glycogen synthase (**Figure 3**) and the transcription of glycolytic enzymes (**Figure 4**). Insulin inhibits gluconeogenesis by decreasing the transcription of several gluconeogenic enzymes (**Figure 4**) and by moderating the peripheral release of gluconeogenic precursors.

Fasting results in a decrease in insulin concentration and a reduction in glucose uptake by the muscle and adipose tissue, which use alternate forms of energy (e.g., free fatty acids). Glucose then becomes available for uptake by the brain, red blood cells, and renal medulla, which are strongly dependent on glucose for energy.

Glucagon Glucagon is a hormone secreted in the bloodstream by the α cells of the pancreas in response to low glucose levels. Glucagon counteracts the action of insulin, and its main role is to stimulate hepatic glucose output and to maintain glucose homeostasis. Glucagon stimulates glycogenolysis by activating glycogen phosphorylase and inhibits glycogen synthesis by inactivating glycogen synthase (**Figure 3**). Furthermore, glucagon stimulates gluconeogenesis by increasing the gene expression of gluconeogenic enzymes and by blocking glycolysis. In the liver, glucagon enhances the rate of gluconeogenesis by lipolysis, resulting in increased concentrations of free fatty acids and glycerol.

Catecholamines Epinephrine and norepinephrine are catecholamines that have a regulatory effect on

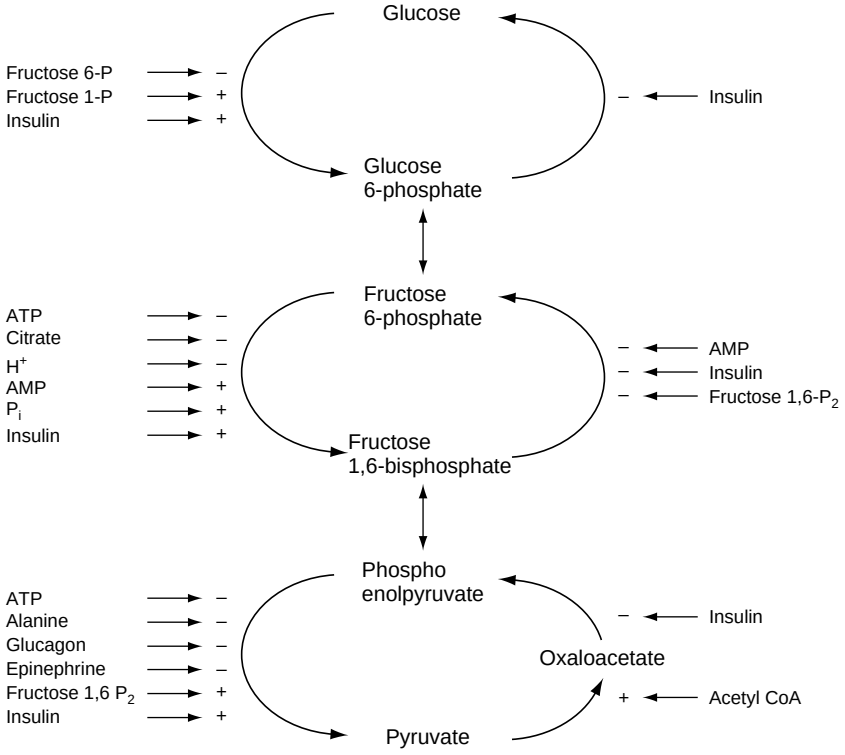


Figure 3 Points of regulation of glycogen synthesis and breakdown.

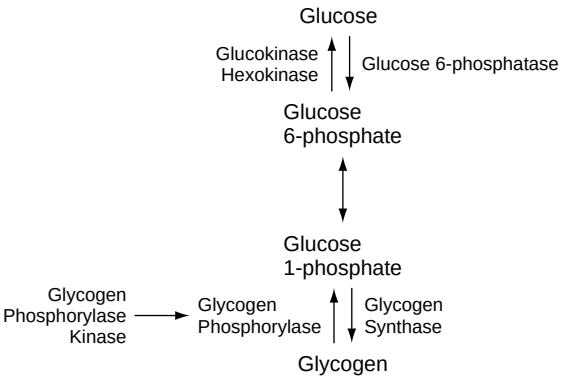


Figure 4 Regulation of glycolysis and gluconeogenesis in the liver.

carbohydrate metabolism. This effect is mainly dependent on the type of receptor present on each cell. Catecholamine receptors are divided into two types: two α receptors and three β receptors. The β and α-1 receptors stimulate catabolic reactions, while the α-2 receptor inhibits them. The presence of different catecholamine receptors on different cell types explains the selective breakdown of stores from certain tissues.

During fasting, catecholamines stimulate gluconeogenesis and glycogenolysis in the liver, as a result of increased secretion of glucagon by epinephrine. Catecholamines normally do not play a central role in maintaining glucose homeostasis during fasting, but they prevent hypoglycaemia when glucagon secretion is low.

Glucocorticoids Cortisol, the principal glucocorticoid, stimulates hepatic glucose output and the expression of genes encoding for gluconeogenic enzymes, thus stimulating gluconeogenesis. Cortisol is essential for the action of several hormones and has a much slower effect on hepatic glucose production than either glucagon or the catecholamines, taking several hours to take place.

Growth hormone Growth hormone, like cortisol, increases hepatic glucose production by changing substrate availability and promoting the expression of gluconeogenic enzymes. Growth hormone secretion is enhanced by starvation. Like cortisol, growth hormone affects hepatic glucose production much more slowly than glucagon or the catecholamines, taking several hours to occur.

Allosteric Enzyme Regulation

Allosteric enzymes are activated or inhibited by substances produced in the pathway in which the enzymes function. These substances are called modulators and can alter the activity of allosteric enzymes by changing their conformation. Adenosine monophosphate (AMP), adenosine diphosphate (ADP), and adenosine triphosphate (ATP) are important modulators of allosteric enzymes in carbohydrate metabolism. The effects of ATP are opposed by those of AMP and ADP. When energy supply is adequate, ATP accumulates and negatively modulates enzymes that catalyze energy-producing or catabolic pathways, e.g., glycolysis. When energy is depleted and ATP concentration is decreased, AMP and ADP accumulate. As a result, allosteric enzymes in catabolic pathways are positively modulated and energy is produced. An increase in ATP inhibits further energy production and blocks glycolytic enzymes, while an increase in AMP or ADP stimulates glycolytic enzymes for energy production (**Figure 3**).

Directional Shifts

The majority of enzymes catalyze reversible reactions, and their action is highly dependent on the concentration of the reactants involved. An increase in the concentration of one reactant will drive the reaction in the direction that results in the breakdown of that reactant so as to achieve homeostasis. An example of a directional shift is the interconversion of glucose 1-phosphate and glucose 6-phosphate. During glucogenolysis, the concentration of glucose 1-phosphate increases and the reaction is driven towards the production of glucose 6-phosphate. During glycogen synthesis and gluconeogenesis, the concentration of glucose 6-phosphate increases and the reaction is driven towards the production of glucose 1-phosphate and, subsequently, towards the formation of glycogen.

Regulation of Gene Expression

Regulation of gene expression enables the human body to respond to changes in nutrient concentration. During increased availability of a specific nutrient, there is no need to express the genes encoding for enzymes involved in the metabolism of that nutrient. Gene expression is highly regulated by hormones, which respond to the concentration of nutrients in the blood. Selective expression of specific genes plays a major role in the regulation of carbohydrate metabolism.

Hormonal and nutrient concentrations affect several regulatory domains of genes that encode for enzymes involved in anabolic and catabolic pathways. High insulin and glucose concentrations increase mRNA levels and the transcription rates

of the glycolytic enzymes and decrease those of the gluconeogenic enzymes. Glucagon has the opposite effect to insulin.

Glycogen Synthesis and Breakdown

The regulatory mechanism of glycogen synthesis and breakdown involves two counteracting enzymes: glycogen synthase and glycogen phosphorylase (**Figure 4**). Insulin activates glycogen synthase and therefore increases glycogen synthesis in the liver and muscle. When blood glucose levels decrease, glucagon inhibits glycogen synthase and activates glycogen phosphorylase in order to break down glycogen in the liver. Epinephrine also activates glycogen breakdown both in the liver and in skeletal muscle.

Peripheral Uptake of Glucose by Skeletal Muscle and Adipose Tissue

Glucose enters the target tissues by facilitated diffusion through a family of transporters known as glucose transporters (GLUTs). Five different isoforms of GLUTs have been isolated and characterized, GLUT1 to GLUT5. GLUT4 is mainly present in skeletal and cardiac muscle and in brown adipose tissue and differs significantly from the other isoforms in that it is stimulated by insulin. The other GLUTs do not require the action of insulin for glucose transport. GLUT1 and GLUT3 are responsible for glucose transport in most body tissues and are found in the brain, kidney, placenta, red blood cells, and fetal tissue. GLUT2 exists mainly in the liver and pancreas, and GLUT5 is responsible for glucose and fructose transport in the small intestine.

The GLUTs are encoded by different genes, and the regulation of their expression is highly tissue specific. GLUT4 is highly regulated by insulin, and its concentration is significantly increased in the presence of this hormone. As a result of the increase in GLUT4 concentration, there is increased glucose uptake by the adipose tissue and skeletal muscle.

Diseases of Carbohydrate Metabolism

Carbohydrate Malabsorption

Fructose intolerance and essential fructosuria Fruc-Fructose intolerance and essential fructosuria are genetic defects of fructose metabolism. Fructose intolerance is an autosomal recessive disease, caused by a genetic defect in fructose 1-phosphate aldolase (aldolase B) in the liver. The symptoms of aldolase B deficiency occur when the infant is exposed to fructose. Aldolase B deficiency results in phosphate depletion and fructose 1-phosphate accumulation in the liver. Consequently, gluconeogenesis and glycogenolysis are blocked, resulting in the inhibition of protein synthesis and subsequent liver failure.

Essential fructosuria is caused by a defect in the fructokinase gene. This disorder is asymptomatic and results in the excretion of fructose in the urine and in the conversion of fructose to fructose 6-phosphate in the muscle and adipose tissue.

Glucose and galactose malabsorption Carbohydrate intolerance is a hereditary disorder that occurs infrequently and poses serious health risks. This disorder is caused by a deficiency in a digestive enzyme (e.g., sucrase-α-dextrinase) and defective glucose-galactose transport. Carbohydrate intolerance presents as the development of profuse infant diarrhoea immediately after birth.

Glycogen Storage Diseases

A lack of the enzyme glucose 6-phosphatase in the liver, kidney, and intestinal mucosa causes a disease known as Von Gierke's disease. This disease results in fasting hypoglycaemia, hepatomegaly, and recurrent acidosis. Genetic defects in glucose 6-phosphatase, glucose 6-phosphatase translocase, or pyrophosphate transporter result in a metabolic imbalance and an inability of the liver to maintain glucose homeostasis by either glycogenolysis or gluconeogenesis.

Gene mutations in the liver and muscle glycogen phosphorylases result in rare autosomal recessive disorders. In the liver, the disease results in glycogen accumulation and is known as Hers' disease. It is characterized by hypoglycaemia, hepatomegaly, and growth delay. In the muscle, the disease results in progressive muscle weakness and glycogen accumulation in the liver, and is known as McArdle's disease. It is characterized by exercise intolerance. A mutation in the debranching enzyme also results in glycogen accumulation in the liver and/or muscle and is known as Cori's disease. It is characterized by fasting hypoglycaemic convulsions, hepatomelagy, and myopathy.

Diabetes Mellitus

Diabetes mellitus is a group of metabolic disorders characterized by high levels of blood glucose (impaired glucose tolerance) and results from defects in insulin secretion, insulin action, or both. Diabetes mellitus is the seventh leading cause of death in the USA and is a major cause of premature mortality, stroke, cardiovascular disease, peripheral vascular disease, congenital malformations, perinatal mortality, and long-term and short-term disability. There

are four principal types of diabetes mellitus: type 1 (formerly known as insulin-dependent diabetes mellitus), type 2 (formerly known as noninsulin-dependent diabetes mellitus), gestational diabetes (GDM), and maturity-onset diabetes of the young (MODY).

Type 1 diabetes Type 1 diabetes is caused by autoimmune pancreatic β cell exhaustion and loss of insulin secretion. Onset of the disease occurs when most of the pancreatic β cells have been destroyed by the immune system. This form of diabetes is generally diagnosed in children and young adults and accounts for between 5% and 10% of all cases of diabetes mellitus.

Type 2 diabetes Type 2 diabetes is a complex heterogenous disorder caused by interactions of various genetic and environmental factors. It is characterized by insulin resistance, obesity, a sedentary lifestyle, and occasionally by decreased insulin secretion. Because obesity and physical inactivity are increasing in children, the prevalence of paediatric type 2 diabetes has increased dramatically over the past 20 years to reach epidemic proportions. More than 85% of the cases of diabetes mellitus are type 2.

Gestational diabetes GDM is a form of glucose intolerance that is diagnosed in some pregnant women. It is usually ameliorated after childbirth, but it increases the risk of developing type 2 diabetes in the future.

Maturity-onset diabetes of the young MODY is an autosomal dominant trait that primarily affects insulin secretion and accounts for between 2% and 5% of the cases of diabetes. MODY can be caused by mutations in the glucokinase genes, leading to a reduced rate of glycolysis in the pancreas, reduced glycogen synthesis, and increased gluconeogenesis in the liver. It is diagnosed mostly in France and in the UK.

Abbreviations

ADP	Adenosine diphosphate
AMP	Adenosine monophosphate
ATP	Adenosine triphosphate
GDM	Gestational diabetes
GLUT1	Glucose transporter 1
GLUT2	Glucose transporter 2
GLUT3	Glucose transporter 3
GLUT4	Glucose transporter 4
GLUT5	Glucose transporter 5
MODY	Maturity-onset diabetes of the young

See also: **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Energy**: Metabolism. **Fructose**. **Galactose**. **Glucose**: Chemistry and Dietary Sources; Metabolism and Maintenance of Blood Glucose Level; Glucose Tolerance. **Glycemic Index**.

Further Reading

Brody T (1999) *Nutritional Biochemistry*, 2nd edn, pp. 57–132 and 157–271. San Diego: Academic Press.

Corssmit EPM, Romijin JA, and Sauerwein HP (2001) Regulation of glucose production with special attention to non-classical regulatory mechanisms: a review. *Metabolism* **50**: 742–755.

Eastwood M (2003) *Principles of Human Nutrition*, 2nd edn, pp. 195–212, 418–426 and 486–509. Oxford: Blackwell.

Gray GM (2000) Digestion and absorption of carbohydrate. In: *Biochemical and Physiological Aspects of Human Nutrition*, 1st edn, pp. 91–106. Philadelphia: WB Saunders Company.

Groff JL and Gropper SS (eds.) (2000) *Advanced Nutrition and Human Metabolism*, 3rd edn, pp. 70–105. Belmont: Wadsworth.

Jiang G and Zhang BB (2003) Glucagon and regulation of glucose metabolism. *American Journal of Physiology. Endocrinology and Metabolism* **284**: E671–E678.

Levin RJ (1999) Carbohydrates. In: Shils ME, Olson JA, Shike M, and Ross AC (eds.) *Modern Nutrition in Health and Disease*, 9th edn, pp. 49–65. Media: Lippincott Williams & Wilkins.

McGrane MM (2000) Carbohydrate metabolism – synthesis and oxidation. In: *Biochemical and Physiological Aspects of Human Nutrition*, 1st edn, pp. 158–210. Philadelphia: WB Saunders Company.

Schlenker ED (2003) Carbohydrates. In: Williams SR and Schlenker ED (eds.) *Essentials of Nutrition & Diet Therapy*, 8th edn, pp. 47–65. St. Louis: Mosby.

Schlenker ED (2003) Digestion, Absorption, and Metabolism. In: Williams SR and Schlenker ED (eds.) *Essentials of Nutrition & Diet Therapy*, 8th edn, pp. 23–45. St. Louis: Mosby.

Stryer L (1995) *Biochemistry*, 4th edn, pp. 463–602. New York: WH Freeman and Company.

Tirone TA and Brunicardi FC (2001) Overview of glucose regulation. *World Journal of Surgery* **25**: 461–467.

Tso P and Crissinger K (2000) Overview of digestion and absorption. In: *Biochemical and Physiological Aspects of Human Nutrition*, 1st edn, pp. 75–90. Philadelphia: WB Saunders Company.

Requirements and Dietary Importance

C L Stylianopoulos, Johns Hopkins University, Baltimore, MD, USA

Introduction

Carbohydrates are an important energy source in the human diet. They generally supply about 45% of the energy requirement in developed countries and up to 85% in developing countries. Carbohydrates have been considered a fundamental source of nourishment and inexpensive and versatile staple of the diet.

The type and composition of dietary carbohydrates varies greatly among different food products. Dietary carbohydrates can be predominantly found in the form of sugar (monosaccharides and disaccharides) and starch or nonstarch polysaccharides. Furthermore, in the food industry they can be used in the form of hydrolyzed cornstarch, high-fructose corn syrups, modified starches, gums, mucilages, and sugar alcohols.

The current global emphasis for healthy eating focuses on increasing carbohydrate consumption, particularly in the form of whole grains, fruit, and vegetables. Epidemiological and clinical studies have shown a positive association between carbohydrate consumption and reduced risk of chronic disease and certain types of cancer.

Dietary Sources and Intakes

The major sources of carbohydrates are cereals, accounting for over 50% of carbohydrate consumed in both developed and developing countries, followed by sweeteners, root crops, pulses, vegetables, fruit, and milk products. Carbohydrate and nutrient intake in general can be estimated using data from food production and balance sheets, household surveys, and individual assessments (**Table 1**). **Figure 1** shows the trends in carbohydrate consumption by food group as a percentage of total carbohydrate in developed and developing countries, obtained from food balance data in 1994.

Sugars

The term 'sugar' includes monosaccharides and disaccharides. The most common monosaccharides are glucose (or dextrose), fructose, and galactose. Glucose is found in fruit, honey, maple syrup, and vegetables. Glucose is also formed from sucrose hydrolysis in honey, maple syrup and invert sugar, and from starch hydrolysis in corn syrups. The properties of glucose are important for improving food texture, flavor, and palatability. Glucose is the major cell fuel and the principal energy source for the brain. Fructose is found in honey, maple sugar, fruit, and vegetables. Fructose is also formed from sucrose hydrolysis in honey, maple syrup, and invert sugar. It is commonly used as a sweetener in soft drinks, bakery products, and candy in the form of high-fructose corn syrups. Galactose is found primarily in milk and dairy products.

The most common disaccharides are sucrose, lactose, and maltose. Sucrose is mostly found in sugar cane and beet, and in lesser amounts in honey, maple sugar, fruit, and vegetables. The properties of sucrose are important in improving viscosity, sweetness, and flavor of baked foods, ice cream, and desserts. Maltose is formed from starch digestion. It is also produced from the germination of grain for malt liquors. Lactose is found in milk and dairy products, and is not as sweet as glucose or sucrose.

In the second part of the twentieth century, sugar intake increased markedly in the US, because of increased consumption of added sugars in beverages and foods. According to the US Food Supply Data, consumption of added sugars has increased from 27 teaspoons/person/day in 1970 to 32 teaspoons/person/day in 1996, which represents a 23% increase. Soft drinks are the most frequently used form of added sugars, and account for one-third of total sugar intake. In Europe the trend of sugar consumption has been a steady one.

Table 1 Approaches for determination of trends in nutrient consumption worldwide

Approach	*Advantages*	*Disadvantages*
Food production	Figures available for every crop	Affected by agricultural practices, weather conditions, external forces
Food balance sheets	Figures available for every food item	Inadequate to determine food waste and spoilage
Household surveys	Figures close to actual food consumption	Inadequate to determine food consumption outside the home, food waste, and spoilage
Individual assessments	Figures close to actual food consumption	Data not available for all countries Diverse methods of assessment

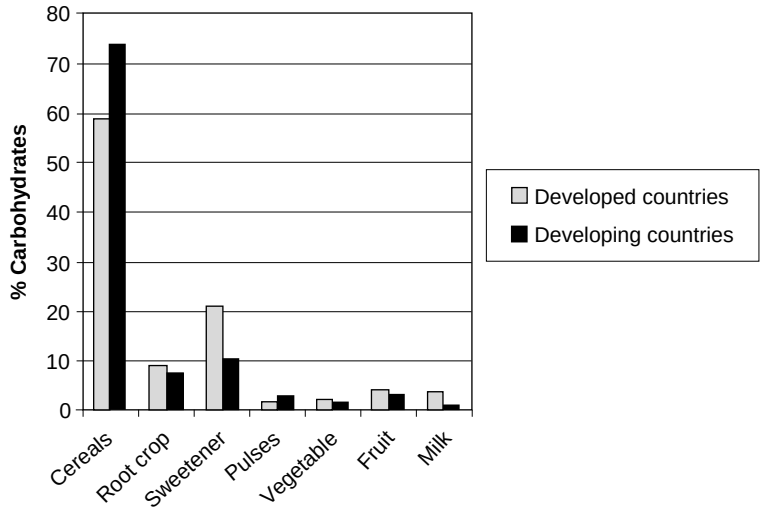


Figure 1 Trends in energy consumption by carbohydrate food group as a percentage of total carbohydrate in developed and developing countries, obtained from food balance data in 1994. Data obtained from FAO/WHO (1998). Carbohydrates in human nutrition. Report of a joint FAO/WHO expert consultation. *FAO Food and Nutrition Papers* 19 **66**: 1–140.

Polysaccharides

Starch Starch is the most important and abundant food polysaccharide. Starch is predominantly derived from plant seed, such as wheat, maize, rice, oats, and rye, and from plant roots, such as potatoes. Legumes and vegetables also contribute to the starch content of the diet. Bread and pasta are popular forms of starch, while tropical starchy foods, such as plantains, cassava, sweet potatoes, and yams are increasingly contributing to carbohydrate intake. Starch accounts for 20–50% of total energy intake, depending on the total carbohydrate consumption.

Nonstarch Nonstarch polysaccharides (NSP), formerly referred to as 'dietary fiber,' can either be soluble or insoluble and are mainly derived from cereals, especially wholegrain. Wheat, rice, and maize contain predominantly insoluble NSP, while oats, rye, and barley contain predominantly soluble NSP. Vegetables are also a source of NSP and contain equal amounts of insoluble and soluble NSP. Intakes of NSP range from about 19 g day^{-1} in Europe and North American countries to 30 g day^{-1} in rural Africa.

Health Effects of Carbohydrates

Carbohydrates are stored in the human body as glycogen mainly in the liver and muscle. The human body has a limited storage capacity for carbohydrates compared to fat. The total amount of carbohydrates stored in tissues and circulating in the blood as glucose is approximately 7.56 MJ (1800 kcal). Diets high in carbohydrate ensure adequate glycogen storage available for immediate energy utilization. Carbohydrates are the preferred energy source for the human brain and have an important role in reducing protein breakdown when energy intake is inadequate.

Dietary carbohydrates are absorbed in their hexose form (glucose, fructose, galactose) and provide 15.6 kJ g^{-1} (3.75 kcal g^{-1}) of energy. Although sugars and polysaccharides provide similar amounts of energy, they differ in their physiological and metabolic properties. The effects of carbohydrate-containing foods on blood glucose levels during digestion and absorption are variable, depending on the type of dietary carbohydrate. Postprandial glucose response is reduced when glucose absorption is slow. Glycemic index (GI) is used for the quantification of blood glucose response after carbohydrate consumption. GI is the area under the curve of the blood glucose increase 2 h after carbohydrate ingestion of a set amount of a particular food (e.g., 50 g) compared to the blood glucose increase 2 h after ingestion of the same amount of a reference food (white bread or glucose). GI is influenced significantly by the carbohydrate types and physical determinants of digestion rate (intact versus ground grains, cooked versus uncooked food, and soluble fiber content). Carbohydrate ingestion in the presence of fat and protein reduces the GI of a meal. The GI of carbohydrate-containing meals has been linked to several health outcomes. The role of

carbohydrates in health is a growing area of research and has received a great amount of interest in the past decade.

Carbohydrates and Nutrient Density

Increased sugar consumption has generated concern in recent years because of the potential to displace the micronutrient content of the diet by increasing 'empty calories' and energy intake. There is some evidence that essential nutrient intake decreases with increasing total sugar intake. However, sugar intake has not been shown to accurately predict micronutrient ingestion. Moderate intakes of sugar coincide with sufficient nutrient intake. The risk of low micronutrient status is increased for individuals with a diet high in sugars and low in total energy intake, as in the case of children or people on restrictive diets. Data analysis on food intake of preschool children suggests that the intake of some micronutrients (calcium, zinc, thiamin, riboflavin, and niacin) is inversely related to sugar intake. However, the dilutional effects of sugars may be somewhat distorted by the fact that some rich sources of added sugars are also fortified with micronutrients, as in the case of breakfast cereals. The Dietary Reference Intake (DRI) Panel on Macronutrients, using national food intake data, reported that a clear dilutional effect on micronutrient intake starts when sugar intake approaches 25% of total calories. The American Heart Association dietary guidelines stress the consumption of fruit, vegetables, grains, and complex carbohydrates so that micronutrient requirements are met by whole rather than supplemented foods.

Several human studies have demonstrated that diets rich in NSP may reduce the bioavailability of minerals, such as iron, calcium, and zinc. Nevertheless, this effect is more likely due to the presence of phytate, which inhibits the absorption of those minerals, than the NSP content of the diet.

Carbohydrates and Obesity

Several studies have been conducted to establish an association between sugar ingestion and total energy intake. There have been consistent reports of a negative association between sugar intake and body mass index in adults and children. However, this observation could be confounded by the correlation of dietary fat and obesity, since high-fat diets are usually low in carbohydrates. Some ad libitum dietary studies have shown that diets low in sugar are associated with weight loss, maybe as a result of reduced calorie intake. Nevertheless, in human metabolic studies, no effect on weight or energy expenditure was observed when carbohydrate was replaced by fat or protein in isocaloric diets.

Foods high in sugars or GI are highly palatable and it has been suggested that they create a potential risk for energy overconsumption and weight gain. However, there is no evidence to support this claim or confirm the role of GI in body weight regulation. Foods high in sugar have high energy density and thus decreasing their consumption can assist in weight reduction. On the contrary, foods rich in NSP are bulky and have less energy density and as a result induce greater satiety when ingested. It follows that diets rich in NSP may be useful for obesity prevention, since they prevent energy overconsumption. However, there is no evidence to indicate that increasing the carbohydrate content of a low-energy diet facilitates weight loss.

The consumption of sugar-sweetened soft drinks may contribute to weight gain because of the low satiety of liquid foods. Short-term human studies have shown that sugar-sweetened soft drink consumption does not result in a decrease of total energy intake. Thus, sugar-sweetened soft drinks can significantly increase the total caloric intake and result in weight gain. Consumption of these drinks has been associated with childhood obesity.

Carbohydrate and Cardiovascular Disease

Dietary factors influence the risk factors, such as obesity, diabetes, and hyperlipidemia, that lead to the development of cardiovascular disease (CVD). A diet rich in carbohydrates in the form of whole grain cereals, fruit, and vegetables may assist in the reduction of saturated fat and may increase the antioxidant content of the diet, thus reducing the risk of heart disease. On the contrary, a high intake of carbohydrates (>65% of total calories), especially in the form of refined sugars and starch, may increase serum triacylglycerol levels and adversely affect plasma lipoprotein profile. Short-term studies show a consistent relationship between sugar consumption and elevation of triacylglycerol levels as well as a decrease in plasma high-density lipoprotein (HDL) levels, which could result in increased atherosclerosis and heart disease risk. However, longitudinal cohort studies have failed to show a consistent association of sugar consumption with CVD, mainly because of the confounding factors associated with increased heart disease risk.

Certain NSP (for example β glycans) have been shown to reduce low-density lipoprotein (LDL) and total cholesterol levels on a short-term basis. Therefore, a protective effect for CVD has been shown with consumption of foods high in NSP. This

protective effect has not been duplicated with NSP supplements. Furthermore, no long-term effect has been established.

High-GI diets have been shown to slightly increase hemoglobin A1c, total serum cholesterol and triacylglycerols, and decrease HDL cholesterol and urinary C-peptide in diabetic and hyperlipidemic individuals. In addition, low-GI diets have been shown to decrease cholesterol and triacylglycerol levels in dyslipidemic individuals. There are insufficient studies performed on healthy individuals and further research on the role of GI in lipid profile and CVD risk factors is warranted.

Carbohydrates and Type 2 Diabetes

There is little evidence from prospective studies to support a positive association between total dietary carbohydrate consumption and risk of type 2 diabetes. Some recent evidence suggests that rapidly digested refined sugars, which have a high GI, may increase the risk of type 2 diabetes. Short-term studies have shown that decreasing the GI of a meal can improve glucose tolerance and insulin sensitivity in healthy people. Furthermore, the substitution of high-GI with low-GI carbohydrates can decrease postprandial glucose and insulin levels. Some epidemiological studies have demonstrated a protective effect of NSP consumption against type 2 diabetes.

Carbohydrates and Dental Caries

The quantity and frequency of sugars in the diet play a significant role in the development of dental caries. Their digestion by salivary amylase provides an acid environment for the growth of bacteria in the mouth, thus increasing the rate of plaque formation. Sucrose is the most cariogenic of the sugars, followed by glucose, fructose, and maltose. The milk sugars (lactose and galactose) are considerably less cariogenic. There is no epidemiological evidence to support a cariogenic role of polysaccharide foods with no added sugars.

Dental caries is a multifaceted disease, affected not only by the frequency and type of sugar consumed, but also by oral hygiene and fluoride supplementation and use. Despite the increase in sugar consumption, the incidence of dental caries has decreased worldwide because of the increased use of fluoride and improvement in oral hygiene.

Carbohydrate and Cancer

Case-control studies have shown that colorectal cancer risk increases with high intakes of sugar-rich foods, while other studies have failed to prove such a relationship. Thus, there is insufficient evidence to support the role of sugar in the risk for colorectal cancer. On the contrary, carbohydrate consumption in the form of fruit, vegetables, and cereals has been shown to be protective against colorectal cancer.

Carbohydrate foods are a good source of phytoestrogens, which may protect against breast cancer. However, studies related to carbohydrate intake and breast cancer have been inconsistent and are insufficient to establish an association between carbohydrates and breast cancer risk.

The Health Professionals Follow-up Study showed a negative association of prostate cancer risk with high fructose intake. Additional data on the role of sugar consumption on prostate cancer risk is lacking. Some evidence suggests that increased fiber intakes are related to decreased prostate cancer risk.

Carbohydrates and Gastrointestinal Health

High intakes of NSP, in the range of 4–32 g day^{-1}, have been shown to contribute to the prevention and treatment of constipation. Population studies have linked the prevalence of hemorrhoids, diverticular disease, and appendicitis to NSP intakes, although there are several dietary and lifestyle confounding factors that could directly affect these relationships. High-carbohydrate diets may be related to bacterial growth in the gut and subsequent reduction of acute infective gastrointestinal disease risk.

Low-Carbohydrate Diets

The recent trend of weight loss diets promotes some level of carbohydrate restriction and increased protein consumption. Some examples are Dr Atkins New Diet Revolution, The South Beach Diet, and The Carbohydrate Addict's Diet. This dietary advice is contrary to that proposed by governmental agencies (US Department of Agriculture/Department of Health Services, National Institutes of Health) and nongovernmental organizations (American Dietetic Association, American Heart Association, American Diabetes Association, and American Cancer Society).

There is consistent evidence that weight loss in low-carbohydrate diets is triggered by negative energy balance resulting from low caloric intake, and that it is not a function of macronutrient composition. There is no scientific evidence to suggest that low-carbohydrate diets are more metabolically efficient than restricted calorie conventional diets. Several studies have shown that low-carbohydrate diets result in weight loss because of reduced caloric intake.

Low-carbohydrate diets promote the lipolysis of stored triacylglycerols known as ketosis, reduce glucose and insulin levels, and suppress appetite. As a result, there is an increase in blood uric acid concentration. Some studies have shown that the consumption of high amounts of nondairy protein results in a decline in kidney functions in individuals with mildly compromised kidney function. However, no such effect has been shown in individuals with normal kidney functions. Furthermore, low-carbohydrate diets can have side effects such as bad taste, constipation, diarrhea, dizziness, headache, nausea, thirst, and fatigue.

Low-carbohydrate diets lack essential vitamins and minerals because of inadequate consumption of fruit, vegetables, and grains, and require supplementation to achieve nutritional adequacy. Controlled trials of low-carbohydrate diets are necessary to establish long-term effectiveness and adverse health effects or benefits.

Requirements and Recommendations

According to the new definition of the expert panel appointed by the Institute of Medicine of the National Academies of Science (IOM), dietary reference intakes (DRIs) are defined as a set of reference values for nutrient intake, and include the estimated average requirement (EAR), recommended dietary allowance (RDA), adequate intake (AI), and tolerable upper intake level (UL). EAR refers to the average daily intake value of a nutrient that is estimated to fulfill the needs of healthy people in a particular lifestage or group. RDA refers to the minimum daily intake that fulfills the need of almost all healthy people in a particular lifestage or group. AI refers to the observed intake of a particular group of healthy people, and is used when there is lack of scientific experimentation for the determination of the EAR or the RDA. UL refers to the maximum daily intake level of a nutrient that is not likely to pose an adverse health effect for almost all people.

The DRIs for carbohydrate consumption of individual groups and lifestages are outlined in **Table 2**. These values are based on the average minimum amount of glucose needed for brain function. A UL for carbohydrates was not set because no studies have shown that excessive consumption of carbohydrates has a detrimental effect on health. Based on the dilutional effect of added sugars on micronutrients, the expert panel suggests a maximal intake of less than 25% of energy from added sugars. Total sugar intake can be decreased by limiting foods high in added sugars and consuming naturally occurring sugar products, like milk, dairy products, and fruit.

The IOM does not specify dietary requirements or recommendations for NSP consumption, but has provided recommended intakes for fiber, which includes NSP. The DRIs for total fiber consumption of individual groups and lifestages are outlined in **Table 3**. It has not been shown that a high fiber intake has a harmful effect in healthy individuals. Therefore, a UL for fiber has not been set.

There is insufficient evidence to support a recommendation by the IOM for the consumption of low-GI foods or the replacement of high-GI foods, like bread and potatoes. Although several studies propose adverse effects of high-GI carbohydrates and beneficial effects of low-GI foods, a recommendation on consumption of low-GI foods is a major dietary change that requires substantial scientific evidence. Therefore, a UL based on GI is not set.

The 1998 report of the Food and Agriculture Organization (FAO) and the World Health Organization (WHO) regarding the role of carbohydrates in human nutrition recommends the consumption of at least 55% of total energy in the form of carbohydrates from a variety of sources. The committee proposes that the majority of carbohydrates consumed should originate from NSP, principally from

Table 2 Carbohydrate requirements and recommendations (DRIs)

Age group/Life stage	*EAR (g day^{-1})*		*RDA (g day^{-1})*		*AI (g day^{-1})*
	Males	*Females*	*Males*	*Females*	
Infants (0–6 months)					60
Infants (6–12 months)					95
Children (1–18 years)	100	100	130	130	
Adults (>18 years)	100	100	130	130	
Pregnancy		135		175	
Lactation		160		210	

DRIs, dietary reference intakes; EAR, estimated average requirement; RDA, recommended dietary allowance; AI, adequate intake.-Data from Institute of Medicine of the National Academies (2002) *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, Amino Acids*. Washington, DC: The National Academies Press.

Table 3 Total fiber recommendations (DRIs)

Age group/Lifestage	*AI (g day^{-1})*	
	Males	*Females*
Children (1–3 years)	19	19
Children (4–8 years)	25	25
Children (9–13 years)	31	26
Children (14–18 years)	38	26
Adults (19–50 years)	38	25
Adults (>51 years)	30	21
Pregnancy		28
Lactation		29

DRIs, dietary reference intakes; AI, adequate intake.Data from Institute of Medicine of the National Academies (2001) *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, Amino Acids*. Washington, DC: The National Academies Press.

cereals, vegetables, legumes, and fruit. Furthermore, it suggests that free sugars should be restricted to less than 10% of total energy. This report recognizes that there is no direct causal link between sugar consumption and chronic disease. However, sugars significantly increase the energy density of the human diet and high-sugar drinks have been associated with childhood obesity.

A 2002 report of the American Heart Association suggests the restriction of sugar consumption. This report recognizes that there are no beneficial effects of increased sugar consumption. On the contrary, some studies suggest that it may have adverse health effects. In order to enhance the nutrient density and reduce the energy density of the diet, increased consumption of high-sugar foods should be avoided.

See also: **Cancer**: Epidemiology and Associations Between Diet and Cancer; Effects on Nutritional Status. **Carbohydrates**: Chemistry and Classification; Regulation of Metabolism; Resistant Starch and Oligosaccharides. **Cereal Grains**. **Dental Disease**. **Diabetes Mellitus**: Etiology and Epidemiology. **Dietary Fiber**: Role in Nutritional Management of Disease. **Energy**: Metabolism. **Fructose**. **Fruits and Vegetables**. **Galactose**. **Glucose**: Chemistry and Dietary Sources. **Glycemic Index**. **Hypertension**: Etiology. **Lipids**: Chemistry and Classification. **Obesity**: Definition, Etiology and Assessment. **Sucrose**: Nutritional Role, Absorption and Metabolism; Dietary Sucrose and Disease. **World Health Organization**.

Further Reading

Anderson GH and Woodend D (2003) Consumption of sugars and the regulation of short-term satiety and food intake. *American Journal of Clinical Nutrition* **78**(supplement): 843S–849S.

Brody T (1999) *Nutritional Biochemistry*, 2nd edn, pp. 57–192 and 457–475. San Diego: Academic Press.

Eastwood M (2003) *Principles of Human Nutrition*, 2nd edn pp. 195–212, 418–426 and 456–509. Oxford: Blackwell.

FAO/WHO (1998) Carbohydrates in human nutrition. Report of a joint FAO/WHO expert consultation. *FAO Food and Nutrition Papers 19* **66**: 1–140.

Groff JL and Gropper SS (2000) *Advanced Nutrition and Human Metabolism*, 3rd edn, pp. 70–115. ch. 4. Belmont: Wadsworth.

Howard BV and Wylie-Rosett J (2002) Sugar and cardiovascular disease. A statement for healthcare professionals from the Committee on Nutrition of the Council on Nutrition, Physical Activity, and Metabolism of the American Heath Association. *Circulation* **106**: 523–527.

Institute of Medicine of the National Academies (2002) *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, Amino Acids*. Washington, DC: The National Academies Press.

James WPT (2001) European diet and public health: the continuing challenge. *Public Health and Nutrition* **4**: 275–292.

Johnson RK and Frary C (2001) Choose beverages and foods to moderate your intake of sugars: the 200 dietary guidelines for Americans – what's all the fuss about? *Journal of Nutrition* **131**: 2766S–2771S.

Levin RJ (1999) Carbohydrates. In: Shils ME, Olson JA, Shike M, and Ross AC (eds.) *Modern Nutrition in Health and Disease*, 9th edn, pp. 49–65 ch. 3. Media: Lippincott Williams & Wilkins.

Ruxton CHS, Garceau FJS, and Cottrell RC (1999) Guidelines for sugar consumption in Europe: is a quantitative approach justified? *European Journal of Clinical Nutrition* **53**: 503–513.

Saris WHM (2003) Sugars, energy metabolism, and body weight control. *American Journal of Clinical Nutrition* **78**(supplement): 850S–857S.

Schlenker ED (2003) Carbohydrates. In: Williams SR and Schlenker ED (eds.) *Essentials of Nutrition and Diet Therapy*, 8th edn, pp. 47–65. St. Louis: Mosby.

Schlenker ED (2003) Digestion, Absorption, and Metabolism. In: Williams SR and Schlenker ED (eds.) *Essentials of Nutrition and Diet Therapy*, 8th edn, pp. 23–45. St. Louis: Mosby.

Touger-Decker R and van Loveren C (2003) Sugars and dental caries. *American Journal of Clinical Nutrition* **78**(supplement): 881S–92S.

Relevant Websites

http://www.usda.gov – US Department of Agriculture and US Department of Health and Human Services (2000) Nutrition and your health: dietary guidelines for Americans, 5th edn.

Resistant Starch and Oligosaccharides

A Laurentin, Universidad Central de Venezuela, Caracas, Venezuela
C A Edwards, University of Glasgow, Glasgow, UK

Introduction

In recent years, there has been increasing interest in those carbohydrates that escape absorption in the small intestine and enter the colon, where they may have specific health benefits due to their fermentation by the colonic microflora and their effect on gut physiology. This entry considers the definition, classification, dietary sources, methods of analysis, colonic fermentation, and health benefits of both resistant starch and oligosaccharides, and compares them with those of dietary fiber.

Resistant Starch

Definition

In 1992, a concerted action of European researchers defined resistant starch as "the sum of starch and the products of starch degradation not absorbed in the small intestine of healthy individuals." This concept completely changed our understanding of the action of carbohydrates in the diet because up until the early 1980s, it was thought that starches were completely digested and absorbed in the human small intestine. Three important considerations are attached to this physiological definition. First, resistant starch is made up not only of high-molecular weight polymers but also can include dextrins, small oligosaccharides, and even glucose, all derived from digested starch that escapes absorption. Second, resistant starches reach the human large intestine where they are metabolized by the complex colonic microflora. Finally, the actual amount of resistant starch in a food (i.e., the amount reaching the colon) depends on the physiology of the individual and it may be affected by age.

Classification and Dietary Sources

Food starches can be classified according to the way they are metabolized by the human small intestine into those that are rapidly digested, those that are slowly digested, and those that are resistant to digestion. Similarly, resistant starch has been classified into three types: physically inaccessible starch, resistant starch granules, and retrograded starch (**Table 1**).

Table 1 Classification of resistant starch

Food source	*Type*[a]	*Content in food (g per 100 g)*	*Contribution to total RS intake*
Cereal products containing whole grains or grain fragments Brown breads Legumes Pastas	RS_1	1–9	Minor
Unripe bananas Uncooked potatoes High amylose starches	RS_2	17–75	Very little
Bread Cornflakes Cooked cooled potatoes Legumes	RS_3	1–10	Major
Amylose–lipid complex Modified starches	Others	Not known	Unknown

[a]RS_1, physically inaccessible starch; RS_2, resistant granules; RS_3, retrograded starch.

Physically inaccessible starch (RS_1) Type I resistant starch is physically inaccessible and is protected from the action of α-amylase, the enzyme that hydrolyzes the breakdown of starch in the human small intestine. This inaccessibility is due to the presence of plant cell walls that entrap the starch, for example, in legume seeds and partially milled and whole grains. RS_1 can also be found in highly compact processed food like pasta. The RS_1 content is affected by disruption of the food structure during processing (e.g., milling) and, to some extent, by chewing.

Resistant granules (RS_2) Starch granules are plant organelles where starch is produced and stored. Each plant has characteristic starch granules that differ in size, shape, amylose to amylopectin ratio, crystalline to amorphous material ratio, starch supramolecular architecture, and amylose–lipid complexes, amongst other features. It is believed that combinations of these factors make some granules more resistant to the attack of digestive enzymes than other granules. Type II resistant starch is found in unripe bananas, uncooked potatoes, and high amylose starches. RS_2 disappears during cooking, especially in water, because a combination of water and heat make the starch gelatinize, giving more access to amylases.

Retrograded starch (RS_3) Type III resistant starch is the most abundant of the resistant starches present in food. It is formed during usual food processing by cooking and then cooling. When starch is cooked in an excess of water, it gelatinizes, i.e., the granular structure is disrupted, the granule swells, and amylose leaks out of the amylopectin matrix. Then, when the food is cooled down, amylose (and more slowly amylopectin) recrystallizes to a new ordered and more compact structure (process known as retrogradation), which decreases access for digestive enzymes. RS_3 production can be affected by the amylose to amylopectin ratio, amount of water, and temperature during cooking, and the number of repeated cooking and cooling cycles. Retrograded starch can be found in bread, some brands of corn flakes, cooked-cooled potatoes, and legumes.

Others sources of resistant starch In recent years, amylose–lipid complex and modified starches have also been recognized as other sources of resistant starches (**Table 1**). Amylose–lipid complexes occur when fatty acids (12–18 carbons) are held within the helical structure of amylose. They are formed naturally during starch biosynthesis, but may also be produced during cooking. Lipids may interfere with amylose retrogradation, impairing the production of retrograded starch during processing. However, these complexes themselves have lower digestibility than cooked starch.

As well as naturally resistant starch complexes, there are different types of modified starches that are manufactured by the food industry for a variety of reasons. They can be defined as native starches that have been submitted to one or more physical, chemical, or enzymatic treatments promoting granular disorganization, polymer degradation, molecular rearrangements, oxidation, or chemical group addition. Modified starches can be classified into four main categories accordingly to their main physicochemical characteristic: pregelatinized, derivatized, cross-linked, and dextrinized starches (**Table 2**). However, they usually are known as physically, chemically, or enzymatically modified starch because of the way they are produced (**Table 3**). The digestibility of these modified starches is variable and depends on the type and extent of the treatment. Some authors have proposed a new category, type IV resistant starch, to include chemically modified starches. Indeed, it has been shown that cross-linked starches have a 15–19% decrease in *in vitro* digestibility when compared with their native starches, and hydroxypropylated starch is only 50% digestible. However, pregelatinized starches produced by drum drying and extrusion have a 3–6% and 5–11% decrease in digestibility, respectively. Part but not all of this reduction in digestibility is due to the formation of retrograded starch; therefore, physically modified starches should also be considered as a category of resistant starch.

In addition to the starch properties already described, several starchy foods (for instance, cereals and legumes) have antinutritional factors, such as lectins, tannins, phytates, and enzyme inhibitors (both protease and amylase inhibitors). Amylase inhibitors present in raw pulses may reduce the activity of amylase in the human small intestine. However, most of these factors, especially enzyme inhibitors, are inactivated during food processing and cooking.

Analysis

The definition of resistant starch is based on its physiological behavior in the human small intestine, i.e., resistant starch is a heterogeneous group of molecules from small monosaccharides to large polymers with different molecular weight, degree of polymerization, and supramolecular architecture.

Table 2 Classification of modified starches

Starch	*Modifying agent*	*Physicochemical characteristic*	*Use in food*
Pregelatinized	Extrusion Drum drying	Soluble in cold water	Cake and instant products
Derivatized	Acetyl Hydroxypropyl Phosphate	Stable at freeze-thawing cycles	Canned and frozen food
Cross-linked	Epiclorhydrine Trimetaphosphate	Stable at higher temperatures, extreme pH, and higher shear forces	Meat sauce thickeners Instant soup Weaning infant food Dressings
Dextrinized	Acid hydrolysis Oxidizing agents Irradiation Heat (pyrodextrins) Amylolytic enzymes	Soluble in cold water Lower or nil viscosity	Chewing gums Jelly Syrups

Table 3 Methods of modified starch production

Treatment	*Modification*	*Description*
Physically modified	Pregelatinization	Starch paste is precooked and dried by extrusion or drum drying
	Dextrinization	Starch polymers are hydrolyzed to smaller molecules by irradiation
Chemically modified	Derivatization[a]	Lateral groups are added to starch lateral chains
	Cross-linking[a]	Multifunctional groups are used to link two different starch molecules together
	Dextrinization	Starch polymers are hydrolyzed by oxidizing agents, acid hydrolysis, pyrodextrinization
Enzymatically modified	Dextrinization	Starch polymers are hydrolyzed to smaller molecules by incubation with amylases

[a]Double-derived starches are produced by combination of these two processes.

This complexity makes it difficult to quantify accurately. All *in vitro* methods therefore need to be corroborated against *in vivo* models; however, *in vivo* models are also very difficult to validate.

In general, *in vitro* methods try to imitate human small intestine digestion using different sample preparation (i.e., milling, chewing, etc.), sample pretreatment (i.e., simulation of oral or stomach digestion), sample treatment (i.e., different enzymes mixtures), sample post-treatment (i.e., different resistant starch solubilizing agents and enzyme mixtures), and incubation conditions (i.e., shaking/stirring, pH, temperature, time) (**Table 4**). The choice of each of these multiple factors represents a huge analytical problem because not only a compromise between physiological conditions and analytical handling has to be achieved, but also because the resistant starch content values must be in agreement with *in vivo* data.

On the other hand, in human *in vivo* methods, samples of digested food that reach the end of the small intestine are taken for analysis, either from ileostomy patients (i.e., where the large intestine has been removed) or from healthy volunteers using special cannulas in the ileum. Animals can also be employed for *in vivo* experiments, such as gnotobiotic (i.e., germ-free) and pseudognotobiotic (i.e., antibiotic treated) rats. In these cases, colonic bacterial fermentation is absent or suppressed by antibiotics and it is assumed that what reaches the end of the small intestine appears in feces. The main difficulty with these methods is that *in vivo* starch digestion may occur during the whole transit through the small intestine, which varies between individuals and the type of meal consumed. Moreover, these studies are difficult to perform in healthy volunteers and the physiological significance of using ileostomy patients is debatable, for example, it may not relate to infants and children who have decreased digestive capacity.

The initial *in vitro* assays were adapted from the enzymatic-gravimetric method used for dietary fiber assessment, but could only measure RS_3. Soon new approaches to assess other types of resistant starch were developed. The Berry method, for instance, measures both RS_3 and RS_2 using an exhaustive incubation (16 h) of milled sample with α-amylase and pullulanase, following by centrifugation to separate the insoluble residue, which contains the resistant starch. This residue is treated with KOH to disperse retrograded and native starches, which are then hydrolyzed to glucose with amyloglucosidase. Finally, released glucose is quantified by a colorimetric assay. The Berry method has been subsequently modified by Faisant *et al*. and Goñi *et al*.: the pullulanase was eliminated from the enzyme mixture and a pretreatment with pepsin added to decrease starch-protein interactions (**Table 4**).

Other methods have been developed to assess all types of resistant starch. Indeed, the Englyst method was developed to assess all nutritionally important starch fractions, such as rapidly digestible and slowly digestible starches, along with the three types of resistant starches described above. In this method, resistant starch fractions are estimated altogether by difference between total and digestible starches. Sample preparation is kept to a minimum in an attempt to mimic the way food is consumed. After pretreatment with pepsin, the sample is incubated with a mixture of amyloglucosidase, invertase, and pancreatic enzymes for 2 h. Glucose released is then used to estimate the digestible starch. Next, total starch is measured as glucose released after solubilization of the nondigestible fractions with KOH, followed by amyloglucosidase hydrolysis. The Englyst method also allows evaluation of RS_1, RS_2, and RS_3. The main problem with this method is its low reproducibility, especially between laboratories, because of the technical difficulties involved. Two other methods include chewing by volunteer subjects as sample preparation. In the Muir method, for instance, the chewed sample is sequentially treated with pepsin and an amyloglucosidase–pancreatic amylase mixture to obtain the nondigestible fraction, which is then boiled with Termamyl (a thermostable α-amylase) and solubilized with dimethyl sulfoxide followed by another amyloglucosidase–pancreatic amylase mixture step to yield finally glucose. The Akerberg method is similar to the Muir method, but it includes other

Table 4 Comparison between different methods to measure resistant starch *in vitro*

Method	*Sample*				*Treatment incubation*	*Types of RS measured*[a]
	Preparation	*Pretreatment*	*Treatment*	*Post-treatment*		
Berry (1986)	Milling	None	Pancreatic α-amylase and pullulanase	KOH[b] Amyloglucosidase	Shaking for 16 h at 37 °C, pH 5.2	Sum of RS_2 and RS_3
Faisant *et al.* (1995)	Same as above	Same as above	Same as above, but without pullulanase	Same as above	Same as above, but pH 6.9	Same as above
Goñi *et al.* (1996)	Same as above	Pepsin	Same as above	Same as above	Same as above	Same as above
Englyst *et al.* (1992)	Minced or as eaten	Pepsin	Pancreatic α-amylase, amyloglucosidase, and invertase	Same as above	Shaking for 2 h at 37 °C, pH 5.2	RS_1, RS_2, RS_3, and total RS
Muir & O'Dea (1992)	Chewing	Salivary α-amylase then pepsin	Pancreatic α-amylase and amyloglucosidase	Thermostable α-amylase Dimethyl sulfoxide[b] Amyloglucosidase and pancreatic α-amylase	Stirring for 15 h at 37 °C, pH 5.0	Total RS
Akerberg *et al.* (1998)	Same as above	Same as above	Same as above	KOH[b] Thermostable α-amylase Amyloglucosidase	Stirring for 16 h at 40 °C, pH 5.0	Same as above
McCleary & Monaghan (2002)	Milling	None	Same as above	KOH[b] Amyloglucosidase	Shaking for 16 h at 37 °C, pH 6.0	Sum of RS_2 and RS_3

[a]RS, resistant starch; RS_1, physically inaccessible starch; RS_2, resistant granules; RS_3, retrograded starch.
[b]KOH and dimethyl sulfoxide are used as resistant starch solubilizing agents.

steps that permit the estimation of available starch and dietary fiber along with resistant starch (**Table 4**).

Recently, the most commonly used *in vitro* methods were extensively evaluated and a simplified version was proposed (McCleary method). Here, samples are treated with an amyloglucosidase–pancreatic amylase mixture only and the insoluble residue, after washing with ethanol, is dispersed with KOH, followed by the amyloglucosidase step to yield glucose. This protocol has been accepted by AOAC International (AOAC method 2002.02) and the American Association of Cereal Chemists (AACC method 32-40) (**Table 4**).

Regarding the quantification of the resistant fractions in modified starches, care must be taken because some nondigestible fractions are soluble in water and they can be lost during washing steps. This is particularly important with pregelatinized starches and pyrodextrins. One suitable way to look at the impact of the modification on the starch availability is measure total starch before and after the modification.

Dietary Intake

It is very difficult to assess resistant starch intake at present, because there are not enough data on the resistant starch content of foods. In addition, as the resistance of the starch to digestion depends on the method of cooking and the temperature of the food as eaten, the values gained from looking at old dietary intake data may be misleading. Despite this, an average value for resistant starch intake across Europe has been estimated as $4.1\,g\,day^{-1}$. Figures comparable with this estimation have been made in other countries, for instance, Venezuela ($4.3\,g\,day^{-1}$). It is very difficult to separate the benefits of slowly, but completely, digestible starches from those that are resistant. In some groups like small children, whose small intestinal digestive capacity is reduced, the very same food may provide more starch that is resistant to digestion than it would in normal adults.

Quantification of modified starch intake is even more difficult. First, food labels do not usually provide information about the nature of the modification used. Second, the commonly used method to estimate resistant starch can underestimate any nondigestible fractions that became soluble in water because of the modification. At present, there is no data available on how much modified starch is eaten.

Fermentation in the Colon

The main nutritional properties of resistant starch arise from its potential fermentation in the colon. The diverse and numerous colonic microflora ferments unabsorbed carbohydrates to short-chain fatty acids (SCFA), mainly acetate, propionate and butyrate, and gases (H_2, CO_2, and CH_4). Acetate is the main SCFA produced (50–70%) and is the only one to reach peripheral circulation in significant amounts, providing energy for muscle and other tissues. Propionate is the second most abundant SCFA and is mainly metabolized by the liver, where its carbons are used to produce glucose (via gluconeogenesis). Propionate has also been associated with reduced cholesterol and lipid synthesis. Finally, butyrate is mainly used as fuel by the colonic enterocytes, but has been shown *in vitro* to have many potential anticancer actions, such as stimulating apoptosis (i.e., programed cell death) and cancer cell differentiation (i.e., increasing expression of normal cell function), and inhibiting histone deacetylation (this protects the DNA). Resistant starch fermentation has been shown to increase the molar proportion of butyrate in the colon.

The main physiological effects of digestion and fermentation of resistant starch are summarized in **Table 5**. However, most of these effects have

Table 5 Physiological effects of resistant starch intake

Energy	8–13 kJ g^{-1}; cf. 17 kJ g^{-1} for digestible starches
Glycemic and insulinemic response	Depends on food, e.g., legumes (high in RS_1) and amylose-rich starchy foods (which tend to produce RS_3 on cooking) increase glucose tolerance, but cornflakes and cooked potatoes, both with high and similar glycemic indexes, have different resistant starch content
Lipid metabolism	Decreases plasma cholesterol and triacylglyceride levels in rat, but not in humans
Fermentability	Complete, although some RS_3 are more resistant
SCFA production	Increased production, especially butyrate
CO_2 and H_2 production	Occurs
Colonic pH	Decreased, especially by lactate production
Bile salts	Deoxycholate, a secondary bile salt with cytotoxic activity, precipitated due to the low pH
Colon cell proliferation	Stimulated in proximal colon, but repressed in distal colon; may be mediated by butyrate
Fecal excretion	At high dose, fecal bulk increases due to an increase in bacteria mass and water retention
Transit time	Increased intestinal transit at high dose
Nitrogen metabolism	Increased bacterial nitrogen and biomass
Minerals	May increase calcium and magnesium absorption in large intestine
Disease prevention	Epidemiological studies suggest prevention against colorectal cancer and constipation

been observed with a resistant starch intake of around 20–30 g day^{-1}, which represents from 5 to 7 times the estimated intake for the European population.

Oligosaccharides

Definition and Classification

Oligosaccharides are carbohydrate chains containing 3–10 sugar units. However, some authors also include carbohydrates with up to 20 residues or even disaccharides. Oligosaccharides can be made of any sugar monomers, but most research has been carried out on fructooligosaccharides (e.g., oligofructose) and galactooligosaccharides (e.g., raffinose, human milk oligosaccharides). Few oligosaccharides are hydrolyzed and absorbed in the small intestine (e.g., maltotriose), but nearly all enter the colon intact (nondigestible oligosaccharides). **Table 6** shows several examples of oligosaccharides (and disaccharides, for comparison purposes), their chemical structure, and source.

Dietary Sources and Intake

The first source of oligosaccharides in the human diet is mother's milk, which contains approximately 12 g l^{-1}. In human breast milk, there are over 100 different oligosaccharides with both simple and complex structures. They are composed of galactose, fucose, sialic acid, glucose, and *N*-acetylglucosamine. Most are of low molecular weight, but a small proportion are of high molecular weight. Ninety per cent of breast milk oligosaccharides are neutral; the remainder are acidic. Interestingly, the nature of these oligosaccharide structures is determined by the mother's blood group. These oligosaccharides may have

Table 6 Chemical structure and source of sugars and oligosaccharides

Common name	*Simplified structure*[a]	*Source*	*NDO*[b]
Sugars (disaccharides)			
Lactose	Galβ1 → 4Glc	Milk, milk products	No
Maltose	Glcα1 → 4Glc	Glucose syrups, hydrolysis of starch	No
Sucrose	Fruβ2 → 1Glc	Table sugar	No
Cellobiose	Glcβ1 → 4Glc	Hydrolysis of cellulose	Yes
Trehalose	Glcα1 → 1Glc	Mushrooms, yeast	No
Melibiose	Galα1 → 6Glc	Hydrolysis of raffinose	Yes
Gentiobiose	Glcβ1 → 6Glcβ	Plant pigments, like saffron	Yes
Trisaccharides			
Maltotriose	Glcα1 → 4Glcα1 → 4Glc	Glucose syrups, hydrolysis of starch	No
Umbelliferose	Galα1 → 2Glcα1 → 2Fruβ	Plant tissues	Yes
Raffinose	Galα1 → 6Glcα1 → 2Fruβ	Legume seeds	Yes
Planteose	Galα1 → 6Fruβ2 → 1Glc	Plant tissues	Yes
Sialylα(2–3)lactose	NeuAcα2 → 3Galβ1 → 4Glc	Human milk	Yes
Tetrasaccharides			
Stachyose	Galα1 → 6Galα1 → 6Glcα1 → 2Fruβ	Legume seeds	Yes
Lychnose	Galα1 → 6Glcα1 → 2Fruβ1 → 1Gal	Plant tissues	Yes
Isolychnose	Galα1 → 6Glcα1 → 2Fruβ3 → 1Gal	Plant tissues	Yes
Sesamose	Galα1 → 6Galα1 → 6Fruβ2 → 1Glc	Plant tissues	Yes
Pentasaccharides			
Verbacose	Galα1 → 6Galα1 → 6Galα1 → 6Glcα1 → 2Fruβ	Plant tissues	Yes
Lacto-*N*-fucopentaose I	Fucα1 → 2Galβ1 → 3GlcNAcβ1 → 3Galβ1 → 4Glc	Human milk	Yes
Lacto-*N*-fucopentaose II	Galβ1 → 3[Fucα1 → 4]GlcNAcβ1 → 3Galβ1 → 4Glc	Human milk	Yes
Fructans			
Oligofructose	[Fruβ2 → 1]Fruβ2 → 1Glc with 1–9 [Fruβ2 → 1] residues	Hydrolysis of inulin or synthesis from sucrose	Yes
Inulin (polysaccharide)	[Fruβ2 → 1]Fruβ2 → 1Glc with 10–64 [Fruβ2 → 1] residues	Artichokes	Yes

[a]Fru, D-fructose; Fuc, L-fucose; Gal, D-galactose; Glc, D-glucose; GlcNAc, *N*-acetylglucosamine; NeuAc, *N*-acetylneuraminic acid (or sialic acid).
[b]NDO, nondigestible oligosaccharides.

important function in the small intestine, where they can bind to the mucosa or to bacteria, interfering with pathogenic bacterial attachment and thus acting as anti-infective agents. As they are nondigestible, they enter the colon and may act as a major energy for the colonic microflora and promote the growth of typical lactic acid bacteria that are characteristic of the normal breast-fed infant. More recently, oligosaccharides have been added to some infant formulas to mimic the actions of those in human milk. Recently, several studies have shown that these promote the growth of bifidobacteria in feces and make the stools more like those of breast-fed infants in terms of consistency, frequency, and pH.

In adults, the main dietary sources of oligosaccharides are chicory, artichokes, onions, garlic, leeks, bananas, and wheat. However, much research has been carried out on purified or synthetic oligosaccharide mixtures, mostly fructooligosaccharides derived from inulin. The normal dietary intake of oligosaccharides is difficult to estimate, as they are not a major dietary component. Around $3\,g\,day^{-1}$ has been suggested in the European diet. However, with the increasing information on the health benefits of isolated oligosaccharide sources (see below) they are being incorporated into functional foods.

Analysis

In general, oligosaccharides are a less heterogeneous group of compounds than resistant starches. Almost all nondigestible oligosaccharides (some fructooligosaccharides are an exception) are soluble in 80% (v/v) ethanol solution, which makes them relatively easy to isolate from insoluble components. Liquid chromatography, more specifically high-performance anion exchange chromatography (HPAEC), has been extensively employed not only to separate mixtures of different oligosaccharides, but also to separate, identify, and quantify individual carbohydrate moieties after appropriate hydrolysis of the oligosaccharide to its individual monomers. A more comprehensive study of the oligosaccharide structure can be achieved using more sophisticated techniques, like nuclear magnetic resonance and mass spectrometry. However, from a nutritional viewpoint, where simpler methods are needed for quality control and labeling purposes, HPAEC is usually applied to quantify the monomers (and dimers) present before and after hydrolysis of the studied oligosaccharide with appropriate enzymes and then the oligosaccharide level is worked out by difference.

Fermentation in the Colon and Health Benefits

Most oligosaccharides escape digestion in the small intestine and are fermented by the colonic bacteria. They are rapidly fermented resulting in a low pH and have been shown to increase the survival of so-called probiotic organisms, i.e., lactobacilli and bifidobacteria. Probiotic bacteria have been show to have strain-specific effects, including reduction in duration of rotavirus and other infective diarrhea and reduction in symptoms of atopic eczema. They may also have some anticarcinogenic effects, but these have not been demonstrated in human *in vivo* studies. This action of oligosaccharides to promote the growth of bifidobacteria and lactobacilli defines them as prebiotics. Some studies are now investigating the synergistic effects of probiotics mixed with prebiotics. These mixtures are termed synbiotics. In addition to these actions, some oligosaccharides have similar health benefits to fermentable dietary fiber and resistant starch by increasing colonic fermentation, production of SCFA (especially butyrate), and reduction in colonic pH.

Resistant Starch, Oligosaccharides, or Just Dietary Fiber?

There has been much debate of the definition of dietary fiber and in particular whether it should include carbohydrates other than nonstarch polysaccharides. Recently, the American Association of Cereal Chemists (AACC) proposed a new definition of dietary fiber, which would include both oligosaccharides and resistant starch as well as associated plant substances. This new definition would also require complete or partial fermentation and demonstration of physiological effects such as laxation, and reduction in blood glucose or blood cholesterol. A similar approach to include beneficial physiological effects is also proposed by the Food and Nutrition Board of the US Institute of Medicine.

Thus, it is being increasingly recognized that oligosaccharides, resistant starch, and nonstarch polysaccharides are very similar especially in their effects on gut physiology and colonic fermentation. A comparison of their actions is summarized in **Table 7**. This inclusion of resistant starch and oligosaccharides in the definition of dietary fiber could have major implications for food labeling.

Table 7 The physiological effects of resistant starch, oligosaccharides, and dietary fiber

Physiological effect	*Resistant starch*	*Oligosaccharides*	*Dietary fiber*
Energy supply	8–13 kJ g^{-1}	8–13 kJ g^{-1}	8–13 kJ g^{-1}
Increased glucose tolerance	Some foods	No	Some NSP[a]
Decreased plasma cholesterol and triacylglyceride levels	No	Not known	Some NSP
Fermentability	Complete	Complete	Variable
Production of SCFA	Yes	Yes	Yes
Increased butyrate production	High	High	Variable
CO_2 and H_2 production	Yes	Yes	Variable
Decreased fecal pH	Yes	Yes	Some NSP
Decreased production of deoxycholate	Yes	Yes	Some NSP
Increased colonocyte proliferation	Yes	Yes	Yes
Increased fecal bulk	At high dose	No	Variable
Faster whole gut transit time	At high dose	No	Yes
Increased bacterial nitrogen and biomass	Yes	Yes	Yes
Reduced mineral absorption in small intestine	No	No	Some NSP
Increased mineral absorption in large intestine	Yes	Yes	Some NSP
Possible prevention of colorectal cancer	Yes	Not known	Yes

[a]NSP, nonstarch polysaccharide.

See also: **Breast Feeding**. **Cereal Grains**. **Colon**: Structure and Function. **Dietary Fiber**: Physiological Effects and Effects on Absorption; Potential Role in Etiology of Disease; Role in Nutritional Management of Disease. **Legumes**. **Microbiota of the Intestine**: Prebiotics.

Further Reading

Blaut M (2002) Relationship of prebiotics and food to intestinal microflora. *European Journal of Nutrition* **41**(supplement 1): i11–i16.

Bornet FRJ, Brouns F, Tashiro Y, and Duvillier V (2002) Nutritional aspects of short-chain fructooligosaccharides: natural occurrence, chemistry, physiology and health implications. *Digestive and Liver Disease* **34**(supplement 2): S111–S120.

Champ M, Langkilde A-M, Brouns F, Kettlitz B, and Collet YL (2003) Advances in dietary fibre characterisation. 1. Definition of dietary fibre, physiological relevance, health benefits and analytical aspects. *Nutrition Research Reviews* **16**: 71–82.

Delzenne NM (2003) Oligosaccharides: state of the art. *Proceedings of the Nutrition Society* **62**: 177–182.

DeVries JW (2003) On defining dietary fibre. *Proceedings of the Nutrition Society* **62**: 37–43.

Ghisolfi J (2003) Dietary fibre and prebiotics in infant formula. *Proceedings of the Nutrition Society* **62**: 183–185.

Greger JL (1999) Non-digestible carbohydrates and mineral bioavailability. *Journal of Nutrition* **129**(supplement S): 1434S–1435S.

Kunz C, Rudloff S, Baier W, Klein N, and Strobel S (2000) Oligosaccharides in human milk: Structural, functional, and metabolic aspects. *Annual Review of Nutrition* **20**: 699–722.

McCleary BV (2003) Dietary fibre analysis. *Proceedings of the Nutrition Society* **62**: 3–9.

Roberfroid M (2002) Functional food concept and its application to prebiotics. *Digestive and Liver Disease* **34**(supplement 2): S105–S110.

Roberfroid MB (2003) Inulin and oligofructose are dietary fibres and functional food ingredients. In: Kritchevsky D, Bonfield CT, and Edwards CA (eds.) *Dietary Fiber in Health and Disease: 6th Vahouny Symposium*, pp. 161–163. Delray Beach, FL: Vahouny Symposium.

Topping DL, Fukushima M, and Bird AR (2003) Resistant starch as a prebiotic and synbiotic: state of the art. *Proceedings of the Nutrition Society* **62**: 171–176.

Carcinogens *see* **Cancer**: Carcinogenic Substances in Food

CAROTENOIDS

Contents

Chemistry, Sources and Physiology

B K Ishida and G E Bartley, Agricultural Research Service, Albany, CA, USA

Published by Elsevier Ltd.

Chemistry

Structure

Most carotenoids are 40-carbon isoprenoid compounds called tetraterpenes. Isoprenoids are formed from the basic five-carbon building block, isoprene (**Figure 1**). In nature, carotenoids are synthesized through the stepwise addition of isopentenyl diphosphate (IPP) units to dimethylallyl diphosphate (DMAPP) to form the 20-carbon precursor geranylgeranyl diphosphate (GGPP). Two molecules of GGPP are combined to form the first carotenoid in the biosynthetic pathway, phytoene, which is then desaturated, producing 11 conjugated double bonds to form lycopene, the red pigment in ripe tomato fruit (**Figure 1**). Nearly all other carotenoids can be derived from lycopene. Lycopene can be cyclized on either or both ends to form α- or β-carotene, and these in turn can be oxygenated to form xanthophylls such as β-cryptoxanthin, zeaxanthin, or lutein (**Figure 1** and **Figure 2**). Carotenoids having fewer than 40 carbons can result from loss of carbons within the chain (norcarotenoids) or loss of carbons from the end of the molecule (apocarotenoids). Longer carotenoids, homocarotenoids (C45–C50), are found in some bacterial species. The alternating double bonds along the backbone of carotenoid molecules form a polyene chain, which imparts unique qualities to this group of compounds. This alternation of single and double bonds also allows a number of geometrical isomers to exist for each carotenoid (**Figure 1**). For lycopene, the theoretical number of steric forms is 1056; however, when steric hindrance is considered, that number is reduced to 72. In nature most carotenoids are found in the all-*trans* form although mutants are known in plants, e.g., *Lycopersicon esculentum* (Mill.) var. Tangerine tomato, and eukaryotic algae that produce poly-*cis* forms of carotenoids. The mutant plant is missing an enzyme, carotenoid isomerase (CRTISO), which catalyzes the isomerization of the *cis* isomers of lycopene and its precursors to the all-*trans* form during biosynthesis. Light can also cause *cis* to *trans* isomerization of these carotenoids depending upon the surrounding environment. The isomeric form determines the shape of the molecule and can thus change the properties of the carotenoid affecting solubility and absorbability. *Trans* forms of carotenoids are more rigid and have a greater tendency to crystallize or aggregate than the *cis* forms. Therefore, *Cis* forms may be more easily absorbed and transported. End groups such as the β or ε rings of α-carotene and β-carotene and the amount of oxygenation will also affect carotenoid properties.

Chemical Properties

In general, carotenoids are hydrophobic molecules and thus are soluble only in organic solvents, having only limited solubility in water. Addition of hydroxyl groups to the end groups causes the carotenoid to become more polar, affecting its solubility in various organic solvents. Alternatively, carotenoids can solubilize in aqueous environments by prior integration into liposomes or into cyclic oligosaccharides such as cyclodextrins.

In general, carotenoid molecules are very sensitive to elevated temperatures and the presence of acid, oxygen, and light when in solution, and are subject to oxidative degradation.

Electronic Properties

What sets carotenoids apart from other molecules and gives them their electrochemical properties is the conjugated double bond system. In this alternating double and single bond system, the π-electrons are delocalized over the length of the polyene chain. This polyene chain or chromophore imparts the characteristic electronic spectra and photophysical and photochemical properties to this group of molecules. The highly delocalized π-electrons require little energy to reach an excited state so that light energy can cause a transition. The length of the conjugated polyene or

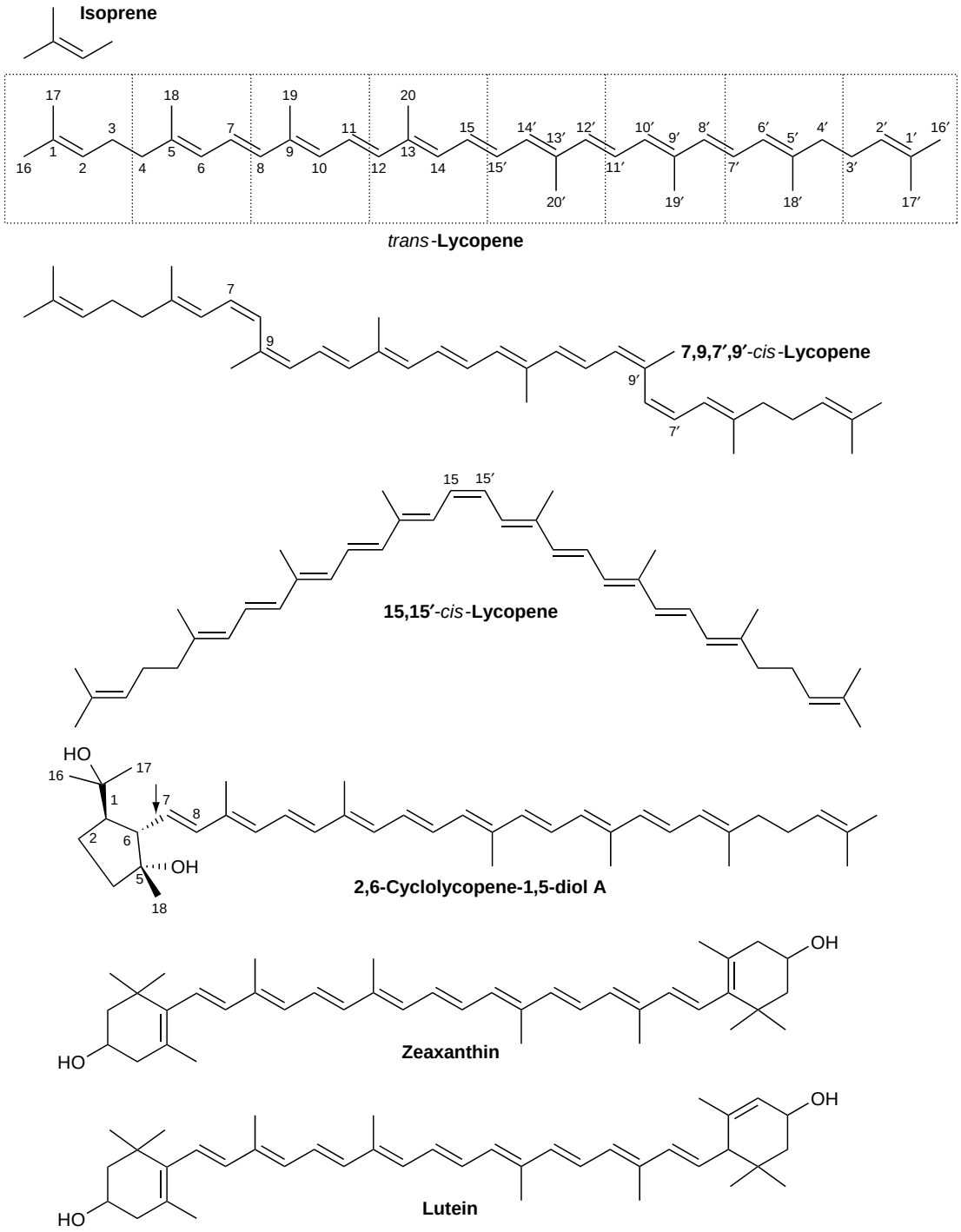


Figure 1 Carotenoid structures. Lycopene is shown with numbered carbons. The down arrow on 2,6-cyclolycopene-1,5-diol A indicates the only difference from the B isomer.

chromophore affects the amount of energy needed to excite the π-electrons. The longer the conjugated system, the easier it is to excite, so longer wavelengths of light can be absorbed. The result is that phytoene, having three conjugated double bonds is colorless, and phytofluene, having five, is colorless, but fluoresces green under UV light. Zeta-carotene has seven, absorbs light at ~400 nm and appears yellow, while neurosporene has nine, absorbs light at ~451, and appears orange, and lycopene has eleven conjugated double bonds, absorbs at ~472, and appears red. The polyene chain also allows transfer of singlet or triplet energy.

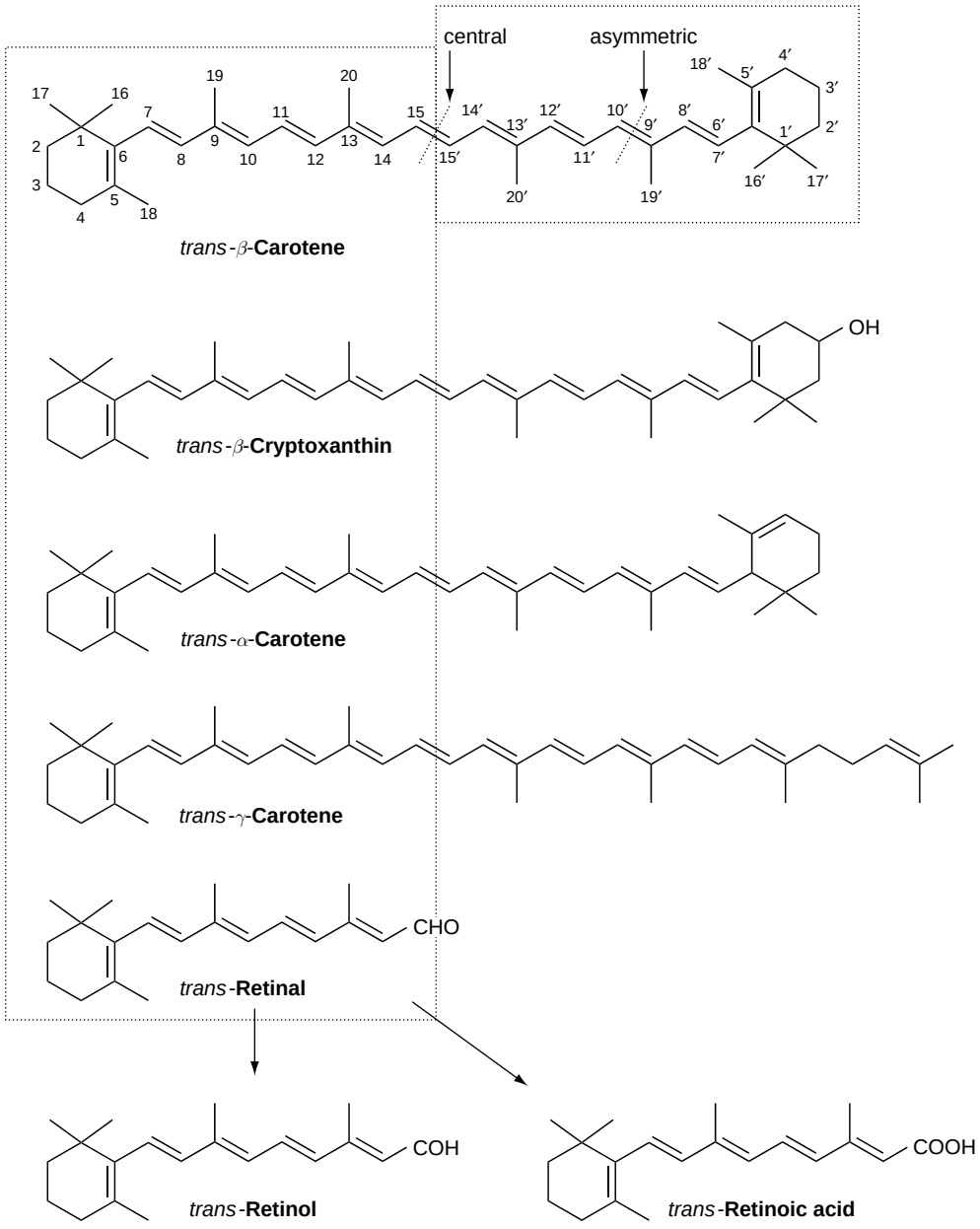


Figure 2 Provitamin A carotenoids. Dotted lines indicate the provitamin A moiety.

Reactions

Light and Chemical Energy

The basic energy-transfer reactions are assumed to be similar in plants and animals, even though environments differ. Excess light can cause excitation of porphyrin molecules (porphyrin triplets). These triplet-state porphyrin molecules can transfer their energy to oxygen-forming singlet oxygen, $^1O_2$. Singlet oxygen can damage DNA and cause lipid peroxidation, thereby killing the cell. Carotenoids, having nine or more conjugated double bonds, can prevent damage by singlet oxygen through: (1) transfer of triplet energy from the excited porphyrin to the carotenoid, forming a carotenoid triplet, which would be too low in energy for further transfer and would simply dissipate as heat; or (2) singlet oxygen energy could transfer to the carotenoid, also forming a triplet carotenoid, dissipating heat, and returning to the ground state. This ability to quench sensitized triplets has been useful in treating protoporphyria (PP) and congenital erythropoietic porphyria (CEP) in humans. Porphyrias are disorders resulting from a defect in heme biosynthesis. Precursor porphyrins accumulate and can be sensitized to the singlet state and drop to the lower triplet state. The triplet state is longer-lived

and thus more likely to react with other molecules such as oxygen to form singlet oxygen, which can cause cellular damage. Because β-carotene can transfer and dissipate either sensitized triplet or singlet oxygen energy it has been used to treat these disorders.

Light absorption and possibly scavenging of destructive oxygen species by the xanthophylls lutein and zeaxanthin are also important in the macula of the primate eye. Lutein and two isomers of zeaxanthin are selectively accumulated in the macula, creating a yellow area of the retina responsible for high visual acuity (smaller amounts are also found in the lens). Both carotenoids absorb light of about 450 nm 'blue light,' thus filtering light to the light receptors behind the carotenoid layer in the macula. Filtering blue light can reduce oxidative stress to retinal light receptors and chromatic aberration resulting from the refraction of blue light. A similar filter effect may occur in the lens, but the concentration of the xanthophylls is much lower, and further protection occurs with age when the lens yellows. Whether scavenging of destructive oxygen species by these carotenoids is useful here is unproven, but the retina is an area of higher blood flow and light exposure than other tissues.

Cleavage to Vitamin A

Provitamin A carotenoids are sources of vitamin A. Of the 50–60 carotenoids having provitamin A activity, β-carotene, β-cryptoxanthin, and α-carotene are the major sources of vitamin A nutrition in humans, β-carotene being the most important (**Figure 2**). Vitamin A (retinol) and its derivatives retinal and retinoic acid perform vital functions in the vertebrate body. Retinal (11-*cis* retinal) combined with opsin functions in the visual system in signal transduction of light reception. Retinol and retinoic acid function in reproduction (spermatogenesis), growth regulation (general development and limb morphogenesis), and cell differentiation. Provitamin A activity requires at least one unsubstituted β-ionone ring, the correct number and orientation of methyl groups along the polyene backbone, and the correct number of conjugated double bonds, preferably in the *trans*-isomer orientation. Two pathways for the formation of retinal from β-carotene have been proposed. First, central cleavage by which β-carotene 15,15′-mono- or dioxygenase catalyzes β-carotene cleavage to form two molecules of retinal, which can then be converted to retinol or retinoic acid (**Figure 2**). Some debate on the mechanism of the β-carotene central cleavage enzyme still exists, but evidence leans towards activity as a monooxygenase, not a dioxygenase. Alternatively, in the eccentric cleavage pathway β-carotene can be cleaved at any of the double bonds along the polyene backbone (other than the 15-15′double bond). Products of these reactions (apocarotenals) are then further metabolized to retinoic acid and retinol. An asymmetric cleavage enzyme has recently been cloned that cleaves β-carotene at the 9′-10′-double bond to form β-ionone and β-apo-10′-carotenal. The discovery of this enzyme indicates at least some eccentric cleavage occurs in vertebrates. This eccentric cleavage process has been proposed to occur during more oxidative conditions, while central cleavage would predominate under normal physiological conditions. Central cleavage is considered to be the major pathway because of the scarcity of eccentric cleavage products detected *in vivo*.

Radical Reactions

Excess amounts of radicals, molecules having unpaired electrons, e.g., peroxyls ($ROO^{\bullet}$), can be created in tissues exogenously, e.g., by light exposure, or endogenously, e.g., by overexercising. Radicals react with lipids, proteins, and DNA causing damage, which possibly contributes to disease symptoms and aging. The special properties of the polyene chain make carotenoids susceptible to electrophilic attack, resulting in formation of resonance-stabilized radicals that are less reactive.

Three possible reactions can occur with carotenoids.

1. Adduct formation ($CAR + R^{\bullet} \rightarrow R\text{-}CAR^{\bullet}$); these products should be stable because of resonance in the polyene structure. If the radical were a lipid peroxyl, this reaction ($CAR + ROO^{\bullet} \rightarrow ROO\text{-}CAR^{\bullet}$) would prevent further propagation (chain-breaking).
2. Hydrogen atom abstraction ($CAR + R^{\bullet} \rightarrow CAR^{\bullet} + RH$), where a hydrogen atom is taken from the carotenoid allylic to the polyene chain, leaving a resonance-stabilized carotenoid radical.
3. Electron transfer ($CAR + R^{\bullet} \rightarrow CAR^{\bullet +} + R^{-}$), which has been reported in plant and cyanobacterial photosystems using laser flash photolysis of Photosystem II.

In many cases, the products formed are colorless, thus revealing the bleaching effect of many oxidants on carotenoids. Further oxidation of the carotenoid or carotenoid radical can occur as in studies of soybean (*Glycine max*) and recombinant pea (*Pisum sativum*) lipoxygenase-mediated cooxidation of carotenoids and polyunsaturated fatty acids. Approximately 50 breakdown products of β-carotene were detected. This large number of products seems to indicate a random attack along the polyene chain of β-carotene by a linoleoylperoxyl radical. Studies using potassium

permanganate, a metalloporphyrin (a P450 enzyme center mimic), and autooxidation have been performed with lycopene, resulting in formation of a number of apo-lycopenals and apo-lycopenones. However, only two metabolites of lycopene have been identified in human plasma, 2,6-cyclolycopene-1,5 diols A and B (**Figure 1**). Additionally, seven metabolites of the carotenoids lutein and zeaxanthin have been detected in human tissues.

Prooxidant Behavior

The ability to quench singlet oxygen, porphyrin triplet energies, and free radical reactions are examples of the antioxidant nature of carotenoids. An *in vitro* study showed that, at low partial pressures of oxygen (pO_2), β-carotene consumed peroxy radicals efficiently as in: $\mathrm{CAR} + \mathrm{ROO}^{\bullet} \rightarrow \mathrm{CAR}^{\bullet+} + \mathrm{ROO}$. At higher pO_2, however, β-carotene became a prooxidant through autooxidation. Recently, experiments in intact murine normal and tumor thymocytes showed that β-carotene lost its antioxidant potency at higher pO_2, and the effect was more pronounced in tumor cells. It is still unclear, however, whether some effects of carotenoid behavior at higher pO_2 are due to prooxidant activity or simply lack of antioxidant ability. Prooxidant effects of β-carotene have also been used to explain results from intervention trials of β-carotene supplementation in diets of smokers or individuals suffering from asbestosis where the incidence of carcinogenesis was higher in those individuals taking the β-carotene supplement. Generation of deleterious oxidation products from β-carotene reaction with reactive oxygen species in tobacco smoke or as a result of asbestosis has been proposed. Interference with retinoid signaling was also considered. However, whether those effects were due to prooxidant behavior or lack of antioxidant ability is still unclear.

Dietary Sources

Carotenoids cannot be synthesized by humans; therefore they must be obtained from dietary sources. These are primarily highly pigmented red, orange, and yellow fruits and vegetables. The carotenoid lycopene is red; however, not all red fruits and vegetables contain lycopene. For example, the red in strawberries, apples, and cherries is a result of their anthocyanin content; whereas, tomatoes, watermelon, and pink grapefruit derive their red color from lycopene. The carotenoids β-carotene, β-cryptoxanthin, lutein, zeaxanthin, and violaxanthin are yellow to orange, and phytoene and phytofluene are colorless. Green, leafy vegetables also contain carotenoids, whose colors are masked by the green color of chlorophyll. **Table 1** lists carotenoids found in fruits and vegetables. Smaller amounts are also available from animal sources

Table 1 Carotenoid content (μg per g fresh weight) of fresh fruit and vegetables

Carotenoid	*Concentration (μg per g fresh weight)*	*Source*
Lycopene	380–3054	Gac (*Momordica cochinchinensis*, Spreng) aril
	179–483	Autumn olive (*Elaeagnus umbellate*)
	27–200	Tomato
	23–72	Watermelon
	53	Guava
	19–40	Papaya
	8–33	Grapefruit, pink
β-Carotene	101–770	Gac aril
	49–257	Carrot, orange
	16–216	Cantaloupe
	15–92	Kale
	0.5–92	Sweet potato
	47–89	Spinach
	46	Turnip greens
	26–64	Apricot
	22–58	Gac mesocarp
	3–70	Tomato
	42	Squash, butternut
	40	Swiss chard
	14–34	Mango
	33	Collards
	4–10	Grapefruit, pink
	0.51–1.2*	Orange (*blood)
Lutein	64–150	Kale
	6–129	Mango
	108	Parsley
	39–95	Spinach
	33–51	Collards
	15–28	Broccoli
	27	Chinese cabbage
	26	Watercress
	25	Pepper, orange
	24	Squash, butternut
	1–7	Tomato
Zeaxanthin	16–85	Pepper, orange
	43	Gou Qi Zi (*Lycium barabarum*)
	9	Gac aril
	22	Pepper, red
	7	Watercress
	1–5	Spinach
	5	Parsley
	5	Japanese persimmon
	1–3	Kale
	3	Squash, butternut
	0.4	Broccoli
	0.03–0.5	Tomato

Continued

Table 1 Continued

Carotenoid	*Concentration (μg per g fresh weight)*	*Source*
Lutein + zeaxanthin	71–3956	Kale
	119	Spinach
	84	Turnip greens
	26	Lettuce
	24	Broccoli
	21	Squash, zucchini
	16	Brussel sprouts
	8	Japanese persimmon
	7	Watercress
	6	Beans, green snap
	5	Tangerine
β-Cryptoxanthin	22	Pepper, sweet red
	14	Japanese persimmon
	11	Starfruit
	0.7–9	Pepper, chili
	2–8	Pepper, orange
	0.5–5	Tangerine
	4	Cilantro
	1.4	Papaya
	1	Watermelon
α-Carotene	20–206	Carrot
	8	Squash, butternut
	2	Collards
	1	Tomato
	0.7–0.9	Beans, green snap
	0.5	Swiss chard

such as ocean fish and dairy products. The pink color of salmon, for example, is derived from the xanthophylls, astaxanthin and canthaxanthin, which they obtain from eating small crustaceans and krill. Lutein imparts its yellow-orange color to eggs, and milk, butter, and cheese contain retinols and β-carotene. Carotenoids, such as lutein from marigolds and bixin (red color) from annatto, are also used widely as colorants in processed foods to make them more attractive.

Concentrations of carotenoids in fruit and vegetable sources vary, resulting from differences in conditions under which they are grown (temperature, amount of sunlight, degrees of stress from extremes in climate such as drought, heat, and cold), genotype, and maturity or ripeness. The carotenoid content in animal sources depends upon amounts contained in animal feeds and seasons of the year, which affect the availability of carotenoid-containing plants eaten by grazing animals.

Human diets and tissues contain six carotenoids in significant amounts (listed in **Table 1**). Lycopene is typically the carotenoid consumed in greatest amounts in Western diets. Per capita intakes in Europe and North America average from 1.6 to more than 18 mg lycopene per day. More than 85% of the lycopene in North American diets comes from tomato products, which also contain significant amounts of other carotenoids (α- and β-carotene and lutein/zeaxanthin), as well as vitamins C, A, and E, and potassium and folic acid. (Flavonoids are also found in tomato skin; thus, cherry tomatoes contain higher concentrations.) In the US, the annual per capita consumption of tomatoes by 1999 averaged about 17.6 pounds of fresh and 72.8 pounds of processed tomatoes.

Effects of Storage and Processing

Carotenoids are susceptible to oxidative degradation and isomerization resulting from storage and processing conditions. These reactions result in both loss of color and biological activity and formation of often unpleasant volatile compounds. Degradation occurs upon exposure to oxygen and is accelerated by the presence of substances such as metals, enzymes, unsaturated lipids, and prooxidants; exposure to light; and conditions that destroy cell wall and ultrastructural integrity. Heating can promote isomerization of the naturally occurring all-*trans* to various *cis* isomers. This process then affects bioavailability of the carotenoid. Processing also affects bioavailability by macerating tissues, destroying or weakening cell ultrastructure, denaturing or weakening complexes with proteins, and cleaving ester linkages, thereby releasing carotenoids from the food matrix.

Processed foods are frequently fortified with carotenoids to increase nutritive value and/or enhance attractiveness. For example, annatto, an extract from the seeds of the *Bixa orella* tree, containing the carotenoids bixin and norbixin, is added to butter, margarine, and processed cheese to give a yellow-orange color to these products. Tomato oleoresin is added to processed tomato products, increasing lycopene content while enhancing their attractive red color.

Physiology

Digestion

Numerous factors affect the intestinal absorption of carotenoids. Digestion of food in the stomach increases accessibility of carotenoids for absorption by maceration in HCl and digestive enzymes. The acidic environment of the stomach helps to disrupt cell walls and other cellular ultrastructure of raw fruits and vegetables and causes further breakdown of cooked foods to release carotenoids from food matrices in which they are contained or bound.

Carotenoids in green leafy vegetables are found in chloroplasts; those in fruit are located in chromoplasts. Absorption studies comparing plasma levels of β-carotene and retinol after consuming fruit vs. green leafy vegetables showed that β-carotene is more efficiently absorbed from fruit, indicating that chloroplasts (or the bonds linking chloroplast proteins and carotenoids) are more resistant to disruption in the digestive tract than chromoplasts. Thus, the location of a carotenoid in the cell affects its accessibility.

Carotenoid isomerization can occur in the acidic gastric milieu. Lycopene present in fruits and vegetables occurs almost exclusively as the all-*trans* isomer, but is converted to *cis* isomers, which seem to be more bioavailable. Plasma and tissue profiles show that *cis* isomers make up more than 50% of the total lycopene present. On the other hand, studies show that no *trans/cis* isomerization of β-carotene occurs in the stomach. In fact, evidence has been found for transfer of a significant portion of both β- and α-carotene to the fat phase of the meal in the stomach, which would increase bioavailability of these carotenoids for absorption. No studies are available relating isomerization to bioavailability of other carotenoids.

Absorption and Transport

Because carotenoids are hydrophobic molecules, they are associated with lipophilic sites in cells, such as bilayer membranes. Polar substituents such as hydroxyl groups decrease their hydrophobicity and their orientation with respect to membranes. Lycopene and β-carotene are aligned parallel to membrane surfaces to maintain a hydrophobic environment, whereas the more polar xanthophylls lutein and zeaxanthin become oriented perpendicular to membrane surfaces to keep their hydroxyl groups in a more hydrophilic environment. These differences can affect the physical nature of a membrane as well as its function. Carotenoids can form complexes with proteins, which would aid them in moving through an aqueous environment. They can also interact with hydrophobic regions of lipoproteins. Carotenoproteins have been found mainly in plants and invertebrates, but intracellular β-carotene-binding proteins have been found in bovine liver and intestine and in livers of the rat and ferret. In addition, a xanthophyll-binding protein has been found in human retina and macula. Carotenoids are also present in nature as crystalline aggregates (lycopene in chromoplasts) or as fine dispersions in aqueous media (β-carotene in oranges).

In the intestinal lumen (**Figure 3**) where carotenoids are released from the food matrix, cleavage of carotenoproteins and fatty acid esters by carboxylic ester hydrolase, which is secreted by the pancreas, can occur. Carotenoids are then solubilized into lipid micelles. These hydrophobic compounds are thus more efficiently absorbed when accompanied by at least a small amount of fat. The amount of fat for optimal carotenoid absorption seems to differ among carotenoids. For example, lutein esters require more fat for optimal absorption than β-carotene. These differences have not been quantified for each carotenoid. In addition, the presence of a nonabsorbable, fat-soluble component was shown to decrease carotenoid absorption. Sucrose polyester, a nonabsorbable fat replacer decreased carotenoid

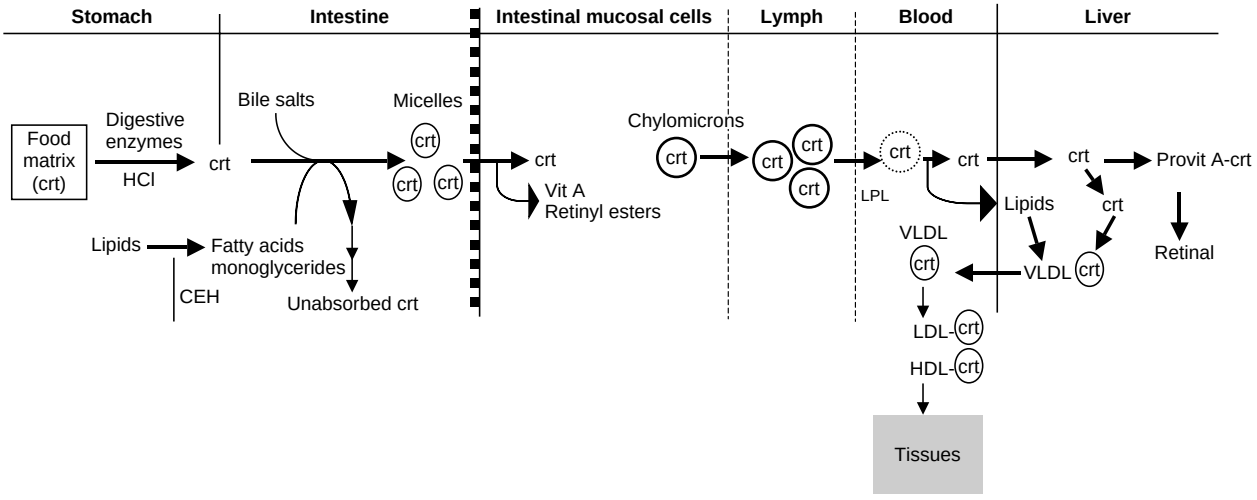


Figure 3 Factors affecting digestion, absorption, metabolism, and transport of carotenoids. crt, carotenoids; CEH, carboxylic ester hydrolase, secreted by the pancreas; LPL, lipoprotein lipase; VLDL, very low-density lipoprotein; LDL, low-density lipoprotein; HDL, high-density lipoprotein.

levels in plasma after ingestion by 20–120%. The extent of this inhibition depends upon the amount of nonabsorbable compound ingested, as well as the particular carotenoid under consideration. The mechanism for this inhibition is apparently similar to the action of fiber, i.e., sequestration. The type of fat that is ingested along with carotenoids will also affect carotenoid absorption. As macerated food passes into the intestinal lumen, carotenoids freed from the food matrix then become incorporated into micelles, consisting of free fatty acids, monoglycerides, phospholipids, and bile acids. Many other factors can affect intestinal absorption such as micelle size, phospholipid composition, solubilization of carotenoids into mixed micelles, and concentration of available bile salts, among others.

The presence of other carotenoids can affect the absorption of carotenoids into intestinal mucosal cells, since carotenoids can compete for absorption or facilitate the absorption of another. Data on carotenoid interactions are not clear. Human studies show that β-carotene decreases lutein absorption, while lutein has either no effect or a lowering effect on β-carotene absorption. Although not confirmed in humans, the inhibitory effect of lutein on β-carotene absorption might be partly attributed to the inhibition of the β-carotene cleavage enzyme by lutein shown in rats. Beta-carotene also seemed to lower absorption of canthaxanthin, whereas canthaxanthin did not inhibit β-carotene absorption. Studies showed that β-carotene increased lycopene absorption, although lycopene had no effect on β-carotene. Alpha-carotene and cryptoxanthin show high serum responses to dietary intake compared to lutein. In addition, *cis* isomers of lycopene seem to be more bioavailable than the all-*trans*, and selective intestinal absorption of all-*trans* β-carotene occurs, as well as conversion of the 9-*cis* isomer to all-*trans* β-carotene. It is clear, then, that selective absorption of carotenoids takes place into the intestinal mucosal cell.

Another complicating factor in the intestinal mucosal cell is the partial conversion of provitamin A carotenoids (β- and α-carotenes and cryptoxanthin) to vitamin A (primarily to retinyl esters). Therefore, in absorption studies these metabolic reactions must be accounted for in measuring intestinal transport. Nonprovitamin A carotenoids such as lycopene, lutein, and zeaxanthin are incorporated intact, although some cleavage can occur. Earlier studies on rats indicated that lycopene and β-carotene are absorbed by passive diffusion. However, recent evidence from the kinetics of β-carotene transport through Caco-2 cell monolayers indicates the involvement of a specific epithelial transporter that facilitates absorption.

In the intestinal mucosa, both carotenoids and retinyl esters are incorporated into chylomicrons and secreted into the lymph for transport to blood. In blood, lipoprotein lipase rapidly degrades the chylomicrons, and the liver sequesters the resulting carotenoid-containing fragments. The liver then secretes carotenoids back into the bloodstream in association with hepatic very low-density lipoproteins (VLDL). Most carotenoids in fasting plasma are carried by low-density lipoproteins (LDL) and high-density lipoproteins (HDL). Seventy-five per cent of the hydrocarbon carotenoids, e.g., lycopene and β-carotene, are associated with LDL, the rest is associated with HDL and, in smaller amounts, with VLDL. More polar carotenoids such as lutein and zeaxanthin are found equally distributed between HDL and LDL. After ingestion, carotenoids first appear in the bloodstream in chylomicrons, resulting from excretion from intestinal mucosal cells (4–8 h). HDL carotenoid levels peak in the circulation between 16 and 28 h; LDL carotenoid levels peak between 24 and 48 h. The bloodstream then transports carotenoids to different tissues (e.g., liver, prostate gland, fat, ocular macula) where they are sequestered by various mechanisms.

Distribution and Impact on Health

In general, carotenoid concentrations in serum reflect concentrations contained in the food that is ingested. Carotenoids have been found in various human organs and tissues. These include human liver, lung, breast, cervix, skin, and adipose and ocular tissues. The major storage organs are adipose tissue (probably because of its volume) and the liver. Tissues containing large amounts of LDL receptors seem to accumulate high levels of carotenoids, probably as a result of nonspecific uptake by lipoprotein carriers. Preferential uptake, however, is indicated in some cases. For example, unusually high concentrations of phytoene in the lung, ζ-carotene and phytofluene in breast tissue, lycopene in the prostate and colon, lycopene, β-carotene, and phytofluene in cervical tissue, and lutein and zeaxanthin in ocular tissues have been found.

The epidemiological findings that the ingestion of tomato and tomato products is strongly correlated with a reduced risk of several types of cancer, particularly prostate cancer, has stimulated a great deal of research on the protective effects of lycopene. Lycopene is the most efficient biological antioxidant. Hence, it has been assumed that it is this antioxidant activity that is responsible for the protection against prostate cancer. However, a recent study in which carcinogenesis was induced in rats using

N-methyl-*N*-nitrosourea showed that a diet containing whole tomato powder inhibited development of prostate cancer, but the same diet to which pure synthetic lycopene was added instead did not. These results indicate that lycopene alone was ineffective in reducing the incidence of prostate cancer. Therefore, either some other element in the tomato powder was the effective agent or the effect was obtained by lycopene working in concert with other tomato constituents. Obviously, more studies are required to determine which elements contained in tomato are responsible for the protective effect.

The finding that lutein and zeaxanthin are accumulated in the macula lutea of the eye has led to the hope that dietary supplementation might reduce the risk of age-related macular degeneration (AMD), which affects the central portion of the retina and is the most common cause of irreversible blindness in the Western world. Some studies have indicated benefits of diets supplemented with lutein and zeaxanthin from spinach in preventing AMD; others found no significant correlation between plasma levels of these carotenoids and reduced risk of AMD. Lutein, zeaxanthin, and a zeaxanthin stereoisomer 3R, 3′S(=meso)-zeaxanthin form the yellow pigment of the macula lutea. 3R, 3′S(=meso)-zeaxanthin is not found in either food or plasma in significant amounts. Also notable is that, in most food consumed in large quantities, the concentration of lutein is much greater than that of zeaxanthin (e.g., see **Table 1**, spinach, kale, broccoli, tomato). The yellow pigment of the macula is located in the center of the macula, covering the central fovea and overlapping the avascular zone. This location would allow the pigment to shield the photoreceptors from blue light. An environmental factor that seems to play a role in the development of age-related macular degeneration is ocular exposure to sunlight, in particular a history of exposure to blue light in the preceding 20 years. Light has been shown to induce oxidative damage in the presence of photosensitizers. Macular carotenoids are distributed in a pattern that is particularly advantageous. The two stereoisomers of zeaxanthin are concentrated in the central area and lutein in higher concentrations in the more peripheral regions. The lutein: zeaxanthin ratio in the center of the macula is about 0.8, in the peripheral regions about 2.4, but in plasma between 4 and 7. Therefore, the macula is able to concentrate lutein and zeaxanthin, change concentration ratios that are normally found in plasma, and invert the ratio to achieve higher zeaxanthin concentrations in the center of the macula lutea. The exact mechanism for this accumulation is not known; however, a specific membrane-associated, xanthophyll-binding protein was recently isolated from the human retina.

Carotenoids are believed to play a significant role in protecting skin from oxidative damage. *In vivo* measurements in humans of lycopene, β-, ζ-, γ-, and α-carotenes, lutein and zeaxanthin, phytoene, and phytofluene have shown that carotenoid concentrations are correlated with the presence or absence of skin cancer and precancerous lesions. Carotenoids are also believed to protect against several other types of cancer, cardiovascular diseases, and cataract formation and aid in immune function and gap-junction communication between cells, which is believed to be a protective mechanism related to their cancer-preventative activities.

Conclusions

Numerous studies indicate that carotenoids and their metabolites play a role in combating degradative reactions that are harmful to human health. Most of these functions seem to be related to their antioxidant nature and ability to dissipate energy from light and free radical-generating reactions. Obviously much research is still required to shed light onto mechanisms involved in these protective functions. Other fascinating roles in nature are also being discovered, for example, the signaling of apparent good health and consequently good potential parenting in birds by the red coloration of beaks, which seems to serve as an attractant to prospective mates.

See also: **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Carotenoids**: Epidemiology of Health Effects. **Vitamin A**: Biochemistry and Physiological Role.

Further Reading

Borel P (2003) Factors affecting intestinal absorption of highly lipophilic food microconstituents (fat-soluble vitamins, carotenoids and phytosterols). *Clinical Chemistry and Laboratory Medicine* **41**: 979–994.

Britton G (1995) Structure and properties of carotenoids in relation to function. *FASEB Journal* **9**: 1551–1558.

Britton G, Liaaen-Jensen S, and Pfander H (eds.) (1995) *Carotenoids: Isolation and Analysis* vol. 1A and *Spectroscopy*, vol. 1B Basel, Boston, Berlin: Birkhåuser Verlag.

During A and Harrison EH (2004) Intestinal absorption and metabolism of carotenoids: insights from cell culture. *Archives of Biochemistry and Biophysics* **430**: 77–78.

Frank HA, Young AJ, Britton G, and Cogdell RJ (1999) *The Photochemistry of Carotenoids*, (*Advances in Photosynthesis*, vol. 8). Dordrecht: Kluwer Academic Publishers.

Holden JM, Eldridge AL, Beecher GR, Buzzard IM, Bhagwat S, Davis CS, Douglass LW, Gebhardt S, Haytowitz D, and Schakel S (1999) Carotenoid content of U.S. foods: An

update of the database. *Journal of Food Composition and Analysis* **12**: 169–196.
Isler O (1971) *Carotenoids*. Basel: Birkhäuser-Verlag.
Khachik F, Carvalho L, Bernstein PS, Muir GJ, Zhao D-Y, and Katz NB (2002) Chemistry, distribution, and metabolism of tomato carotenoids and their impact on human health. *Experimental Biology and Medicine* **227**: 845–851.
Krinsky NI, Mayne ST, and Sies H (eds.) (2004) *Carotenoids in Health and Disease* (Oxidative Stress and Disease Series vol. 15). New York: Marcel Dekker.
Rodriguez-Amaya B (1999) In *A Guide to Carotenoid Analysis in Foods*. Washington, DC: ILSI Press.
Schalch W (2001) Possible contribution of lutein and zeaxanthin, carotenoids of the macula lutea, to reducing the risk for age-related macular degeneration: a review. *HKJ Ophthalmology* **4**: 31–42.
Yeum K-J and Russell RM (2002) Carotenoid bioavailability and bioconversion. *Annual Review of Nutrition* **22**: 483–504.

Epidemiology of Health Effects

S A Tanumihardjo and Z Yang, University of Wisconsin-Madison, Madison, WI, USA

Introduction

The colors of many fruits and vegetables are due to a class of compounds known as carotenoids. Over 600 carotenoids have been identified in nature. Humans are unique in that they can assimilate carotenoids from the foods that they eat whereas many other animals do not. Thus, carotenoids are an important class of phytochemicals. Phytochemicals are compounds derived from plants that may or may not have nutritional value. While many carotenoids circulate in humans, the most commonly studied ones are β-carotene, α-carotene, β-cryptoxanthin, lycopene, lutein, and zeaxanthin (**Figure 1**). The nutritional significance of carotenoids is that some are used by the body to make vitamin A. Indeed, approximately 50 carotenoids can be converted by the body into vitamin A and are known as provitamin A carotenoids. The three most abundant provitamin A carotenoids in foods are β-carotene, α-carotene, and β-cryptoxanthin. Provitamin A carotenoids, especially β-carotene, provide less than one-half of the vitamin A supply in North America but provide more than one-half in Africa and Asia.

Dietary recommendations for the intake of specific carotenoids have not been established due to lack of an adequate evidence base. To date, carotenoids are not considered essential nutrients. Dietary recommendations for vitamin A exist: 900 retinol activity equivalents (RAE) for men and 700 RAE for women. An RAE is equivalent to 1 μg of retinol. The recommendations for infants and children are less and range from 300 to 600 RAE depending on age. Consumers need to eat sufficient amounts of carotenoid-rich fruits and vegetables to meet their daily vitamin A requirement, and to achieve optimal dietary carotenoid intake to lower the risk of certain chronic diseases. In 2001, the Institute of Medicine revised the amount of carotenoids needed to provide vitamin A from foods as being approximately 12 μg of β-carotene or 24 μg of other provitamin A carotenoids to yield 1 RAE. Currently, high-dose pharmacological supplementation with carotenoids is not advised. Despite this, a tolerable upper intake level, the maximum daily amount of a nutrient that appears to be safe, has not been established for any individual carotenoid; however, supplemental β-carotene at $20\,\mathrm{mg\,day^{-1}}$ or more is contraindicated for use in current heavy smokers by the European Commission.

Because many factors affect the assimilation of carotenoids from foods (**Figure 2**), conversion factors need to be considered. This is especially important when most sources of vitamin A are from provitamin A carotenoids in the population. Bioavailability of preformed vitamin A, i.e., retinol and retinyl esters, is not a major concern because 80–95% of them are absorbed. However, foods that are high in preformed retinol (liver, eggs, and fortified milk) are not necessarily consumed by everybody. When discussing carotenoids from food, four terms need to be defined (see **Table 1**):

- bioaccessibility refers to how much carotenoid can be extracted from the food and is available for absorption;
- bioavailability is how much carotenoid is absorbed from the food and is available for physiological function;
- bioconversion relates to the provitamin A carotenoids and is defined as the amount of retinol that is formed from absorbed provitamin A carotenoids; and
- bioefficacy encompasses all of the biological processing of provitamin A carotenoids and is the amount of retinol formed from the amount of carotenoid contained in the food.

The study of carotenoid bioefficacy from foods is important in international health as the most frequently consumed sources of vitamin A are fruit and vegetables. A 100% bioefficacy means that 1 μmol of dietary β-carotene provides 2 μmol of retinol in the body; however, 100% bioefficacy does not actually occur in the process of digestion and carotenoid uptake by the body.

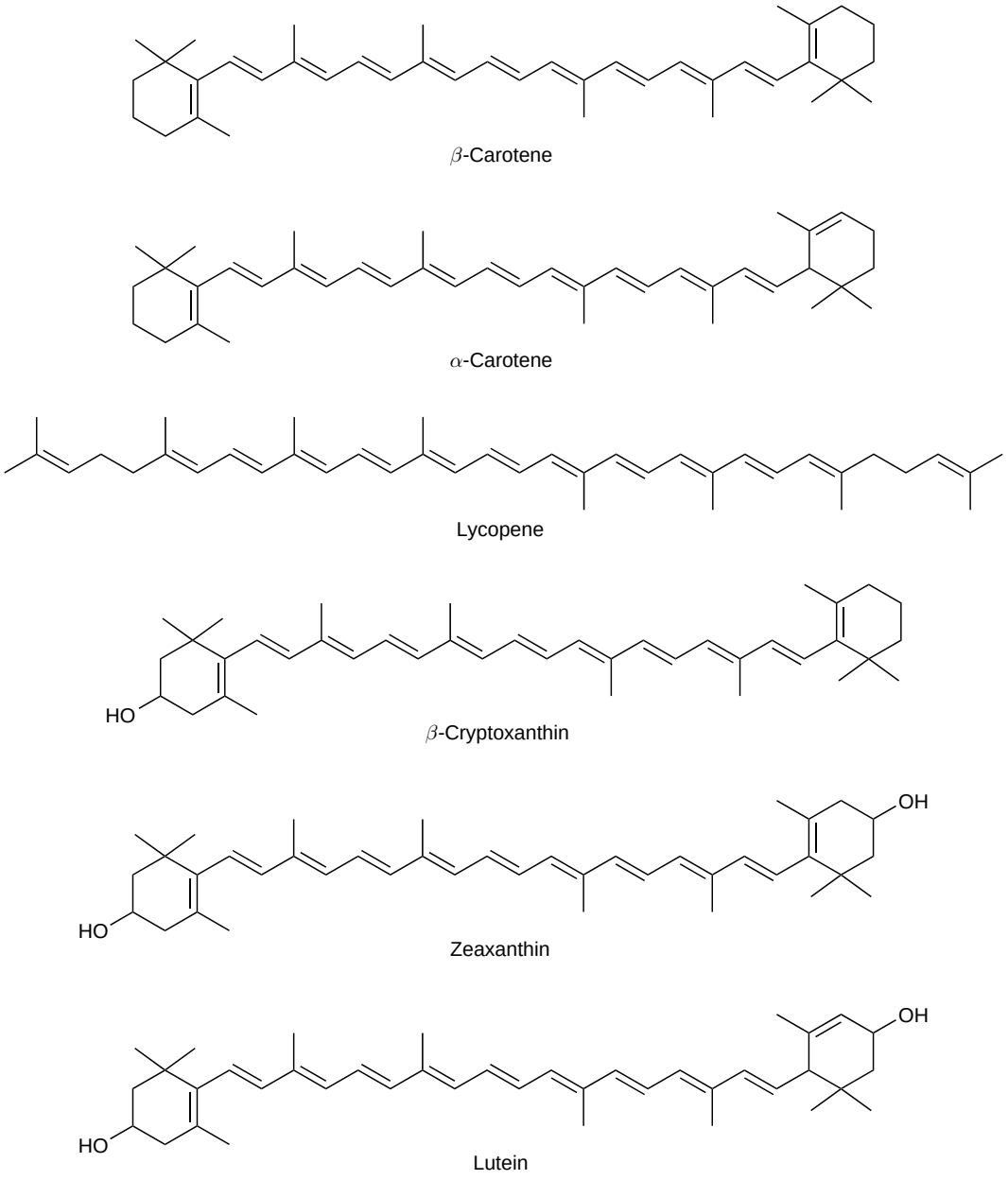


Figure 1 The structures of the most common carotenoids found in the human body. Three of them, β-carotene, α-carotene and β-cryptoxanthin, can be used by the body to make vitamin A. All carotenoids are antioxidants found in fruits and vegetables.

Once in the body, carotenoids can act as potent antioxidants, which are substances that neutralize free radicals formed from the natural metabolic processes of cells. Free radicals damage tissues and cells through oxidative processes. While free radical formation is a natural process in the body, environmental factors such as smoking and pollution can increase free radical load and thus disease risk. Carotenoids may counter these influences by functioning as an antioxidant and quenching oxygen-containing free radicals. In high- and low-density lipoproteins and cell membranes, carotenoids may also regenerate the antioxidant form of vitamin E as well as protect vitamin E from oxidation.

At the whole-body level, some population studies have indicated that certain carotenoids from either dietary intake or blood concentration data are associated with better immune response, lower rates of age-related macular degeneration (AMD) and cataract, as well as lower risk for certain cancers and cardiovascular disease. β-Carotene may increase immunological functions by enhancing lymphocyte proliferation independent of its provitamin A functions. The associations between specific carotenoids

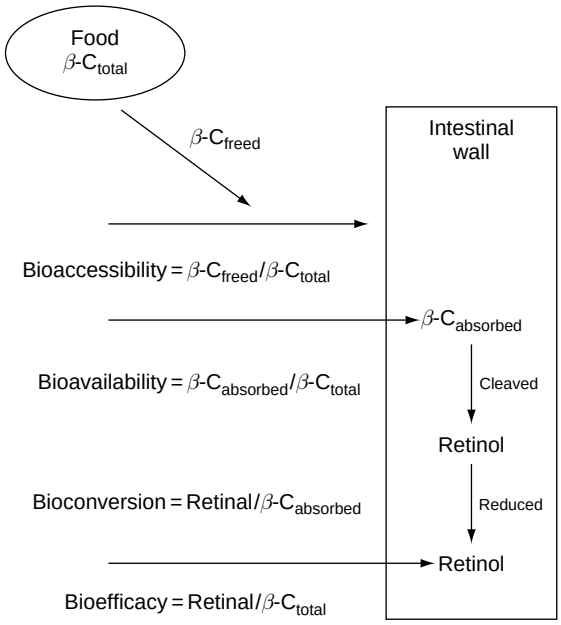


Figure 2 A schematic outlining the path of β-carotene (β-C) as it moves out from the food into the intestinal wall. The definition of terms associated with understanding β-carotene release, absorption, and conversion to retinol are illustrated: bioaccessibility, bioavailability, bionconversion, and bioefficacy. (Reproduced with permission from Tanumihardjo SA (2002) Factors influencing the conversion of carotenoids to retinol: Bioavailability to bioconversion to bioefficacy. *International Journal of Vitamin and Nutrition Research* **72**: 40–45.)

Table 1 Terms that are associated with the β-carotene vitamin A value of foods and subsequent utilization as retinol

Term	*Definition*	*100%*
Bioaccessibility	$\frac{\beta\text{-Carotene freed}}{\beta\text{-Carotene in food}}$	$\frac{1\,\mu\text{mol freed}}{1\,\mu\text{mol in food}}$
Bioavailability	$\frac{\beta\text{-Carotene absorbed}}{\beta\text{-Carotene in food}}$	$\frac{1\,\mu\text{mol absorbed}}{1\,\mu\text{mol in food}}$
Bioconversion	$\frac{\text{Retinol formed}}{\beta\text{-Carotene absorbed}}$	$\frac{2\,\mu\text{mol formed}}{1\,\mu\text{mol absorbed}}$
Bioefficacy	$\frac{\text{Retinol formed}}{\beta\text{-Carotene in food}}$	$\frac{2\,\mu\text{mol formed}}{1\,\mu\text{mol in food}}$

and decreased risk of various diseases are summarized in **Table 2**.

Blood levels of specific carotenoids are often used as biomarkers of fruit and vegetable intake to strengthen or replace dietary intake data. A wide variation in analytical methods exists and standardization between laboratories does not routinely occur. Nonetheless, higher blood concentrations have been favorably correlated with certain disease states. For example, vitamin A and carotenoid concentrations in serum were measured in middle-aged women who later developed breast cancer. Median concentrations of β-carotene, lycopene, lutein, and total carotenoids were significantly lower in women with breast cancer compared with case-control women who had not developed breast cancer. In contrast, vitamin A concentrations were either not different or showed a mixed response between cohorts, suggesting that carotenoids may be protective against breast cancer. Furthermore, the Nurses' Health Study, which included a cohort of over 83 000 women, also showed a significant inverse association between dietary β-carotene intake and breast cancer risk. This was especially strong for premenopausal women with a family history of breast cancer or high alcohol consumption. However, other prospective studies have had mixed results.

Hydrocarbon Carotenoid: β-Carotene

β-Carotene is one of the most widely studied carotenoids – for both its vitamin A activity and its abundance in fruits and vegetables. Epidemiological studies have often pointed to an abundance of carotenoids in the diet being protective against many diseases. Diets rich in fruits and vegetables are recommended to reduce the risk of cardiovascular disease and some forms of cancer. However, when β-carotene is removed from the plant matrix and administered as a supplement, these benefits sometimes disappear. For example, because lung cancer is

Table 2 A summary of epidemiologic and/or clinical studies where carotenoids and a significant association to a specific disease risk has been shown in at least one study[a]

Carotenoid	*Cardiovascular disease*	*Cataract*	*Macular degeneration*	*Lung cancer*	*Prostate cancer*
β-Carotene	Yes	–	–	Yes[b]	–
α-Carotene	Yes	–	–	Yes	–
β-Cryptoxanthin	–	–	–	Yes	–
Lycopene	Yes	–	–	Yes	Yes
Lutein/zeaxanthin	Yes	Yes	Yes	Yes	–

[a]For a more complete discussion of the association of specific carotenoids to disease please refer to: Krinsky NI, Mayne SI, and Sies H, (eds.) (2004) *Carotenoids in Health and Disease*. New York: Marcel Dekker.
[b]The opposite finding has been observed in clinical trials.

the leading cause of cancer death in many developed countries, the β-Carotene and Retinol Efficacy Trial (CARET) in the 1990s set out to test whether β-carotene conferred protection against cancer. CARET was based on a number of observational studies that showed high levels of β-carotene from food sources were protective against lung cancer. However, the CARET trial was halted for showing an increased risk for lung cancer in the treatment group over the control. Subsequent studies in ferrets showed that the amounts of β-carotene commonly consumed from fruit and vegetables were protective against lung damage but higher amounts, equivalent to those in CARET, increased the formation of abnormal tissue in the lung.

A similar outcome was observed among smokers in the α-Tocopherol β-Carotene (ATBC) Study Group. Although evidence clearly exists showing an association between β-carotene and enhanced lung function, as in the CARET study, the ATBC trial also found an increase in lung cancer rates among smokers. It is plausible that the lung cancer had already been initiated in the smokers and supplementation with β-carotene could not prevent the development of cancer. The ATBC study also showed an increased incidence of angina pectoris, a mild warning sign of heart disease characterized by chest pain, among heavy smokers. This may have been due to low blood levels of vitamin C in the study group leading to the inability of the individuals to quench β-carotene radicals, but this relationship requires more research.

In both the CARET and ATBC intervention trials, much higher doses of β-carotene were used than could be obtained from the diet, and the blood levels attained were two to six times higher than the 95th percentile level of β-carotene in a survey of a representative sample of the US population. Thus, it remains unclear whether β-carotene is a procarcinogen or an anticarcinogen. The associations for lower disease risk observed in epidemiologic studies may reflect other protective dietary agents or an interaction between dietary components. Furthermore, people with higher intake of fruits and vegetables may have healthier lifestyles that contribute to their lower risk of chronic diseases. The higher disease risk observed in the clinical trials may be correlated with the use of high doses of β-carotene where the mechanisms have not yet been identified, the limited duration of treatment, and/or the timing of the interventions was too late for cancers that were already present due to a history of heavy smoking. More research on β-carotene's biological actions is needed to explore the mechanisms involved. Current consensus is that the beneficial effects of β-carotene are associated with dietary consumption, whereas the harmful effects in some subpopulations are with pharmacological supplements.

Another explanation for a lack of a beneficial outcome with β-carotene supplementation may be that not all people respond to the same degree to β-carotene treatment, some being low- or non-responders. Some researchers believe that individuals who do not respond to β-carotene supplementation may be better at converting it to vitamin A. Blood response to β-carotene supplementation is also inversely related to body mass index (BMI), which may be due to increased sequestration of lipophilic β-carotene by the larger amount of fat stores present in people with larger BMI. This theory may not hold true as individuals with larger BMIs do not necessarily have a high body fat percentage, but rather increased lean muscle mass.

Excellent food sources of β-carotene include carrots, winter squash, red-orange sweet potato, and various types of dark green leafy vegetables. No deficiency or toxicity has been observed from dietary β-carotene intake, although very rarely high intakes can be associated with yellow pigmentation of the skin as carotenoids are stored in adipose tissue. Supplements containing β-carotene are common. In the Women's Health Initiative, the largest observational/intervention study in postmenopausal women to date, approximately 50% reported using a supplement containing β-carotene. This trial included both a clinical trial and observational study involving more than 160 000 women. The Physicians' Health Study II also included β-carotene as one of its interventions to determine the balance of risks and benefits of this carotenoid with cancer, cardiovascular disease, and eye disease.

Hydrocarbon Carotenoid: α-Carotene

α-Carotene, another carotenoid frequently present in food, also has provitamin A activity. Based on its structure, it is only converted to one molecule of biologically active retinol after central cleavage. Like other carotenoids, it has antioxidant and possibly anticarcinogenic properties, and may enhance immune function as well. Some, but not all, epidemiological studies observed that higher α-carotene intake was associated with lower risk of cardiovascular disease and cancer, whereas others did not. Clinical trials to test α-carotene influences in humans have not been conducted to date. This is probably because α-carotene is usually associated with ample amounts of β-carotene when found in fruits and vegetables and singling out α-carotene is difficult.

α-Carotene's concentration is especially high in orange carrots. Low or high dietary intake of α-carotene alone has not been associated with any specific disease outcome or health condition.

Xanthophyll: β-Cryptoxanthin

β-Cryptoxanthin is one of the lesser known carotenoids that also has provitamin A activity and appears to have a protective health role. Several epidemiological studies suggest that dietary β-cryptoxanthin is associated with lower rates of lung cancer and improved lung function in humans. A large prospective study on dietary intake and cancer, which included an interview on dietary habits and life style, identified β-cryptoxanthin as protective against lung cancer after correcting for smoking. However, the beneficial effects for β-cryptoxanthin suggested by these results could be merely an indicator for other antioxidants and/or a measure of a healthy life style that are more common in people with high dietary intakes of β-cryptoxanthin. In tissue culture, β-cryptoxanthin has a direct stimulatory effect on bone formation and an inhibitory effect on bone resorption. Studies of these beneficial effects in humans have not been conducted.

No deficiency or toxicity has been observed from dietary β-cryptoxanthin intake. The best food sources for β-cryptoxanthin are oranges, papaya, peaches, and tangerines. Tropical fruit intake is directly proportional to β-cryptoxanthin blood concentrations.

Hydrocarbon Carotenoid: Lycopene

Lycopene, while having no provitamin A activity, is a potent antioxidant with twice the activity of β-carotene for quenching singlet oxygen and 10 times the antioxidant activity of α-tocopherol in some model systems. The antioxidant potential of food chemicals varies widely according to location in the body and the presence of other body chemicals. The primary sources of dietary lycopene are tomatoes and tomato products. Epidemiological evidence shows an inverse association between lycopene consumption and the incidence and development of certain cancers. This association is especially strong for prostate cancer, which is the most common cancer among men in Western countries and the second leading cause of cancer death in American men. Prostate cancer rates in Asian countries are much lower, but appear to be increasing rapidly. Lycopene is localized in prostate tissue. The current consensus is that a high consumption of tomatoes or high circulating concentrations of lycopene are associated with a 30–40% risk reduction for prostate cancer, especially the most aggressive forms. Recent studies in rats show that tomato products are more protective against prostate cancer than isolated lycopene.

Epidemiologic studies have also observed lower rates of bladder, cervical, and breast cancers as well as cancers of the gastrointestinal tract among people with high intake of lycopene. The discovery of significant concentrations of lycopene in specific tissues in the body, i.e., plasma, testes, adrenal glands, liver and kidney, suggests that lycopene may play a role in these tissues.

While the body of evidence seems strong, several studies have found either no or weak associations between lycopene consumption and disease. Some of this may be explained by the fact that blood lycopene concentrations were much lower in these studies than in those that showed a beneficial effect. Thus, future dietary based studies need to include blood sampling to further define the range of blood concentrations of lycopene in the population, ideally with method standardization so that studies can be directly compared. The prostate cancer association is usually stronger for cooked tomato products rather than raw tomatoes or total lycopene intake. This too supports the idea that it is the whole food, with a broad array of nutrients and nonnutritive bioactive components, that is important for overall health rather than isolated compounds. It is possible that the beneficial effects of tomatoes are increased by preparing a concentrated product that enhances the nutrient bioavailability, as processed and cooked tomatoes are more closely associated to decreased risk of disease than either raw tomatoes or tomato juice.

Because lycopene is a potent antioxidant, it may be protective against heart disease by slowing down the oxidation of polyunsaturated fats in the low-density lipoprotein particles in the blood. Epidemiological and clinical studies show that higher blood lycopene concentrations are associated with lower risk and incidence of cardiovascular disease. Higher fat stores of lycopene have also been associated with lower risk of myocardial infarction. The most profound protective effect is in nonsmokers. The evidence for protective cardiovascular effects is compelling, as studies have shown a 20–60% improvement in cardiovascular parameters with higher blood concentrations of lycopene. Furthermore, higher intake of fruits and vegetables is associated with better lung function. In particular, high tomato intake is associated with higher timed expiratory volume.

The major food source of lycopene globally is tomatoes and tomato products. In the US, more than 80% of dietary lycopene comes from tomatoes.

Other sources include watermelon, pink grapefruit, and red carrots.

Xanthophylls: Lutein and Zeaxanthin

The structural isomers lutein and zeaxanthin are non-provitamin A carotenoids that are also measurable in human blood and tissues. Lutein and zeaxanthin have been identified as the xanthophylls that constitute the macular pigment of the human retina. The relative concentration of lutein to zeaxanthin in the macula is distinctive. Zeaxanthin is more centralized and lutein predominates towards the outer area of the macula. A putative xanthophyll-binding protein has also been described, which may explain the high variability among people to accumulate these carotenoids into eye tissues. Increased lutein intake from both food sources and supplements is positively correlated with increased macular pigment density, which is theorized to lower risk for macular degeneration. AMD is the leading cause of irreversible blindness in the elderly in developed countries. AMD adversely affects the central field of vision and the ability to see fine detail. Some, but not all, population studies suggest lower rates of AMD among people with higher levels of lutein and zeaxanthin in the diet or blood. Possible mechanisms of action for these carotenoids include antioxidant protection of the retinal tissue and the macular pigment filtering of damaging blue light.

Free radical damage is also linked to the development of cataracts. Cataracts remain the leading cause of visual disability in the US and about one-half of the 30–50 million cases of blindness throughout the world. Although cataracts are treatable, blindness occurs because individuals have either chosen not to correct the disease or do not have access to the appropriate medical treatment. Several epidemiological studies have shown inverse associations between the risk of cataracts and carotenoid intake. However, these studies also present inconsistencies with regard to the different carotenoids and their association with cataract risk. Lutein and zeaxanthin are found in the lens and are thought to protect cells in the eye against oxidative damage, and consequently prevent formation of cataracts. However, to date, there is no evidence that any carotenoid supplement can protect against cataract development. Eating plenty of fruits and vegetables, good sources of many antioxidants including carotenoids, is a preventative measure for many diseases.

Because lutein and zeaxanthin may be involved in disease prevention, much needs to be learned regarding human consumption of these carotenoids. One complicating factor that requires better understanding is the bioavailability of lutein from food sources and supplements. The food matrix is an important factor influencing lutein bioavailability and the amount and type of food processing generally influences the bioavailability of all carotenoids. For example, the processing of spinach does not affect bioavailability of lutein, but it does enhance that of β-carotene. Such studies have been conducted with lutein supplements and/or foods containing lutein fed to human subjects. In humans, lutein from vegetables seems to be more bioavailable than β-carotene; however, this may be partially explained by bioconversion of β-carotene to vitamin A. Competition between carotenoids, such as lutein and β-carotene, for incorporation into chylomicra has been noted in humans consuming vegetables and supplements. The amount of fat consumed with the lutein source also affects bioavailability, as higher fat increases the bioavailability of lipid-soluble carotenoids. Decreased plasma lutein concentrations are noted when alcohol is consumed, but the mechanism is poorly defined.

Lutein may also protect against some forms of cancer and enhance immune function. Lutein may work in concert with other carotenoids such as β-carotene to lower cancer risk due to their antimutagenic and antitumor properties. Because of these potential health benefits, lutein supplements are sold commercially and incorporated into some multivitamins. However, the amount provided in multivitamins (about 10–20% of the level in an average diet) is likely to be too low for a biological influence. Levels of lutein available as a single supplement vary widely and neither benefit nor safety of lutein supplements has been adequately studied. Major dietary sources of both lutein and zeaxanthin in the diet include corn, green leafy vegetables, and eggs. Lutein tends to be the predominant isomer in foods. Lutein supplements are often derived from marigold flowers.

Summary

Most of the epidemiological evidence points to carotenoids being a very important class of phytochemicals. While some of the effects may be attributable to a diet high in fruits and vegetables, and an overall healthy lifestyle, the presence of specific carotenoids localized in different areas of the human body lend evidence to their overall importance in the human diet. As methods are developed to assess carotenoid levels noninvasively in humans, large-scale studies that determine carotenoid levels in blood, skin, and the eye may lead to a better understanding of their importance in human health and disease prevention. Additional epidemiologic studies to further strengthen the associations that have been observed in populations are needed.

It must be kept in mind that study design and statistical analyses vary across published work and no one study can give conclusive evidence. An integrated multidisciplinary approach to study the functions and actions of carotenoids in the body is necessary to understand fully the role of carotenoids in health and disease prevention. This includes comparisons of carotenoids in whole fruits and vegetables and their effect on human health and well being. High fruit and vegetable intake is associated with a decreased risk of cancer, cardiovascular disease, diabetes, AMD, and osteoporosis. Removing any one class of phytochemicals from the intricate matrix of the whole plant may not give the same beneficial outcome in terms of human health. Considering that the average intake of fruits and vegetables is still less than that recommended by health professionals, programs that promote the consumption of more fruit and vegetables may be more effective at preventing disease in the long-term than using individual pharmacological carotenoid supplements.

A question that remains is whether or not carotenoids can be considered nutrients. A variety of phytochemicals contained in fruits and vegetables including carotenoids are assumed to be needed for optimal health and reduction of chronic disease risk, but have not been classified as nutrients. Indeed, in 2000, the Institute of Medicine was unable to recommend a daily reference intake for any carotenoid. Several factors have been defined that categorize substances as nutrients: substances that must be obtained from the diet because the body cannot synthesize the active form, and are used in the body for growth, maintenance, and tissue repair. In addition, to being classified as a nutrient, further studies must be done to determine the essentiality of the substance and its specific function in the body. Other criteria for defining a nutrient include concentration in specific tissues, consumption, and/or supplementation resulting in tissue concentration increases and improved tissue function. Lastly, a daily established dosage needs to be defined and a biomarker identified to assess status.

A large body of observational studies suggests that high blood concentrations of carotenoids obtained from food are associated with chronic disease risk reduction. However, there is little other evidence of their specific role in the body. Lutein and zeaxanthin are the only carotenoids found in a specific tissue (the macular region of the retina) that seem to have a specific function. Providing lutein in the diet increases macular pigment in humans. Animal studies show that a diet low in lutein can deplete macular pigment, but the influence on the health of the eye is not yet well understood. To further our understanding, large randomized prospective intervention trials need to be conducted to explore the essentiality of lutein supplementation for reducing ocular disease risk in humans. Thus, to date, no one specific carotenoid has been classified as an essential nutrient.

See also: **Antioxidants**: Diet and Antioxidant Defense; Observational Studies; Intervention Studies. **Bioavailability**. **Cancer**: Epidemiology and Associations Between Diet and Cancer; Epidemiology of Lung Cancer; Effects on Nutritional Status. **Carotenoids**: Chemistry, Sources and Physiology. **Coronary Heart Disease**: Prevention. **Fruits and Vegetables**. **Lycopenes and Related Compounds**. **Older People**: Nutrition-Related Problems. **Phytochemicals**: Epidemiological Factors. **Supplementation**: Dietary Supplements.

Further Reading

Alves-Rodrigues A and Shao A (2004) The science behind lutein. *Toxicology Letters* **150**: 57–83.

Christen WG, Gaziano JM, and Hennekens CH (2000) Design of Physicians' Health Study II – a randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. *Annals of Epidemiology* **10**: 125–134.

Giovannucci E (2002) A review of epidemiologic studies of tomatoes, lycopene, and prostate cancer. *Experimental Biology of Medicine* **227**: 852–859.

Hwang ES and Bowen PE (2002) Can the consumption of tomatoes or lycopene reduce cancer risk? *Integrative Cancer Therapies* **1**: 121–132.

Institute of Medicine, Food and Nutrition Board (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. Washington, DC: National Academy Press.

Johnson EJ (2002) The role of carotenoids in human health. *Nutrition in Clinical Care* **5**: 56–65.

Krinsky NI, Landrum JT, and Bone RA (2003) Biologic mechanisms of the protective role of lutein and zeaxanthin in the eye. *Annual Review of Nutrition* **23**: 171–201.

Krinsky NI, Mayne SI, and Sies H (eds.) (2004) *Carotenoids in Health and Disease*. New York: Marcel Dekker.

Mares-Perlman JA, Millen AE, Ficek TL, and Hankinson SE (2002) The body of evidence to support a protective role for lutein and zeaxanthin in delaying chronic disease. *Journal of Nutrition* **132**: 518S–524S.

Rapola JM, Virtamo J, Haukka JK *et al.* (1996) Effect of vitamin E and beta-carotene on the incidence of angina pectoris. A randomized, double-blind, controlled trial. *Journal of the American Medical Association* **275**: 693–698.

Tanumihardjo SA (2002) Factors influencing the conversion of carotenoids to retinol: Bioavailability to bioconversion to bioefficacy. *International Journal of Vitamin and Nutrition Research* **72**: 40–45.

The Alpha-Tocopherol Beta Carotene Cancer Prevention Study Group (1994) The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *New England Journal of Medicine* **330**: 1029–1035.

CEREAL GRAINS

R W Welch, University of Ulster, Coleraine, UK

Introduction

Cereal grains are dietary staples that provide a very substantial proportion of the dietary energy, protein, and micronutrients for much of the world's population. The major cereal crops are rice, maize (corn), wheat, barley, sorghum, millets, oats, and rye. Worldwide, these cereals are subjected to a very diverse range of traditional and technologically advanced processes before consumption. Thus, cereal-based foods vary enormously in their structural, storage, and sensory characteristics. Cereal-based foodstuffs also vary in nutritional value owing to inherent differences in nutrient content and to changes as a result of processing, which may be beneficial or detrimental. Cereals are also the raw materials for the production of alcoholic beverages and food ingredients including starches, syrups, and protein and fiber isolates. Furthermore, very substantial quantities of cereal enter the food chain as livestock feed.

Types of Cereal and Their Role in the Diet

Cereal grains are the seeds of cultivated annual species of the grass family (Gramineae). Cultivated cereal species have evolved with humankind and include a range of types differing widely in their environmental adaptation and their utility for food or other uses. Some cereals are adapted to tropical or subtropical regions, others are adapted to temperate climates, while some can withstand sub-zero temperatures. The type of cereal grown is largely determined by climatic and edaphic factors, although economic and cultural factors are also important. Total world cereal production is about 2 billion tonnes (**Table 1**). The major cereals produced are rice, maize (corn), wheat, barley, sorghum, millets, oats, and rye. Some of these are single species; others include a number of species with different agronomic and utilization characteristics. Each species comprises a range of cultivars (varieties or genotypes), which also differ in characteristics. Minor cereals include triticale (*Triticosecale*; a wheat–rye hybrid), buckwheat (*Fagopyrum esculentum*), and quinoa (*Chenopodium quinoa*). Buckwheat and quinoa are not Gramineae and are thus pseudocereals. All cereals are used for human nutrition. However, the forms in which they are consumed and their dietary significances vary substantially across cereal types and regions.

Grain Characteristics

The harvested grain of some cereals (wheat, maize, rye, sorghum, and some millets) comprises, botanically, a caryopsis. In other cereals (barley, oats, rice, and some millets) the harvested grain generally includes a hull (or husk) that encloses the caryopsis. The hull is tough and very high in fiber. It is unsuitable for human nutrition and is removed in primary

Table 1 World production of cereals 2000; figures are in thousands of tonnes

	Total	*Rice*[a]	*Maize*	*Wheat*	*Barley*	*Sorghum*	*Millets*	*Oats*	*Rye*	*Other*[b]
World	2 063 578	602 610	591 988	585 297	134 421	56 044	27 636	26 394	19 642	19 546
Asia (excluding China and the Russian Federation)	592 016	359 375	42 707	154 299	15 021	8 538	10 999	516	389	172
North and Central America	427 528	10 917	279 404	90 780	20 826	18 265	166	6 063	473	634
China	407 328	189 814	106 180	99 636	3 346	2 608	2 126	650	650	2 318
Europe (excluding the Russian Federation)	320 442	2 567	61 782	149 039	70 383	685	439	10 677	12 544	12 326
Africa	112 149	17 649	44 409	14 328	2 081	18 630	12 679	194	30	2 149
South America	104 275	20 583	55 407	20 199	1 543	5 083	48	1 128	135	149
Russian Federation	64 342	586	1 500	34 500	14 100	116	1 122	6 000	5 400	1 018
Oceania	35 498	1 119	599	22 516	7 121	2 119	57	1 166	21	780

[a]Paddy (rough) rice.
[b]Includes other cereals, pseudocereals, mixed grains, and their products not accounted for elsewhere.
Source: Food and Agriculture Organisation, FAOSTAT data at http://faostat.fao.org/accessed May 2003.

processing. The caryopsis is the edible part of the cereal grain. Cereal caryopses all have the same basic structure. The major part is the endosperm (63–91% of the total). The endosperm is high in starch and contains nutritionally significant amounts of protein. Enclosing the endosperm are cell layers, amounting to 5–20% of the caryopsis, which form the bran that is often separated from the endosperm during milling processes. The embryo, which is found at one end of the caryopsis, accounts for 2.5–12% of its weight. The embryo is the major component of the germ fraction, which is separated in some milling processes.

At harvest, cereal grains have a low moisture content (12–16%), and they are hard and inedible without processing. Some grain types may be subjected to simple milling procedures and made into palatable unleavened products; others are subjected to more complex milling procedures and further processed into leavened, extruded, or fermented products using technologically advanced processes.

Food use of cereals is shown in **Table 2**. From this it can be seen that rice and wheat are the major world food cereals, and maize is important in some regions. However, the data in **Table 2** represent food use of the crops. Since varying proportions of these crops will be lost in milling, the actual consumption levels may be substantially less.

Rice

World rice production exceeds 600 million tonnes annually (**Table 1**), and over 85% of the crop is used for human nutrition. Rice is a dietary staple for over half the world's population, and over 90% of the crop is produced and consumed in Asia and China. The per capita supply of paddy rice exceeds 600 g day^{-1} in six countries (Myanmar, 847 g day^{-1}; Laos, 707 g day^{-1}; Viet Nam 686 g day^{-1}; Cambodia, 668 g day^{-1}; Bangladesh, 643 g day^{-1}; and Indonesia, 611 g day^{-1}), and exceeds 400 g day^{-1} in Thailand (448 g day^{-1}), Guinea-Bissau (422 g day^{-1}), the Philippines (418 g day^{-1}), Côte d'Ivoire (412 g day^{-1}), Nepal (409 g day^{-1}), and Madagascar (402 g day^{-1}). Rice makes a modest contribution to the diets in most industrialized countries, except Japan where rice is the major dietary cereal with a per capita supply of 244 g day^{-1}. Rice (*Oryza sativa*) can be grown in a wide range of environments, under water or on dry land. The subspecies *indica* is grown in the tropics, and the subspecies *japonica* is grown mainly in warm temperate regions. Harvested rice, known as paddy (or rough) rice, has a hull. Removal of the hull yields brown rice, which can be rendered edible by prolonged boiling. However, the type of rice preferred in most regions is white polished rice, in which the outer bran layers are removed by milling. The yield of edible milled rice from paddy rice is about 66%. Types of rice vary in grain morphology and in cooking characteristics. When boiled, short-grain *indica* types generally become soft and sticky and aggregate, whereas long-grain *japonica* types remain relatively firm and separated. A small proportion of rice is processed to flour. However, the absence of gluten precludes its general use in leavened products.

Maize

World production of maize (corn; *Zea mays*) is nearly 600 million tonnes annually (**Table 1**). About 20% is used directly for human food, and about 10% is used in industrial processing to yield oil and other products including starch, syrups, flour, and grits for food and other uses. Maize is grown in tropical and warm temperate regions.

Table 2 Food use of cereals, 2000; figures are per capita supply in grams per day

	Total	*Rice*[a]	*Maize*	*Wheat*	*Barley*	*Sorghum*	*Millets*	*Oats*	*Rye*	*Other*[b]
World	503	236	51	187	3.0	10.8	9.8	1.3	2.9	2.2
Asia (excluding China and the Russian Federation)	584	350	35	181	2.1	6.4	8.3	0.2	0.4	1.3
North and Central America	364	46	111	197	1.1	3.1	0.0	2.9	0.5	3.0
China	622	366	44	203	1.2	2.5	2.5	0.4	0.5	2.1
Europe(excluding the Russian Federation)	365	19	19	294	4.3	0.0	1.7	4.3	21.7	1.1
Africa	421	78	115	125	8.3	51.2	35.9	0.4	0.1	7.3
South America	347	128	60	150	1.3	0.0	0.0	5.9	0.1	1.0
Russian Federation	415	20	1	355	2.5	0.0	6.5	3.3	25.5	0.3
Oceania	251	65	9	168	0.2	1.5	0.0	3.3	1.0	2.5

[a]Paddy (rough) rice.
[b]Includes other cereals, pseudocereals, mixed grains, and their products not accounted for elsewhere.
Source: Food and Agriculture Organisation, FAOSTAT data at http://faostat.fao.org/accessed May 2003.

Maize is a dietary staple in parts of Africa. Food use is highest in Lesotho (427 g day^{-1} per person) and is 300–400 g day^{-1} per person in other southern and eastern African countries (Malawi, Zambia, Zimbabwe, and South Africa) and in Mexico, where it has a pre-Colombian tradition. Food use is 200–300 g day^{-1} per person in Moldova, Bosnia and Herzegovina, Guatemala, Kenya, El Salvador, and Honduras and 100–200 g day^{-1} per person in Benin, Bolivia, Botswana, Burkina Faso, Cameroon, Columbia, Egypt, Ethiopia, Georgia, Ghana, North Korea, Mozambique, Nepal, Nicaragua, Paraguay, Romania, Swaziland, Tanzania, Togo, and Venezuela. A harvested grain of maize is a caryopsis. There is a wide range of types varying in grain size, color, and endosperm characteristics. These include white, yellow, and red types and types with endosperms ranging from soft and floury to hard and flinty textures. Sweetcorn is generally regarded as a vegetable; it is harvested before the grain is mature and has a high water content (70–90%).

Wheat

World wheat production is nearly 600 million tonnes annually (**Table 1**). Worldwide, over 70% of the wheat supply is used for food, and it is the major dietary cereal in all regions except Asia and China (**Table 2**). Consumption levels are highest in western Asia (Turkmenistan, Kyrgyzstan, Azerbaijan, and Turkey) and North Africa (Algeria and Tunisia), where supplies for food use exceed 500 g day^{-1} per person. Although other species are grown in some regions, the major species are common or bread wheat (*Triticum aestivum*) and durum wheat (*Triticum durum*). Of these two species, common wheat accounts for about 95% of total production. Common wheat can be grown across widely diverse geographical regions including subtropical, warm temperate, and cool temperate climates, where it can withstand frosts. There is an extensive range of genotypes, varying in agronomic adaptability and grain quality. Quality is usually assessed in terms of suitability for milling and baking. The whole grain comprises about 82% endosperm and 18% bran. The aim of milling white flour is to separate the starchy endosperm from the darker coarser bran. Wheat bran is used as a fiber source in some foodstuffs including ready-to-eat cereals. The yield of white flour will depend on milling efficiency, but extraction rates of 75–81% are usually achieved. Wheat grains are classified as hard or soft and as strong or weak. The terms hard and soft indicate milling quality; hard wheats have superior milling characteristics. The terms strong and weak refer to bread-baking quality. Strong flours yield bread with a large loaf volume and good crumb structure. The quantity and quality of the viscoelastic gluten proteins are important for bread quality. Weak flours produce poor bread products but can be blended with strong flours in bread making or used to make other products. Durum wheat is hard and vitreous. It is used almost exclusively for the production of pasta in Europe, the Americas, and elsewhere, but in the Middle East and North Africa 85% of durum wheat is used to produce breads, couscous, and other non-pasta products.

Barley

World barley production is over 130 million tonnes annually (**Table 1**). Barley (*Hordeum vulgare*) is grown mainly in temperate regions and can withstand sub-zero temperatures. About 17% of the total crop is used industrially, primarily to produce malt, which is used mainly in brewing and distilling. Only about 5% of the total is used for food. The highest food usage is in Morocco (104 g day^{-1} per person), followed by the Baltic States (Estonia, Latvia, and Lithuania) and Moldova, where food usage is 86, 57, 55, and 55 g day^{-1} per person, respectively. Most barley grain has a fibrous hull that adheres closely to the caryopsis. However, naked or hull-less types are grown in some regions. Barley is processed for human use by removing the hull and polishing to yield pearl barley, which is used in soups and other foods. Pearl barley is also ground to a coarse meal and cooked as a gruel or ground to barley flour for making flat breads. Barley flours and brans, for ingredient use, are produced from pearl or hull-less barley.

Sorghum

World production of sorghum (*Sorghum bicolor*) is 56 million tonnes per year (**Table 1**), and about 40% of this is used for food. Sorghum is grown in semi-arid zones and is especially important in tropical and sub-tropical regions. Food use of sorghum is highest in Africa. The food supply in grams per head per day in the major consuming countries is: Sudan, 248 g day^{-1}; Burkino Faso, 192 g day^{-1}; Nigeria, 143 g day^{-1}; Chad, 142 g day^{-1}; Eritrea, 124 g day^{-1}, and Mali, 124 g day^{-1}. Food use is between 60 g day^{-1} and 75 g day^{-1} per person in a further six African countries (Botswana, Cameroon, Ethiopia, Mauritania, Niger, and Togo). Both sorghum and millets are generally milled by traditional methods to yield grits and flours, which are used to

make a variety of traditional foodstuffs including porridges, steamed products, breads, and pancake products.

Millets

Annual production of millets is over 27 million tonnes. About 77% of production is used for food, and millets are very important food crops in semi-arid regions of Africa. There are at least nine species of millet. In production terms, the most important is pearl or bulrush millet (*Pennisetum glaucum*) (about 53% of the total). Foxtail millet (*Setaria italica*), proso millet (*Panicum miliaceum*), and finger millet (*Eleusine coracana*) contribute about 17%, 14%, and 12%, respectively, to world production. Other species, which include Japanese or barnyard millet (*Echinochloa crus-galli*), kodo millet (*Paspalum scrobiculatum*), teff (*Eragrostis tef*), fonio (*Digitaria exilis* and *D. iburua*), and little millet (*Panicum sumatrense*), contribute the remaining 4% of production. Pearl millet is a dietary staple in Niger, where the supply for food use is 426 g day^{-1} per person, providing over 70% of the total dietary cereal. High consumption levels are found in other African countries (Burkina Faso, Mali, Gambia, Namibia, Nigeria, Senegal, and Chad), where the supply for food use ranges from 100 g day^{-1} to 200 g day^{-1} per person.

Oats

World oat production is about 26 million tonnes annually (**Table 1**). Oats (*Avena sativa*) require a temperate climate and thrive in cool wet conditions but are less cold tolerant than rye, wheat, or barley. Approximately 11% of world production is used for food. Formerly a dietary staple in Northern and Western Europe, oats now contribute modestly to diets worldwide (**Table 2**). Dietary intakes are highest in Belarus, Estonia, and Finland, where supplies for food use are 24, 20, and 18 g day^{-1} per person, respectively. However, these data are for oat grain, which includes 18–36% of fibrous inedible hull, which is removed in the first stages of milling to yield caryopses (groats). Naked or hull-less types also exist. Groats are further processed by cutting, rolling, or grinding to yield a range of oatmeal, oat-flake, and oat-flour products. These are wholemeal products with a composition similar to the groats. Oat bran, which is less structurally distinct than wheat bran, is made by sieving coarse-milled groats. Oat-mill products can be used for traditional porridge and oatcakes, as an ingredient in baby foods, and in breakfast cereals. Oats cannot be used to make good quality bread, but oat flour can be incorporated at 20–30% in wheat breads.

Rye

Rye (*Secale cereale*) is grown in temperate regions and is the most cold-tolerant of the cereals. World rye production is about 20 million tonnes annually (**Table 1**). About 33% of rye production is used for food and about 15% for industrial use, including whisky production. Food intakes are highest in Eastern and Central Europe, the Baltic states, and Scandinavia. Per capita supplies for food use in these countries are: Belarus, 110 g day^{-1}; Poland, 108 g day^{-1}; Austria, 39 g day^{-1}; Ukraine, 36 g day^{-1}; Slovakia, 31 g day^{-1}; Lithuania, 57 g day^{-1}; Estonia, 54 g day^{-1}; Finland, 46 g day^{-1}; Denmark, 36 g day^{-1}, and Sweden, 35 g day^{-1}. Harvested rye is a caryopsis, and it is milled to extractions ranging from 65% to wholemeal (100%). Rye flour is used to make crispbreads and yeast-leavened breads, where it is often mixed with wheat flour.

Energy, Macronutrient, and Fiber Content

Tables 3–6 show the water, energy, macronutrient, and fiber contents of cereals and cereal products. The macronutrients (carbohydrate, protein, and fat) and dietary fiber comprise the bulk of the dry

Table 3 Water, energy, macronutrient, and fiber contents of rice and maize (corn) products; representative values per 100 g

	Brown rice, uncooked	*White rice, uncooked*	*Brown rice, boiled*	*White rice, boiled*	*White rice, flour*	*Maize meal*	*Cornflour*	*Cornflakes*[a]
Water (g)	13.9	11.7	56.0	69.9	11.9	12.2	12.5	3.0
Energy (kJ)	1518	1536	599	522	1531	1517	1508	1515
Energy (kcal)	357	361	141	123	366	357	354	355
Carbohydrate (g)	81.3	86.8	32.1	29.6	80.1	77.2	92.0	84.9
Protein (g)	6.7	6.5	2.6	2.2	6.0	9.4	0.6	7.9
Fat (g)	2.8	1.0	1.1	0.3	1.4	3.3	0.7	0.6
Dietary fiber (g)	1.9	0.5	0.8	0.2	2.4	2.2	0.1	0.9

[a]Ready-to-eat cereal.

Table 4 Water, energy, macronutrient, and fiber contents of wheat products; representative values per 100 g

	Wholemeal flour	*White bread-making flour*	*Bran*	*White bread*	*Wholemeal bread*	*Uncooked pasta*[a]	*Boiled pasta*[a]
Water (g)	14.0	14.0	8.3	37.3	38.3	9.7	75.9
Energy (kJ)	1320	1452	872	1002	915	1473	411
Energy (kcal)	310	341	206	236	215	346	97
Carbohydrate (g)	63.9	75.3	26.8	49.3	41.6	74.9	20.9
Protein (g)	12.7	11.5	14.1	8.4	9.2	12.0	3.2
Fat (g)	2.2	1.4	5.5	1.9	2.5	1.9	0.6
Dietary fiber (g)	9.0	3.1	36.4	1.5	5.8	3.0	1.0

[a]Mean of lasagne, macaroni, and white spaghetti.

Table 5 Water, energy, macronutrient, and fiber contents of barley, oat, and rye products; representative values per 100 g

	Pearl barley, uncooked	*Pearl barley, boiled*	*Oatmeal*	*Oatmeal porridge*[a]	*Oat bran*	*Wholemeal rye flour*	*Rye crispbread*	*Rye bread*
Water (g)	10.6	69.6	8.5	87.4	9.5	15.0	6.4	37.4
Energy (kJ)	1535	510	1644	210	1478	1268	1367	937
Energy (kcal)	360	120	388	50	349	298	321	220
Carbohydrate (g)	83.6	27.6	69.4	9.0	53.5	65.9	70.6	45.8
Protein (g)	7.9	2.7	11.8	1.5	19.8	8.2	9.4	8.3
Fat (g)	1.7	0.6	9.0	1.1	7.7	2.0	2.1	1.7
Dietary fiber (g)	5.9	2.0	7.0	0.8	15.1	11.7	11.7	4.4

[a]Made with water and salt.

Table 6 Water, energy, macronutrient, and fiber contents of sorghum and millets; representative values per 100 g

		Millets					
	Sorghum	*Pearl*	*Foxtail*	*Proso*	*Finger*	*Japanese*	*Fonio*
Water (g)	12.0	11.0	11.3	13.5	11.7	11.1	10.0
Energy (kJ)	1422	1468	1364	1491	1377	1382	1541
Energy (kcal)	335	346	321	351	323	326	363
Carbohydrate (g)	69.9	68.1	67.8	70.7	76.0	65.4	75.8
Protein (g)	10.7	11.8	9.9	11.6	6.4	10.4	8.5
Fat (g)	3.3	4.8	3.0	4.4	1.4	4.3	3.5
Dietary fiber (g)	7.5	6.9	n/a	n/a	n/a	n/a	8.5

n/a, no data available.

matter of cereals. Carbohydrates are the major constituent, and there is a nutritionally significant amount of protein. Cereals can also be an important source of dietary fiber. However, most cereals are low in fat.

Dietary Energy

Dietary energy values are inversely related to water contents and also depend on the relative amounts of the macronutrients (digestible carbohydrate, protein, and fat). Fat has over twice the energy per gram of carbohydrate or protein, and thus differences in dietary energy are largely determined by variations in the levels of fat, water, and indigestible components (principally fiber). Energy values are higher when fat is higher and lower when water or fiber contents are higher (**Tables 3–6**). Higher fiber contents are found in whole-grain and bran-rich products, while water and fat contents may be changed during processing.

Carbohydrates

Digestible carbohydrate, in the form of starch, is the major dry-matter component of all cereals (**Tables 3–6**). Sugars, which usually account for much less than 1% of cereal grain, may be added in

processing; cornflakes, for example, have about 7% sugars. Cornflour is a milling product with a very high starch content (**Table 3**). Most cereal starches are 20–30% amylose, the rest being amylopectin. However, there are types of rice, maize, and barley with up to 80% amylose or with up to 100% amylopectin (waxy types). In some cereal products a small proportion of the starch (up to 3%) exists as resistant starch, which resists enzymatic digestion. Digestible starch yields glucose. However, the rate of digestion and absorption is influenced by the degree of processing among other factors. Thus, there can be substantial variations in the post-prandial blood glucose responses following ingestion of equivalent digestible-carbohydrate loads from different cereal products.

Protein

Protein is the major nitrogen-containing component of cereal grains, and most protein data are based on nitrogen determination, followed by multiplication by a nitrogen-to-protein conversion factor. The general factor is 6.25, but factors vary from 5.7 to 6.31 for cereal products. Representative values for protein content (**Tables 3–6**) show that levels are lowest in rice, barley, and finger millet and highest in wheat, oats, pearl millet, and proso millet. However, the protein content of cereals can vary substantially, and greater than 2-fold ranges in protein content are found in crops of the same species. This variation is due partly to genetic differences, but agronomic factors are of greater importance. This variation is of little significance when crops are handled in bulk, as in modern milling and processing, but it may be important in less-developed regions. Although not usually considered a good protein source, many cereals provide an adequate amount, relative to energy, for adults. However, protein quality must also be considered, since cereal-based diets tend to be deficient in one or more essential amino-acids (see below).

Protein quality Cereal protein consists predominantly of endosperm storage proteins, which are low in dietary essential (indispensable) amino-acids. These amino-acids are required in differing amounts, and thus quality can be considered only in relation to requirements. Furthermore, requirements can differ. The most notable differences are related to age: the young have higher requirements for both protein and essential amino-acids than do adults. The first limiting essential amino-acid in cereals is generally lysine. However, there are variations between cereals. In oats, rice, and finger millet the deficiency in lysine may be only marginal, while in sorghum, maize, and other millets it is more pronounced (**Tables 7** and **8**). Tryptophan is also limiting in maize and some millets, while threonine and methionine may also be limiting in some cereals. Protein quality must be considered in relation to total protein content. Furthermore, as protein content is increased, for example by the use of nitrogenous fertilizer, the relative amounts of the essential amino-acids tend to decline as percentages of the protein. High-lysine types of maize, barley, and sorghum have been identified. However, their lower grain yield precludes their wide use.

Table 7 Amino-acid composition of rice, maize, wheat, barley, oats, and rye; representative values in grams per 100 g protein

Amino-acid	*Rice*	*Maize*	*Wheat*	*Barley*	*Oats*	*Rye*
Essential						
Histidine	2.4	2.6	2.3	2.1	2.1	2.2
Isoleucine	3.8	3.6	3.5	3.5	3.8	3.5
Leucine	8.2	11.1	6.7	6.7	7.2	6.2
Lysine	3.7	2.3	2.7	2.6	3.7	3.4
Methionine	2.1	1.6	1.2	1.6	1.8	1.4
Cysteine	1.6	2.0	2.5	2.2	2.7	1.9
Phenylalanine	4.8	4.4	4.6	5.1	5.0	4.5
Tyrosine	4.0	3.5	1.7	3.0	3.4	1.9
Threonine	3.4	3.3	2.8	3.4	3.4	3.4
Tryptophan	1.3	0.7	1.5	1.6	1.3	1.1
Valine	5.8	4.0	4.3	5.0	5.1	4.8
Non-essential						
Alanine	5.8	8.2	3.5	4.2	4.5	4.3
Arginine	7.5	4.4	4.3	4.8	6.2	4.6
Aspartic acid	9.6	7.2	4.9	5.6	7.7	7.2
Glutamic acid	19.2	18.6	32.1	23.5	21.0	24.2
Glycine	4.3	3.9	4.0	3.8	4.6	4.3
Proline	4.6	8.8	10.7	10.9	5.1	9.4
Serine	4.6	4.6	4.5	4.0	4.6	3.8

Fat

Cereals are generally very low in fat, most containing only 2–4% (**Tables 3–6**). However, some types of maize and oats have more than 10% fat. The distribution of fat within the grain is variable. In oats the fat is distributed throughout the endosperm; in maize it is concentrated in the germ, from which it can be extracted after separation, while rice bran contains 15–20% oil depending on production conditions. Fat is often added in processing, for example in baked products.

Fatty acid composition Cereal fat is liquid at room temperature; it is high in unsaturates and is more correctly described as oil. The major fatty acids in cereal oils are oleic acid (monounsaturated), linoleic acid (polyunsaturated), and palmitic acid

Table 8 Amino-acid composition of sorghum and millets; representative values in grams per 100 g protein

		Millets					
Amino-acid	*Sorghum*	*Pearl*	*Foxtail*	*Proso*	*Finger*	*Japanese*	*Fonio*
Essential							
Histidine	2.2	2.2	2.3	2.2	2.6	1.9	2.2
Isoleucine	4.1	4.4	5.0	4.5	5.1	4.5	4.1
Leucine	14.6	12.2	13.3	12.9	13.5	11.5	10.8
Lysine	2.2	3.3	2.1	2.2	3.7	1.7	2.2
Methionine	1.4	2.2	2.6	2.0	2.6	1.8	4.3
Cysteine	1.7	1.5	1.4	1.7	1.6	1.5	2.5
Phenylalanine	5.0	5.2	5.3	5.2	6.2	5.9	5.9
Tyrosine	3.2	3.2	2.7	3.9	3.6	2.7	3.7
Threonine	3.3	3.9	3.9	3.4	5.1	2.7	3.7
Tryptophan	1.1	1.6	1.5	0.9	1.3	1.0	1.6
Valine	5.4	5.7	5.2	5.1	7.9	6.1	5.5
Non-essential							
Alanine	9.1	8.5	8.9	9.3	8.0	9.2	9.4
Arginine	4.3	4.8	6.1	4.4	5.2	3.2	3.6
Aspartic acid	6.4	8.7	6.9	5.5	7.9	6.3	9.0
Glutamic acid	22.6	21.2	18.8	20.5	27.1	20.7	22.3
Glycine	3.2	3.6	2.9	2.2	4.8	2.7	3.0
Proline	7.6	7.2	10.6	7.2	6.7	10.3	7.2
Serine	4.2	4.9	5.8	6.3	6.9	5.8	5.4

(saturated), and representative values for the fatty-acid compositions of cereals are given in **Table 9**. Stearic and linolenic acids are present in small but significant amounts, and a range of other fatty acids are present in trace amounts.

Dietary Fiber

Although it is not a nutrient, dietary fiber is being increasingly recognized as important in the prevention or alleviation of disease. Fiber can also yield some dietary energy from short-chain fatty acids produced by fermentation in the large intestine. Fiber is concentrated in the outer bran layers of cereals, and thus levels are higher in bran and whole-grain products than in refined milling products (**Tables 3–5**). Dietary fiber includes cellulose and other insoluble and soluble non-starch polysaccharides. Resistant starch, lignin, and other minor components are included in some definitions. A significant amount of soluble fiber (3–5%) occurs as β-glucan gum in oats and barley. A minimum of 5.5% β-glucan is found in oat bran. This gum is the major factor responsible for the reductions in serum cholesterol that result from diets high in these cereals. Wheat bran, which improves gut function, is high in total fiber (40%) but contains only 3–4% soluble fiber.

Table 9 Fatty-acid composition of cereals; representative values in grams per 100 g total fatty acids (total includes 2–3% of trace fatty acids)

	Palmitic acid (16:0)	*Stearic acid (18:0)*	*Oleic acid (18:1)*	*Linoleic acid (18:2)*	*Linolenic acid (18:3)*
Rice	22	2	34	38	2
Maize	12	2	32	50	2
Wheat	18	2	18	56	3
Barley	22	1	13	56	5
Sorghum	13	2	34	46	2
Pearl millet	20	4	26	44	3
Foxtail millet	10	3	17	64	3
Proso millet	9	2	21	64	2
Finger millet	24	2	46	24	1
Kodo millet	18	2	36	40	2
Oats	19	2	36	38	2
Rye	15	1	17	58	7

Micronutrient Content

Micronutrients comprise the inorganic mineral elements and the vitamins. Ash (inorganic mineral matter) comprises 1–3% of grain dry matter. Major mineral elements (potassium, sodium, calcium, phosphorus, and magnesium) and minor or trace elements (iron, zinc, copper, manganese, etc.) are found in all cereals. However, there are significant variations due to processing and other factors (**Tables 10–13**). There can also be substantial variations in the levels of trace minerals between crops due primarily to differences in their availability from the soil.

All cereals provide vitamin E (tocopherols and tocotrienols; tocols), thiamin, riboflavin, niacin, vitamin B_6, pantothenate, folate, and biotin (**Tables 10–13**). Vitamin A (retinol) is not found in cereals. However, carotenes and cryptoxanthins, which yield retinol and thus have provitamin A activity, are found in maize, pearl millet, and sorghum. Levels of provitamin A are variable, with the highest amounts in yellow endosperm types and negligible amounts in white endosperm types. Typical values for retinol equivalents in maize, pearl millet, and sorghum are 44, 42, and 8 μg $100\,g^{-1}$, respectively. Brown rice contains a trace amount (0–11 μg $100\,g^{-1}$, and most of this is lost on milling. Vitamin A deficiency can be a major problem in areas where rice is a dietary staple. In an effort to combat this, rice has recently been genetically modified to produce so called 'golden rice' with provitamin A levels of about 160 μg $100\,g^{-1}$. Vitamins B_{12}, C, and D are not found in unfortified cereals.

Effects of Processing

Vitamin and mineral contents may be profoundly influenced by processing. Vitamins and minerals are found at the highest concentrations in the outer bran layers. Comparison of the various whole-grain and milled products in **Tables 10–12** shows that bran is richer in vitamins and minerals, while flour and meal fractions are depleted. Sodium may be substantially increased by the addition of salt (**Tables 10–12**), leavening agents, or other additives. Other minerals and vitamins are often added as fortification to replace, standardize, or augment the levels naturally present. Although white bread made from fortified flour has a lower mineral and vitamin content than wholemeal bread (**Table 11**), wholemeal bread contains higher levels of phytic acid, which will influence availability (see below). Cornflour contains low levels of minerals and only traces of vitamins. Cornflakes are fortified with a number of minerals and vitamins, including vitamins B_{12} and D (**Table 10**). The fortification of breads and other cereal foodstuffs with folic acid has recently become increasingly common.

Table 10 Mineral and vitamin contents of rice and maize (corn) products; representative values per 100 g fresh weight (water contents as per **Table 3**)

	Brown rice, uncooked	*White rice, uncooked*	*Brown rice, boiled*[a]	*White rice, boiled*[a]	*White rice, flour*	*Maize meal*	*Cornflakes*[b]
Sodium (mg)	3	6	1	2	5	40	1110
Potassium (mg)	250	110	99	38	76	350	100
Calcium (mg)	10	4	4	1	10	20	15
Magnesium (mg)	110	13	43	4	35	140	14
Phosphorus (mg)	310	100	120	34	98	290	38
Iron (mg)	1.4	0.5	0.5	0.2	0.4	3	6.7
Zinc (mg)	1.8	1.3	0.7	0.5	0.8	2	0.3
Copper (mg)	0.85	0.18	0.33	0.06	0.13	0.40	0.03
Manganese (mg)	2.30	0.87	0.90	0.30	1.20	0.60	0.08
Vitamin E (mg)	0.80	0.10	0.30	0.02	0.13	0.50	0.40
Thiamin (mg)	0.59	0.08	0.14	0.01	0.14	0.40	1.00
Riboflavin (mg)	0.07	0.02	0.02	0.01	0.02	0.11	1.50
Niacin (mg)	5.3	1.5	1.3	0.3	2.6	2.2	16.0
Vitamin B_6 (mg)	0.70	0.30	0.30	0.10	0.44	0.53	1.80
Pantothenate (mg)	1.2	0.6	0.4	0.2	0.8	0.6	0.3
Folate (μg)	40	20	10	3	4	40	250
Biotin (μg)	7	3	2	1	n/a	10	2

[a]Unsalted water.
[b]Ready-to-eat cereal, fortified.
n/a, no data available.

Table 11 Mineral and vitamin content of wheat products; representative values per 100 g fresh weight (water contents as per **Table 4**)

	Wholemeal flour[a]	*White flour*[a]	*Bran*	*White bread*[b]	*Wholemeal bread*[a]	*Uncooked pasta*[c]	*Boiled pasta*[c,d]
Sodium (mg)	3	3	28	520	550	8	1
Potassium (mg)	340	130	1160	110	230	237	24
Calcium (mg)	38	15	110	110	54	24	6
Magnesium (mg)	120	31	520	24	76	52	14
Phosphorus (mg)	320	120	1200	91	200	190	45
Iron (mg)	3.9	1.5	12.9	1.6	2.7	1.8	0.5
Zinc (mg)	2.9	0.9	16.2	0.6	1.8	1.5	0.5
Copper (mg)	0.45	0.18	1.34	0.20	0.26	0.30	0.09
Manganese (mg)	3.14	0.68	9.00	0.45	1.90	0.87	0.25
Vitamin E (mg)	1.40	0.30	2.60	trace	0.20	trace	trace
Thiamin (mg)	0.50	0.10	0.90	0.21	0.34	0.30	0.03
Riboflavin (mg)	0.09	0.03	0.36	0.06	0.09	0.05	0.01
Niacin (mg)	5.7	0.7	29.6	1.7	4.1	2.8	0.5
Vitamin B_6 (mg)	0.50	0.15	1.38	0.07	0.12	0.13	0.01
Pantothenate (mg)	0.8	0.3	2.4	0.3	0.6	0.3	trace
Folate (μg)	57	22	260	30	40	30	4
Biotin (μg)	7	1	45	1	6	1	trace

[a]Unfortified.
[b]Made from UK fortified white flour containing 140 mg calcium, 2.1 mg iron, 0.32 mg thiamin, 2.0 mg niacin per 100 g.
[c]Means of lasagne, macaroni, and white spaghetti.
[d]Unsalted water.

Availability

The presence of micronutrients does not ensure availability for metabolic processes. Mineral availability is reduced by the presence of phytic acid and phytates. Phytic acid (myoinositol 1,2,3,4,5,6-hexakis dihydrogen phosphate) accounts for a substantial proportion (usually over 50%) of the total phosphorus in cereals, and this phosphorus is not fully available for digestion and absorption. Phytic acid accounts for about 1% of whole-grain cereals. However, phytic

Table 12 Mineral and vitamin contents of barley, oat and rye products; representative values per 100 g fresh weight (water contents as per **Table 5**)

	Pearl barley, uncooked	*Pearl barley, boiled*[a]	*Oatmeal*	*Oatmeal, porridge*[b]	*Oat bran*	*Wholemeal rye flour*	*Rye crispbread*	*Rye bread*
Sodium (mg)	3	1	21	560	4	1	220	580
Potassium (mg)	270	92	360	46	586	410	500	190
Calcium (mg)	20	7	54	7	79	32	45	80
Magnesium (mg)	65	22	110	18	241	92	100	48
Phosphorus (mg)	210	71	380	47	723	360	310	160
Iron (mg)	3.0	1.0	4.0	0.5	6.1	2.7	3.5	2.5
Zinc (mg)	2.1	0.7	3.3	0.4	4.2	3.0	3.0	1.3
Copper (mg)	0.40	0.14	0.36	0.03	0.31	0.42	0.38	0.18
Manganese (mg)	1.30	0.44	3.80	0.46	5.80	0.68	3.50	1.00
Vitamin E (mg)	0.40	0.10	1.60	0.21	3.30	1.60	0.50	1.20
Thiamin (mg)	0.12	0.02	0.70	0.06	1.10	0.40	0.28	0.29
Riboflavin (mg)	0.05	0.01	0.10	0.01	0.18	0.22	0.14	0.05
Niacin (mg)	2.5	0.5	0.9	0.1	0.9	1.0	1.1	2.3
Vitamin B_6 (mg)	0.22	0.04	0.23	0.01	0.15	0.35	0.29	0.09
Pantothenate (mg)	0.5	0.1	1.1	0.1	1.0	1.0	1.1	0.5
Folate (μg)	20	3	60	4	37	78	35	24
Biotin (μg)	n/a	trace	21	2	38	6	7	n/a

[a]Unsalted water.
[b]Made with water and salt.
n/a, no data available.

Table 13 Mineral and vitamin contents of sorghum and millets; representative values per 100 g fresh weight (water contents as per **Table 6**)

		Millets					
	Sorghum	*Pearl*	*Foxtail*	*Proso*	*Finger*	*Japanese*	*Fonio*
Sodium (mg)	20.5	7.4	6.8	9.4	15.9	n/a	15.0
Potassium (mg)	285	432	243	215	367	n/a	160
Calcium (mg)	28	39	22	14	321	32	30
Magnesium (mg)	156	125	116	104	129	n/a	40
Phosphorus (mg)	291	335	268	220	251	330	175
Iron (mg)	5.1	8.4	5.3	4.7	4.6	4.3	6.0
Zinc (mg)	2.2	3.2	1.9	1.6	1.3	n/a	3.0
Copper (mg)	0.98	0.50	0.71	1.15	0.53	n/a	1.6
Manganese (mg)	1.84	1.45	1.87	1.64	1.19	n/a	3.0
Vitamin E (mg)	1.13	1.69	2.75	1.94	n/a	n/a	n/a
Thiamin (mg)	0.35	0.34	0.51	0.39	0.35	0.33	0.47
Riboflavin (mg)	0.15	0.17	0.10	0.19	0.10	0.10	0.10
Niacin (mg)	3.8	2.0	3.1	1.3	4.0	n/a	1.9
Vitamin B_6 (mg)	0.50	n/a	n/a	n/a	n/a	n/a	n/a
Pantothenate (mg)	1.2	1.1	0.7	1.0	n/a	n/a	n/a
Folate (μg)	19	63	18	n/a	n/a	n/a	n/a
Biotin (μg)	42	n/a	n/a	n/a	n/a	n/a	n/a

n/a, no data available.

acid is concentrated in the bran and germ fractions. Wheat bran and germ contain 3–4% phytic acid, while white endosperm flour contains 0.1–0.2%. Phytic acid reacts with calcium, magnesium, iron, zinc, and copper to form phytates, a process that renders these minerals unavailable for absorption.

Vitamin B_6 and niacin in cereals are also of limited availability. The niacin deficiency disease pellagra may develop where maize grits are the dietary staple. However, niacin, which is present in a bound form, is made available by alkali treatment in traditional maize tortilla production. Niacin can also be synthesized in the body from tryptophan; 60 mg of tryptophan yields 1 mg of niacin.

Dietary Contribution

Data on the relative amounts of minerals and vitamins present in cereals are useful for comparative purposes. However, data on availability and requirements are needed to give a fuller evaluation. Relative to energy content, whole-grain cereals have the potential to contribute significantly to intakes of potassium, magnesium, phosphorus, iron, copper, zinc, vitamin E, thiamin, riboflavin, niacin, vitamin B_6, and folic acid. Significant dietary selenium may also be provided, although levels depend on its availability in the soil.

Non-Nutrients of Potential Benefit

In addition to dietary fiber, cereals contain a number of other non-nutrient components, present in minor amounts, that have the potential to exert beneficial physiological effects. Some of these phytochemicals, which include phytic acid, sterols, phenolics, and flavonoids, have been shown to have *in vitro* antioxidant and oestrogen-like activities. Fiber and phytochemicals are found at their highest levels in bran and unrefined grain. Thus, these components may play an important role in the protective effects against heart disease and certain cancers that are conferred by diets rich in whole grains.

Potential Adverse Effects

Cereals do not have any intrinsic non-specific toxins. However, acrylamide, a carcinogen and potential neurotoxin, has recently been found at levels up to 120 μg $100\,g^{-1}$ in baked and fried foods, including breads and processed cereals. Research is ongoing, but the early indications are that acrylamide from these sources is unlikely to increase cancer risk. Detrimental effects may be caused by antinutrients in cereals and, in susceptible individuals, by adverse immune responses (celiac disease, food allergies). Cereals may also be a source of toxins of fungal origin (mycotoxins) or of toxic environmental, agricultural, or industrial contaminants.

Antinutrients

Phytic acid and phytates are antinutrients that are found in all cereals. They reduce mineral availability

(see above). Tannins are polyphenolic compounds that are found in most cereals. Tannins can bind to protein, reducing its digestibility. Tannins can also inhibit the activity of digestive enzymes. In addition, cereals contain specific protease inhibitors, but the levels are low in comparison with those found in some seed legumes. The tannins and protease inhibitors are unlikely to have any significant adverse effects in human nutrition. However, pearl millet contains phenolic flavonoids, which have been implicated in the onset of goitre, a symptom of iodine deficiency.

Adverse Immune Responses

Many natural products, including cereals and other common foodstuffs, induce allergic responses in susceptible individuals. In such cases, after appropriate diagnosis, the individual should avoid the foodstuff responsible. Celiac disease (gluten enteropathy) is a condition characterized by a severe adverse immunological gastrointestinal reaction to gliadin, which is a component of gluten, the viscoelastic protein found in wheat and other cereals. Celiac disease is prevalent in all regions where wheat is commonly consumed, and its incidence may reach 0.5% of the population. Celiac patients must exclude gluten from their diets. Thus, products containing wheat, rye, barley, and triticale are not permitted. Although oats were originally proscribed, it is now becoming increasingly clear that they are safe for celiac patients.

Mycotoxins

Mycotoxins are produced by fungi, which may infect growing crops and stored grain. Ergot (*Claviceps purpurea*) infects rye and other temperate cereals and produces alkaloid toxins. If ingested in sufficient amounts, these alkaloids induce mental derangement, gangrene, and other symptoms. Aflatoxin, a toxin and potent carcinogen, is produced by the fungus *Aspergillus flavus*, which may occur in maize crops and stored grain. Other fungal toxins include ochratoxin, produced by *Penicillium* species, and trichothecenes, produced by *Fusarium* species. Mycotoxins are less of a problem in cereals than in seed legumes and nuts. Mycotoxin levels can be controlled by correct agronomy and storage, and crops are monitored to ensure that safe levels are not exceeded.

Contaminants

Cereals have the potential to be contaminated with toxic environmental, agricultural, or industrial chemicals in the field and during storage and processing. However, the use of these chemicals is strictly controlled, and incidents of hazardous contamination are rare.

See also: **Bioavailability**. **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Celiac Disease**. **Coronary Heart Disease**: Prevention. **Dietary Fiber**: Physiological Effects and Effects on Absorption. **Folic Acid**. **Food Fortification**: Developed Countries; Developing Countries. **Food Intolerance**. **Food Safety**: Other Contaminants. **Legumes**. **Niacin**. **Nuts and Seeds**. **Pellagra**. **Phytochemicals**: Classification and Occurrence. **Protein**: Deficiency. **Vitamin A**: Biochemistry and Physiological Role. **Whole Grains**.

Further Reading

Anderson JW (2002) Whole-grains intake and risk for coronary heart disease. In: Marquart L, Slavin JL, and Fulcher RG (eds.) *Whole Grain Foods in Health and Disease*, pp. 187–200. St Paul: American Association of Cereal Chemists.

Betschart AA (1988) Nutritional quality of wheat and wheat foods. In: Pomeranz Y (ed.) *Wheat: Chemistry and Technology*, 3rd edn, vol. II, pp. 91–130. St Paul: American Association of Cereal Chemists.

Bhatty RS (1993) Nonmalting uses of barley. In: MacGregor AW and Bhatty RS (eds.) *Barley: Chemistry and Technology*, pp. 355–417. St Paul: American Association of Cereal Chemists.

Dendy DAV (2001) Sorghum and millets. In: Dendy DAV and Dobraszczyk BJ (eds.) *Cereals and Cereal Products: Chemistry and Technology*, pp. 341–366. Gaithersburg: Aspen Publishers Inc.

Haq N and Ogbe FD (1995) Fonio (*Digitaria exilis* and *D. iburua*). In: Williams JT (ed.) *Cereals and Pseudocereals*, pp. 225–245. London: Chapman & Hall.

Juliano BO (1985) Production and utilization of rice. In: Juliano BO (ed.) *Rice: Chemistry and Technology*, 2nd edn, pp. 1–16. St Paul: American Association of Cereal Chemists.

Juliano BO and Bechtel DB (1985) The rice grain and its gross composition. In: Juliano BO (ed.) *Rice: Chemistry and Technology*, 2nd edn, pp. 17–57. St Paul: American Association of Cereal Chemists.

Klopfenstein CF and Hoseney RC (1995) Nutritional properties of sorghum and the millets. In: Dendy DAV (ed.) *Sorghum and Millets: Chemistry and Technology*, pp. 125–168. St Paul: American Association of Cereal Chemists.

Lorenz KJ (1991) Rye. In: Lorenz KJ and Kulp K (eds.) *Handbook of Cereal Science and Technology*, pp. 331–371. New York: Marcel Dekker.

McIntosh GH and Jacobs DR (2002) Cereal-grain foods, fibers and cancer prevention. In: Marquart L, Slavin JL, and Fulcher RG (eds.) *Whole Grain Foods in Health and Disease*, pp. 201–232. St Paul: American Association of Cereal Chemists.

Obilana AB and Manyasa E (2002) Millets. In: Belton PS and Taylor JRN (eds.) *Pseudocereals and Less Common Cereals*, pp. 177–217. Berlin: Springer Verlag.

Pomeranz Y (1988) Chemical composition of kernel structures. In: Pomeranz Y (ed.) *Wheat: Chemistry and Technology*,

3rd edn, vol. I, pp. 97–158. St Paul: American Association of Cereal Chemists.
Poutanen K, Liukkonen K, and Adlercreutz H (2002) Whole grains, phytoestrogens, and health. In: Marquart L, Slavin JL, and Fulcher RG (eds.) *Whole Grain Foods in Health and Disease*, pp. 259–268. St Paul: American Association of Cereal Chemists.
Rooney LW and Serna-Saldivar SO (1987) Food uses of whole corn and dry-milled fractions. In: Watson SA and Ramstad PE (eds.) *Corn: Chemistry and Technology*, pp. 399–429. St Paul: American Association of Cereal Chemists.
Serna-Saldivar S and Rooney LW (1995) Structure and chemistry of sorghum and millets. In: Dendy DAV (ed.) *Sorghum and Millets: Chemistry and Technology*, pp. 69–124. St Paul: American Association of Cereal Chemists.
Shewry PR and Bechtel DB (2001) Morphology and Chemistry of the Rye Grain. In: Bushuk W (ed.) *Rye: Production, Chemistry and Technology*, pp. 69–127. St Paul: American Association of Cereal Chemists.
Taylor JRN and Belton PS (2002) Sorghum. In: Belton PS and Taylor JRN (eds.) *Pseudocereals and Less Common Cereals*, pp. 25–91. Berlin: Springer Verlag.
Watson SA (1987) Structure and composition. In: Watson SA and Ramstad PE (eds.) *Corn: Chemistry and Technology*, pp. 53–82. St Paul: American Association of Cereal Chemists.
Welch RW (1995) Oats in Human Nutrition and Health. In: Welch RW (ed.) *The Oat Crop – Production and Utilization*, pp. 433–479. London: Chapman & Hall.
Welch RW (1995) The Chemical Composition of Oats. In: Welch RW (ed.) *The Oat Crop – Production and Utilization*, pp. 279–320. London: Chapman & Hall.

Cheese *see* **Dairy Products**

CHILDREN

Contents

Nutritional Requirements

M Lawson, Institute of Child Health, London, UK

Nutrition is particularly important in childhood because nutrients are required not only for general health and maintenance of body composition but also for linear growth, neurological development, body maturation, and as a basis for long-term health. There is considerable evidence of an association between early nutrition and later risk of diseases such as cardiovascular disease, obesity, and type 2 diabetes.

Two major factors affect nutrient requirements during infancy, childhood, and adolescence: body size and growth velocity. Growth velocity varies according to age, and age is used as a general proxy for weight velocity. Nutritional requirements for children should therefore be expressed in terms of units of body weight for each age throughout childhood. Nutrient requirements per unit of body weight are highest at birth and reduce as growth velocity decreases.

The pattern of nutrient requirement changes and is not constant per unit of body weight as body composition changes throughout the growing period from birth until the end of puberty. This is illustrated in **Figure 1**, which shows the changes in energy and protein requirements per kilogram body weight between birth and 15 years. Nutrient requirements are therefore both quantitatively and qualitatively different from those of adults.

For infants, breast milk is used as a model for assessing nutrient requirements. However, although breast milk is the 'gold standard' on which infant formulas are based, the physiological actions of nutrients and other components of breast milk and

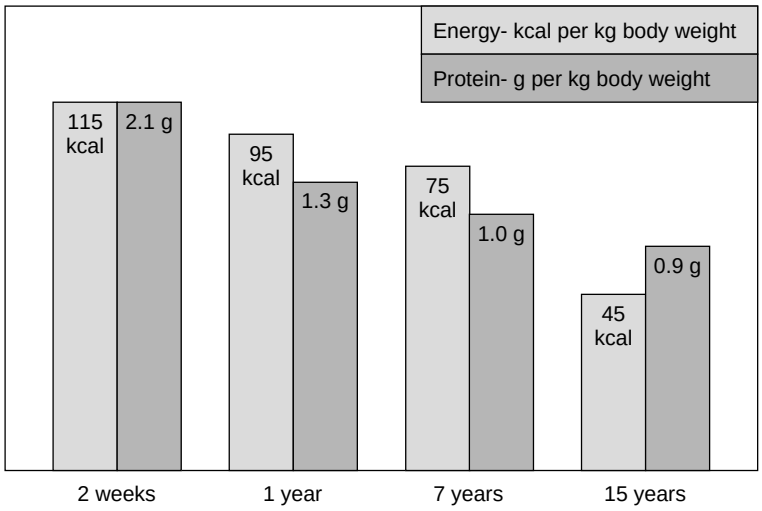


Figure 1 Changes in protein and energy requirements during infancy and childhood.

bioavailability of nutrients are quite different between breast milk and infant formula. In addition, breast milk composition changes during the duration of a feed and it is difficult to estimate the usual volumes of breast milk consumed. Dietary reference values are therefore only applicable to artificially fed infants: An adequate quantity of breast milk is assumed to meet all nutrient requirements for the majority of infants up until the age of approximately 6 months.

Estimates of physiological requirements are used to make dietary recommendations, which depend on the usual diet of the country or area. Again, estimates cannot be 'scaled down' versions of adult requirements because bioavailability of many nutrients varies with age, physiological state, and nutritional status. For example, iron absorption is poor during the first 6 months of life, increases during later infancy, and is greatest during adolescence.

Requirements for children are estimated from a limited number of direct studies of body composition and body content of children at different ages. Most of these were carried out on children living in unfavorable environments and very little data exist on normally nourished children, particularly those aged between 1 and 5 years. Recommendations are often extrapolated from adult studies. For most nutrients, energy and protein requirements are not known with any great precision and are expressed as units per day.

Discrepancies between recommendations published by different countries or authorities are due to a number of factors: Different assumptions are made about weight at a particular age, growth velocity, and age of onset of puberty, and different age groupings are used by different authorities. Some countries give recommendations for each year of life, whereas others aggregate several years together, giving an overestimation for the youngest ages and underestimation for older children in each age band. Some countries separate recommendations for males and females as early as the second year; others separate them at a later age or only during adolescence. **Table 1** illustrates the framework on which different recommendations and reference values are based. Diets of varying composition will affect bioavailability and, as a consequence, recommendations for dietary intake. Therefore, it is difficult to compare recommendations from different authorities and very large discrepancies exist in recommendations for some nutrients.

Energy

Most energy recommendations for infants and children are based on the 1985 FAO/WHO/UNU report and are shown in **Tables 2** and **3**. The FAO/WHO data have been reviewed recently and it is likely that a number of revisions to the 1985 document are required in light of additional information and the development of new assessment techniques such as doubly labeled water. Energy requirements expressed in terms of body weight are highest during the first few months of life, decrease fairly sharply after the age of 1 year, and then show a gradual decline until the onset of puberty, when they increase. This mirrors the growth velocity seen at different ages.

Table 1 Terminology for nutritional recommendations

Authority	*Mean − 2SD*	*Mean*	*Mean + 2SD*	*Less evidence-based data*	*Upper limit of intake*
United Kingdom	Lower Reference Nutrient Intake (LNRI)	Estimated Average Requirement (EAR)	Reference Nutrient Intake (RNI)	Estimated Safe + Adequate Dietary Intake (ESADI)	
European Union	Lowest Threshold of Intake (LTI)	Average Requirement (AR)	Population Reference Intake (PRI)	Acceptable Ranges	
USA/ Canada	—	Estimated Average Requirement (EAR)	Recommended Dietary Allowance (RDA)	Adequate Intake (AI)	Tolerable Upper Intake Level (TUL)
FAO/WHO	—	Estimated Average Requirement (EAR)	Recommended Dietary Intake (RDI)	—	Upper Tolerable Nutrient Intake (TUL)

The review suggests that the 1985 recommendations for dietary energy are too high for children younger than 5 years and possibly those younger than 7 years of age, whereas recommendations for adolescent boys and pubertal girls appear to be set too low, particularly in developing countries. The 1985 recommendations increased reported energy intakes by a factor of 5% to accommodate a 'desirable' level of physical activity. This may not be realistic in the increasingly sedentary environment seen in industrialized countries, and it has been suggested that recommendations for energy intake should be accompanied by recommendations for activity levels.

The proportion of dietary energy that is required to sustain normal growth varies according to the growth velocity at that particular age. **Figure 2** shows the percentage of the energy requirement that is needed for maintenance and growth at different ages. When estimating the likely energy requirement for a child who is unusually inactive (e.g., children with mobility difficulties), the estimate, based on the Estimated Average Requirement for age, should be reduced by the percentage of dietary energy normally required for activity (e.g., 29% at age 4 or 5 years). There is evidence that formula-fed

Table 2 Recommended energy intakes for infants (kJ/kg body weight/day)

Age (months)	*UK EAR*	*Europe AR*	*USA EAR*	*FAO/WHO EAR*
0–3	480	465	450	480
4–6	420	440	450	420
7–9	400	410	410	400
10–12	400	405	410	400

Table 3 Recommended energy intakes for Children (MJ/day)

Age (years)	*Sex*	*UK EAR*	*Europe AR*	*USA EAR*	*FAO/ WHO EAR*
2	M	4.9	5.0	5.4	5.9
	F	4.9	4.8	5.4	5.5
3	M	6.2	6.0	5.4	6.5
	F	5.7	5.6	5.4	6.0
4	M	6.7	6.6	7.5	7.1
	F	6.1	6.2	7.5	6.4
5	M	7.2	7.1	7.5	7.6
	F	6.5	6.8	7.5	6.8
6	M	7.6	7.7	7.5	7.9
	F	6.8	7.1	7.5	7.1
7	M	7.9	81	8.4	8.3
	F	7.0	7.3	8.2	7.4
8	M	8.2	8.3	8.4	8.7
	F	7.3	7.4	8.2	7.6
9	M	8.5	8.6	8.4	9.0
	F	7.5	7.4	8.2	0.79
10	M	8.2	8.7	8.4	10.5
	F	7.3	7.6	8.2	9.6
11	M	9.3	9.2	10.5	10.9
	F	7.7	8.0	9.0	9.8
12	M	9.3	9.8	10.5	11.3
	F	7.7	8.3	9.0	10.0
13	M	9.3	10.6	10.5	11.7
	F	7.7	9.0	9.0	10.2
14	M	9.3	10.9	10.5	12.1
	F	7.7	8.7	9.0	10.4
15	M	11.5	11.4	12.5	12.5
	F	8.8	8.9	9.2	10.5
16	M	11.5	11.9	12.5	12.8
	F	8.8	9.0	9.2	10.1
17	M	11.5	12.0	12.5	13.0
	F	8.8	9.0	9.2	9.8
18	M	11.5	11.9–12.5[a]	12.5	13.0
	F	8.8	8.3–10.6[a]	9.2	9.8

[a]Depends on physical activity level.
F, female; M, male.

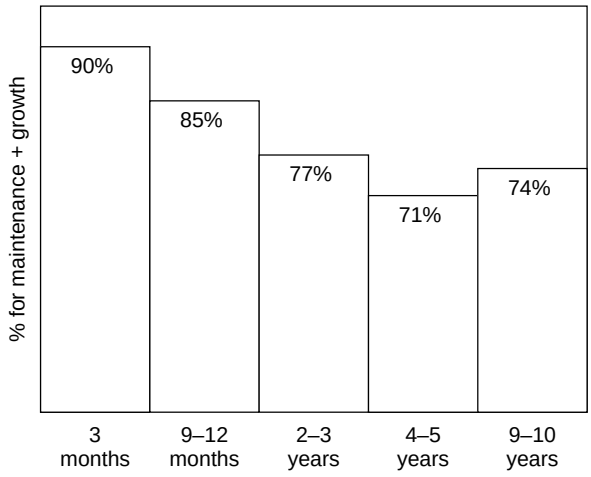


Figure 2 Energy required for maintenance and growth as percentage of total energy expenditure.

infants require a slightly increased energy intake to achieve the same growth velocity as breast-fed infants; children suffering from malnutrition have also been reported as having an increased energy expenditure and therefore may have a higher energy requirement than normal for their age.

Protein

Most recommendations for the protein intake of children are based on the 1985 FAO/WHO report, although much of the data were collected from studies on formula-fed infants. Because of differences in growth rates, efficiency of protein utilization, and the amino acid pattern of breast and cow's milk, these recommendations are likely to be an overestimation of true requirements. More recent data compiled on breast-fed infants give lower estimates of requirements. The Dewey estimates of minimum and safe protein intakes plus recommendations from other authorities are shown in **Tables 4** and **5**. The minimum and safe intakes refer to a diet based on high biological value proteins such as breast milk or eggs. Adjustments need to be made for infants receiving an alternative source of protein (e.g., soya or a protein hydrolysate) and for older children consuming a mixed diet.

Protein requirements are highest during the first month of life and decrease thereafter. The proportion of protein intake that is required for growth decreases from 64% of requirement during the first month of life to 35% at age 3–6 months, 16% at age 1–2 years, and 11% at age 2–5 years.

Essential amino acid requirements are dependent on the growth rate of the infant and the rate of protein deposition, which changes throughout infancy. Some nonessential amino acids, including creatine, taurine, glycine, cysteine histidine, and arginine, cannot be synthesized in adequate quantities to meet the demands of the very rapid protein deposition that occurs during the first month of life, and these are considered to be semiessential during early infancy. It is likely that part of the nonprotein nitrogen portion of breast milk, such as choline, carnitine, and nucleotides, is used in metabolism and for amino acid synthesis and may also be conditionally essential.

Some countries express protein as a percentage of the estimated energy requirements, with a range of 8–15% energy from protein (i.e., 2–3.75 g protein per 100 total kilocalories). The protein: energy ratio is approximately 7.5% in human milk and 8–8.5% in infant formulas. These ratios are adequate for a normal rate of growth, although it may be argued that the more rapid growth rate seen in young formula-fed infants indicates that the ratio has been set too high in infant formulas.

Table 4 Recommended protein intakes for infants (g/kg body weight/day)

Age (months)	*UK RNI*	*Europe PRI*	*USA Recommended Dietary Allowance*	*FAO/WHO*[a]	
				Minimum requirements[b]	*Safe minimum intake*[c]
0–1	2.11		2.16	1.99	2.69
1–2	2.11		2.16	1.54	2.04
2–3	2.11		2.16	1.19	1.53
3–4	1.65		2.16	1.06	1.37
4–5	1.65	1.81	2.16	0.98	1.25
5–6	1.65	1.81	2.16	0.92	1.19
7–9	1.55	1.64	1.55	0.85	1.09
10–12	1.54	1.50	1.55	0.78	1.02

[a]From Dewy *et al.* (1996).
[b]Statistically will only meet requirements of 50% of the population.
[c]Should meet the requirements of 97.5% of the population.

Table 5 Recommendations for protein intake for children

Age (years)	Sex	UK RNI (g/day)	Europe PRI (g/day)	USA RDA (g/day)	WHO[a]	
					Minimum requirements (g/kg/day)[b]	Safe minimum intake (g/kg/day)[c]
2	M+F	14.5	15.5	16.0	0.74	0.92
3	M+F	14.5	17.0	16.0	0.72	0.90
4	M+F	19.7	18.5	24	0.71	0.88
5	M+F	19.7	20.0	24	0.69	0.86
6	M+F	19.7	22.0	24	0.69	0.86
7	M+F	28.3	24.5	28	0.69	0.86
8	M+F	28.3	27.5	28	0.69	0.86
9	M+F	28.3	29.5	28	0.69	0.86
10	M	28.3	32.5	28	0.69	0.86
	F	28.3	34.0	28	0.69	0.86
11	M	42.1	36.0	45	0.69	0.88
	F	41.2	37.0	46	0.69	0.86
12	M	42.1	41.0	45	0.71	0.88
	F	41.2	41.5	46	0.69	0.86
13	M	42.1	45.5	45	0.71	0.88
	F	41.2	45.0	46	0.69	0.86
14	M	42.1	51.0	45	0.69	0.86
	F	41.2	45.5	46	0.68	0.84
15	M	55.2	53.5	59	0.69	0.86
	F	45.4	45.5	44	0.66	0.82
16	M	55.2	46.5	59	0.68	0.84
	F	45.4	45.0	44	0.66	0.81
17	M	55.2	55.5	59	0.67	0.83
	F	45.4	43.5	44	0.63	0.78
18	M	55.2	56.0	59	0.65	0.81
	F	45.4	47.0	44	0.63	0.78

[a]From Dewey *et al.* (1996).
[b]Statistically will only meet the requirements of 50% of the population.
[c]Should meet the requirements of 97.5% of the population.
F, female; M, male.

An increase in the protein:energy ratio is necessary for infants and children who are wasted or stunted and who need to grow at an increased velocity in order to catch up. Malnutrition during the first year of life, whether a result of a poor environment or because of conditions such as malabsorption or cystic fibrosis, is more serious than wasting and stunting later in childhood. Although the potential for catch-up growth remains until the end of puberty, deficits during early life can lead to permanent impairment of cognitive function. Estimates of requirements for malnourished infants and children should be based on the height age (i.e., the age at which the child's measured height falls on the 50th centile) because estimations based on normal requirements for chronological age are likely to be difficult to achieve and may lead to obesity. The increase in protein requirement for catch-up growth is proportionally greater than the increase in energy requirement and is dependent on age and growth velocity. For example, a child aged 1 year who is growing at twice the normal rate has a 5% increase in energy requirements and a 32% increase in protein requirements. When designing regimens for infants and children who need to catch up, a protein:energy ratio of at least 10% and possibly up to 15% is necessary. Ratios less than this will result in changes in body composition, with greater amounts of fat and water and lower amounts of lean tissue being deposited.

Water

The water content of the body is highest at birth (70%) and declines gradually to the adult value of 60% of body composition. The proportion of the fluid requirement that is needed for growth is small—approximately 5% soon after birth, decreasing to 1% at age 1 year. Fluid intake in infants is particularly important: They are unable to signify thirst, renal function is immature during the first few months of life, resulting in high obligatory losses of water, and extrarenal losses are high due to the high surface area of infants and young children. Physiological

requirements for water are quite variable, are likely to be higher than adult requirements, and depend on climate, physical activity, and habitual diet.

Water is rarely on the list of nutrients for which dietary recommendations exist, and only one authority (Austria/Germany/Switzerland) recommends intakes that are based on energy intake and urine osmolality. Where recommendations do exist, they are usually approximately 1 ml water per kilocalorie of energy intake and are the same for both children and adults.

Fat, Carbohydrate, Fiber, and Recommendations for Healthy Eating

The transition from the high-fat, milk-based diet of the young infant to the generally accepted adult recommendations for healthy eating should be gradual, beginning from the onset of weaning. There is little consensus as to when qualitative adult intakes should be achieved, although dietary modification is not recommended by any authority for children younger than the age of 2 years and most agree that an adult-type diet is appropriate from the age of 5 years. Some authorities recommend a gradual change between the ages of 3 and 5 years, whereas others suggest that for most children a low-fat, cereal/vegetable predominant diet is suitable from age 2 years. If changes toward 'healthy eating' are made at too young an age, there is a danger that an inadequate energy and nutrient intake will result because infants and young children find it difficult to consume adequate quantities of such a bulky diet.

Fat

Fat contributes more than 50% of the energy intake in infants and is necessary to deliver the infant's high energy requirement in a small volume. Fat continues to be important during the first and second years of life because of the relatively high energy requirements per unit of weight. The United Kingdom, United States/Canada, and FAO/WHO do not recommend a specific fat:energy ratio for young children; some countries allow for higher fat intakes in young children, whereas other European countries suggest 25–30% energy from fat from the age of 2 years. Guidance on the intakes of fatty acids varies considerably throughout the world, and there are considerable differences in the way that recommendations are expressed where they do exist. Saturated fatty acids, where recommendations exist, are usually limited to a maximum of 10% total energy intake. For polyunsaturated fatty acids, recommendations for children are similar to those for adults (5–10% of energy intake), except for infants, for whom a higher percentage is often deemed desirable, similar to quantities in breast milk. Many countries do not make a recommendation for omega-3 and omega-6 fatty acids. Where recommendations exist, they are usually the same as for adults: 2–4% and 0.5% of total energy intake for omega-6 and omega-3, respectively. Some authorities recommend a ratio of <5:1 omega-6:omega-3 for adult intakes, but it is unclear whether this recommendation is appropriate for children.

Carbohydrate

Few countries set reference values for carbohydrates for children. Young infants, particularly those born preterm, have a limited capacity for glucose synthesis and require approximately 40% of their energy from simple sugars in order to maintain adequate blood glucose levels. Breast milk and infant formulas normally contain approximately 7 g carbohydrate per 100 ml. For older children there is no reason to believe that requirements for carbohydrates are different from those for adults. Where recommendations for total carbohydrates and for simple or nonmilk extrinsic sugars are set, they are usually the same as those for adults.

Fiber

A lack of agreement on the definition of fiber and differences in analytical techniques make it difficult to compare recommendations from different sources, and there appears to be a 10-fold variation in recommendations worldwide. Some authorities (United Kingdom, European Union (EU), and FAO/WHO) set no reference values for children. Estimations of desirable fiber intakes are based on adult data corrected for body weight and energy requirement. A popular concept (which is not evidence-based) is the 'age + 5' concept. This recommendation states that children older than the age of 2 years consume an amount of fiber equivalent to their age in years plus 5 g daily. Thus, fiber recommendations increase by 1 g per year until adult values are reached at age 15–18 years. Infants consume a very low-fiber diet, although oligosaccharides in breast milk are thought to have fiber-like properties. fiber should be introduced gradually into the weaning diet from age 6 months, but the use of large quantities of whole grain cereals and pulses or nuts is not recommended in infancy because they are likely to affect the bioavailability of micronutrients and result in a bulky low-energy diet.

Mineral Elements

Requirements for mineral elements should be expressed per unit of body weight since excess minerals are excreted by the kidney. In addition, requirements are directly proportional to growth rates: Sodium and potassium increments reflect the amount of water retained during cellular growth; calcium, phosphorus, and magnesium requirements are in proportion to bone growth and maturation. However, few countries express recommendations in terms of body weight, reflecting the lack of direct data for infants and children.

Sodium and Potassium

Recommendations for sodium and potassium intakes are given in **Table 6**. Young infants are less efficient at excreting sodium than older children, and hypernatremic dehydration can occur in infants being fed a high-sodium milk (such as unmodified cow's milk) if they have any extrarenal water losses such as diarrhea. Many countries do not give recommendations for sodium or potassium. Recommendations that do exist are often based on usual sodium intakes, which in Western countries are far in excess of requirement. Because of the relationship between hypertension and sodium intake in populations, most Western countries recognize that a decrease in sodium and a concurrent increase in potassium intake are desirable. In the United Kingdom, in 2003 the Scientific Advisory Committee on Nutrition set target recommendations for salt intake in children. Infants younger than the age of 1 year should take 17 mmol or less per day; for ages 1–3, 4–6, 7–10, and 11–14 years, the targets are 34, 51, 85, and 102 mmol per day, respectively. The target recommendations are higher than the UK Reference Nutrient Intake but below estimated intakes of infants and children in the United Kingdom and below the upper limit of the EU acceptable range.

Table 6 Recommendations for sodium and potassium intake, infants and children (mmol/day)

Nutrient	*Age*	*UK LNRI*	*EU acceptable range*	*USA minimum requirement*
Sodium	0–6 months	6	—	—
	6–12 months	9	—	—
	1–3 years	9	—	13
	4–6 years	12	25–152	13
	7–10 years	15	25–152	17.5
	11–14 years	20	25–152	22
	15–18 years	25	25–152	22
		RNI	*EU PRI*	*USA RDA*
Potassium	0–3 months	20	—	12.8
	4–6 months	22	—	18
	7–12 months	08	20	26
	1–3 years	20	20	36
	4–6 years	28	28	36
	7–10 years	50	41	41
	11–14 years	80	51	51
	15–18 years	90	51	51

Calcium, Phosphorus, and Magnesium

Recommendations for calcium, phosphorus, and magnesium intakes are given in **Table 7**. A low calcium intake is rarely a cause of rickets, although calcium deficiency has been reported in children who have no milk or dairy products in their diet. An adequate calcium intake in childhood is particularly important in order to maximize bone density. There is some controversy over the optimum calcium intake during childhood in order to achieve maximum bone density, with US and FAO/WHO

Table 7 Recommendations for minerals for infants and children (mg/day)

Nutrient	*Age*	*UK RNI*	*Europe PRI*	*USA RDA*	*FAO/WHO RDI*
Calcium	0–6 months	525	—	440	500
	6–12 months	525	400	600	600
	1–3 years	350	400	500	500
	4–6 years	450	450	800	600
	7–10 years	550	550	800	700
	11–14 years				
	M	1000	1000	1300	1300
	F	800	800	1300	1300
	15–18 years				
	M	1000	1000	1300	1300
	F	800	800	1300	1300
Phosphorus	0–6 months	400	—	300	—
	6–12 months	400	300	500	—
	1–3 years	270	300	460	—
	4–6 years	350	350	500	—
	7–8 years	200	450	500	—
	9–10 years	450	450	1250	—
	11–18 years				
	M	775	775	1250	—
	F	625	625	1250	—
Magnesium	0–6 months	60		40	
	6–12 months	80		60	
	1–3 years	85	85	80	60
	4–6 years	120	120	130	76
	7–8 years	200	200	130	100
	9–10 years	200	200	240	100
	11–14 years				
	M	280	280	240	230
	F	280	280	240	220
	15–18 years				
	M	300	300	410	230
	F	300	300	350	220

F, female; M, male.

recommendations set higher than EU and UK ones. Most countries differentiate between boys and girls at adolescence to account for greater bone mass in males.

Recommendations for phosphorus intakes show a similar variation. Estimates of requirements assume that there is an optimum ratio of calcium:phosphorus in the diet, so phosphorus recommendations are based on those for calcium. The optimum molar calcium:phosphorus ratio during childhood is assumed to be 1:1. The recommended ratio in infant formulas ranges from 1.3:1 to 2.1:1; the ratio in human milk is approximately 2.3:1.

There are limited data on magnesium requirements during infancy and childhood, and this is reflected in the variation of recommendations worldwide. Magnesium deficiency is rarely reported in healthy children and recommendations are largely based on normal intakes.

Trace Elements

Recommendations for trace element intakes in infancy and childhood are shown in **Tables 8** and **9**. In general, there are few data on trace element requirements in young children, although some studies have been carried out in infants. In the manufacture of breast milk substitutes, quantities available in breast milk are considered to be adequate, although the bioavailability of trace elements added to cow's milk-based formula has not been fully elucidated. Preterm infants are born with low stores of trace elements. They require supplements of iron until age 1 year. Deficiencies of selenium and copper have also been described in preterm and low-birthweight infants. Several authorities do not set dietary recommendations for molybdenum, manganese, and chromium, although all are essential for growth development and health. Recommendations for selenium have been set by the United Kingdom, EU, United States, and WHO. Where data for infants exist, recommendations for children are set between those for infants and those for adults. No reference values have been set for fluorine by any authority, although the effect of fluorine on the inhibition and reversal of caries progression is well accepted and is particularly important during tooth eruption and growth during infancy and childhood.

Iron

Recommendations for iron intake in infancy and childhood are shown in **Table 8**. Term infants are born with sufficient stores of iron to last until they have approximately doubled their birth weight.

Table 8 Recommendations for iron intake in infants and children (μmol/day)

Age	*Sex*	*UK RNI*	*EU PRI*	*USA RDA*	*FAO/WHO RNI*
0–3 months	M + F	30	—	105	—
4–6 months	M + F	80	—	105	150
7–12 months	M + F	140	105	180	150
1–2 years	M + F	120	70	130	90
2 years	M	120	70	120	70
	F	120	70	128	70
3 years	M	120	70	135	70
	F	120	70	130	70
4 years	M	110	70	140	75
	F	110	70	145	75
5 years	M	110	70	145	75
	F	110	70	150	75
6 years	M	110	70	170	75
	F	110	70	160	75
7 years	M	160	107	185	105
	F	160	107	180	105
8 years	M	160	107	200	105
	F	160	107	190	105
9–10 years	M + F	160	107	140	105
11–14 years	M	200	180	140	270
	F	260	320–390	140	285
15–18 years	M	200	230	195	225
	F	260	320–390	270	370

F, female; M, male.

There is controversy over whether healthy infants have a requirement for dietary iron until the age of 4–6 months. Although most infant formulas are fortified with iron, there are some unfortified ones reflecting the uncertainty over requirements in very young infants. After the age of 4–6 months or when birth weight has doubled, iron requirements are very high. Levels of iron in breast milk are low, although bioavailability is high. Nevertheless, infants exclusively breast-fed after the age of 6 months have lower iron stores than those who receive a fortified formula or iron-containing complementary foods. Cow's milk is low in iron and the iron is poorly absorbed. Infants fed on cow's milk as a main drink younger than the age of 1 year or who consume large quantities of cow's milk after the age of 1 year are at risk of developing iron deficiency. Iron is required in early life not only for adequate growth but also because it is important in brain growth, and iron deficiency during infancy may lead to irreversible changes in mental and motor development. It is estimated that 43% of infants and children worldwide suffer from iron deficiency in infancy and childhood, most commonly between the ages of 6 and 24 months. The problem is worse in developing countries: The prevalence in Western industrialized countries is

Table 9 Recommendations for trace minerals

Nutrient	*Age*	*UK RNI*	*Europe PRI*	*USA RDA*	*FAO/WHO RDI*
Zinc (mg/day)	0–6 months	4.0	—	5.0	3.1–5.3[a]
	6–12 months	4.0	4.0	5.0	5.6
	1–3 years	5.0	4.0	3.0	5.5
	4–6 years	6.5	6.0	5.0	6.5
	7–10 years	7.0	7.0	5.0–8.0	7.5
	11–14 years				
	M	9.0	9.0	8.0	12.1
	F	9.0	9.0	8.0	10.3
	15–18 years				
	M	9.5	9.0	11.0	13.1
	F	7.0	7.0	9.0	10.2
Copper (mg/day)	0–6 months	0.2–0.3	—	0.4–0.6[b]	0.33–0.62
	6–12 months	0.3	0.3	0.6–1.0	0.6
	1–3 years	0.4	0.4	1.0–1.5	0.56
	4–6 years	0.6	0.6	1.0–2.0	0.57
	7–10 years	0.7	0.7	1.0–2.0	0.75
	11–14 years				
	M	0.8	0.8	1.5–2.5	1.0
	F	0.8	0.8	1.5–2.5	1.0
	15–18 years				
	M	1.0	1.0	1.5–3.0	1.33
	F	1.0	1.0	1.5–3.0	1.15
Selenium (μg/day)	0–6 months	10–13	—	10	6–9
	6–12 months	10	8	15	12
	1–3 years	15	10	20	20
	4–6 years	20	15	20	24
	7–10 years	30	25	30	25
	11–14 years				
	M	45	35	40	36
	F	45	35	45	30
	15–18 years				
	M	70	45	50	40
	F	60	45	50	30
Iodine (μg/day)	0–6 months	56–60	—	40	50
	6–12 months	60	50	50	50
	1–3 years	70	70	70	90
	4–6 years	100	90	90	90
	7–10 years	110	100	120	120
	11–14 years				
	M	130	120	150	150
	F	130	120	150	150
	15–18 years				
	M	140	130	150	150
	F	140	130	150	150

[a]Assuming diets of moderate bioavailability.
[b]USA Adequate Intake (not RDA).
F, female; M, male.

approximately 10% in children younger than the age of 2 years. Immigrant groups to Western countries, particularly those of Asian origin, have a higher prevalence rate than Caucasian children. This may in part be accounted for by differences in sources of dietary iron. Absorption varies according to the composition of the diet. Recommendations from FAO/WHO take this into consideration and give reference values for four absorption levels: 5, 10, 12, and 15%, with a 3-fold difference in recommendations between the lowest and highest level. Throughout the world, there is considerable variation in iron intake recommendations, part of which is due to differences in dietary composition. Most countries include two reference intakes for adolescent girls, depending on whether or not they have reached menarche.

Zinc

Breast milk provides adequate zinc for term infants until birth weight has approximately doubled. Requirements for zinc are particularly high in infancy because of the demands of growth, and zinc-containing complementary foods or a fortified infant formula is needed to meet requirements after the age of approximately 6 months. Zinc deficiency in childhood is common in developing countries, leading to slow weight gain in infancy and impaired linear growth in children. In unfavorable environments, zinc supplementation of infants and young children is associated with improvements in growth and a reduction of up to 25% in the incidence of diarrhea and 40% reduction in the incidence of pneumonia. Recommendations for zinc intakes vary considerably throughout the world. Zinc absorption is dependent on the composition of the diet and, as with iron, FAO/WHO reference intakes give three values for each age group for low, medium, and high bioavailability.

Copper

Although levels of copper in breast milk appear to be low, it is readily absorbed. Breast-fed term infants obtain adequate copper to meet requirements from stores in the liver and from breast milk until they have approximately doubled their birth weight. Unmodified cow's milk is particularly low in copper and does not meet the requirements for infants if used as a main drink under the age of 1 year; modified infant formulas are fortified with copper. Deficiency has been reported in infants and young children who consume a diet containing large quantities of cow's milk and few complementary foods. Most recommendations for copper intake in children have been extrapolated from infant and adult data and recommendations vary considerably worldwide.

Iodine

Adequate thyroid function is essential for optimal growth and development, and hypothyroidism due to iodine deficiency is seen in many developing countries and some areas of Eastern Europe. The breast milk of mothers who are iodine deficient does not provide an adequate intake for infants, and WHO has described childhood iodine deficiency as the most common cause of preventable brain damage in the world. Infants and children consuming vegetarian and vegan diets have a lower intake than children who consume dairy products. There is little recent data on iodine requirements in infants and children, and most data are extrapolated from adult requirements. Recommendations worldwide for desirable iodine intakes are similar with little variation.

Water-Soluble Vitamins

Recommendations for dietary intakes of water-soluble vitamins are shown in **Table 10**. Recommendations vary throughout the world. There are few direct data on infants and children and most recommendations are extrapolated from adult data. Recommendations for children for thiamine and for some other B vitamins are set relative to energy intakes; vitamin B_6 recommendations are expressed in terms of protein intake and are the same as for adults. The ratio of vitamins to macronutrients is set lower for infants younger than the age of 1 year, reflecting the relative quantities in breast milk. However, there are very few data on the bioavailability and efficiency of utilization of these nutrients in fortified infant formula and in complementary foods. Few countries set recommendations for biotin and pantothenic acid. Recommendations for some other B vitamins are expressed as safe intakes because there is insufficient data on which to base estimates of requirements or recommendations.

Deficiencies of water-soluble vitamins are rare in European and other Westernized countries. Infants are born with small stores of folate and can quickly become depleted if breast milk levels are low. A deficiency of folic acid is the most common cause of megaloblastic anemia in childhood. Infants of vegan mothers also have small stores of vitamin B_{12} and breast milk levels are likely to be low. Children consuming a macrobiotic or strict vegan diet are at risk of not meeting requirements for vitamin B_{12} unless they receive a supplement or a fortified infant soya formula.

Vitamin C is particularly important in childhood not only because of its functions as a vitamin but also because it improves the absorption of non-heme iron. Failure to consume foods rich in vitamin C at the same time as vegetable sources of iron plays a part in the etiology of iron deficiency in childhood. Levels of vitamin C in cow's milk are low and infantile scurvy has been reported in infants receiving unmodified cow's milk as a main drink. Infant formulas are fortified with vitamin C. Estimates of requirements are generally extrapolated from adult data. Intakes of vitamin C that were low (but that met current recommendations) have been associated with increased prevalence of asthma in childhood.

Table 10 Recommendations for water-soluble vitamins

Nutrient	*Age*	*UK RNI*	*Europe PRI*	*USA RDA*	*FAO/WHO RDI*
Thiamine (mg/day)	0–6 months	0.2	—	0.3	0.3
	6–12 months	0.3	0.3	0.4	0.3
	1–3 years	0.5	0.5	0.7	0.5
	4–6 years	0.7	0.7	0.9	0.6
	7–10 years	0.7	0.8	1.0	0.9
	11–14 years				
	M	0.9	1.0	1.3	1.2
	F	0.7	0.9	1.2	1.1
	15–18 years				
	M	1.1	1.1	1.1	1.2
	F	0.8	0.9	1.0	1.1
Riboflavin (mg/day)	0–6 months	0.4	—	0.4	0.5
	6–12 months	0.4	0.4	0.5	0.5
	1–3 years	0.6	0.8	0.5	0.5
	4–6 years	0.8	1.0	0.6	0.6
	7–10 years	1.0	1.2	0.6–0.9	0.9
	11–14 years				
	M	1.2	1.4	0.9	1.3
	F	1.1	1.2	0.9	1.0
	15–18 years				
	M	1.3	1.6	1.3	1.3
	F	1.1	1.3	1.0	1.0
Niacin (NE/day)[a]	0–6 months	3	—	5	5.4
	6–12 months	4–5	5	6	5.4
	1–3 years	8	9	6	6
	4–6 years	11	11	8	8
	7–10 years	12	13	8–12	12
	11–14 years				
	M	15	15	12	16
	F	12	14	12	16
	15–18 years				
	M	18	18	16	16
	F	14	14	14	16
Vitamin B_6 (mg/day)	0–6 months	0.2	—	0.3	0.3
	6–12 months	0.3–0.4	0.4	0.6	0.6
	1–3 years	0.7	0.7	0.5	0.5
	4–6 years	0.9	0.9	0.6	0.6
	7–10 years	1.0	1.1	0.6–1.0	1.0
	11–14 years				
	M	1.2	1.3	1.0	1.3
	F	1.2	1.1	1.0	1.2
	15–18 years				
	M	1.5	1.5	1.3	1.3
	F	1.5	1.1	1.2	1.2
Vitamin B_{12} (μg/day)	0–6 months	0.3	—	0.3	0.3
	6–12 months	0.4	0.5	0.5	0.5
	1–3 years	0.5	0.7	0.9	009
	4–6 years	0.8	0.9	1.2	1.2
	7–10 years	1.0	1.0	1.2–1.8	1.8–2.4
	11–14 years				
	M	1.2	1.3	1.8	2.4
	F	1.2	1.2	1.8	2.4
	15–18 years				
	M	1.5	1.4	2.4	2.4
	F	1.5	1.4	2.4	2.4
Folate (μg/day)	0–6 months	50	50	50	50
	6–12 months	50	50	50	50
	1–3 years	70	100	150	100
	4–6 years	100	130	200	130
	7–10 years	150	150	200–300	150

Continued

Table 10 Continued

Nutrient	*Age*	*UK RNI*	*Europe PRI*	*USA RDA*	*FAO/WHO RDI*
	11–14 years				
	M	200	180	300	180
	F	200	180	300	180
	15–18 years				
	M	200	200	400	200
	F	200	200	400	200
Vitamin C (mg/day)	0–6 months	25	—	20	20
	6–12 months	25	20	20	20
	1–3 years	30	25	15	30
	4–6 years	30	25	25	30
	7–10 years	30	30	25–45	35
	11–14 years				
	M	35	35	45	40
	F	35	35	45	40
	15–18 years				
	M	40	45	75	40
	F	40	40	65	40

[a]NE, nicotinic acid equivalent: 1 mg = 60 mg tryptophan.
F, female; M, male.

Fat-Soluble Vitamins

Recommendations for fat-soluble vitamins are shown in **Table 11**. There are few direct data on requirements for infants and children, and many of the recommendations are derived from adult data.

Vitamin A

Breast milk content of vitamin A is dependent on the mother's status, and term infants who are breast fed by a well-nourished mother will meet their requirements until approximately 6 months of age. The United Kingdom recommends a supplement of vitamin A for breast-fed infants between the ages of 6 months and 5 years because it is difficult to obtain adequate quantities from weaning foods. Clinical vitamin A deficiency is a major cause of childhood blindness in developing countries but is rarely seen in Western countries. There is growing evidence that mild or subclinical vitamin A deficiency is associated with increased susceptibility to infection.

Vitamin D

Vitamin D requirements in adults and older children are met by the exposure of the skin to sunlight, but infants and young children and those who may not receive adequate exposure have a dietary requirement for vitamin D. Breast milk content depends on maternal status, but breast milk and body stores should provide sufficient vitamin D to meet requirements until the age of 6 months. After this age, several countries recommend a supplement for breast-fed infants. Formula milks are fortified with vitamin D. Deficiency of vitamin D in childhood causes rickets. In Western countries dark-skinned immigrant groups, particularly preschool children and adolescent Asian groups, are especially at risk of low levels of vitamin D because they require more sunlight exposure to synthesise adequate amounts of vitamin D compared to fairer skinned groups. The use of sun-block creams for children in Western countries reduces skin synthesis of vitamin D and may contribute to low levels. Recommendations for dietary intakes of children vary throughout the world. Some authorities make a recommendation but indicate that it may only apply to those without access to sunlight. Others make no general recommendations for children older than the age of 2 years except for those at risk of low skin synthesis.

Vitamin E

Vitamin E is particularly important for preterm infants because placental transfer is low and stores at birth are poor. Preterm infants are subjected to high levels of oxidative stress, and deficiency of vitamin E is associated with hemolytic anemia and bronchopulmonary dysplasia. Recommendations for vitamin E intakes are often expressed in relation to dietary polyunsaturated fatty acid intake. Recommendations for dietary intakes for infants and children are fairly consistent throughout the world.

Vitamin K

There are few recommendations for vitamin K since the major source of the vitamin is gastrointestinal

Table 11 Recommendations for fat-soluble vitamins

Nutrient	*Age*	*UK RNI*	*Europe PRI*	*USA RDA*	*FAO/WHO RDI*
Vitamin A (μg retinol equivalent/day)	0–6 months		—	350	350
	6–12 months		—	350	350
	1–3 years		400	300	400
	4–6 years		400	400	450
	7–10 years		500	400–600	500
	11–14 years				
	M		600	600	600
	F		600	600	600
	15–18 years				
	M		700	900	600
	F		600	700	600
Vitamin D (μg/day)	0–6 months		10–25	5[a]	5
	6–12 months		10	5	5
	1–3 years		10	5	5
	4–6 years		0–10	5	5
	7–10 years		1–10	5	5
	11–14 years				
	M		0–15	5	5
	F		0–15	5	5
	15–18 years				
	M		0–15	5	5
	F		0–15	5	5
Vitamin E (mg/day)	0–6 months	0.4 PUFA[c]	0.4 PUFA	6	—
	6–12 months	0.4 PUFA	0.4 PUFA	6	
	1–3 years	—	0.4 PUFA	6	—
	4–6 years	—	—	7	—
	7–10 years	—	—	7–11	—
	11–14 years				—
	M	>4	>4	11	
	F	>3	>3	11	
	15–18 years				—
	M	>4	>4	15	
	F	>3	>3	15	

[a]USA Adequate Intake (not RDA).
[b]FAO/WHO Recommended Safe Intake.
[c]UK Safe Intake.
F, female; M, male.

bacteria. Infants are born with a limited ability to synthesize vitamin K from this source, and hemorrhagic disease of the newborn, due to vitamin K deficiency, has a prevalence of 1 in 200–400 live births in Western countries. A single intravenous or intramuscular injection of vitamin K or an oral dose is usually offered to all neonates in many countries.

See also: **Amino Acids**: Chemistry and Classification. **Breast Feeding**. **Calcium**. **Carbohydrates**: Requirements and Dietary Importance. **Children**: Nutritional Problems. **Copper**. **Dietary Fiber**: Physiological Effects and Effects on Absorption. **Infants**: Nutritional Requirements. **Iodine**: Physiology, Dietary Sources and Requirements. **Iron**. **Magnesium**. **Phosphorus**. **Potassium**. **Protein**: Requirements and Role in Diet. **Sodium**: Physiology. **United Nations Children's Fund**. **Vitamin A**: Biochemistry and Physiological Role; Deficiency and Interventions. **Vitamin B_6**. **Vitamin D**: Rickets and Osteomalacia. **Vitamin E**: Metabolism and Requirements; Physiology and Health Effects. **Vitamin K**.

Further Reading

Aggett P, Bresson J, Haschke F *et al.* (1997) Recommended Dietary Allowances (RDAs), Recommended Dietary Intakes (RDIs), Recommended Nutrient Intakes (NRIs) and Population Reference Intakes (PRIs) are not "recommended intakes." *Journal of Pediatric Gastroenterology and Nutrition* **25**: 236–241.

Butte NF (1996) Energy requirements of infants. *European Journal of Clinical Nutrition* **50**(supplement 1): S24–S36.

Department of Health (1991) *Report on Health and Social Subjects, No. 41. Dietary Reference Values for Food Energy and Nutrients for the United Kingdom.* London: HMSO.

Dewey KG, Beaton G, Fjeld C *et al.* (1996) Protein requirements of infants and children. *European Journal of Clinical Nutrition* **50**(supplement 1): S119–S150.

FAO/WHO/UNI Expert Consultation (1985) *Energy and Protein Requirements*, Technical Report Series 724. Geneva: WHO.
Jackson AA (1990) Protein requirements for catch-up growth. *Proceedings of the Nutrition Society* **49**: 507–516.
Torun B, Davies PS, and Livingstone MB (1996) Energy requirements for 1–18 year olds. *European Journal of Clinical Nutrition* **50**(supplement 1): S37–S81.

Nutritional Problems

E M E Poskitt, London School of Hygiene and Tropical Medicine, London, UK

Introduction

Good nutrition is needed for normal growth and development. However, it is not only provision of food that is necessary for energy and nutrients to be consumed, absorbed, and utilized optimally by the body: physical and mental health, emotion, and the overall environment in which children are fed affect what is eaten and how food is utilized. As a result complex interactions involving 'nature' and 'nurture' influence nutrition in children who are essentially healthy and those who are ill.

'Child' and 'children' are not specific terms, as the age at which individuals are defined as adults rather than children varies with the circumstances. Children under 1 year of age, or infants, are considered separately in this encyclopedia, as is nutrition in adolescents. In the following discussion, the term 'children' refers to individuals between infancy and the onset of puberty, i.e., roughly between 1 and 10 years of age.

Feeding Young Children

Nutritional needs for growth are unique to infants, children, and adolescents. Growth rates fall rapidly in early life. By 1 year of age, energy needs for growth have fallen to $<3\%$ of total energy requirements but relatively high proportions of protein and micronutrients per unit of energy are still required for growth. These needs must be reflected in the quantity and quality of foods offered to young children.

High nutrient and energy needs per kilogram body weight make it difficult for young children to consume sufficient food with only two meals a day. Frequent feeding, perhaps three significant meals and two to three snacks interposed between these meals, should overcome the problems posed by the large volumes of food per kilogram body weight needed daily. As children grow, mature, and their growth rates slow, the volumes of food needed decline. Growing independence and, with this, the ability to obtain and consume frequent snacks increase with age. By the end of their first year children are consuming foods similar to the rest of the family. The consistency of their food will still be different since their inexperience of chewing and biting off appropriately sized pieces of food means that most food requires chopping or mashing lightly.

Table 1 lists some feeding skills that develop after infancy. Young children need encouragement to practice the skills that enable them to progress from breast- and bottle-feeding to soft malleable foods in early weaning and to foods that require chewing by 9 months to 1 year. After 1 year they should be taking fluids (other than breast milk) predominantly from cups rather than bottles. Continuing to offer drinks from feeding bottles after 12 months can discourage children from accepting foods that need to be chewed. Persistent bottle-feeders may have excessively high fluid intakes because these provide their main nutrition. If the fluid is milk or infant formula, obesity may result. If fruit juices or carbonated drinks are fed, fluids may

Table 1 Developmental skills associated with feeding acquired after first year of life

Age	*Relevant skills*
12–15 months	Sits well and can move around in chair when feeding
	Keen to feed self
	Tries to feed with spoon but cannot manage rotation of wrist so food falls off spoon before reaching mouth
	Finger feeds well with thumb-first finger apposition
	Interested in emptying cups and throwing food and food containers from chair
	Will search for them after throwing
18 months	Manages to feed with rotation of wrist so food reaches mouth
	Hand preference beginning to show
	Drinks well without spilling
	Hands cup back to adult
21 months	Will ask for drink or food
	Follows family eating habits
2½ years	Very self willed and determined
	May be stubborn and rebellious
	Active – may not stay sitting throughout meal
3 years	Feeds with spoon and fork – rather messily
	Will carry utensils and can help lay table
5 years	Feeds with spoon and fork neatly
	Has clear and persistent likes and dislikes for foods
	Understands consequences of choice
	Greatly influenced by peer group preferences for foods

substitute for other more energy-dense foods. Failure to thrive may then result.

Everyday Feeding problems

Young children are usually determined to show their independence. This can be frustrating for carers. The children want to do things for themselves yet do not have the skills to succeed. Approaches such as letting young children attempt spoon-feeding whilst their carers feed them unobtrusively from other spoons are 'feeding skills' parents and carers develop. Young children are also easily distracted. It is wise to feed them away from active television sets and brothers and sisters at play if the family is not eating together. At family meals young children are likely to be slow eaters. They may eat in 'fits and starts' and continue to eat even when food has gone cold and is no longer palatable to adult tastes. Removing food as soon as a young child stops eating may be inappropriate since stopping eating is sometimes a temporary respite and not a sign that the child has had 'enough.' After a pause, eating may continue. Telling children to 'eat up,' nagging them to 'hurry up', or trying to force them to eat are not helpful and can lead to mealtime defiance, frantic carers, and impaired nutrition. Most parents become skilled at anticipating and forestalling mealtime problems in diverse ways.

If children refuse food at mealtimes or show no intention of finishing their meals, they probably do not need the food. This is particularly likely if they have recently had a snack. Small children are readily sated by amounts of food that seem very small to adults. Snacks should be timed closer to previous meals than to following meals – perhaps 2 h before the next meal. Children who eat poorly at one meal often eat much better at the next meal because by then they are hungry. Offering biscuits or confectionery in exchange for an unfinished meal ('because they have not had enough') is neither helpful nor usually necessary. Children learn very quickly that if they do not eat meals they may get foods that are, to them, more enjoyable. Mealtime organization begins to collapse. However, very young children have slight risk of hypoglycemia if they go for prolonged periods without food so it is advisable to feed them before bed if they have exhibited persistent food refusal earlier in the day.

The appetites of young children are very variable. Low intakes on one day are usually compensated by excellent intakes on other days. Carers should adopt organized, but relaxed, approaches to meals and eating. Mealtimes should be enjoyable occasions for positive parent–child interaction, not the battles that sometimes develop from parents' understandable anxieties because their children 'don't eat.'

Nutrient needs that are easily met in healthy children may be more difficult to achieve if children are offered, or accept, only a limited variety of foods or have poor appetites because of illness. Vegetarian diets for young children can provide adequate nutrition but some nutritional knowledge is advisable for those managing children on such diets. Plant proteins do not individually contain all the amino acids so mixing of protein sources is important for the provision of the amino acids needed for optimal nutrition and growth. Provided breast-feeding continues, or children take significant amounts of other milk or formula (cows' milk-based or soy-based infant formulas or, after 1 year, neat cows' milk), amino acid requirements can be met from milk or formula and little other protein is needed from plant or animal sources.

WHO recommends that breast-feeding continues as part of a mixed diet into the second year of life. Milk in some form is recommended for young children. It provides a ready source of calcium throughout childhood. Before 1 year of age, European Union (EU) recommendations are for infant formula rather than neat cows' milk. After the first year of life whole cows' milk is appropriate. In the UK National Diet and Nutrition Survey 1992/3, 83% of children aged $1\frac{1}{2}$–$4\frac{1}{2}$ years were taking some whole cows' milk and for 68% of children this was as a drink. Twenty-six per cent of food energy, on average, came from milk and milk products whereas in the $3\frac{1}{2}$–$4\frac{1}{2}$ years age group this figure was only 16%. In a similar study by Gregory and Lowe, 7–10-year-old children were still consuming, on average, close to 1 l of whole milk a week providing around 12% of total daily energy. Fat-reduced milks should only be used as drinks in children under 5 years if the rest of the diet is varied and 'balanced' with other sources of fat-soluble micronutrients, in which case semi-skimmed milk may be used as a drink from the age of 2 years onwards.

Nutrient Interactions

Nutrient–nutrient interaction in the process of digestion and absorption is more important for child than adult nutrition. The requirements for micronutrients are high in childhood because of the need to form new tissues in growth. However, phytates from cereal and vegetable foods bind minerals, particularly calcium, vitamin D and iron, in the bowel and reduce absorption. Asian children in northern latitudes on high-phytate traditional diets are at risk of developing vitamin D deficiency rickets. This is partly due to poor absorption of both calcium and vitamin D because they are bound with phytates in the small intestine

and absorption is reduced. Inadequate synthesis of vitamin D from precursors in the skin because of low sunlight exposure will exacerbate poor vitamin D nutrition and further impair calcium absorption. Thus, whilst nonstarch polysaccharides (NSP) from unrefined cereals, whole fruits, and vegetables should be an increasing proportion of the diets of children as they grow, NSP intakes should only be gradually increased to perhaps 15 g day^{-1} by 10 years of age.

Nutrient–nutrient interactions are not necessarily disadvantageous. Vitamin C, through its reducing power, maintains iron in the ferrous state in the gastrointestinal tract and thus facilitates absorption of this important micronutrient. Vitamin C also facilitates absorption of a number of other micronutrients. Each meal should contain a good source of vitamin C to optimize utilization of other micronutrients.

Assessment of Nutrition in Children: Anthropometry

After the first year of life, children usually follow very predictable gains in weight and height over time. Growth as gain in weight and height remains, with activity, the aspect of energy consumption that the body can reduce if energy intakes are inadequate for all needs. The wide range of normal weights for age in a population means that a single weight in an individual child is not a good indicator of over or under nutrition. Nevertheless, weight change over time is the most widely used parameter for judging nutritional status. Failure to gain weight at the expected rate is often the first evidence of declining nutritional status. Where inadequate nutrition is prolonged, linear growth faltering also occurs. Growth curves showing weights and heights plotted against age with trajectories for mean and standard deviation or centile distributions of a population are the basis of growth assessment in childhood. In infancy, crossing the centiles upwards or downwards is quite common as infants express their genetic potential for growth in a postnatal environment of different constraints from those *in utero*. In adolescence, growth may diverge from population patterns because of the timing of the pubertal growth spurt in relation to the reference population. Between these two periods of apparent growth instability, the majority of children follow very stable growth trajectories in relation to population distribution for age. This is particularly so for height. The obesogenic environment enveloping so many children in westernized societies today may be causing children's weights to move up across the centiles much more frequently than in the past.

Failure to Thrive

Failure to thrive is failure to gain in height and weight at the expected rate, the expected rate usually being that indicated by charts for height and weight related to age and sex in reference populations. Since it is often easier to measure weight than height in small children and since weight can be lost as well as not gained, whereas height gain can only be absent or slowed, assessment of failure to thrive is frequently made on weight progress alone. Although a child may be low weight for height and age, this does not necessarily imply failure to thrive since some 'normal' children are always small, perhaps because of genetic endowment. They grow with normal velocity but at the lower extreme of normal population distribution. Thus, following weight gain over time is essential for diagnosis of failure to thrive (**Figure 1**).

Failure to thrive can result from a wide range of underlying medical problems as outlined in **Table 2**.

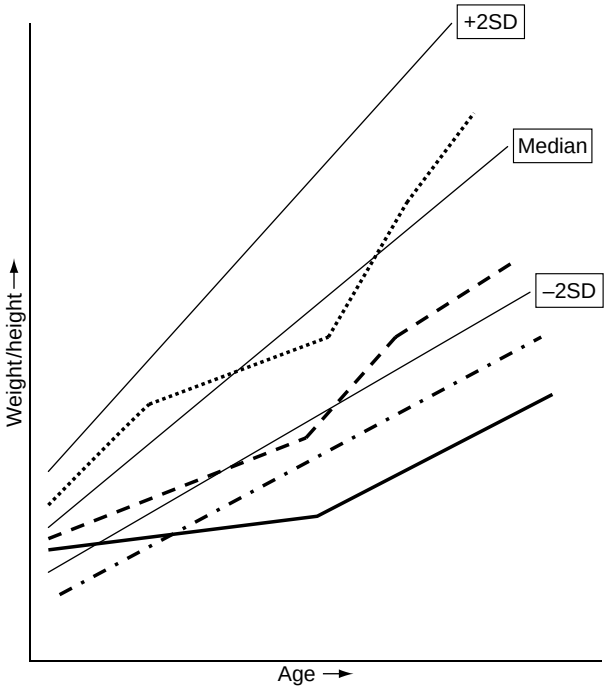


Figure 1 Failure to thrive: figurative growth chart for increase in either weight or height of four children with abnormal growth showing normal population growth as median and ± standard deviations (SD). ·····, failure to thrive in above average child with growth falling below the median and then showing complete catch up growth; – – – – –, failure to thrive in a child growing below the median with failure to thrive, falling below −2SD and then showing complete catch up; ———, child with failure to thrive and failure to complete catch up so growth continues at about average rate but child remains below previous SD position; – · – · – ·, child gaining throughout at normal velocity but starting below −2SD and remaining below −2SD in size, i.e., 'small normal'.

Table 2 Some causes of failure to thrive according to pathophysiology

Basic cause of FTT	*Clinical situation*	*Medical condition*
Too little energy taken in	Inadequate food energy offered	Poverty; ignorance of child's needs; lack of understanding of progression of weaning process
	Significant feeding difficulties	Neurological and other conditions affecting motor coordination of chewing and swallowing, especially hypertonic cerebral palsy
	Energy density of food inappropriately low	Strict vegetarian diet; excessive and inappropriate parental concern to feed 'healthy diet'
	Poor appetite	Anorexia due to infection or other illness
	Vomiting	Esophageal reflux, infection, metabolic disturbance
Too much energy lost from the body	Energy lost as sugar in urine	Diabetes mellitus
	Protein lost as protein in urine	Nephrotic syndrome
	Protein lost through skin	Burns
		Severe eczema
Failure to absorb	Malnutrition syndromes	Gluten-sensitive enteropathy/celiac syndrome
		Cystic fibrosis
		Protein-losing enteropathy
		Lactose intolerance
Failure to utilize	Chronic illness	Chronic infection
		Urinary tract infection
		Other hidden infections
		Cyanotic congenital heart disease
		Inborn errors of metabolism
		Severe mental retardation
		Deficiency of other essential nutrients
Increased requirements	Elevated BMR	Thyrotoxicosis
		Congenital heart disease with left to right shunts and high output state
	Increased growth rates	Catch up growth following period of failure to thrive
	Increased activity as increased respiratory rate	Cystic fibrosis
		Congenital heart disease
	Increased activity	Uncontrolled hyperactive behavior

BMR, basal metabolic rate.

Some children have no recognizable pathological abnormalities and yet they fail to thrive. In some children inappropriate management of the feeding problems described earlier underlies poor growth. Others may be underfed because of poverty, ignorance, or incompetence amongst carers. Some parents, anxious to forestall obesity or cancer and cardiovascular conditions in later life, feed diets that are too restricted in nutrient quantity or variety for normal growth, or which follow rigidly the dietary recommendations intended for adults. In the so-called muesli belt syndrome, skimmed milk, wholemeal cereals, and low-sugar low-fat foods dominate.

Psychosocial Deprivation

Some children fail to thrive despite apparently adequate intakes and no evidence of underlying illness. Some are subject to physical and/or emotional abuse and have psychosocial deprivation (PSD), 'non-organic failure to thrive,' or 'emotional deprivation syndrome.' If children are not stimulated by face to face contact and communication with parents and carers, if they are verbally and/or physically abused, or if they are neglected and ignored, the silent response may be growth failure. Affected children may show other signs of emotional distress. They are withdrawn, developmentally delayed, rarely smile, and avoid eye to eye contact. Sometimes they seek attention through destructive, aggressive, and disturbing behavior or steal food, raid dustbins for food, or eat the leftovers on their schoolmates' plates, and yet show no weight gain.

The process inhibiting normal growth in PSD is likely to be multifactorial and to vary with individual circumstances. In some cases, affected children have not received sufficient food for normal growth. In other cases, food may have been offered but misery and fear prevented the children eating it thus subjecting them to further ridicule or punishment because of unfinished meals. Some children, who appear to have been given adequate food and yet fail to grow, may have elevated metabolic rates and increased energy needs secondary to anxiety and

stress. A few studies suggest the growth problem in PSD can be linked to neuroendocrine abnormalities secondary to stress. Levels of growth hormone and adrenocorticotrophic hormone have been low in some children studied in their adverse environments. Changes to loving, caring, and stimulating environments result in rapid resolution of the endocrine abnormalities.

The treatment of nonorganic failure to thrive is to change the caring environment either by support and help for affected families or, in difficult cases, removal of affected children from their homes on a temporary or permanent basis. Proof of the diagnosis comes with rapid onset of vigorous catch up growth and positive changes in behavior following improvement in the home environment without other specific treatment.

Toddler Diarrhea

One troublesome diet-related problem occurring in preschool children is 'toddler diarrhea.' Young children are susceptible to gastrointestinal infections because of immature, inexperienced immune systems and poor hygiene from their habit of 'mouthing' almost everything they handle. Children are also prone to develop loose stools in response to minor nongastrointestinal infections. However, some children suffer frequent episodes of loose watery stools with or without increased stool frequency and without evidence of infection. These episodes of diarrhea may last weeks or months. Since it may be difficult to distinguish this diarrhea from other significant gastrointestinal pathology, affected children may be subjected to a lot of unrewarding clinical investigation.

Children with toddler diarrhea usually grow normally, unlike most children with significant gastrointestinal pathology. Typically, they are untroubled by their diarrhea although their parents are understandably very concerned – parents have to cope with loose stools in incontinent children! Investigation of toddler diarrhea should include stool microscopy and stool culture to look for evidence of infection and fat maldigestion. Results are usually normal although parents frequently state that the children's stools contain 'undigested food.' On questioning, 'undigested food' means tomato skins, bean husks, and other substances always present in stools but not obvious in formed stools. The fluid stools make these normal food residues more noticeable than usual.

The balance of macronutrients in affected children's diets may be at least partly responsible for toddler diarrhea. Young children who drink a lot of sugared fluids can have loose stools due to sugars reaching the large bowel and causing osmotic diarrhea particularly when fructose corn starch is used since the fructose is poorly absorbed without associated glucose. Sorbitol, sometimes used as a artificial sweetener, is also unabsorbed and can cause loose stools. In the 1992 National Diet and Nutrition Survey children aged $1\frac{1}{2}$–$4\frac{1}{2}$ years drank on average 1.5 l of sugar-containing drinks and 1 l of 'diet' drinks per week. Some drank much more. On average, 6% of food energy came from soft drinks in children $1\frac{1}{2}$–$4\frac{1}{2}$ years old. High intakes of food-derived refined carbohydrates add to osmotic diarrhea. Sugar, preserves, confectionery, and sweetened drinks contribute around 20% of the total energy intake of 4–10-year-olds in the UK.

Diets high in refined carbohydrates are commonly low in nonstarch polysaccharides (NSP) and in the proportion of energy derived from fat. Reducing the intake of refined sugars and carbohydrates, increasing the fat content of the diet to 30–35% total energy, and increasing the intake of unrefined carbohydrates and NSP usually result in less fluid stools. The improved continence that comes with age also helps resolution of toddler diarrhea. Most affected children are free of major symptoms by school age. Some retain tendencies to loose stools in response to infection and stress throughout life.

Anemia

Anemia is common in young children. It has many causes but nutritional deficiencies, i.e., inadequate nutrient intakes and anemia secondary to disease processes, explain much childhood anemia. Acceptable hemoglobin levels in children are lower on average than in adults. The World Health Organization (WHO) accepted lower limit of normal hemoglobin is $110\,g\,l^{-1}$ for children 1–6 years old and $120\,g\,l^{-1}$ for those over 6 years. Infants are born with relatively high hemoglobin levels which, whilst appropriate for the low oxygen tensions of intrauterine life, are unnecessary for postnatal life. The bone marrow therefore becomes quiescent and hemoglobin levels fall. Although erythropoeisis increases again after about 2 months of postnatal life, hemoglobin levels remain lower than in adults for most of childhood perhaps partly because erythropoeisis cannot keep up with the rapid expansion of the blood compartment as body size increases.

Nutritional anemia is most commonly due to iron deficiency. **Table 3** outlines some causes of iron deficiency. Iron absorption from foods other than breast milk is never very efficient although deficiency increases the proportion of iron absorbed from the

Table 3 Some causes of iron deficiency in children

Categorization of problem	*Causative condition*
Too little iron ingested	Diet poor in meats, dark green leaves, iron-fortified cereals
	Anorexia and low iron intakes
Too much iron lost from body	Hemorrhage from any cause if severe or chronic
	Insidious intestinal blood loss, e.g., cows' milk protein intolerance
	Crohn's disease
Failure to absorb	Vegetarian diets where no heme iron in diet
	Low fruit and vegetable intake so ferric iron not reduced in stomach by dietary vitamin C
	Lack of gastric acid: achlorhydria
	Pyrexia reducing absorption
	Malabsorption involving jejunum and upper ileum, e.g., celiac syndrome
Failure to utilize	Deficiency of other essential nutrients for formation of hemoglobin, e.g., vitamin A, riboflavin
	Chronic inflammatory conditions
Increased requirements	After blood loss
	In rapid catch up growth

jejunum. Absorption of iron in heme from meat is more efficient than as inorganic ferrous or ferric iron. Thus, vegetarian diets present increased risk of iron deficiency. However, most diets, even in affluent westernized countries, are marginal in the amount of iron in relation to population needs. All children are at risk of developing iron deficiency with minor disturbances in dietary quality, iron absorption and metabolism, or with blood loss. Iron absorption takes place in the jejunum and upper ileum so conditions such as gluten-sensitive enteropathy (celiac syndrome), where the brunt of intestinal damage is in the jejunum and upper ileum, may present as severe iron deficiency. Reduced iron absorption during pyrexial illness contributes to iron deficiency in children who suffer frequent infections.

Anemia is not easy to recognize clinically since pallor is a very nonspecific sign. Koilonychia (spoon-shaped nails) although fairly specific is not obvious in the small finger nails of children. Iron deficiency affects mentality and behavior making children irritable, uncooperative and anorexic, or tired and apathetic. Severe iron deficiency is associated with pica or desire for abnormal foods particularly those with metallic earthy tastes such as clay and coal.

Hypochromic, microcytic anemia is the end point of iron deficiency when stores have been exhausted, tissue iron levels are falling, and there is insufficient iron to meet the needs of red cell production. If severe, iron deficiency anemia (IDA) causes breathlessness, tiredness, poor appetite, and failure to thrive. Anemia develops slowly so there is physiological adaptation to the developing anemia and hemoglobin levels may be very low ($<40\,g\,l^{-1}$) before symptoms are noticed. Oral iron therapy allows gradual return to normal physiology and is safer than blood transfusion even when hemoglobin levels are very low.

Worldwide, iron deficiency is possibly the commonest nutritional deficiency. Studies show that IDA is associated with poor outcomes in growth and intellectual capacity. The evidence that impaired growth and intellectual development can also result from iron deficiency without anemia is less definite and remains an area of research.

Circulating iron-binding factors such as transferrin are reduced in children with protein energy malnutrition. Normally, iron circulating in plasma and tissues is bound to proteins such as transferrin. When these proteins are reduced in quantity, iron in the tissues may be inadequately 'bound.' Unbound tissue iron encourages free radical damage in tissues and cells, counteracting immunological resistance and facilitating overwhelming infection. Iron supplements should not be given to malnourished children until they are showing signs of recovery by which time they will need the iron for catch up growth.

Zinc Deficiency

Zinc deficiency is associated with impaired growth and maturation. Zinc deficiency is common in those with frequent diarrhea since zinc concentrations in gastrointestinal secretions are high. Nevertheless, there is little consistent population evidence that children who show recurrent acute diarrhea or who have prevalent growth faltering are improved by zinc supplementation. Supplementation of children with diarrhea in zinc-deficient environments may reduce the duration of diarrhea and thus the risk of persistent diarrhea, but does not reduce the risk of children developing diarrhea.

The original reports of zinc deficiency referred to short boys in the Middle East who had delayed puberty that responded to zinc supplementation with maturation and growth acceleration. However, few studies provide good evidence that zinc has a significant effect on maturation in the vast majority of children with pubertal delay. Clinical signs of zinc deficiency are unusual apart from the nonspecific signs of poor growth. In children fed artificial diets deficient in zinc or in those with acrodermatitis enteropathica (congenital lack of zinc-binding

intestinal ligand), zinc deficiency results in severe diarrhea, peeling eczematous skin, and death from overwhelming infection and/or malnutrition without supplementary zinc. Breast milk contains a zinc-binding ligand that facilitates the absorption of zinc, so clinical evidence of acrodermatitis enteropathica does not appear until breast-feeding ceases.

Calcium and Vitamin D

These micronutrients are discussed elsewhere (see 00033 and 00051). Childhood is an important time for deposition of bone mineral and the development of peak bone mass (PBM). Seventy-five per cent of bone mineral is deposited in childhood. Low PBM in late adolescence is a significant precursor of later osteoporosis. Much of the population variation in PBM is genetically determined, but low calcium and vitamin D together with a relatively sedentary lifestyle seem factors likely to contribute to low PBM and risks of osteoporosis later in life. Data from The Gambia show that in preadolescent children adequate calcium deposition takes place despite very low intakes of calcium. In some studies where milk-derived calcium phosphate was fed to children there was accelerated growth and maturation in the supplemented children. The Gambian studies showed no change in growth in the supplemented children although there was increased bone mineralization.

Vigorous calcium deposition in children in the tropics and subtropics despite very low calcium intakes may relate to higher circulating 25 hydroxy vitamin D derived from the action of sunlight on 7-dehydrocholesterol in the skin and to high levels of physical activity amongst these children. Deposition of calcium in bone is significantly affected by the weight-bearing activity of the individual. Where children are less active and vitamin D levels lower, dietary calcium intakes may have a greater determining effect on bone mineralization. In northerly temperate climates, such as the UK, children's diets should provide a good source of calcium and there should be reasonable exposure to summer sunshine (a slightly controversial area in view of current concerns about increased skin cancer risk from UVL). An active lifestyle is also important for optimal bone mineralization and high PBM, particularly in those where there is reason to suspect genetic predisposition to low PBM.

Other Micronutrients

Table 4 and **Table 5** outline some of the micronutrient deficiency problems that can arise in childhood. Most of these conditions are rare in their overt form in westernized countries because of the variety of children's diets. There remain questions as to whether levels of micronutrient nutrition that do not lead to clinical deficiency syndromes may nevertheless have effects on immunity, growth, and development in children. This seems possible for iron deficiency and psychoneurological functioning and is proven for folic acid and embryological development. For other micronutrients the evidence is less strong. Similarly, there is currently no good evidence that raising children's micronutrient intakes above recommended levels incurs any particular benefit to health or intellectual development.

Allergy

Food allergy is discussed elsewhere (see 00122 and 00123). It is a frequent diagnosis in childhood. Diarrhea, rashes, and wheezing are common symptoms caused by infection probably more commonly than by food allergy. Parental desire to explain a child's frequent illness may lead to food being wrongly blamed for recurrent symptoms. Vague associations between food and the development of symptoms can result in many foods being unnecessarily excluded and children reduced to diets of very limited variety. For example, whilst 14% of children may be described as allergic to some food, as few as 5% may have had this diagnosis confirmed by their medical practitioners.

Much publicity has surrounded the idea that 'fast foods' and carbonated drinks influence behavior. Even so, there is very little hard scientific evidence to support the widely held view that food additives are harmful to children's behavior. A few studies have shown a minority of children where food does seem to be directly associated with behavioral change, particularly when tartrazine is considered. However, the organization and discipline necessary to eliminate certain foods from young children's diets may be as influential in improving behavior as removal of a food from the diet. The area remains a confused one but there seems little justification for the use of most food colorants, texturizers, and even preservatives, irrespective of their effects on behavior.

Promoting Good Nutrition for Children

Good nutrition is only one aspect of the positive life style that promotes sustainable physical, mental, and social well-being. But the diets and life styles learnt in childhood are likely to influence adult life and

Table 4 Vitamin deficiency problems that may occur in childhood

Vitamin	*Condition*	*Features*
Vitamin A – retinol, carotene	Xerophthalmia	Minor: night blindness Major: total destruction of eye and blindness Nutritional stunting Increased infection; severe infection
Vitamin B_1 – thiamine	Beriberi	Mental changes: apathy, irritability, depression Cardiac failure Peripheral neuritis
Vitamin B_2 – riboflavin	Anemia	Hypochromic, microcytic anemia Dry skin; seborrhoeic dermatitis in skinfolds; angular stomatitis; raw red tongue
Vitamin B_3 – niacin	Pellagra	Cracked, dry, peeling, or blistering light-sensitive dermatitis Apathy, depression, confusion Diarrhea
Vitamin B_6 – pyridoxine	Convulsions	Unusual except in inborn errors of metabolism involving pyridoxine Peripheral neuritis Convulsions Anemia
Vitamin B_{12} – cyanocobalamin	Pernicious megaloblastic anemia Subacute combined degeneration of the spinal cord	Megaloblastic macrocytic anemia Anesthesia and loss of position sense and motor weakness in limbs Encephalopathy
Folic acid	Anemia Growth faltering	Megaloblastic anemia Folic acid deficiency may develop whenever there is rapid growth. It may also cause impaired catch up growth or growth faltering
Vitamin C – ascorbic acid	Scurvy	Bleeding, bruising with painful bleeding subperiosteally Anemia Osteoporosis
Vitamin D – calciferol	Rickets	Painful deformity of weight-bearing bones with bowed legs; swollen ends to shafts of long bones and ribs giving Harrison's sulcus at attachment of diaphragm to lower ribs and 'rickety rosary' effect; bossed skull Poor growth and increased infection Rarely, hypocalcemic tetany
Vitamin E – tocopherol	Hemolytic anemia Neurological degeneration	Unusual except in fat malabsorption conditions such as cystic fibrosis and abetalipoproteinemia Loss of sensation and motor power in limbs similar to that with vitamin B_{12} deficiency

Table 5 Mineral deficiencies that can occur in childhood

Mineral	*Condition*	*Features*
Iron	Anemia	Hypochromic, microcytic anemia
	Impaired growth	Role of iron, as separate from anemia, in relation to poor growth not clear.
	Impaired cognition	Likewise role of iron independent of anemia in impaired cognition not clear
Iodine	Cretinism	Severity of iodine deficiency and age at which it occurs affect the damage caused
	Stunting	The younger the child the more likely there is permanent neurological damage
	Goiter	Goiter common when growth rapid in early life and at adolescence
Zinc	Loss of taste	Anorexia
	Growth impairment	Zinc deficiency is associated with growth faltering and delayed maturation
	Gastrointestinal effects	Acute diarrhea may be more likely to persist when there is zinc deficiency
Selenium	Keshan disease	Cardiomyopathy which may be precipitated by associated deficiency of vitamin E or perhaps by Coxsackie virus infection
	Kachin-Beck syndrome	Osteoarthritis affecting particularly lower limbs
Copper	Anemia	Microcytic hypochromic anemia
	Poor bone mineralization	Osteoporosis
Calcium	Rickets-like syndrome	Very low calcium diets can be associated with rickets-like condition but more usual for associated vitamin D deficiency as well
Magnesium	Often silent deficiency	Deficiency is very unusual except in severe malnutrition when it is universal Occasionally magnesium deficient convulsions

Table 6 Principles of good nutrition in childhood

Policy	*Practice*
Feed children appropriate to their developmental skills	Progress diet from soft mushy weaning diet to increasingly lumpy and then chewable foods Encourage biting and chewing Encourage new foods, textures, and tastes
Encourage recognition of satiety and hunger	Organize regular meals and snacks Eat as a family where possible Eat away from activities such as watching television Avoid snacking whilst occupied with television or computer or other absorbing activity
Encourage varied nutrient intake	Home prepared foods when possible so nutrient content is recognized Vary diet and recognize that children may not accept new foods on first tasting If possible use other members of the family to show how specific foods are enjoyable
Achieve nutrient needs by varied diet	Plan for an energy-dense staple, protein source, micronutrient source at each meal Aim for about 500 ml of milk or formula daily to provide a good calcium source Provide a source of vitamin C at each meal Aim towards five small portions of fruit and vegetable/day and gradually increase portion sizes (e.g., from few segments of an orange work up to a whole orange as the child's appetite grows)
Aim for diet that will protect as far as possible against chronic noncommunicable diseases as adult	Work towards recommendations for a healthy adult diet as child matures but avoid high fiber, low-fat diets in early childhood
Maintain active life style	Encourage activity both outside and inside the home Encourage children to take part in household chores Encourage children to take an interest in and understand food preparation and food choice

should optimize opportunities for health and social well being throughout life.

It is unlikely that the Western world, with its working mothers and busy parents, can return to a society where all meals are consumed as families. Eating and socializing over meals are nevertheless important for child development since they may encourage children to recognize satiety during the time spent eating. Further, the communication and stimulation that should take place during family meals develops children's social skills. Even when family meals are not practical, eating should be organized, and to some extent formalized, preferably away from distractions such as television, which may detract from awareness of the quantity of food consumed. Snacking should be at specific times rather than an activity that happens because food is available or because there is nothing else to do. **Table 6** outlines some of the principles to be followed when trying to develop good eating and nutritional habits in childhood.

A healthy diet needs to be accompanied by plenty of exercise, out of doors whenever possible. Young children are naturally extremely active and should be encouraged to explore and develop their physical skills under supervision. Sedentary interests, particularly watching television, should occupy only a very small portion of the day although as children grow older, homework, which is usually fairly sedentary, inevitably occupies part of the after school hours. At home children can be kept active helping around the house, playing games, and pursuing varied hobbies. Overcoming sedentariness may involve parents as well as children, but that should be good for parents' health also.

Eating and drinking opportunities at school can run counter to the good nutritional practices parents may be pursuing. For example, in schools, water dispensers could replace those selling carbonated drinks although drinks sales are often significant sources of revenue for schools. Many countries now have guidelines, such as making fruit available at lunch, for healthy eating programs in schools. Nevertheless, diet is only one, admittedly important, area for action when developing life styles that should sustain normal nutrition throughout the life cycle.

Conclusion

The common nutritional problems in children on varied Western diets are mostly fairly minor feeding-related difficulties in the very young, failure to thrive, or the concerning modern problem of obesity (see 00210). However, childhood nutrition can influence health in later life. Good nutritional practice should be a continuum running from infancy through to old age. Childhood is an important stage in setting such practice.

See also: **Anemia**: Iron-Deficiency Anemia. **Behavior**. **Calcium**. **Children**: Nutritional Requirements. **Dietary Fiber**: Role in Nutritional Management of Disease.

Food Allergies: Etiology; Diagnosis and Management. **Growth and Development, Physiological Aspects**. **Iron**. **Malnutrition**: Secondary, Diagnosis and Management. **Obesity**: Childhood Obesity. **Vegetarian Diets**. **Vitamin D**: Rickets and Osteomalacia. **Zinc**: Physiology.

Further Reading

Anon (1997) Guidelines for school health programs to promote lifelong healthy eating. *Journal of School Health* **67**: 9–26.

Black RE and Sazawal S (2001) Zinc and childhood infectious disease morbidity and mortality. *British Journal of Nutrition* **85** (supplement) 2: S125–S129.

Dibba B, Prentice A, Ceesay M *et al.* (2000) Effect of calcium supplementation on bone mineral accretion in Gambian children accustomed to a low calcium diet. *American Journal of Clinical Nutrition* **71**: 544–549.

Gregory JR, Collins DL, Davies PSW, Hughes JM, and Clarke PC (1995) *National Diet and Nutrition Survey: Children Aged 1½–4½*, pp. 27–48. London: HMSO.

Gregory J and Lowe S (2000) *National Diet and Nutrition Survey: Young People Aged 4 to 18 Years, Volume 1: Report of the Diet and Nutrition Survey*, pp. 45–105. London: HMSO.

Prasad AS, Miale A, Farid Z, Schulert A, and Sandstead HH (1963) Zinc metabolism in patients with iron deficiency anaemia, hypogonadism and dwarfism. *Journal of Laboratory and Clinical Medicine* **61**: 537–549.

Rayner PH and Rudd BT (1973) Emotional deprivation in three siblings associated with functional pituitary growth hormone deficiency. *Australian Paediatric Journal* **9**: 79–84.

Rowe KS and Rowe KJ (1994) Synthetic food coloring and behavior: a dose response effect in a double blind, placebo controlled, repeated measure study. *Journal of Pediatrics* **125**: 691–698.

WHO (1972) *Nutritional Anemia*. WHO Technical Report Series Number 3. Geneva: WHO.

Cholecalciferol *see* **Vitamin D**: Physiology, Dietary Sources and Requirements; Rickets and Osteomalacia

CHOLESTEROL

Contents

Sources, Absorption, Function and Metabolism

D J McNamara, Egg Nutrition Center, Washington, DC, USA

The intestinal absorption of cholesterol, its transport through the vascular system, its utilization for synthesis of steroids and bile acids, and its fecal excretion represent a complex interplay between two metabolic processes: the absorption of exogenous dietary cholesterol and endogenous cholesterol synthesis by body tissues. From the initial mixing of dietary and biliary cholesterol in the small intestine to the excretion of fecal neutral and acid steroids, there are two sources of cholesterol, which combine to determine the overall dynamics of whole-body cholesterol metabolism and the ability of the metabolic regulatory processes to maintain tissue cholesterol homeostasis as well as a steady-state plasma cholesterol level. It is the balance between dietary cholesterol intake and absorption, endogenous cholesterol synthesis, and fecal steroid excretion that determine overall cholesterol homeostasis in the body (**Table 1**).

Absorption, Transport, and Storage

Cholesterol Absorption

Cholesterol in the intestinal lumen typically consists of one-third dietary cholesterol and two-thirds biliary cholesterol. The average daily diet contains 300–500 mg of cholesterol obtained from animal

Table 1 Average cholesterol metabolism values for a 70-kg adult

Cholesterol pools and flux	*Mass*
Cholesterol pool (70-kg adult)	160 g
Plasma cholesterol pool	8 g
Dietary cholesterol intake	300 mg/day
Absorption (average 60%)	180 mg/day
Synthesis (12 mg/kg/day)	840 mg/day
Total cholesterol input	1020 mg/day
Bile acid synthesis (= fecal excretion)	250 mg/day
Neutral steroid excretion	770 mg/day

products. The bile provides an additional 800–1200 mg of cholesterol throughout each day as gallbladder contractions provide a flow of bile acids, cholesterol, and phospholipids to facilitate lipid digestion and absorption. Dietary cholesterol is a mixture of free and esterified cholesterol, whereas biliary cholesterol is nonesterified and is introduced into the small intestine as a cholesterol–bile salt–phospholipid water-soluble complex. The only other source of intraluminal cholesterol is mucosal cell cholesterol, derived from either sloughed mucosal cells or cholesterol secreted by the mucosal cells into the intestinal lumen. Measurements of exogenous and endogenous cholesterol absorption in humans indicate that there is probably very little direct secretion of newly synthesized cholesterol from mucosal cells into the intraluminal contents.

Cholesterol absorption occurs primarily in the duodenum and proximal jejunum of the small intestine and is dependent on the presence of bile salts. In the absence of bile secretion, or in the presence of bile acid-binding resins, there is virtually no intestinal absorption of cholesterol. On average, humans absorb 50–60% of the intestinal contents of cholesterol, but there is a large interindividual variance in absorption, with values ranging from as low as 20% to as high as 80%. Intestinal transit time is related to cholesterol absorption, with slower transit times resulting in higher fractional absorption rates. Dietary factors that affect the relative percent absorption of cholesterol include the total mass of dietary cholesterol, the concentration of plant sterols in the diet, and the type and amount of dietary fiber. Studies suggest that the ratio of polyunsaturated to saturated fat (P:S) in the diet has little effect on cholesterol absorption rates in humans, nor does the amount of dietary fat.

Two interesting, and as yet undefined, aspects of cholesterol absorption are that it decreases as the mass of cholesterol increases above an intake of 1500 mg per day, and that the fractional absorption below this level is relatively constant for an individual. For example, at a daily cholesterol intake of 800 mg a subject may absorb 60% or 480 mg per day, whereas at a daily intake of 400 mg the absorption remains at 60%, equaling 240 mg per day absorbed. The quandary is that if the system can accommodate absorption of 480 mg at the high cholesterol intake, then why is the amount absorbed 240 mg at the low intake? Clearly, the upper value of cholesterol absorption is achievable, yet at the lower intake level the absorption rate stays at a fixed fractional value. The mechanisms controlling this aspect of cholesterol absorption have not been defined.

Experimental evidence indicates that biliary cholesterol and dietary cholesterol are absorbed equally; however, the pattern of exogenous and endogenous cholesterol absorption differs along the length of the intestinal lumen. Dietary cholesterol enters the small intestine solubilized in the oil phase of the stomach digest, whereas the binary cholesterol enters in the micelle phase of the bile. This differential distribution results in a greater absorption of biliary cholesterol in the upper portion of the small intestine, with dietary cholesterol absorption increasing as the oil phase of the intestinal contents are hydrolyzed. As the oil phase is reduced, dietary cholesterol moves from the oil phase to the aqueous micelle phase and becomes available for absorption. In the case of cholesteryl esters in the diet, it is necessary that the esters are hydrolyzed by pancreatic cholesterol esterase (CEase) before the cholesterol is available for absorption. Pancreatic CEase requires the presence of bile salts for activity and may play a key role in the actual absorption process.

The process, and selectivity, of sterol absorption involves a complex interplay of regulated transporters, transporting sterols into and out of the enterocyte, and the assembly and secretion of chylomicrons into the lymph. The enterocyte takes up both cholesterol and phytosterols from the intestinal lumen by what appears to be a common sterol transporter or permease in the brush border membrane. Preliminary studies suggest that the Neiman–Pick C1 like 1 (NPC1L1) protein is involved in this process. Once the sterols enter the enterocyte, the ATP binding cassette (ABC) hemitransporters ABCG5 and ABCG8 function in the apical excretion of sterols back into the intestinal lumen. The selectivity of this process accounts for the higher absorption rates of cholesterol (50–60%) compared to the phytosterols, which are very poorly absorbed. Loss of ABCG5/G8 function results in excessive absorption of both cholesterol and phytosterols. Studies in mice have shown that ABCG5/G8 are expressed primarily in the liver and intestine, are coordinately upregulated at the transcriptional level by dietary cholesterol intake,

and require the liver X receptor-α (LXR-α), a nuclear receptor that regulates the expression of a number of key genes involved in lipid metabolism.

Evidence is accumulating that the fractional cholesterol absorption rates are regulated by one or more genetic determinants. The apolipoprotein (apo) E phenotype has a significant effect on fractional cholesterol absorption and appears to play a major role in determining the plasma lipoprotein response to changes in dietary cholesterol intake. Men with the *apoE4* allele have a high cholesterol absorption rate, whereas those with the *apoE2* allele have a low cholesterol absorption efficiency. The absorption values for the more common *apoE3/3* fall between the *apoE2* and *apoE4* patterns. Polymorphisms of the apolipoprotein A-IV and of the low-density lipoprotein (LDL) receptor gene have also been related to differences in fractional cholesterol absorption. These genetic variants affecting cholesterol absorption no doubt play a significant role in determining an individual's fractional absorption of cholesterol as well as accounting for much of the heterogeneity of plasma lipid responses to changes in dietary cholesterol intakes.

Exogenous Cholesterol Transport

Cholesterol is absorbed in the unesterified state, whereas the cholesterol secreted into the lymph is 70–80% esterified. This esterification process generates a concentration gradient of free cholesterol within the mucosal cell that may facilitate absorption rates. Cholesterol is esterified in intestinal mucosal cells by acyl-coenzyme A:cholesterol acyltransferase-2 to form cholesteryl esters, which are secreted from the basolateral surface of the enterocyte as part of the chylomicrons. At this stage, it is assumed that cholesterol molecules from exogenous and endogenous sources are indistinguishable and have similar effects on endogenous cholesterol and lipoprotein metabolism. Chylomicrons are large particles (>70 nm in diameter) composed mainly of triacylglycerols (95% by weight) and contain 3–7% cholesterol by weight, with the esterified cholesterol localized in the hydrophobic core and the free cholesterol primarily in the hydrophilic outer layer. The data indicate that the amount of dietary cholesterol consumed has little effect on the cholesterol content of chylomicrons. The chylomicrons are released from the intestinal cells, enter the lymphatic system, and are transported via the lymphatics (thoracic duct) to the bloodstream. Since chylomicrons are too large to pass through the capillaries, this is the only mechanism by which they can enter the bloodstream.

In the plasma compartment, the chylomicrons pick up a number of apolipoproteins, which are required for intravascular metabolism of the particles. The initial metabolism of chylomicrons involves hydrolysis of the associated triacylglycerols by endothelial cell lipoprotein lipase (LPL) located in adipose, muscle, and heart tissues, which results in production of chylomicron remnants. The chylomicron remnants, depleted of triacylglycerol and enriched with cholesteryl ester, are taken up by the liver via the LDL receptor-related protein (LRP). The ligand for hepatic uptake of the chylomicron remnant appears from various transgenic mouse studies to be the apo-E moiety of the particle. The clearance of chylomicrons from the bloodstream is rapid, with particles having a half-life of less than 1 h. The liver cannot take up native chylomicrons but rather takes up the chylomicron remnant, which has lost approximately 90% of its triacylglycerol content and become relatively enriched in free and esterified cholesterol through the actions of the plasma cholesteryl ester transfer protein (CETP), which transfers cholesteryl ester from high-density lipoprotein (HDL) to the apo-B-containing lipoproteins.

The chylomicron remnants taken up by the liver are subjected to lysosomal hydrolysis, resulting in the release of the absorbed dietary and biliary cholesterol into the hepatocyte as free cholesterol. The influx of cholesterol contained in the chylomicron remnant has the ability to affect a number of regulatory sites of hepatic cholesterol metabolism that function to maintain cholesterol homeostasis in the liver. The liver has four primary fates for the newly delivered cholesterol: catabolism to bile acids, secretion as biliary cholesterol, storage in lipid droplets as cholesteryl ester, or incorporation into very low-density lipoprotein (VLDL) for secretion from the liver.

Tissue Uptake and Storage

The body pool of cholesterol is approximately 145 g, with one-third of this mass localized in the central nervous system. The remainder of the metabolically active cholesterol pool exists in the plasma compartment (7.5–9 g) and as constituents of body tissues. In humans, tissue cholesterol levels are relatively low, averaging 2 or 3 mg/g wet weight. Little information exists regarding changes in hepatic and extrahepatic tissue cholesterol concentrations with changes in dietary cholesterol intake. Animal studies, which are usually carried out using very high levels of dietary cholesterol, have shown that hepatic cholesterol can increase from 2-fold up to 10-fold, depending on the species and other dietary constituents, when dietary cholesterol is increased.

Biosynthesis

Tissue Cholesterol Synthesis

Cholesterol biosynthesis occurs in every nucleated cell in the body. Although it is often thought that the majority of cholesterol synthesis occurs in the liver, studies have shown that the bulk tissues of the body account for the overwhelming majority of endogenous cholesterol production. Hepatic cholesterol synthesis in humans is thought to contribute 10–20% of the total daily synthesis rate. Since the majority of cholesterol synthesis in the body occurs in extrahepatic tissues, and since the only quantitatively significant site for excretion and catabolism of cholesterol is the liver, approximately 600–800 mg of cholesterol each day must be transported from peripheral tissues through the plasma compartment to the liver to account for daily cholesterol catabolism and binary secretion. Approximately 9 mg cholesterol per kilogram body weight is synthesized by peripheral tissues every day and must be moved to the liver for catabolism via a process termed 'reverse cholesterol transport' (RCT).

RCT describes the metabolism, and important antiatherogenic function, of the HDL-mediated efflux of cholesterol from nonhepatic cells and its subsequent delivery to the liver and steroidogenic organs for use in the synthesis of lipoproteins, bile acids, vitamin D, and steroid hormones. A cellular ABC transporter (ABCA1) mediates the first step of RCT involving the transfer of cellular cholesterol and phospholipids to lipid-poor apolipoproteins. Lecithin:cholesterol acyltransferase-mediated esterification of cholesterol generates spherical particles that continue to expand with ongoing cholesterol esterification and phospholipid transfer protein-mediated particle fusion and surface remnant transfer. Larger HDL_2 particles are converted into smaller HDL_3 particles when CETP facilitates the transfer of cholesteryl esters from HDL onto apoB-containing lipoproteins. The scavenger receptor B1 (SR-B1) promotes selective uptake of cholesteryl esters into liver and steroidogenic organs, whereas hepatic lipase- and LPL mediate hydrolysis of phospholipids and triglycerides. SR-B1 mediates the selective uptake of cholesteryl esters from HDL and also LDL into hepatocytes and steroid hormone-producing cells without internalizing HDL proteins, which can recycle through the RCT sequence moving cholesterol from peripheral tissues to the liver.

Regulation of Synthesis

The rate-limiting enzyme in cholesterol biosynthesis is 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase, a microsomal enzyme that converts HMG-CoA to mevalonic acid in the polyisoprenoid synthetic pathway. Peripheral tissue cholesterol synthesis is much less responsive to regulatory factors compared to the liver, which is controlled by a variety of dietary, hormonal, and physiological variables. Studies indicate that endogenous cholesterol synthesis is significantly increased in obesity and in patients with the metabolic syndrome. Obesity, insulin resistance, and diabetes have pronounced effects on both cholesterol absorption and synthesis. Findings in type 1 diabetes appear to be related to low expression of ABCG5/G8 genes, resulting in high absorption and low synthesis of cholesterol. Cholesterol absorption efficiency is lower and cholesterol synthesis is higher in obese subjects with type 2 diabetes compared to obese subjects without diabetes, suggesting that diabetes modulates cholesterol metabolism to a greater extent than obesity alone. Similarly, low cholesterol absorption and high synthesis appear to be part of the insulin resistance (metabolic) syndrome.

Research shows that in most individuals, dietary cholesterol alters endogenous cholesterol synthesis and that this feedback regulation can effectively compensate for increased cholesterol input from dietary sources. The precision of these regulatory responses depends on a number of genetic factors, and data suggest that multiple genetic loci are involved. For example, family studies have shown that in siblings of low cholesterol absorption families, cholesterol absorption percentages are significantly lower and cholesterol and bile acid synthesis, cholesterol turnover, and fecal steroids are significantly higher than in siblings of high absorption families.

Metabolism and Excretion

The body's metabolic processes cannot break the sterol rings of cholesterol and therefore must either catabolize cholesterol to other products, which can be excreted in the urine or feces, or directly excrete cholesterol in the bile, with a fraction of the biliary cholesterol lost daily as fecal neutral sterols. In humans, the major route of excretion is as biliary cholesterol (two-thirds of the total lost each day), with catabolism to bile acids and bile acid excretion being the second most important route, accounting for approximately one-third of the daily turnover.

For all practical purposes, the body must excrete daily an amount of neutral and acidic sterols equivalent to the combined inputs of total dietary and newly synthesized cholesterol. Given an average fecal excretion of 1020 mg per day with 250 mg as acidic sterols, it can be calculated that the 770 mg per day excreted as neutral steroids comes from unabsorbed biliary

(650 mg) and unabsorbed dietary (120 mg) cholesterol (**Table 1**) It is easy to see that even small changes in the daily balance between a cholesterol input and output value of 1020 mg per day could, over years, result in significant tissue cholesterol accumulation.

Bile Acid Synthesis

The results from numerous sterol balance studies carried out in subjects fed diets low or high in cholesterol indicate that in humans, dietary cholesterol has little effect on fecal bile acid excretion rates. This finding is in striking contrast to results from studies of some rodent models that show that intake of pharmacological doses of dietary cholesterol can result in severalfold increases in bile acid synthesis and excretion. In contrast, some rodent species and nonhuman primates have little, if any, increase in bile acid excretion with increased intakes of cholesterol. Although there have been a few reports of enhanced bile acid excretion on a high-cholesterol diet in some patients, this does not appear to be a major regulatory response in humans.

Biliary Cholesterol Secretion

The majority of cholesterol entering the intestinal tract is biliary cholesterol. Biliary cholesterol secretion averages 1000 mg per day as part of the bile system and enters as free cholesterol already solubilized with bile acids and phospholipids. Both cholesterol absorption by enterocyte and biliary cholesterol secretion by hepatic cells are regulated by expression of the half-transporters ABCG5 and ABCG8. Studies in animals have shown that treatment with a LXR agonist decreases cholesterol absorption, increases biliary cholesterol secretion, and increases fecal neutral sterol excretion. Studies in transgenic mouse models demonstrate that increased expression of ABCG5 and ABCG8 increases biliary neutral sterol secretion and reduces intestinal cholesterol absorption, leading to increased neutral sterol excretion and cholesterol synthesis.

Fecal Excretion

The only route of significant cholesterol excretion is through fecal excretion of neutral sterols. The combination of unabsorbed biliary and dietary cholesterol accounts for the total neutral sterol output and under most conditions is 750–850 mg per day. Dietary patterns or drugs that interfere with intestinal cholesterol absorption result in increased fecal neutral steroid excretion. In the colon, intestinal bacteria are able to metabolize cholesterol to a variety of neutral steroids as well as to nonsteroid end products. Some studies have suggested that the intestinal metabolism of cholesterol by bacteria, which can be altered by diet and drugs, can influence endogenous cholesterol metabolism as well as plasma cholesterol levels. What these relationships may be and the mechanisms involved have not been defined.

Metabolic Function

Steroid Hormones

Daily production of steroid hormones is quantitatively a very small fraction of the daily turnover of dietary and newly synthesized cholesterol in the body. For men, the average daily excretion of steroid hormones is approximately 50 mg per day, whereas for women the value can be substantially higher depending on the menstrual phase.

Bile Acid Synthesis

The enterohepatic circulation of bile acids is essential for fat and cholesterol digestion and absorption. Each day, the bile acid pool (approximately 3–5 g) cycles through the intestine 6–10 times. The absorption of bile acids by the ileum is very effective and 98 or 99% of bile acids secreted in the bile are returned to the liver via the portal vein. The small amount of bile acids lost each day as fecal acidic steroids are replaced through the conversion of hepatic cholesterol to the primary bile acids, cholic acid and chenodeoxycholic acid. This catabolism of cholesterol can be as little as 250 mg per day up to 500 mg per day depending on the diet. The bile acids represent the only major catabolic product of cholesterol metabolism and in humans account for approximately 25–30% of the daily loss of cholesterol from the body.

Very Low-Density Lipoprotein Synthesis

The endogenous pathway for cholesterol transport focuses on the liver with the synthesis and secretion of VLDL particles. Cholesterol in these triacylglycerol-rich particles comes from multiple sources: endogenous synthesis, diet, and plasma lipoproteins. Catabolism of VLDL by LPL leads to formation of intermediate-density lipoproteins (IDLs), which can either be taken up by the liver or undergo further metabolism to form LDL. Low-density lipoproteins contain apo-B_{100} and account for 60–80% of the plasma cholesterol in most individuals. During lipolysis of VLDL triacylglycerol, the lipoproteins containing apo-B become enriched with cholesteryl ester through the plasma CETP-catalyzed net transfer of cholesteryl ester from HDL. This process, called reverse cholesterol transport, moves cholesterol from extrahepatic tissues to HDL to VLDL–IDL–LDL and eventual uptake by the liver.

Approximately 70% of the LDL degraded each day is degraded by the hepatic LDL receptor pathway.

Dietary Cholesterol and Plasma Cholesterol

The effect of dietary cholesterol on plasma cholesterol levels has been an area of considerable debate. In 1972, the American Heart Association recommended that dietary cholesterol intake should average less than 300 mg per day as part of a 'heart-healthy,' plasma cholesterol-lowering diet. Since that initial recommendation, a number of other public health dietary recommendations in the United States have endorsed the 300 mg daily limit. Interestingly, few dietary recommendations from other countries contain a dietary cholesterol limitation. The evidence for a relationship between dietary cholesterol and plasma cholesterol indicates that the effect is relatively small, and that on average a change of 100 mg per day in dietary cholesterol intake results in a 0.057 mmol l^{-1} (2.2 mg dl^{-1}) change in plasma cholesterol concentrations. Studies have also shown that the majority of individuals are resistant to the plasma cholesterol-raising effects of dietary cholesterol 'nonresponders' and have less than the predicted response. In contrast, a segment of the population (estimated to be between 15 and 25%) is sensitive to dietary cholesterol 'responders' and exhibits a greater than expected plasma cholesterol response to a change in dietary cholesterol intake. To date, there are no defined physiological or clinical characteristics to differentiate responders from nonresponders, but studies suggest that the apoE phenotype plays a role, as does the clinical condition of combined hyperlipidemia. Data also suggest that sensitivity to dietary cholesterol is associated with sensitivity to dietary fat, and that overall adiposity may also play a role. Although on a population basis the plasma cholesterol response to dietary cholesterol is relatively small, and in most epidemiological analyses not related to hypercholesterolemia, some individuals are sensitive to dietary cholesterol changes and, if hypercholesterolemic, would experience plasma cholesterol reduction with dietary cholesterol restrictions. For the majority, however, dietary cholesterol restrictions have little effect on plasma cholesterol levels.

Dietary Sources

Dietary Cholesterol Intake Patterns

Dietary cholesterol intakes in the United States have been declining, from an average of 500 mg per day in men and 320 mg per day in women in 1972 to levels in 1990 of 360 mg per day in men and 240 mg per day in women. This decline is due in part to dietary recommendations to the US public to reduce total and saturated fat intake and to reduce dietary cholesterol daily intake to less than 300 mg and in part from the increased availability of products with reduced fat and cholesterol content. Major efforts in the early 1970s by public health agencies and advertising emphasized reducing dietary cholesterol as a means to lower plasma cholesterol levels, leading to a high degree of consumer concern regarding cholesterol-containing foods and demand for low-cholesterol products. Today, practically all foods sold in the United States are labeled for their cholesterol content and their percentage contribution to the daily value of 300 mg for cholesterol.

Major Dietary Sources

The major sources of cholesterol in the diet are eggs, meat, and dairy products. A large egg contains approximately 215 mg of cholesterol and contributes approximately 30–35% of the total dietary cholesterol intake in the United States. Meat, poultry, and fish contribute 45–50%, dairy products 12–15%, and fats and oils 4–6%. In the United States, the range of dietary cholesterol intake is 300–400 mg per day for men and 200–250 mg per day for women; thus, for much of the population the national goal of a dietary cholesterol intake of less than 300 mg per day has been met.

See also: **Coronary Heart Disease**: Hemostatic Factors; Lipid Theory; Prevention. **Fats and Oils**. **Fatty Acids**: Metabolism. **Hyperlipidemia**: Overview; Nutritional Management. **Lipids**: Chemistry and Classification; Composition and Role of Phospholipids.

Further Reading

Dietschy JM, Turley SD, and Spady DK (1993) Role of liver in the maintenance of cholesterol and low density lipoprotein homeostasis in different animal species, including humans. *Journal of Lipid Research* **34**: 1637–1659.

Gylling H and Miettinen TA (2002) Inheritance of cholesterol metabolism of probands with high or low cholesterol absorption. *Journal of Lipid Research* **43**: 1472–1476.

McNamara DJ (1987) Effects of fat-modified diets on cholesterol and lipoprotein metabolism. *Annual Review of Nutrition* **7**: 273–290.

McNamara DJ (1990) Relationship between blood and dietary cholesterol. *Advances in Meat Research* **6**: 63–87.

McNamara DJ (2000) Dietary cholesterol and atherosclerosis. *Biochimica et Biophysica Acta* **1529**: 310–320.

Millatt LJ, Bocher V, Fruchart JC, and Staels B (2003) Liver X receptors and the control of cholesterol homeostasis: Potential therapeutic targets for the treatment of atherosclerosis. *Biochimica et Biophysica Acta* **1631**: 107–118.

Oram JF (2003) HDL apolipoproteins and ABCA1—Partners in the removal of excess cellular cholesterol. *Arteriosclerosis Thrombosis and Vascular Biology* **23**: 720–727.

Ordovas JM and Tai ES (2002) The babel of the ABCs: Novel transporters involved in the regulation of sterol absorption and excretion. *Nutrition Reviews* **60**: 30–33.

Sehayek E (2003) Genetic regulation of cholesterol absorption and plasma plant sterol levels: Commonalities and differences. *Journal of Lipid Research* **44**: 2030–2038.

Thompson GR, Naumova RP, and Watts GF (1996) Role of cholesterol in regulation of apolipoprotein B secretion by the liver. *Journal of Lipid Research* **37**: 439–447.

Wilson MD and Rudel LL (1994) Review of cholesterol absorption with emphasis on dietary and biliary cholesterol. *Journal of Lipid Research* **35**: 943–955.

Yu LQ, Li-Hawkins J, Hammer RE *et al.* (2002) Overexpression of ABCG5 and ABCG8 promotes biliary cholesterol secretion and reduces fractional absorption of dietary cholesterol. *Journal of Clinical Investigation* **110**: 671–680.

Factors Determining Blood Levels

S M Grundy, University of Texas Southwestern Medical Center, Dallas, TX, USA

Introduction

A high blood (serum) cholesterol level is a major risk factor for atherosclerotic coronary heart disease (CHD). Consequently, there has been much interest in the causes of elevated serum cholesterol concentrations. Although the serum cholesterol can be measured as a single entity, in fact cholesterol is carried in the bloodstream by several independent entities called lipoproteins. Each lipoprotein has its own characteristics, and the concentrations of each are affected by different factors. Several of these factors are related to diet, i.e., dietary cholesterol, certain fatty acids, and energy imbalance resulting in obesity. Other factors also modify lipoprotein metabolism, including advancing age, the postmenopausal state in women, and genetics. Consideration of each of the factors regulating serum cholesterol concentrations first requires a description of the different lipoprotein species.

Serum Lipoproteins

Lipoproteins are macromolecular complexes that consist of discrete particles and are composed of both lipids and proteins. The lipids include cholesterol, phospholipids, and triacylglycerols (TAG). A portion of serum cholesterol is esterified with a fatty acid; the remainder is unesterified. The protein components go by the name of apolipoproteins. The major forms of apolipoproteins and their functions are listed in **Table 1**. Four categories of lipoproteins that carry cholesterol in the serum are chylomicrons, very low-density lipoproteins (VLDL), low-density lipoproteins (LDL), and high-density lipoproteins (HDL). The characteristics and metabolism of each lipoprotein will be reviewed briefly.

Table 1 Apolipoproteins of serum lipoproteins

Apolipoprotein	*Function*
A-I	Major apolipoprotein of HDL Activator of LCAT
A-I	Structural apolipoprotein of HDL (other functions unknown)
A-IV	Apolipoprotein of chylomicrons (other functions unknown)
B-48	Chylomicron assembly and secretion
B-100	VLDL assembly and secretion Ligand for LDL receptor unknown
C-I	Unknown
C-II	Activator of LPL
C-III	Inhibitor of LPL
E	Apolipoprotein of remnant lipoproteins Ligand for LDL receptor Promotes hepatic uptake of remnants

Abbreviations: HDL, high-density lipoproteins; LDL, low-density lipoproteins; VLDL, very low-density lipoproteins; LCAT, lecithin cholesterol acyl transferase; LPL, lipoprotein lipase.

Chylomicrons

Dietary cholesterol enters the intestine together with fat, which is predominantly TAG. The latter undergoes hydrolysis by pancreatic lipase and releases fatty acids and monoacylglycerols. In the intestine, these mix with bile acids, phospholipids, and cholesterol from the bile. The mixture of hydrolyzed lipids associates with phospholipids and bile acids to form mixed micelles. Fatty acids, monoacylglycerols, and cholesterol are taken up by the intestinal mucosa. In the mucosal cells, the fatty acids and monoacylglycerols are recombined by enzymatic action to form TAG, which are incorporated with the cholesterol into lipoprotein particles called chylomicrons. Most of the cholesterol in chylomicrons is esterified with a fatty acid. The major apolipoprotein of chylomicrons is apo B-48; other apolipoproteins – apo Cs, apo Es and apo As – attach to the surface coat of chylomicrons and aid in metabolic processing. In the mucosal cells, microsomal lipid transfer protein (MTP) facilitates the transfer of TAG and cholesterol ester into chylomicron particles. The presence of MTP is required for the secretion of chylomicrons from mucosal cells.

Chylomicrons are secreted by intestinal mucosal cells into the lymphatic system; from here they pass through the thoracic duct into the systemic circulation. When chylomicrons enter the peripheral

circulation they come into contact with an enzyme, lipoprotein lipase (LPL), which is located on the endothelial surface of capillaries. LPL is activated by apo C-II on chylomicrons; this process is modulated by apo C-III, an inhibitor of LPL activity. Nonetheless most chylomicron TAG is hydrolyzed by LPL; a residual lipoprotein particle, named chylomicron remnant, is released into the circulation and is rapidly removed by the liver. Hepatic uptake of chylomicron remnants is believed to be mediated by binding of remnants with a glycoprotein on the surface of liver cells. Almost all newly absorbed cholesterol thus enters the liver in association with chylomicron remnants.

Very Low-Density Lipoproteins

The liver also secretes a TAG-rich lipoprotein called VLDL. Fatty acids used in synthesis of TAG in the liver are normally derived from circulating nonesterified fatty acids (NEFA); even so, the liver has the capacity to synthesize fatty acids when the diet contains mainly carbohydrate. MTP inserts TAG into newly forming VLDL particles. The surface coat of VLDL contains unesterified cholesterol, phospholipids, and apolipoproteins. The major apolipoprotein of VLDL is apo B-100. Other apolipoproteins, notably apo Cs and apo Es, also are present. As VLDL circulate they acquire cholesterol esters from HDL. Circulating VLDL particles lose TAG through interaction with LPL in the peripheral circulation; in this process, VLDL are transformed into VLDL remnants. The latter can have two fates: hepatic uptake or conversion to LDL. Hepatic uptake of VLDL remnants may occur via two mechanisms: interaction with glycoproteins or interaction with LDL receptors. Both glycoproteins and LDL receptors are located on the surface of liver cells.

Low-Density Lipoproteins

Conversion of VLDL remnants to LDL appears to be largely the result of hydrolysis of remaining TAG by hepatic triacylglycerol lipase (HTGL). Normally about two-thirds of cholesterol is carried by LDL, most of this LDL cholesterol existing in the form of esters. The only apolipoprotein in LDL is apo B-100. LDL is removed from the circulation largely by hepatic LDL receptors. The level of expression of LDL receptors is a major determinant of serum LDL cholesterol concentrations. The synthesis of LDL receptors is regulated in large part by the liver's content of cholesterol. An increase in hepatic cholesterol content suppresses LDL receptor synthesis and raises serum LDL cholesterol; conversely, a decrease in hepatic cholesterol stimulates receptor synthesis and lowers serum LDL cholesterol. The mechanism whereby hepatic cholesterol controls LDL receptor synthesis is through a regulatory protein called sterol regulatory element-binding protein (SREBP). When hepatic cholesterol content falls, SREBP is activated and stimulates the synthesis of LDL receptors.

The regulatory form of cholesterol in the liver cell is unesterified cholesterol, not cholesterol ester. The hepatic content of unesterified cholesterol depends on several factors including the amounts of cholesterol derived from chylomicrons and other lipoproteins, hepatic synthesis of cholesterol, secretion of cholesterol into bile, conversion of cholesterol into bile acids, esterification of cholesterol, and secretion of cholesterol into serum with VLDL. Factors that influence each of these processes can alter serum LDL cholesterol concentrations by modifying the hepatic content of unesterified cholesterol and thereby expression of LDL receptors.

High-Density Lipoproteins

HDL consist of a series of lipoprotein particles of relatively high density, all of which contain apo A-I. A proportion of HDL particles also contain apo A-II. Some HDL species (HDL_3) are denser than others (HDL_2). HDL particles are composed largely of by-products of catabolism of TAG-rich lipoproteins. The surface coats of HDL particles contain phospholipids and unesterified cholesterol, apo A-I with or without apo A-II, and other apolipoproteins (apo Cs and apo Es). Their particle cores consist largely of cholesterol esters, although small amounts of TAG are also present. The cholesterol esters of HDL are formed by esterification with a fatty acid through the action of an enzyme, lecithin cholesterol acyl transferase (LCAT); the substrates for this reaction derive either from unesterified cholesterol released during lipolysis of TAG-rich lipoproteins or from the surface of peripheral cells. After esterification of cholesterol, the cholesterol esters of HDL are transferred back to TAG-rich lipoproteins and eventually are removed by the liver through direct uptake of remnant lipoproteins or LDL. Whether whole HDL particles can be directly removed from the circulation is uncertain. Some investigators believe that the HDL components are dismantled and removed piecemeal.

Dietary Regulation of Serum Lipoproteins

A large body of research has shown that diet has a major impact on the concentrations and composition of serum lipoproteins, and hence on serum

cholesterol concentrations. Three major factors affect cholesterol and lipoprotein concentrations: (1) dietary cholesterol, (2) the macronutrient composition of the diet, particularly dietary fatty acids, and (3) energy balance, as reflected by body weight. The influence of each of these factors can be considered.

Dietary Cholesterol

All dietary cholesterol is derived from animal products. The major sources of cholesterol in the diet are egg yolks, products containing milk fat, animal fats, and animal meats. Many studies have shown that high intakes of cholesterol will increase the serum cholesterol concentration. Most of this increase occurs in the LDL cholesterol fraction. When cholesterol is ingested, it is incorporated into chylomicrons and makes its way to the liver with chylomicron remnants. There it raises hepatic cholesterol content and suppresses LDL receptor expression. The result is a rise in serum LDL cholesterol concentrations. Excess cholesterol entering the liver is removed from the liver either by direct secretion into bile or by conversion into bile acids; also, dietary cholesterol suppresses hepatic cholesterol synthesis. There is considerable variability in each of these steps in hepatic cholesterol metabolism; for this reason the quantitative effects of dietary cholesterol on serum LDL cholesterol levels vary from one person to another. For every 200 mg of cholesterol per day in the diet, serum LDL cholesterol is increased on average by about 6 mg dl^{-1} (0.155 mmol l^{-1}).

Macronutrient Composition of the Diet

Dietary fat and fatty acids Most of the fat in the diet consists of TAG that are composed of three fatty-acid molecules bonded to glycerol. The contribution of TAG to total energy intake varies among individuals and populations, ranging from 15% to 40% of total nutrient energy. The fatty acids of TAGs are of several types: saturated, *cis*-monounsaturated, *trans*-monounsaturated, and polyunsaturated fatty acids. All fatty acids affect lipoprotein levels in one way or another. **Table 2** lists the major fatty acids of the diet and denotes their effects on serum lipoproteins. Also shown are the effects of carbohydrates, which also influence serum lipoprotein metabolism. It should be noted that all lipoprotein responses are compared with and related to those of *cis*-monounsaturated fatty acids, which are widely accepted to be neutral, or baseline.

Saturated fatty acids The saturated fatty acids are derived from both animal fats and plant oils. Rich sources of dietary saturated fatty acids include butter fat, meat fat, and tropical oils (palm oil, coconut oil, and palm kernel oil). Saturated fatty acids are straight-chain organic acids with an even number of carbon atoms (**Table 2**). All saturated fatty acids that have from eight to 16 carbon atoms raise the serum LDL cholesterol concentration when they are consumed in the diet. In the USA and much of Europe, saturated fatty acids make up 12–15% of total nutrient energy intake.

Table 2 Macronutrient effects on serum lipoprotein cholesterol

Nutrient	*Symbol*[a]	*VLDL cholesterol*[a]	*LDL cholesterol*	*HDL cholesterol*
Fatty acids				
Saturated				
Palmitic	$C_{16:0}$	–[b]	↑↑	–
Myristic	$C_{14:0}$	–	↑↑↑	↓
Lauric	$C_{12:0}$	–	↑	–
Caproic	$C_{10:0}$	–	↑	–
Caprilic	$C_{8:0}$	–	↑	–
Stearic	$C_{18:0}$	–	–	or ↓
trans-Monounsaturated	*trans* $C_{18:1\ n\text{-}9}$	–	↑ or ↑↑	↓
cis-Monounsaturated	*cis* $C_{18:1\ n\text{-}9}$	–	–	–
Polyunsaturated				
n-6[d]	$C_{18:2\ n\text{-}6}$	– or ↓	– or ↓	– or ↓
n-3[d]	DHA, EPA[e]	↓↓↓	– or ↓	–
Carbohydrate		↑↑↑	–	↓↓

[a]First number denotes number of carbon atoms; second number denotes number of double bonds.
[b]The dash (–) indicates that there is no change in level compared with that produced by *cis*-monosaturated fatty acids (oleic acid) ($C_{18:1\ n\text{-}9}$). All the lipoprotein responses to oleic acid are considered 'neutral', i.e., no effect.
[c]The letter 'n' and number indicates at which carbon atom, numbered from the terminal methyl group, the first double bond appears.
Abbreviations: VLDL, very low-density lipoproteins; LDL, low-density lipoproteins; HDL, high-density lipoproteins; DHA, docosahexanoic acid ($C_{22:6\ n\text{-}3}$); EPA, eicosapentanoic acid ($C_{20:5\ n\text{-}3}$).

The mechanisms whereby saturated fatty acids raise LDL cholesterol levels are not known, although available data suggest that they suppress the expression of LDL receptors. The predominant saturated fatty acid in most diets is palmitic acid ($C_{16:0}$); it is cholesterol-raising when compared with *cis*-monounsaturated fatty acids, specifically oleic acid ($C_{18:cis1\ n\text{-}9}$), which is considered to be 'neutral' with respect to serum cholesterol concentrations. In other words, oleic acid is considered by most investigators to have no effect on serum cholesterol or lipoproteins. Another saturated fatty acid, myristic acid ($C_{14:0}$), apparently raises LDL cholesterol concentrations somewhat more than does palmitic acid, whereas other saturates – lauric ($C_{12:0}$), caproic ($C_{10:0}$), and caprylic ($C_{8:0}$) acids – have a somewhat lesser cholesterol-raising effect. On average, for every 1% of total energy consumed as cholesterol-raising saturated fatty acids, compared with oleic acid, the serum LDL cholesterol level is raised about 2 mg dl^{-1} (0.025 mmol l^{-1}).

One saturated fatty acid, stearic acid ($C_{18:0}$), does not raise serum LDL cholesterol concentrations. The main sources of this fatty acid are beef tallow and cocoa butter. The reason for its failure to raise LDL cholesterol concentrations is uncertain, but may be the result of its rapid conversion into oleic acid in the body.

***Trans*-monounsaturated fatty acids** These fatty acids are produced by hydrogenation of vegetable oils. Intakes of *trans*-monounsaturates vary from one country to another depending on consumption of hydrogenated oils. In many countries they contribute between 2% and 4% of total nutrient energy intake. A series of *trans* acids are produced by hydrogenation: most are monounsaturated. For many years, it was accepted that *trans*-monounsaturated fatty acids were neutral with respect to LDL cholesterol concentrations. However, recent studies have shown that they raise LDL cholesterol concentrations to a level similar to that of palmitic acid when substituted for dietary oleic acid. In addition, they cause a small reduction in serum HDL cholesterol concentrations. Thus, *trans*-monounsaturates must be placed in the category of cholesterol-raising fatty acids.

***Cis*-monounsaturated fatty acids** The major fatty acid in this category is oleic acid ($C_{18:cis1\ n\text{-}9}$). It is found in both animal and vegetable fats, and typically is the major fatty acid in diet. Intakes commonly vary between 10% and 20% of total energy. Oleic acid intake is particularly high in the Mediterranean region where large amounts of olive oil are consumed. Other sources rich in oleic acid are rapeseed oil (canola oil) and high-oleic forms of safflower and sunflower oils. Peanuts and pecans also are high in oleic acid. Animal fats likewise contain a relatively high percentage of oleic acid among all their fatty acids; even so, these fats also tend to be rich in saturated fatty acids. When high-carbohydrate diets are consumed, the human body can synthesize fatty acids; among these, oleic acid is the predominant fatty acid produced.

As indicated before, oleic acid generally is considered to be the 'baseline' fatty acid with respect to serum lipoproteins levels, i.e., it does not raise (or lower) LDL cholesterol or VLDL cholesterol concentrations, nor does it lower (or raise) HDL cholesterol concentrations. It is against this 'neutral' fatty acid that responses of other fatty acids are defined (**Table 2**). For example, when oleic acid is substituted for cholesterol-raising fatty acids, the serum LDL cholesterol concentration will fall. Nonetheless, oleic acid is not designated a cholesterol-lowering fatty acid, but instead, this response defines the cholesterol-raising potential of saturated fatty acids.

Polyunsaturated fatty acids There are two categories of polyunsaturated fatty acids: n-6 and n-3. The major n-6 fatty acid is linoleic acid ($C_{18:2,n\text{-}6}$). It is the predominant fatty acid in many vegetable oils, e.g., corn oil, soya bean oil, and high linoleic forms of safflower and sunflower seed oils. Intakes of linoleic acid typically vary from 4% to 10% of nutrient energy, depending on how much vegetable oil is consumed in the diet. The n-3 fatty acids include linolenic acid ($C_{18:3,n\text{-}3}$), docosahexanoic acid (DHA) ($C_{22:6,n\text{-}3}$), and eicosapentanoic acid (EPA) ($C_{20:5,n\text{-}3}$). Linolenic acid is high in linseed oil and present in smaller amounts in other vegetable oils. DHA and EPA are enriched in fish oils.

For many years, linoleic acid was thought to be a unique LDL cholesterol-lowering fatty acid. Recent investigations suggest that earlier findings overestimated the LDL-lowering potential of linoleic acid. Even though substitution of linoleic acid for oleic acid in the diet may reduce LDL cholesterol levels in some people, a difference in response is not consistent. Only when intakes of linoleic acid become quite high do any differences become apparent. At high intakes, however, linoleic acid also lowers serum HDL cholesterol concentrations. Moreover, compared with oleic acid, it may reduce VLDL cholesterol levels in some people. Earlier enthusiasm for high intakes of linoleic acid to reduce LDL cholesterol levels has been dampened for several reasons: for example, its LDL-lowering ability does not offset

potential disadvantages of HDL lowering, and other concerns include possible untoward side effects such as promoting oxidation of LDL and suppressing cellular immunity to cancer.

The n-3 fatty acids in fish oils (DHA and EPA) have a powerful action to reduce serum VLDL levels. This action apparently results from suppression of the secretion of VLDL by the liver. The precise mechanism for this action is not known. However, these fatty acids do not reduce LDL cholesterol concentrations relative to oleic acid. They have been used for treatment of some patients with elevated VLDL concentrations, although drug treatment generally is employed when it is necessary to lower serum VLDL levels.

Carbohydrate When carbohydrates are substituted for oleic acid in the diet, serum LDL cholesterol levels remain unchanged. However, VLDL cholesterol concentrations usually rise and HDL cholesterol concentrations fall on high-carbohydrate diets. Thus, a lack of difference in total serum cholesterol concentrations during the exchange of carbohydrate and oleic acid is misleading. The two categories of nutrients have different actions on lipoprotein metabolism. The differences in response to dietary carbohydrate and oleic acid provide a good example of how measurements of serum total cholesterol fail to reveal all of the changes that are occurring in the lipoprotein fractions.

Energy Balance

Obesity When energy intake exceeds energy expenditure, the balance of energy is stored in adipose tissue in the form of TAG. When the TAG content of adipose tissue becomes excessive (body mass index 30 or above), a state of obesity is said to exist. In some obese persons, excessive accumulations of TAG occur in other tissues than adipose tissue. Two such tissues are skeletal muscle and liver. High contents of TAG in muscle and liver arise primarily because of continuous leakage of excessive quantities of NEFA from adipose tissue. In the presence of desirable body weight, normal insulin levels are sufficient to suppress hydrolysis of TAG in adipose tissue, and NEFA release is low. On the other hand, in obese persons NEFA release is excessive, and skeletal muscle and liver are flooded with high serum NEFA concentrations. The result is engorgement of these organs with TAG. When skeletal muscle is overloaded with TAG, insulin-mediated glucose uptake is impaired. This condition is called insulin resistance. When liver is packed with TAG, hepatic metabolism is altered and insulin action on the liver is deranged. As a result, there is an overproduction of VLDL; this leads to high VLDL cholesterol concentrations and, because LDL is a product of VLDL, to higher LDL cholesterol levels. In addition, obesity is accompanied by a reduction in HDL cholesterol concentrations. Thus obesity is responsible for multiple alterations in lipoprotein metabolism; it has significant effects on three major lipoprotein species – VLDL, LDL, and HDL. These changes appear to be the result of a combination of excessive hepatic TAG as a substrate for VLDL formation and failure of insulin to exert its usual action to curtail VLDL secretion.

Exercise Many of the adverse metabolic effects of obesity are reversed by exercise. Increased energy expenditure through regular and sustained exercise helps to prevent accumulation of excessive quantities of TAG in adipose tissue. In addition, increased muscle metabolism produced by exercise burns off NEFA and prevents TAG accumulation in the liver. Hence, increased and sustained energy expenditure favourably modifies the lipoproteins, particularly by lowering VLDL cholesterol concentrations and raising serum HDL cholesterol. Effects of exercise on LDL cholesterol concentrations are more modest, but in some people exercise produces a reduction.

Other Factors Affecting Serum Lipoproteins

Advancing Age

Between the ages of 20 and 50 years, there is a gradual rise in serum cholesterol concentrations. In the USA, for example, the serum cholesterol increases on average about $50\,mg\,dl^{-1}$ ($1.295\,mmol\,l^{-1}$). This change may be related in part to increasing obesity, according to the mechanisms described above. However, even in people who do not gain weight with advancing age, serum cholesterol concentrations usually rise to some extent. Available evidence indicates that this rise results from a decrease in expression of LDL receptors. The reasons for a decline in receptor synthesis with aging are not known, but may reflect 'metabolic' aging. However, in men, after age 50 years, there is little further rise in serum cholesterol. This observation suggests that the impact of weight gain, which occurs mostly between the ages of 20 and 50 years, may be greater than generally recognized.

Postmenopausal State in Women

In women, there is a further rise in serum cholesterol concentrations after age 50 years. This rise is believed to be due largely to loss of oestrogens after the

menopause. Oestrogens are known to stimulate the synthesis of LDL receptors, and, consequently, receptor expression declines after the menopause. This increment in cholesterol levels can be largely reversed by oestrogen replacement therapy.

Genetics

Family studies and research in twins indicate that about 50% of the variation of serum cholesterol concentrations in the general population can be explained by genetic polymorphisms. Presumably this variation is related to factors that regulate lipoprotein concentrations. In some cases, specific genetic defects are severe, resulting in marked changes in lipoprotein concentrations. When this occurs, the affected individual is said to have a monogenic disorder. In other cases, multiple genetic modifications are present that combine to alter lipoprotein concentrations. When a few modifications are present, the condition is called oligogenic, but when many modifications combine to change lipoprotein concentrations, the condition is named polygenic. Several monogenic disorders have been identified; a few oligogenic conditions have been described, but there are very few instances in which complex polygenic traits have been unravelled. A question of great interest is whether nutritional and genetic factors ever interact synergistically to alter lipoprotein concentrations. Undoubtedly, dietary factors and genetic changes can be additive in their effects on serum lipoproteins, but synergistic interaction has been difficult to prove. In what follows, consideration will be given to the impact of modification of some of the key gene products regulating lipoprotein metabolism.

LDL receptors The most severe elevations in LDL cholesterol levels occur in patients who have mutations in the gene encoding for LDL receptors. About one in 500 people are heterozygous for these mutations. Their condition is called heterozygous familial hypercholesterolemia. LDL cholesterol concentrations are essentially twice the normal level in this condition. Very rarely patients are homozygous for mutation in the LDL receptor gene and thus have homozygous familial hypercholesterolemia. Their LDL cholesterol levels are approximately four times normal. Individuals with this condition develop severe premature atherosclerosis.

Many other people appear to have a reduction in LDL receptor expression on a genetic basis, but they do not have as severe elevations of serum LDL cholesterol as patients with familial hypercholesterolemia. Presumably, these people have genetic modifications in factors that regulate transcription of the LDL receptor gene. Although such genetic modifications may be relatively common, they are poorly defined. Again, an important but unanswered question is whether some people are genetically susceptible to the cholesterol-raising effects of dietary cholesterol and saturated fatty acids. If so, they may possess modifications in the genetic control of LDL receptor expression.

Apolipoprotein B-100 structure About one in 500 people also have a mutation in the primary structure of apo B that interferes with its binding to LDL receptors. This mutation gives rise to the disorder called familial defective apolipoprotein B-100. The consequence is an elevation of LDL cholesterol concentrations, and the clinical pattern resembles that of familial hypercholesterolemia.

Apolipoprotein B synthesis Rare patients have mutations in the gene encoding for apo B that impair the synthesis of this apolipoprotein. Such patients usually have very low LDL cholesterol concentrations. These individuals are said to have familial hypobetalipoproteinemia. In other rare cases, the intracellular TAG transport protein called MCT is genetically absent; when this occurs, no lipoprotein particles containing apo B can be formed. LDL cholesterol is absent from serum, and the disorder is called familial abetalipoproteinemia.

Some researchers speculate that serum elevations in VLDL cholesterol and LDL cholesterol can result from excessive synthesis and/or secretion of apo B-containing lipoproteins by the liver. When this occurs on a genetic basis, the disorder is designated familial combined hyperlipidemia. However, a monogenic basis of this clinical phenotype has not yet been identified. Therefore, most investigators have concluded that familial combined hyperlipidemia probably represents an oligogenic or polygenic disorder. In this disorder, lipoprotein elevations appear to be worsened by nutritional factors – particularly by obesity.

Apolipoprotein E This apolipoprotein is present on TAG-rich lipoproteins and it facilitates removal of remnant lipoproteins by LDL receptors in the liver. When apo E is affected by mutation, this enabling action is curtailed and hepatic uptake of remnant lipoproteins is impaired. The result is an accumulation of chylomicron remnants and VLDL remnants in the circulation. The accumulation is accentuated by the coexistence of other disorders of metabolism of TAG-rich lipoproteins. When remnant

accumulation occurs on a genetic basis, the disorder is called familial dysbetalipoproteinemia.

Apolipoprotein C There are two forms of apo C – apo C-II and apo C-III. Apo C-II is required for activation of LPL; when it is genetically absent, affected patients develop severe elevations of TAG-rich lipoproteins. Apo C-III inhibits the activity of LPL. In certain metabolic disorders, notably insulin resistance, synthesis of apo C-III is increased; an elevated apo C-III can lead to impaired function of LPL and increases in serum concentrations of TAG-rich lipoproteins.

Apolipoprotein A-I This is the major apolipoprotein of HDL. Rare patients have mutations in apo A-I that result in very low concentrations of HDL cholesterol. However, most people in whom HDL cholesterol concentrations are moderately reduced show increased catabolism of apo A-I. The mechanism for this change has not been fully determined, but one important cause may be an overexpression of HTGL.

Lipoprotein lipase This enzyme is required for lipolysis of TAG in TAG-rich lipoproteins. Rare patients are homozygous for mutations in LPL that impair its function. In such patients, serum concentrations of chylomicrons are markedly increased. The accumulation of chylomicrons in serum is greatly accentuated by the presence of fat in the diet. Only by severe dietary fat restriction is it possible to prevent severe TAG elevations in serum.

Genetic regulation of HDL cholesterol Family and twin studies reveal that about 50% of the variation in serum HDL cholesterol levels in the general population is explained by genetic factors. However, the regulation of HDL cholesterol concentrations is complex, and HDL cholesterol levels are determined by many factors, e.g., serum TAG concentrations, activity of HTGL, production rates of apo A-I, and activities of cholesterol ester transfer protein and LCAT. Genetic factors undoubtedly affect each of these regulating factors.

See also: **Eggs**. **Exercise**: Beneficial Effects; Diet and Exercise. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Lipoproteins**. **Meat, Poultry and Meat Products**. **Obesity**: Definition, Etiology and Assessment.

Further Reading

Bonanome A and Grundy SM (1988) Effect of dietary stearic acid on plasma cholesterol and lipoprotein levels. *New England Journal of Medicine* **318**: 1244–1248.

Cater NB, Heller HJ, and Denke MA (1997) Comparison of the effects of medium-chain triacylglycerols, palm oil, and high oleic sunflower oil on plasma acylglycerol fatty acids and lipid and lipoprotein concentrations in humans. *American Journal of Clinical Nutrition* **65**: 41–45.

Connor WE (1988) Effects of omega-3 fatty acids in hypertriglyceridemic states. *Seminars in Thrombosis and Hemostasis* **14**: 271–284.

Denke MA, Sempos CT, and Grundy SM (1993) Excess body weight: an underrecognized contributor to high blood cholesterol levels in white American men. *Archives of Internal Medicine* **153**: 1093–1103.

Dietschy JM, Turley SD, and Spady DK (1993) Role of liver in the maintenance of cholesterol and low density lipoprotein homeostasis in different animal species, including humans. *Journal of Lipid Research* **34**: 1637–1659.

Ericsson S, Eriksson M, Vitols S *et al.* (1991) Influence of age on the metabolism of plasma low density lipoproteins in healthy males. *Journal of Clinical Investigation* **87**: 591–596.

Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (S M Grundy, chairman) (1994) National Cholesterol Education Program: Second Report of the Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol (Adult Treatment Panel II). *Circulation* **89**: 1329–1445.

Grundy SM (1986) Comparison of monounsaturated fatty acids and carbohydrates for lowering plasma cholesterol. *New England Journal of Medicine* **314**: 745–748.

Grundy SM (1991) Multifactorial etiology of hypercholesterolemia: implications for prevention of coronary heart disease. *Arteriosclerosis and Thrombosis* **11**: 1619–1635.

Grundy SM and Denke MA (1990) Dietary influences on serum lipids and lipoproteins. *Journal of Lipid Research* **31**: 1149–1172.

Hegsted DM, McGandy RB, Myers ML, and Stare FJ (1965) Quantitative effects of dietary fat on serum cholesterol in man. *American Journal of Clinical Nutrition* **17**: 281–295.

Innerarity TL, Mahley RW, Weisgraber KH *et al.* (1990) Familial defective apolipo-protein B-100: a mutation of apolipoprotein B that causes hypercholesterolemia. *Journal of Lipid Research* **31**: 1337–1349.

Keys A, Anderson JT, and Grande F (1965) Serum cholesterol response to changes in the diet. IV. Particular saturated fatty acids in the diet. *Metabolism* **14**: 776–787.

Mattson FH and Grundy SM (1985) Comparison of effects of dietary saturated, monounsaturated, and polyunsaturated fatty acids on plasma lipids and lipoproteins in man. *Journal of Lipid Research* **26**: 194–202.

Mensink RP and Katan MB (1990) Effect of dietary *trans* fatty acids on high-density and low-density lipoprotein cholesterol levels in healthy subjects. *New England Journal of Medicine* **323**: 439–445.

CHOLINE AND PHOSPHATIDYLCHOLINE

X Zhu and S H Zeisel, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Introduction

Choline, an essential nutrient for humans, is consumed in many foods. It is part of several major phospholipids (including phosphatidylcholine – also called lecithin) that are critical for normal membrane structure and function. Also, as the major precursor of betaine it is used by the kidney to maintain water balance and by the liver as a source of methyl groups for the removal of homocysteine in methionine formation. Finally, choline is used to produce the important neurotransmitter acetylcholine (catalyzed by choline acetyltransferase in cholinergic neurons and in such non-nervous tissues as the placenta). Each of these functions for choline is absolutely vital for the maintenance of normal function.

Although there is significant capacity for biosynthesis of the choline moiety in the liver, choline deficiency can occur in humans. Male adults deprived of dietary choline become depleted of choline in their tissues and develop liver and muscle damage. Premenopausal women may not be sensitive to dietary choline deficiency (unpublished data). No experiments have been conducted to determine if this occurs in similarly deprived pregnant women, infants, and children.

Endogenous Formation of Choline Moiety as Phosphatidylcholine

Unless eaten in the diet, choline can only be formed during phosphatidylcholine biosynthesis through the methylation of phosphatidylethanolamine by phosphatidylethanolamine *N*-methyltransferase (PEMT) using *S*-adenosylmethionine as the methyl donor. This enzyme is most active in the liver but has been identified in many other tissues including brain and mammary gland. At least two isoforms of PEMT exist: PEMT1, localized to the endoplasmic reticulum and generating the majority of PEMT activity, and PEMT2, which resides on mitochondria-associated membranes. Both enzymes are encoded by the same gene but differ either because of post-translational modification or alternative splicing. This gene is very polymorphic and functional SNPs (single nucleotide polymorphisms) in humans may exist and, if so, would influence dietary requirements for choline. In mice in which this gene is knocked out, the dietary requirement for choline is increased and they get fatty liver when eating a normal choline diet. Estrogen induces greater activity of PEMT perhaps explaining why premenopausal women require less choline in their diets. In addition to formation of choline, this enzyme has an essential role in lipoprotein secretion from the liver.

Choline, Homocysteine, and Folate are Interrelated Nutrients

Choline, methionine, methyltetrahydrofolate (methyl-THF), and vitamins B_6 and B_{12} are closely interconnected at the transmethylation metabolic pathways that form methionine from homocysteine. Perturbing the metabolism of one of these pathways results in compensatory changes in the others. For example, as noted above, choline can be synthesized *de novo* using methyl groups derived from methionine (via *S*-adenosylmethionine). Methionine can be formed from homocysteine using methyl groups from methyl-THF, or using methyl groups from betaine that are derived from choline. Similarly, methyl-THF can be formed from one-carbon units derived from serine or from the methyl groups of choline via dimethylglycine. When animals and humans are deprived of choline, they use more methyl-THF to remethylate homocysteine in the liver and increase dietary folate requirements. Conversely, when they are deprived of folate, they use more methyl groups from choline, increasing the dietary requirement for choline. There is a common polymorphism in the gene for methyltetrahydrofolate reductase that increases dietary requirement for folic acid; 15–30% of humans have this mutation. In mice in which this gene is knocked out, the dietary requirement for choline is increased and they get fatty liver when eating a normal choline diet.

Choline in Foods

Choline, choline esters, and betaine can be found in significant amounts in many foods consumed by humans (see **Figure 1** and **Figure 2**); some of the choline and betaine is added during processing (especially in the preparation of infant formula).

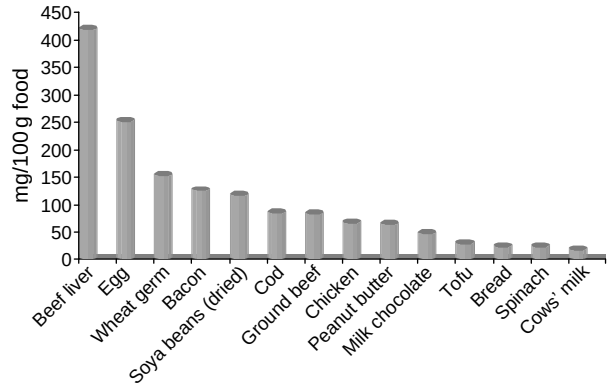


Figure 1 Total choline content of some common foods. Foods, which had been prepared as normally eaten, were analyzed for choline, phosphocholine, glycerophosphocholine, phosphatidylcholine, and sphingomyelin content using an HPLC mass spectrometric method. (Modified from Zeisel SH, Mar M-H, Howe JC, and Holden JM (2003) Concentrations of choline-containing compounds and betaine in common foods. *Journal of Nutrition* 133: 1302–1307.)

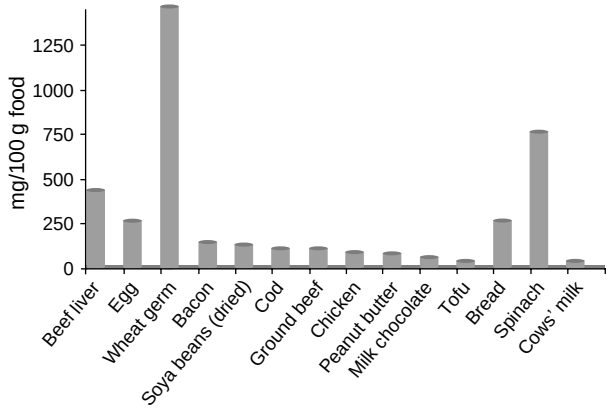


Figure 2 Total choline plus betaine content of some common foods. For methyl donation choline must be converted to betaine, thus the methyl donor capacity is best expressed as total choline and betaine content, assayed as in **Figure 1**. Several vegetable and grain products contain significant amounts of betaine (Modified from Zeisel SH, Mar M-H, Howe JC, and Holden JM (2003) Concentrations of choline-containing compounds and betaine in common foods. *Journal of Nutrition* 133: 1302–1307.)

Though the different esters of choline have different bioavailability, it is likely that choline in all forms is fungible; therefore, total choline content is probably the best indicator of food choline content. Betaine should also be considered, as it spares the use of choline for methyl donation.

A number of epidemiologic studies have examined the relationship between dietary folic acid and cancer or heart disease. It may be helpful to also consider choline intake as a confounding factor because folate and choline methyl donation can be interchangeable.

Dietary Recommendations

The Institute of Medicine, USA National Academy of Sciences, recommended an adequate intake (I) of 550 mg/70 kg body weight for choline in the diet. This amount may be influenced by gender, and it may be influenced by pregnancy, lactation, and stage of development (**Table 1**).

Amino acid-glucose solutions used in total parenteral nutrition of humans lack choline. The lipid emulsions that deliver extra calories and essential fatty acids during parenteral nutrition contain choline in the form of lecithin (20% emulsion contains 13.2 mmol l^{-1}). Humans treated with parenteral nutrition require 1–1.7 mmol of choline-containing phospholipid per day during the first week of parenteral nutrition therapy to maintain plasma choline levels.

Human milk, which contains approximately 200 mg l^{-1} choline and choline esters, is an especially good source of choline. An infant consuming 500 ml breast milk in a day ingests 50 mg choline. Human milk is not a static food; its choline composition changes over time postnatally. The choline composition of infant formulas can differ greatly from that present in human milk. It is essential that variations in the bioavailability and utilization of choline, phosphocholine, glycerophosphocholine, and lecithin in milk be considered when milk substitutes are developed.

Functional Effects of Varying Choline in the Diet

Fatty Liver

The triacylglycerol (TG) produced by the liver is mainly delivered to other tissues as very low-density

Table 1 Recommended adequate intakes (AI) for choline

Population	*Age*	*AI (mg day^{-1})*
Infants	0–6 months	125
	6–12 months	150
Children	1 through 3 years	200
	4 through 8 years	250
	9 through 13 years	375
Males	14 through 18 years	550
	19 years and older	550
Females	14 through 18 years	400
	19 years and older	425
Pregnancy	All ages	450
Lactation	All ages	550

From Institute of Medicine, National Academy of Sciences USA (1998) Dietary reference intakes for folate, thiamin, riboflavin, niacin, vitamin B_{12}, panthothenic acid, biotin, and choline, vol. 1. Washington DC: National Academy Press.

lipoprotein (VLDL) of which lecithin is a required component. In choline deficiency, the diminished ability of liver cells to synthesize new lecithin molecules results in the intracellular accumulation of TG. Treating malnourished patients with high-calorie total parenteral nutrition (TPN) solutions that contain little or no choline will deplete choline stores and cause fatty liver and hepatic dysfunction that can be reversed by treatment with phosphatidylcholine.

Liver Cell Death

When deprived of dietary choline, healthy male subjects have diminished plasma concentrations of choline and phosphatidylcholine, and they develop liver cell death (elevated plasma alanine aminotransferase). In similarly deprived animal models, the liver cell death is caused by apoptosis, a regulated form of cell suicide. In an ongoing study of choline deficiency in humans, muscle cell death (elevated plasma creatine phosphokinase, MM form) has also been noted.

Liver Cancer

Dietary deficiency of choline in rodents causes development of hepatocarcinomas in the absence of any known carcinogen. Choline is the only single nutrient for which this is true. It is interesting that choline-deficient rats not only have a higher incidence of spontaneous hepatocarcinoma but also are markedly sensitized to the effects of administered carcinogens. Several mechanisms are suggested for the cancer-promoting effect of a choline-devoid diet. A progressive increase in cell proliferation that is related to regeneration after parenchymal cell death occurs in the choline-deficient liver. Cell proliferation and its associated increased rate of DNA synthesis could be the cause of the heightened sensitivity to chemical carcinogens. Methylation of DNA is essential to the regulation of expression of genetic information, and the undermethylation of DNA observed during choline deficiency (despite adequate dietary methionine) may be responsible for carcinogenesis. Choline-deficient rats experience increased lipid peroxidation in the liver. Lipid peroxides in the nucleus are a possible source of free radicals that could modify DNA and cause carcinogenesis. Choline deficiency activates protein kinase C signaling, usually involved in growth factor signaling in hepatocytes. Finally, a defect in cell suicide (apoptosis) mechanisms may contribute to the carcinogenesis of choline deficiency.

Kidney Function

Renal function is compromised by choline deficiency, which leads to abnormal concentrating ability, free-water reabsorption, sodium excretion, glomerular filtration rate, renal plasma flow, and gross renal hemorrhage. The deterioration in renal function may be related to changes in acetylcholine release by nerves that regulate blood flow to the kidney. Additionally, the renal glomerulus uses the choline-metabolite betaine as an osmolyte to assist cells in maintaining their volume in the presence of concentrated salts in urine.

Brain Development

During the fetal and neonatal period, the availability of choline to tissues fluctuates because of the varied dietary intake of choline among neonates and the slower oxidation of choline during the first weeks of life. However, ensured availability of this amine appears to be vital to infants because organ growth, which is extremely rapid in the neonate, requires large amounts of choline for membrane biosynthesis. Choline is also particularly important during the neonatal period because it appears to change brain function. There are two sensitive periods in rat brain development during which treatment with choline produces long-lasting enhancement of spatial memory that is lifelong and has been detected in elderly rats. The first occurs during embryonic days 12–17 and the second, during postnatal days 16–30. Choline supplementation during these critical periods elicits a major improvement in memory performance at all stages of training on a 12-arm radial maze.

The choline-induced spatial memory facilitation correlates with altered distribution and morphology of neurons involved in memory storage within the brain, with biochemical changes in the adult hippocampus and with electrophysiological changes in the adult hippocampus. It also correlates with changes in proliferation, apoptosis, and migration of neuronal precursor cells in the hippocampus during fetal brain development. When pregnant rats were treated with varying levels of dietary choline between day 12 and 18 of gestation, it was found that choline deficiency significantly decreased the rate of mitosis in the neuroepithelium of fetal brain adjacent to the hippocampus. An increased number of apoptotic cells were found in the region of the dentate gyrus of choline-deficient hippocampus compared to controls. Modulation of dietary choline availability changed the distribution and migration of precursor cells produced on embryonic

day 16 in the fimbria, primordial dentate gyrus, and Ammon's horn of the fetal hippocampus. Choline deficiency also decreased the migration of newly proliferating cells from the neuroepithelium into the lateral septum, thus indicating that the sensitivity of fetal brain to choline availability is not restricted to the hippocampus. The expression of TOAD-64 protein, an early neuronal differentiation marker, increased in the hippocampus of choline-deficient day E18 fetal brains compared to controls. These findings show that dietary choline availability during pregnancy alters the timing of mitosis, apoptosis, migration, and the early commitment to neuronal differentiation by progenitor cells in fetal brain hippocampus and septum, two regions known to be associated with learning and memory.

A disruption in choline uptake and metabolism during neurulation produces neural tube defects in mouse embryos grown *in vitro*. Exposing early somite staged mouse embryos *in vitro* with an inhibitor of choline uptake and metabolism, 2-dimethylaminoethanol (DMAE) causes craniofacial hypoplasia and open neural tube defects in the forebrain, midbrain, and hindbrain regions. Embryos exposed to an inhibitor of phosphatidylcholine synthesis, 1-*O*-octadecyl-2-*O*-methyl-rac-glycero-3-phosphocholine (ET-18-OCH_3) exhibit similar defects or expansion of the brain vesicles and a distended neural tube at the posterior neuropore as well as increased areas of cell death. Thus, choline like folic acid is important during neural tube closure.

Are these findings in rats likely to apply to humans? We do not know. Human and rat brains mature at different rates; rat brain is comparatively more mature at birth than is the human brain, but in humans synaptogenesis may continue for months after birth. Are we varying the availability of choline when we substitute infant formulas for human milk? Does choline intake in infancy contribute to variations in memory observed between humans? These are good questions that warrant additional research.

Brain Function in Adults

It is unlikely that choline acetyltransferase in brain is saturated with either of its substrates, so that choline (and possibly acetyl-CoA) availability determines the rate of acetylcholine synthesis. Under conditions of rapid neuronal firing acetylcholine release by brain neurons can be directly altered by dietary intake of choline. Based on this observation, choline has been used as a possible memory-improvement drug. In some patients with Alzheimer's disease, choline or phosphatidylcholine has beneficial effects, but this effect is variable. Both verbal and visual memory may be impaired in other patients who require long-term intravenous feeding and this may be improved with choline supplementation.

Measurement of Choline and Choline Esters

Radioisotopic, high-pressure liquid chromatography, and gas chromatography/isotope dilution mass spectrometry (GC/IDMS) methods are available for measurement of choline. However, these existing methods are cumbersome and time consuming, and none measures all of the compounds of choline derivatives. Recently, a new method has been established for quantifying choline, betaine, acetylcholine, glycerophosphocholine, cytidine diphosphocholine, phosphocholine, phosphatidylcholine, and sphingomyelin in liver, plasma, various foods, and brain using liquid chromatography/electrospray ionization-isotope dilution mass spectrometry (LC/ESI-IDMS).

Acknowledgments

Supported by the National Institutes of Health (DK55865, AG09525, DK56350).

See also: **Brain and Nervous System**. **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Folic Acid**. **Homocysteine**. **Liver Disorders**. **Vitamin B_6**.

Further Reading

Albright CD, Tsai AY, Friedrich CB, Mar MH, and Zeisel SH (1999) Choline availability alters embryonic development of the hippocampus and septum in the rat. *Brain Research. Developmental Brain Research* **113**: 13–20.

Buchman AL, Dubin M, Jenden D, Moukarzel A, Roch MH, Rice K, Gornbein J, Ament ME, and Eckhert CD (1992) Lecithin increases plasma free choline and decreases hepatic steatosis in long-term total parenteral nutrition patients. *Gastroenterology* **102**: 1363–1370.

da Costa KA, Badea M, Fischer LM, and Zeisel SH (2004) Elevated serum creatine phosphokinase in choline-deficient humans: mechanistic studies in C2C12 mouse myoblasts. *Am J Clin Nutr* **80**: 163–70.

Institute of Medicine, National Academy of Sciences USA (1998) *Dietary Reference Intakes for Folate, Thiamin, Riboflavin, Niacin, Vitamin B12, Panthothenic Acid, Biotin, and Choline*, vol. 1. Washington DC: National Academy Press.

Koc H, Mar MH, Ranasinghe A, Swenberg JA, and Zeisel SH (2002) Quantitation of choline and its metabolites in tissues and foods by liquid chromatography/electrospray ionization-isotope dilution mass spectrometry. *Analytical Chemistry* **74**: 4734–4740.

Meck WH and Williams CL (1997) Simultaneous temporal processing is sensitive to prenatal choline availability in mature and aged rats. *Neuroreport* **8**: 3045–3051.

Montoya DA, White AM, Williams CL, Blusztajn JK, Meck WH, and Swartzwelder HS (2000) Prenatal choline exposure alters hippocampal responsiveness to cholinergic stimulation in adulthood. *Brain Research. Developmental Brain Research* **123**: 25–32.

Niculescu MD and Zeisel SH (2002) Diet, methyl donors and DNA methylation: interactions between dietary folate, methionine and choline. *Journal of Nutrition* **132**: 2333S–2335S.

Zeisel SH and Blusztajn JK (1994) Choline and human nutrition. *Annual Review of Nutrition* **14**: 269–296.

Zeisel SH, daCosta K-A, Franklin PD, Alexander EA, Lamont JT, Sheard NF, and Beiser A (1991) Choline, an essential nutrient for humans. *FASEB Journal* **5**: 2093–2098.

Zeisel SH, Mar M-H, Howe JC, and Holden JM (2003) Concentrations of choline-containing compounds and betaine in common foods. *Journal of Nutrition* **133**: 1302–1307.

CHROMIUM

R A Anderson, US Department of Agriculture, Beltsville, MD, USA

Chromium (Cr) in the trivalent form is an essential nutrient that functions primarily in sugar and fat metabolism. Dietary intake of Cr by humans and farm animals is often suboptimal. Insufficient dietary intake of Cr is associated with increased risk factors associated with type 2 diabetes mellitus (DM) and cardiovascular diseases. Chromium functions in glucose and insulin metabolism primarily via its role in the improvement of insulin activity. Improved insulin function is also associated with an improved lipid profile. People with type 2 diabetes have a more than twofold increased incidence of cardiovascular diseases compared to control subjects.

Chromium in foods and dietary supplements is trivalent, whereas Cr often found in paints, welding fumes, and other industrial settings is hexavalent and is severalfold more toxic than the trivalent nutritional Cr. Trivalent Cr is one of the safest nutrient supplements based on the ratio of the amount that is needed relative to the amount that can be consumed over a lifetime with no adverse effects. An expert panel of the US Food and Nutrition Board was unable to set an upper level of safe intake since none of the levels of intake tested showed any signs of toxicity. Toxicity is also alleviated by the low level of absorption, usually less than 2%.

Essentiality and Metabolic Functions of Chromium

The essentiality of trivalent Cr in human nutrition was documented in 1977 when diabetic signs and symptoms of a patient on total parenteral nutrition (TPN) were reversed by supplemental Cr. Diabetic symptoms, including elevated blood glucose, weight loss, impaired nerve conduction, brain disorders, and abnormal respiratory quotient, that were refractory to exogenous insulin were reversed following increased intake of the essential nutrient Cr. Upon daily addition of supplemental Cr to the patient's TPN solution for 2 weeks, diabetic symptoms were alleviated and exogenous insulin requirement declined from 45 units per day to zero. These findings have been repeated and documented in the scientific literature on several occasions.

Signs and symptoms of Cr deficiency listed in **Table 1** are not limited to subjects on TPN. Improvements in glucose and/or lipid concentrations have been reported in children with protein calorie malnutrition; the elderly; people with type 1 and type 2 DM, hypoglycemia, and marginally impaired glucose tolerance; and numerous animal species.

The hallmark sign of marginal Cr deficiency is impaired glucose tolerance. The effects of Cr on people with high, low, and normal glucose tolerance as well as diabetes are illustrated in **Figure 1**. Chromium leads to a decrease in blood glucose in people with elevated blood sugar and an increase in those with low blood sugar due to its role in normalizing insulin. In the presence of Cr in a physiologically active form, insulin is more efficient and much lower levels of insulin are required. During periods of elevated blood glucose, more efficient insulin leads to a decrease in blood glucose. In people with low blood sugar, reactive hypoglycemia, more efficient insulin leads to a rapid rise in response to a glucose challenge and a more rapid return to baseline values. This leads to less of a decline in

Table 1 Signs and symptoms of Cr deficiency

Function	*Animals*
Impaired glucose tolerance	Human, rat, mouse, squirrel monkey, guinea pig, cattle
Elevated circulating insulin	Human, rat, pig, cattle
Glycosuria	Human, rat
Fasting hyperglycemia	Human, rat, mouse
Impaired growth	Human, rat, mouse, turkey
Hypoglycemia	Human
Elevated serum cholesterol and triglycerides	Human, rat, mouse, cattle, pig
Increased incidence of aortic plaques	Rabbit, rat, mouse
Increased aortic intimal plaque area	Rabbit
Nerve disorders	Human
Brain disorders	Human
Corneal lesions	Rat, squirrel monkey
Ocular eye pressure	Human
Decreased fertility and sperm count	Rat
Decreased longevity	Rat, mouse
Decreased insulin binding	Human
Decreased insulin receptor number	Human
Decreased lean body mass	Human, pig, rat
Elevated percent body fat	Human, pig
Impaired humoral immune response	Cattle
Increased morbidity	Cattle
Gestational diabetes	Human
Steroid-induced diabetes	Human
Atypical depression	Human

Adapted from Anderson RA (1998) Chromium, glucose intolerance and diabetes. *Journal of the American College of Nutrition* **17**: 548–555.

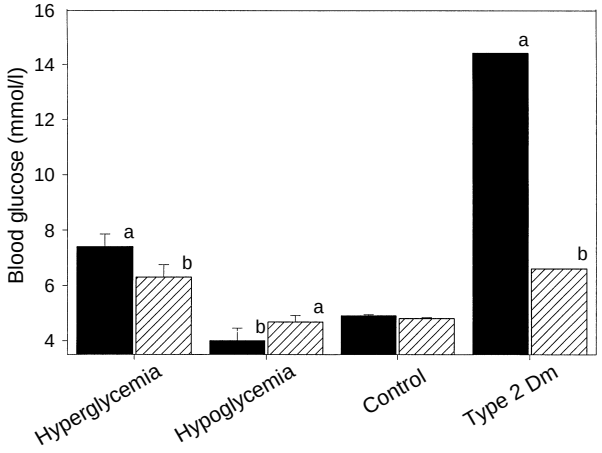


Figure 1 Response to supplemental Cr of people with hyperglycemia, hypoglycemia, optimal glycemia (control), and type 2 diabetes mellitus (DM). The minimal amount of Cr usually showing beneficial effects in people with high or low blood sugar is 200 μg per day. People with diabetes require 400–600 μg per day or more. Bars with different superscripts denote differences at $p < 0.05$.

hypoglycemic glucose values. Supplemental Cr also leads to increased insulin binding and increased insulin receptor number, and evidence suggests that Cr may be involved in the phosphorylation–dephosphorylation of the insulin receptor proteins. Chromium activates insulin receptor kinase, the enzyme that phosphorylates the insulin receptor, leading to activation of insulin function, and it appears to inhibit the phosphatase enzyme that deactivates insulin function.

Recent advances in Cr nutrition research include the demonstration of an inverse relationship between toenail Cr and cardiovascular disease (CVD) in studies from the United States and Europe, supporting studies indicating that people with CVD tend to have lower levels of serum and tissue Cr and also substantiating the beneficial effects of supplemental Cr on blood cholesterol, triglycerides, and high-density lipoprotein cholesterol. Supplemental Cr as chromium picolinate (the most common form of supplemental Cr) was shown to be effective in the treatment of depression. Preliminary studies suggest that the effects of Cr are greater than those of any drugs used in the treatment of atypical depression. Supplemental Cr is also free of side effects associated with drugs, which are often quite serious in the treatment of depression. Studies also show that Cr is beneficial in the reversal of polycystic ovarian syndrome, gestational diabetes, and steroid-induced associated with administration of steroids such as prednisone given as antiinflammatory agents in the treatment of arthritis, asthma, allergies, and related diseases.

Response to Cr is due to not only the Cr status of the subjects but also the forms and amount of Cr consumed. Subjects with diabetes or glucose intolerance who consume 200 μg daily of supplemental Cr or less often do not respond to supplemental Cr, but they may respond to 400–600 μg daily or more. A dose response to Cr for subjects with type 2 DM is shown in **Figure 2**. Subjects had been diagnosed with diabetes for approximately 5 years and had taken no Cr supplements. There was a progressive decline in the hemoglobin A1c after 2 and 4 months of consuming 200 or 1000 μg daily of Cr as Cr picolinate, respectively. There were also dose-dependent improvements in glucose, insulin, and cholesterol. These results were confirmed in a separate double-blind, placebo-controlled study.

The responses to Cr are difficult to predict and the phenotypic characteristics of the individual may be important. Phenotype is also important in insulin signaling and may explain, in part, the wide range

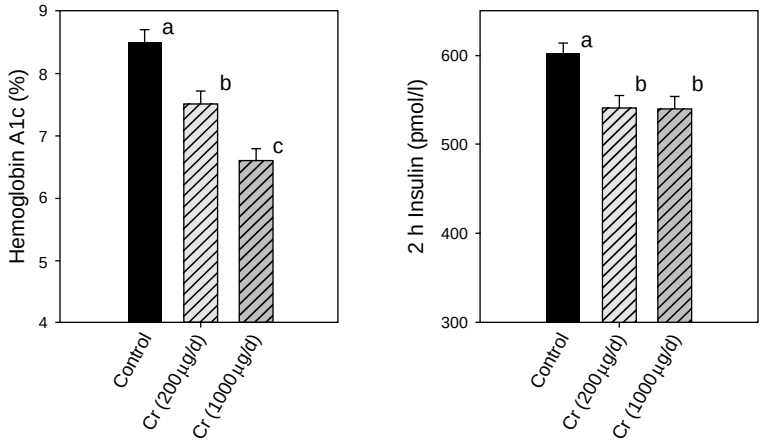


Figure 2 Chromium effects on hemoglobin A1C and 2-h insulin. The study involved chromium supplementation (200 or 1000 μg/day) for 4 months in 180 people with type 2 diabetes mellitus. Bars with different superscripts denote differences at $p < 0.05$. (Adapted from Anderson RA, Cheng N, Bryden NA *et al.* (1997) Elevated intakes of supplemental chromium improve glucose and insulin variables in individuals with type 2 diabetes. *Diabetes* **46**: 1786–1791.)

of individual responses to Cr supplementation. Data demonstrate that Cr has beneficial effects in insulin-resistant obese but not lean JCR:LA corpulent rats, which are used as a model for insulin resistance. Although Cr had no effect on the weight of lean and obese rats, obese rats consuming supplemental Cr displayed lower insulin, improved glucose control, and increased phosphoinositol-3-kinase activity. This work documents a specific effect of Cr in a key control site in the insulin signaling pathway, which is the system responsible for the overall control of sugar, fat, and energy metabolism.

Chromium and Stress

Stresses that have been shown to alter Cr metabolism in humans are glucose loading, high simple sugar diets, lactation, infection, acute exercise, chronic exercise, and physical trauma. Urinary losses can be used as a measure of the response to stress since once Cr is mobilized in response to stress it is not reabsorbed by the kidney but is lost in the urine. The degree of stress as measured by the stress hormone, cortisol, is correlated with the amount of Cr lost in the urine.

The administration of glucocorticoids also leads to increased urinary Cr losses as well as insulin resistance. These steroids are often employed as antiinflammatory agents in the treatment of common chronic diseases, such as asthma, allergies, and arthritis, and they are also administered following organ transplantation, but a side effect of glucocorticoid administration is steroid-induced diabetes. The mechanisms responsible for steroid-induced diabetes are unknown, but decreased insulin sensitivity is an overlying cause. Impaired Cr metabolism also appears to be a related cause since supplementation of 50 people who had uncontrolled steroid-induced diabetes with Cr for 10 days resulted in the reversal of the steroid-induced diabetes in 47 of these patients, with no adverse side effects, and a decrease of at least 50% of the medication needed prior to the supplementation of Cr.

Dietary Intake and Requirements of Chromium

A panel on micronutrients convened by the Institute of Medicine defined an adequate intake for Cr of 25 μg for women and 35 μg for men 19–50 years old and 20 μg for women and 30 μg for men older than age 50 years. Adequate intake is "the recommended average daily intake based on observed or experimentally determined approximations or estimates of nutrient intake by a group (or groups) of apparently healthy people that are assumed to be adequate—used when an RDA cannot be determined." The adequate intake for Cr is nearly identical to the average Cr intake but much lower than earlier committee recommendations.

It is unclear why the adequate intake for Cr should be lower for people older than 50 years when the primary function of Cr is to combat problems associated with insulin and glucose metabolism, which tend to increase with age. Indices of Cr status, such as the Cr content of hair, sweat, and

urine, were shown to decrease with age in a study of more than 40 000 people. The recommended intakes in France are higher and more in line with studies demonstrating that a large segment of the population may not be consuming adequate Cr. The French Conseil National d'Etudes et de Recherche sur la nutrition et l'Alimentation has proposed daily intakes of 55 μg for adult French women 19–65 years old and 60 μg/day for those older than 65 years and 65 and 70 μg, respectively, for men. Since Cr losses are increased by high intake of simple sugars such as glucose, sucrose, and fructose, modern diets high in these sugars appear to be increasing the requirements for Cr.

More than 34 studies have reported beneficial effects of supplemental Cr for people with blood glucose values ranging from hypoglycemia to diabetes when consuming diets of average Cr content. In a controlled diet study, consumption of diets in the lowest quartile of normal Cr intakes, but near the new adequate intakes, led to detrimental effects on glucose and insulin in subjects with marginally impaired glucose tolerance (90-minute glucose between 5.6 and 11.1 mmol/l or 100–200 mg/dl) following an oral glucose load of 1 g/kg body weight. The average person older than 25 years of age has blood glucose in this range. Consumption of these same diets by people with good glucose tolerance (90-minute glucose less than 5.6 mmol/l) did not lead to changes in glucose and insulin variables. This is consistent with previous studies demonstrating that the requirement for Cr is related to the degree of glucose intolerance and demonstrates that an intake of 20 μg Cr per day is not adequate for people with decreased insulin sensitivity, such as those with marginally impaired glucose tolerance, and certainly not for those with impaired glucose tolerance or diabetes.

Absorption, Transport, Storage, and Excretion

Absorbed Cr is excreted primarily in the urine, and only small amounts of Cr are lost in the hair, perspiration, and bile. Therefore, urinary Cr excretion can be used as an accurate estimation of absorbed Cr. At normal dietary Cr intakes (10–40 μg per day), Cr absorption is inversely related to dietary intake. Chromium absorption is approximately 0.5% at a daily intake of 40 μg and increases to 2% when intake decreases to 10 μg. Therefore, the amount of absorbed Cr over this range is approximately 0.2 μg and is reflected in the urinary Cr losses of approximately 0.2 μg/day. This inverse relationship of Cr intake and absorption appears to be a basal control mechanism to maintain a minimal level of absorbed Cr. Intakes higher than 40 μg result in corresponding increases in total Cr absorbed. There is no direct evidence that Cr absorption involves active transport.

Chromium absorption in young and old normal subjects is similar, but people with type 1 DM absorb two- to fourfold more Cr than other groups tested. People with diabetes appear to have an impaired ability to convert inorganic Cr to a useable form. Diabetic mice also lose the ability to convert Cr to a useable form. People with diabetes require additional Cr and the body responds with increased absorption, but the absorbed Cr cannot be utilized effectively and is excreted in the urine. Chromium content in tissues of people with diabetes is also lower.

Chromium absorption and incorporation into tissues are also dependent on the form of Cr ingested. An accurate estimation of Cr absorption and utilization in animal studies can be achieved by measuring Cr incorporation into tissues. The tissue with the greatest Cr concentration is the kidney, followed by the spleen, liver, lungs, heart, and skeletal muscle.

Tissue Cr is an accurate method to assess Cr absorption and utilization and is also a measure of Cr storage. The kidney, which is one of the primary sites of tissue Cr storage, is also one of the best sources of insulin potentiating forms of Cr. Chromium is transported to the tissues primarily bound to transferrin, the same protein that transports iron. There are two metal binding sites on transferrin—one primarily for iron and a second involved in Cr transport. During conditions of high iron excess or iron overload, such as in iron storage diseases (hemosiderosis and hemochromatosis), all the metal transport sites on transferrin are occupied by iron. This may explain the high incidence of diabetes in hemochromatosis patients that may be due in part to Cr deficiency.

Dietary Sources

Dietary Cr content of foods varies widely, and there are no comprehensive databases to calculate dietary Cr intake. Chromium content of foods is often erroneously high due to Cr contamination during collection and analyses. For example, stainless-steel blender blades are often used in the homogenization of foods, but stainless steel is approximately 18% Cr. In the presence of acidic foods, more Cr may leach from the blender blades than is present originally in the foods.

Table 2 Daily intakes of chromium from various food groups

Food group	*Average daily intake (μg/day)*	*Comment*
Cereal products	3.7	55% from wheat
Meat	5.2	55% from pork; 25% from beef
Fish and seafoods	0.6	
Fruits, vegetables, nuts, and mushrooms	6.8	70% from fruits and berries
Dairy products, eggs, and margarine	6.2	85% from milk
Beverages, confectionaries, sugar, and condiments	6.6	45% from beer, wine, and soft drinks
Total	29.1	

From Anderson RA (1988) Chromium In: Smith K (ed.) *Trace Minerals in Foods*, pp. 231–247. New York: Marcel Dekker.

Chromium is present in several food groups but at low levels. The distribution is similar among fruits, vegetables, dairy products, beverages, and meats, with lesser amounts from cereal products and small amounts from fish and seafood (**Table 2**). Chromium content of foods is a combination of the endogenous Cr present in the foods and the Cr introduced during the various stages of growing and processing. For example, fruit juices are often high in Cr since Cr may leach from containers during processing and storage under acidic conditions.

Safety of Chromium

Trivalent Cr, the form of Cr found in foods and nutrient supplements, is one of the least toxic nutrients. The reference dose established by an expert panel of the US Environmental Protection Agency (EPA) is 350 times the upper limit of the estimated safe and adequate daily dietary intake. The newer adequate intakes are lower and would have a safety ratio of approximately 2000. The reference dose is defined as "an estimate (with uncertainty spanning perhaps an order of magnitude) of a daily exposure to the human population, including sensitive subgroups, that is likely to be without an appreciable risk of deleterious effects over a lifetime." This conservative estimate of safe intake has a much larger safety factor for trivalent Cr than almost any other nutrient. The ratio of the EPA reference dose to the requirements is approximately two to five for other trace elements, such as zinc and manganese, and five to seven for selenium. Chromium in the form of both Cr chloride and Cr picolinate fed to rats at several thousand times the current adequate intake (based on body weight) resulted in no detectable signs of toxicity.

Conclusion

Dietary intake of Cr may be suboptimal for most humans and farm animals. Increased intake of trivalent Cr often leads to improved glucose and lipid metabolism. The physiological role of Cr appears to be primarily through the improved function of insulin. Increased intake of Cr leads to increased insulin binding, increased insulin receptor number, and increased phosphorylation of the insulin receptor proteins, leading to increased insulin sensitivity and function and lower glucose and lipids. Chromium is a nutrient and not a therapeutic agent, and only subjects whose impaired glucose and insulin function is related to suboptimal intake of Cr will benefit from additional Cr. Although a significant number of subjects often respond to supplemental Cr, there are also a significant number of subjects who do not respond to improved Cr nutrition. This is likely due to the amount and form of Cr consumed and glucose tolerance and Cr status of the subjects. No negative effects of supplemental Cr have been reported in any of the Cr supplementation studies involving daily Cr intakes of up to 1000 μg per day.

See also: **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Glucose**: Chemistry and Dietary Sources. **Lipoproteins**.

Further Reading

Althuis MD, Jordan NE, Ludington EA, and Wittes JT (2002) Glucose and insulin responses to dietary chromium supplements: A meta-analysis. *American Journal of Clinical Nutrition* **76**: 148–155.

Anderson RA (1994) Stress effects on chromium nutrition of humans and farm animals. In: Lyons TP and Jacques KA (eds.) *Proceedings of Alltech's Tenth Symposium on Biotechnology in the Feed Industry*, pp. 267–274. Nottingham, UK: University Press.

Anderson RA (1998a) Chromium, glucose intolerance and diabetes. *Journal of the American College of Nutrition* **17**: 548–555.

Anderson RA (1998b) Effects of chromium on body composition and weight loss. *Nutrition Review* **56**: 266–270.

Anderson RA (2000) Chromium in the prevention and control of diabetes. *Diabetes and Metabolism* **26**: 22–27.

Anderson RA (2003) Chromium and insulin resistance. *Nutrition Research Review* **16**: 267–275.

Anderson RA, Cheng N, Bryden NA *et al.* (1997) Elevated intakes of supplemental chromium improve glucose and insulin variables in individuals with type 2 diabetes. *Diabetes* **46**: 1786–1791.

Anonymous (2001) *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron,*

Manganese, Molybdenum, Nickel, Silicon, Vanadium and Zinc, pp. 197–223. Washington, DC: National Academy Press.
Cefalu WT, Wang ZQ, Zhang XH, Baldor LC, and Russell JC (2002) Oral chromium picolinate improves carbohydrate and lipid metabolism and enhances skeletal muscle Glut-4 translocation in obese, hyperinsulinemic (JCR-LA corpulent) rats. *Journal of Nutrition* **132**: 1107–1114.
Gunton JE, Hams G, Hitchman R, and McElduff A (2001) Serum chromium does not predict glucose tolerance in late pregnancy. *American Journal of Clinical Nutrition* **73**: 99–104.
Heimbach JT and Anderson RA (2005) Chromium: Recent studies regarding nutritional roles and safety. *Nutrition Today*.
Morris BW (1999) Chromium action and glucose homeostasis. *Journal of Trace Element in Experimental Medicine* **12**: 61–70.
Ravina A, Slezak L, Mirsky N, and Anderson RA (1999) Control of steroid-induced diabetes with supplemental chromium. *Journal of Trace Elements in Experimental Medicine* **12**: 375–378.
Vincent JB (2000) The biochemistry of chromium. *Journal of Nutrition* **130**: 715–718.

COBALAMINS

R Green, University of California, Davis, CA, USA

Introduction

The cobalamins are a group of closely related and interconvertible compounds with a complex structure that are collectively known by the common name of vitamin B_{12}. Recommended biochemical nomenclature restricts the term 'vitamin B_{12}' for the particular form of cobalamin known as cyanocobalamin. All cobalamins belong to the broader family of corrinoids, which share the characteristic of consisting of a planar four-member pyrrole ring (corrin ring) containing a central cobalt atom. Cobalamins are distinguished from other corrinoids by possessing both alpha (lower) and beta (upper) axial ligands that are attached to the central cobalt atom (**Figure 1**). The lower ligand consists of a base (5,6-dimethylbenzimidazole) attached to a sugar (ribose), which in turn is attached to a phosphate and an amino-propyl group that ultimately is tethered back to the corrin ring. In the naturally occurring cobalamins the upper ligand is variably a cyano-, hydroxo-, aquo-, methyl-, or adenosyl- group, giving rise to the correspondingly named chemical forms of the vitamin. Of these, methylcobalamin and deoxyadenosylcobalamin are the forms that function as coenzymes for metabolic reactions. These are sensitive to destruction by light. Cyanocobalamin is a stable form and is therefore used in therapeutic preparations. Hydroxo- or

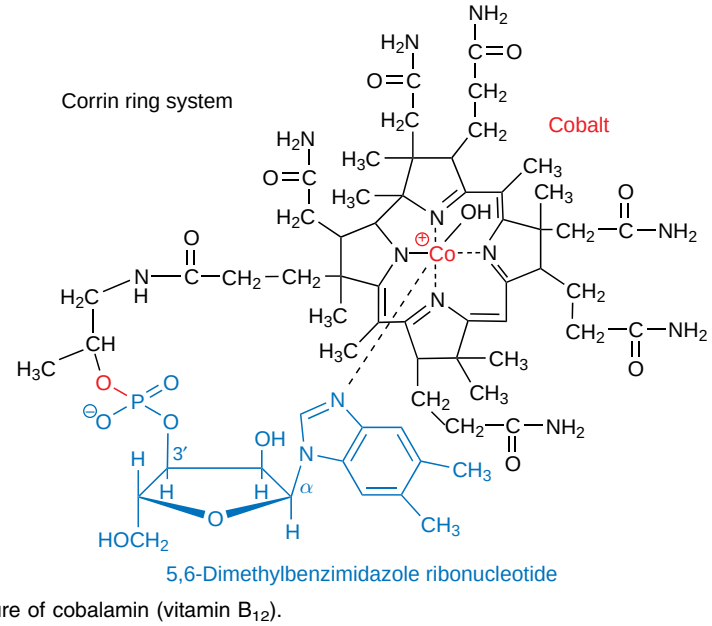


Figure 1 Chemical structure of cobalamin (vitamin B_{12}).

aquocobalamin are intermediates formed during the synthesis of the coenzyme forms. Other forms including sulfito-, nitrite-, and glutathionyl- derivatives of cobalamin have also been described but their role in metabolism is not known.

Biochemistry and Metabolic Functions

Only two reactions in humans and other animals are known to require cobalamin (**Figure 2**). One is isomerization of methylmalonyl coenzyme A (CoA), which requires deoxyadenosylcobalamin, is catalyzed by the enzyme methylmalonyl CoA mutase, and is mitochondrial. The other reaction is the transmethylation of homocysteine by 5-methyl-tetrahydrofolate to methionine, catalyzed by the enzyme methionine synthase (N^5-methyltetrahydrofolate:homocysteine methyltransferase) which requires methylcobalamin as coenzyme and is located in the cytosol. It is through their essential roles in this important metabolic reaction that cobalamin and folate interact and are linked with respect to their importance in nutrition. In addition, there are major similarities in the effects of their deficiencies in humans. These will be discussed below. Considering this 'metabolic crossroad' for the two vitamins, it may be pointed out that without adequate supplies of both nutrients, the synthesis of methionine and its derivative *S*-adenosylmethionine (SAM) is disrupted, with consequent profound effects on normal cellular function. Methionine is a key and essential amino acid and normal supply depends critically on recycling through the remethylation pathway (**Figure 3**). Moreover, SAM is the universal methyl donor, essential for over 100 transmethylation reactions involving amino acid, nucleotide, neurotransmitter, and phospholipid metabolism as well as detoxification reactions.

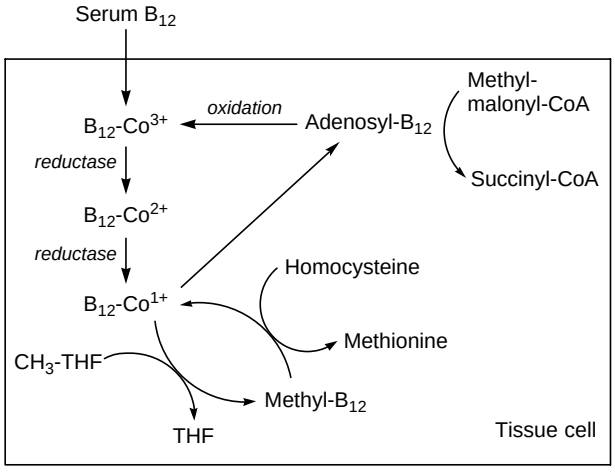


Figure 2 Reactions in humans and other animals known to require cobalamin.

Apart from methionine the other product of the methionine synthase reaction, which is almost completely irreversible, is tetrahydrofolate (THF); this constitutes the first step by which folate enters bone marrow and other cells from plasma, for its conversion into the various intracellular forms of reduced folate containing a series of one-carbon substituents (see 00119 and **Figure 3**). The active forms of these

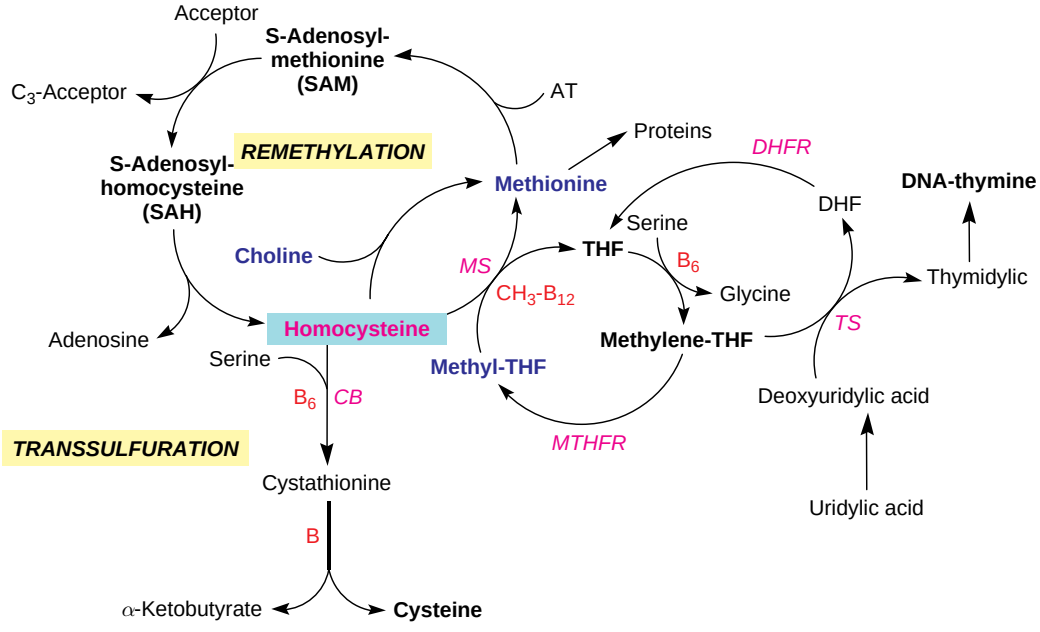


Figure 3 The remethylation pathway.

folate congeners are all polyglutamated by an enzyme, folate polyglutamate synthetase, that cannot use methyl-THF as substrate. Tetrahydrofolate is the obligate substrate for polyglutamate addition. Consequently, when the methionine synthase reaction is blocked as a result of cobalamin deficiency, there is 'THF starvation.' Methyl-THF accumulates in the plasma, while intracellular folate concentrations fall due to failure of formation of the critical intracellular folate polyglutamates because of 'methyl-folate trapping.' This theory explains the abnormalities of folate metabolism that occur in cobalamin deficiency (high concentrations of serum folate, low red cell folate) and also why the anemia that occurs in cobalamin deficiency will temporarily or partially respond to folic acid in large doses. The explanation of why the serum cobalamin falls in folate deficiency may also be related to impairment of the methionine synthase reaction resulting in reduced formation of methylcobalamin, the predominant circulating form of cobalamin in plasma.

Physiology

The recommended daily allowance (RDA) for cobalamin in adults proposed by the Food and Nutrition Board of the National Academy of Sciences/National Research Council in 1997 is 2.4 μg. Cobalamins do not occur in plants but are synthesized by certain bacteria, fungi, and algae, which constitute the ultimate source of all cobalamin found in nature. Cobalamins enter the food chain through herbivorous animals that harbor cobalamin-producing microorganisms in their upper gastrointestinal tract (e.g., the 'first stomach' of ruminants). Consumption of the meat or products of these animals supplies cobalamin in the diet for other animals. Dietary sources of cobalamin in humans are restricted to meat, poultry, fish, shellfish, eggs, and dairy products. Cobalamin is resistant to destruction by cooking, unlike the heat labile folates. On account of the exceedingly small daily requirement for cobalamin, in the order of 2 to 3 μg, and the relatively large body store of the vitamin (3000–5000 μg in developed countries) complete absence of intake or absorption of cobalamin is preceded by a long lag of up to 3–5 years before depletion of body cobalamin stores reaches a critical point that results in the manifestations of cobalamin deficiency. This is not the case in developing countries where the onset of depletion may be much more rapid because of initially lower stores. Still, complete lack of dietary intake of cobalamin is somewhat rare and occurs only in strict vegans who shun all animal foods, including dairy products and eggs.

Cobalamin absorption is a complex mechanism that consists of several steps and involves several protein chaperones and receptors, defects of which can result in reduced or absent uptake of dietary cobalamin. The related and continuous processes of ingestion, digestion, and absorption of cobalamin comprising the assimilation of the vitamin are arbitrarily divided into six steps. During the first step of mastication and swallowing of food, dietary cobalamin becomes mixed with a binding protein derived from saliva belonging to the family of cobalamin-binding proteins known as haptocorrins. Cobalamin in foods is generally complexed to proteins that must first be digested to release the bioavailable vitamin. In the second step release of cobalamin takes place largely in the stomach, under the influence of gastric hydrochloric acid and proteolytic digestion by pepsin. It is during this process and in the acid environment of the stomach that salivary haptocorrin preferentially binds and protects food cobalamin. Another specific cobalamin-binding protein, known as intrinsic factor, is secreted by the parietal cells of the stomach, but is unable to bind the cobalamin still tightly complexed to haptocorrin. During the third step, which occurs in the duodenum, cobalamin is released from its complex with haptocorrin through the combined effects of pancreatic bicarbonate, which neutralizes the gastric acid, and the proteolytic action of the enzymes trypsin and chymotrypsin that digest haptocorrin and thus enable the binding of the free cobalamin by gastric intrinsic factor. In the fourth step, the intrinsic factor–cobalamin complex, having traversed the full length of the small intestine, arrives at the luminal surface of the terminal ileum. There, it comes in contact with specialized receptors. In the presence of calcium, the complex attaches to the receptor consisting of two distinct proteins, cubulin and a recently identified protein known as amnionless that is necessary to complete the assimilation process. Both proteins are essential for the internalization of the intrinsic factor–vitamin B_{12} complex through a process known as receptor-mediated endocytosis. Through this process, cobalamin together with intrinsic factor and escorted by the receptor are taken into lysosomes where the intrinsic factor–cobalamin complex is released and intrinsic factor is degraded through the action of acid hydrolysis of peptidases. The final fifth step is poorly understood but involves first the release from lysosomes and then the metabolism of cobalamin to its methyl and deoxyadenosyl derivatives. It is primarily in the form of methylcobalamin that the vitamin finally enters the plasma. The assimilation of food vitamin

B_{12} is a lengthy process, as evidenced by the 6–8 h taken for orally administered cobalamin to first appear in the plasma and several additional hours for the process to be completed.

When cobalamin enters the plasma, it is bound to the cobalamin-binding protein transcobalamin (previously known as transcobalamin II to distinguish it from transcobalamins I and III, which, together with the salivary cobalamin-binding protein and other binders present in secretions are now referred to collectively as the haptocorrins). The properties of transcobalamin and the haptocorrins are summarized in **Table 1**. The fraction of cobalamin bound to transcobalamin accounts for only 20–30% of the total plasma cobalamin. The major residual fraction of the plasma cobalamin is attached to haptocorrins. The function of haptocorrins is not known, but rapidly proliferating cells including bone marrow precursors can obtain cobalamin only from transcobalamin. Consequently, the critical fraction of the serum cobalamin is the transcobalamin-bound portion, known as holo-transcobalamin. Conditions that alter the amount or distribution of cobalamin on these binding proteins can critically affect its delivery and transport. Therefore, conditions that lead to an increase in haptocorrins, such as chronic granulocytic leukemia (haptocorrins are produced in granulocytes), can give rise to an apparently normal serum cobalamin level even in patients who have pernicious anemia or some other cause of cobalamin deficiency. Conversely, a decrease in holo-transcobalamin can result in cobalamin deficiency even if the serum cobalamin level is apparently normal. This occurs in infants and children affected by congenital transcobalamin deficiency, which is associated with severe megaloblastic anemia. Levels of transcobalamin, which is produced mainly in endothelial cells, may be affected by a number of factors. Lowering of holo-transcobalamin can result in tissue cobalamin deficiency with a normal serum cobalamin level. Although holo-transcobalamin has been measured, and appears to correlate well with serum metabolite levels and other parameters of cobalamin deficiency, routine assays sufficiently sensitive to measure the small fraction of cobalamin that occurs as holo-transcobalamin have only recently become available.

Table 1 Properties of human plasma cobalamin binding proteins

	Haptocorrins (transcobalamin I + III)	*Transcobalamin II*
Source	Granulocytes	Endothelial cells
Transport functions	Storage, excretion of vitamin B_{12} analogs, antimicrobial	Cellular vitamin B_{12} uptake
Binding specificity	Low specificity, binds vitamin B_{12} analogs	Binds vitamin B_{12} with higher specificity
Membrane receptors	Nonspecific asialoglycoprotein receptors on hepatocytes	Specific receptors on most cells
Saturation	High (mainly 'holo')	Low (mainly 'apo')
Fraction of total vitamin B_{12}	80–90%	10–30%
Plasma clearance	Slow ($t_{1/2} \approx 10$ days)	Rapid ($t_{1/2} \approx 6$ min)
Molecular weight (Daltons)	60 000	38 000–45 000

Causes and Effects of Cobalamin Deficiency and Mechanisms

There are a number of causes of cobalamin deficiency that range in severity and frequency of occurrence. These are summarized in **Table 2**. In general, causes of cobalamin deficiency can be divided into those caused by absent or markedly reduced dietary intake and chemical inactivation (both rare) and those caused by malabsorption, either gastric or ileal. The most frequent cause of clinically important cobalamin deficiency is malabsorption of the vitamin and, in particular, pernicious anemia, caused by an autoimmune

Table 2 Causes of cobalamin deficiency

1. Dietary
   Veganism
2. Gastric
   - Atrophic gastritis and food cobalamin malabsorption
   - Autoimmune gastritis/gastric atrophy (classical pernicious anemia)
   - Extensive gastric disease or resection
3. Ileal
   - Extensive ileal disease (Crohn's, inflammatory bowel disease, tuberculous enteritis) or resection for these diseases
   - Luminal disturbances (chronic pancreatic disease, gastrinoma) and parasites (giardiasis, bacterial overgrowth, fish tapeworm)
4. Chemical/drug
   - Nitrous oxide
   - PAS, oral antidiabetic agents, colchicine
5. Congenital/inherited
   - Intrinsic factor deficiency/defect ('juvenile' pernicious anemia)
   - Intrinsic factor receptor deficiency/defect (Immerslund-Grasbeck syndrome)
   - Transcobalamin II deficiency
   - Cobalamin mutants (C-G)

destruction of the gastric mucosa with consequent failure of intrinsic factor production. Another more common cause of malabsorption of cobalamin is so-called food cobalamin malabsorption, caused by atrophy of the stomach lining that often occurs with increasing age. The mechanism of malabsorption in this condition is caused by a failure of complete food digestion and hence release of cobalamin for binding to intrinsic factor rather than a failure of intrinsic factor production itself. Food cobalamin malabsorption *per se* does not appear to result in clinically important cobalamin deficiency but may set the stage for greater susceptibility to the onset of deficiency when other conditions completely abrogate cobalamin absorption.

In all situations resulting from impairment of cobalamin absorption, the time to onset of deficiency depends on several factors, including the size of the body store, the extent of impairment of absorption (partial or complete), and, in diseases like pernicious anemia and others affecting all of the intestine, the rate of progression of the disease. In general, however, cobalamin deficiency resulting from malabsorption develops sooner (2–5 years) than is the case in the dietary deficiency encountered among vegans (10–20 years). This difference may be explained by the existence of a considerable enterohepatic recirculation of cobalalmin. Biliary cobalamin is efficiently reabsorbed in vegans compared with patients who have pernicious anemia or other forms of malabsorption.

Deficiency of cobalamin, when severe, affects all rapidly growing (DNA synthesizing) tissues, particularly in the bone marrow, with resulting production of larger abnormal cells with nuclei that bear evidence of impaired maturation. These marrow precursor cells then give rise to reduced numbers of blood cell progeny that are, in turn, abnormally large and show the stigmata of disrupted development. Resulting macrocytic anemia together with neutrophil leucocytes that have, on average, more nuclear lobes than normal (hypersegmented neutrophil leucocytes) are the usual telltale signs of megaloblastic anemia seen in cobalamin, as well as folate, deficiency. After the marrow, the next most affected tissues are the epithelial cell surfaces of the mouth, stomach, and the small intestine. Affected cells are also large, with increased numbers of multinucleated and dying cells. The gonads are also affected and infertility is common in patients with deficiency. Cobalamin deficiency may also be associated with skin hyperpigmentation and has been described in association with reduced bone-derived alkaline phosphatase and osteocalcin in the plasma.

Cobalamin deficiency may also cause nervous system complications including bilateral peripheral neuropathy or degeneration (demyelination) of the posterior and pyramidal tracts of the spinal cord, and, less frequently, atrophy of the optic nerve or cerebral symptoms. Cobalamin-deficient patients typically display sensory disturbances or paraesthesiae, muscle weakness, difficulty in walking, and sometimes dementia, psychotic disturbances, or visual impairment. Long-term nutritional cobalamin deficiency in infancy leads to poor brain development and impaired intellectual development. The effects of cobalamin deficiency on the blood and on the nervous system may occur separately or in combination and their severity is often inversely rather than directly correlated. The biochemical basis for cobalamin neuropathy, however, remains obscure. Its occurrence in the absence of methylmalonic aciduria in transcobalamin II deficiency, and in monkeys given the anesthetic agent nitrous oxide, suggests that the neuropathy is related to a defect in homocysteine–methionine conversion. Accumulation of *S*-adenosylhomocysteine in the brain resulting in inhibition of transmethylation reactions has been suggested.

Psychiatric disturbance is common in cobalamin deficiencies. Like the neuropathy, this has been attributed to a failure of the synthesis of SAM, due to reduced conversion of homocysteine to methionine. SAM is needed in the methylation of biogenic amines (e.g., dopamine), as well as of proteins, phospholipids, and neurotransmitters in the brain. A reduced ratio of SAM to *S*-adenosyl-homocysteine is postulated to result in reduced methylation. In cobalamin deficiency there is an intriguing inverse correlation between the degree of anemia, on the one hand, and the severity of the myeloneuropathy, on the other hand.

Diagnosis of Cobalamin Deficiency

Cobalamin deficiency is suspected in individuals who display the typical manifestations of deficiency of the vitamin as described in the section above on the effects of deficiency. In addition to the symptoms that may be experienced by individuals that are related to anemia (easy fatigue, shortness of breath, palpitations) and neuropathy (sensory and motor disturbances and memory loss) there are features that may be detected by a physician, including skin pallor (from anemia), abnormalities in neurological examination (sensory loss, abnormal balance and reflexes, mental changes), and epithelial changes (skin pigmentation, smooth tongue). On the basis of any combination of such changes, cobalamin

deficiency may be suspected but confirmation is necessary using laboratory tests because other conditions may give rise to effects that closely resemble cobalamin deficiency. This need to confim suspected cobalamin deficiency applies also in individuals who have abnormalities in their blood count results with anemia and macrocytosis (larger than normal red blood cells).

The standard screening test for cobalamin deficiency consists of direct measurement of circulating levels of cobalamin. Serum levels less than $150\,\text{pmol}\,\text{l}^{-1}$ are considered deficient and $150–250\,\text{pmol}\,\text{l}^{-1}$ are considered borderline. Serum or plasma cobalamin concentration can be measured in several ways and this has evolved from early microbial growth assays through competitive binding assays that were first radioisotopic and are now enzyme-linked or based on chemiluminescence detection. The sensitivity and specificity of these assays is imperfect, such that measurement of serum cobalamin levels does not always detect the presence of deficiency, nor does the finding of a low serum cobalamin always connote true deficiency. There are several reasons for this including the distribution of cobalamin between the binding proteins in circulation (**Table 1**), imperfections in the assays for its measurement, and various poorly understood factors relating to exchange of cobalamin between cellular and circulatory compartments. Regarding the distribution of cobalamin between plasma-binding proteins, since transcobalamin is responsible for cobalamin delivery to cells, the fraction of the total cobalamin that is associated with transcobalamin (holoTC), even though small in comparison with the haptocorrin-associated fraction, is more likely to be indicative of cobalamin status than is the total serum cobalamin. Some studies bear this out, although technical difficulties with measuring holoTC levels have only recently been overcome.

The other approach to identification of cobalamin deficiency is indirect, based on the detection of raised levels of compounds in the blood or urine that require adequate tissue levels of cobalamin for their metabolic disposal. The compounds most commonly measured for identification of possible cobalamin deficiency are methylmalonic acid and homocysteine. These are the substrates in two cobalamin-dependent reactions shown in **Figure 2**. Since the identification of these metabolic roles for cobalamin, it has been apparent that deficiency of cobalamin or disturbances in its metabolism would result in accumulation of these substances and a variety of assays for these metabolites is now available. Of the two compounds, elevation of the levels of methylamalonic acid is the more specific for identification of cobalamin deficiency; however, renal insufficiency can cause raised levels of methylmalonate in the blood. In addition to cobalamin deficiency, several other conditions also can cause raised homocysteine levels in the blood, including deficiencies of folate and of vitamin B_6, lack of thyroid hormone, and renal insufficiency (see 00151).

Table 3 shows the idealized usefulness of the various tests commonly available for detection of cobalamin deficiency.

Inborn Errors of Cobalamin Metabolism

There are several known but rare inherited molecular defects resulting in absence or structural defects in proteins required for normal absorption, transport. or metabolism of cobalamin. These include the intestinal binding proteins gastric intrinsic factor and its ileal receptor complex cubulin and amnionless, the plasma binders transcobalamin and haptocorrin, the enzymes that are required for conversion of cobalamin to its coenzymatically active methyl and deoxyadenosyl forms, and enzyme complexes involved in the catalysis of the two cobalamin-dependent reactions responsible for conversion of homocysteine to methionine and methylmalonate to succinate, respectively. Individuals who inherit a defective gene from each parent for any one of these proteins that are critical for cobalamin metabolism suffer from varying degrees of impairment of normal cobalamin-related status (see **Table 4**), closely mimicking the various

Table 3 Laboratory identification of cobalamin deficiency

Test	*Finding*	*Major limitations*
Serum/plasma cobalamin concentration	Low ($<150\,\text{pmol}\,\text{l}^{-1}$)	Normal levels in some deficient subjects; slight to moderately low levels may not connote deficiency
Serum/plasma holotranscobalamin (holo TC II)	Low ($<35\,\text{pmol}\,\text{l}^{-1}$)	Test not yet widely available; insufficient validation of usefulness
Serum/plasma or urine methylmalonic acid	Raised ($>350\,\text{nmol}\,\text{l}^{-1}$)	Levels raised in renal insufficiency
Plasma homocysteine	Raised ($>12\,\mu\text{mol}\,\text{l}^{-1}$)	Levels raised in folate and in vitamin B_6 deficiencies, renal insufficiency, hypothyroidism

Table 4 Inherited disorders affecting cobalamin metabolism and their effects

Cobalamin protein	*Effects of deletion or mutation*
Intrinsic factor	Cobalamin malabsorption (juvenile pernicious anemia)
Cubulin/amnionless complex	Cobalamin malabsorption (Immerslund-Grasbeck syndrome)
Transcobalamin	Severe cobalamin deficiency
Haptocorrin	No apparent abnormality
Cobalamin reducing and activating enzymes (mut^+ and mut^-, cobalamin mutants C-G)	Varying degrees of disruption in one or both cobalamin-dependent pathways

manifestations of cobalamin deficiency described above. These disorders usually become manifest at an early age.

See also: **Amino Acids**: Chemistry and Classification; Metabolism; Specific Functions. **Anemia**: Iron-Deficiency Anemia; Megaloblastic Anemia. **Folic Acid**. **Homocysteine**. **Inborn Errors of Metabolism**: Classification and Biochemical Aspects.

Further Reading

Carmel R (2000) Current concepts in cobalamin deficiency. *Annual Review of Medicine* **51**: 357–375.

Carmel R, Green R, Rosenblatt DS, and Watkins D (2003) Update on cobalamin, folate, and homocysteine. *Hematology (American Society of Hematology Education Program)*: 62–81.

Green R and Kinsella LJ (1995) Current concepts in the diagnosis of cobalamin deficiency. *Neurology* **45**: 1435–1440.

Stabler SP and Allen RH (2004) Vitamin B12 deficiency as a worldwide problem. *Annual Review of Nutrition* **24**: 299–326.

Stover PJ (2004) Physiology of folate and vitamin B12 in health and disease. *Nutrition Reviews* **62**: S3–12.

CELIAC DISEASE

V Nehra, E Marietta and J Murray, The Mayo Clinic College of Medicine, Rochester, MN, USA

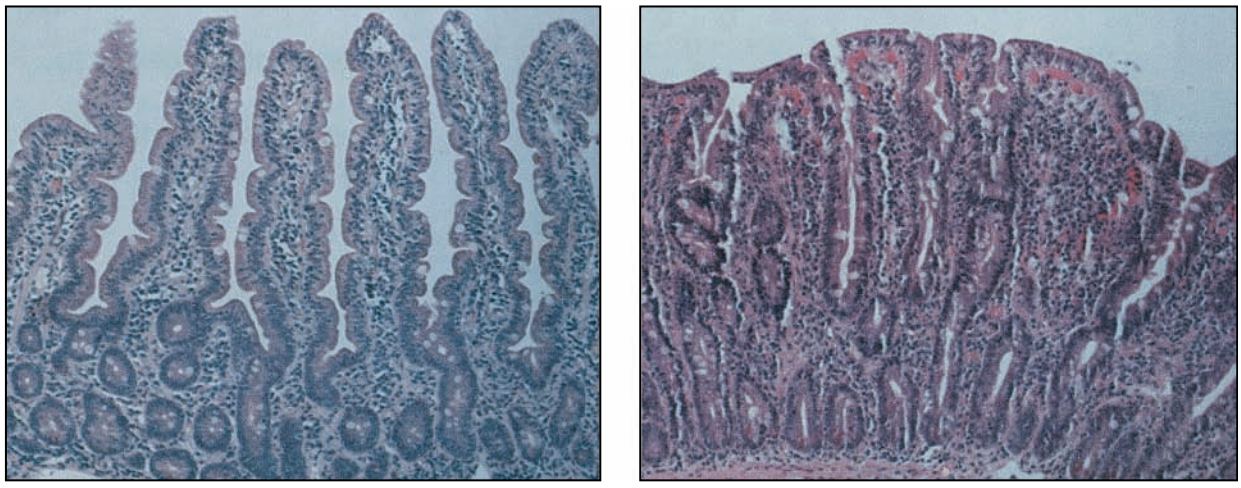
Introduction

Celiac disease is the end result of a collision between the human immune system and the widespread cultivation of wheat, where the point of contact is the lining of the small intestine. This collision results in inflammatory and architectural changes of the absorptive mucosa in those susceptible to celiac disease. The inflammation leads to the destruction and eventual loss of the absorptive surface (villi), increased net secretion, and malabsorption, leading to a multitude of consequences. Celiac disease predominantly affects Caucasians, and it is relatively rare in peoples from sub-Saharan Africa and the Far East, which may be due to different genetic backgrounds and/or the absence of wheat from the diet. The disease occurs in people who carry the particular tissue types HLA-DQ2 or HLA-DQ8, which appear to play an essential role in the disease pathogenesis. The inflammation usually resolves completely with the exclusion of gluten from the diet, will recur if gluten is reintroduced, and, as such, is regarded as permanent. While once thought to be a rare disease, it is recognized as a common chronic disorder that affects as many as 1% of some Western populations. Indeed, in some populations, it is regarded as the most common genetic disease that affects the gastrointestinal tract. It is now frequently detected by the presence of circulating autoantibodies against tissue transglutaminase, which is released in the damaged intestine. The final diagnosis of celiac disease is defined by biopsy evidence of the characteristic inflammatory changes in the small intestine and ultimately a response to the gluten-free diet.

Pathogenesis

Established celiac disease is characterized by an inflammatory response in the proximal small intestine. This inflammation consists of increased numbers of lymphocytes, plasma cells, and macrophages in the lamina propria and increased lymphocytes in the surface layer of the epithelium, called intraepithelial lymphocytes. The surface enterocytes are shorter and wider than normal and have poorly ordered nuclei. The normally tall thin villi are shortened and flattened. The cryptal layer is increased in depth. These changes may be patchy and affect variable lengths of the proximal small intestine (**Figure 1**).

The lamina propria is packed with T cells, many of which are CD4+ cells that respond to gliadin molecules in a manner that is DQ2 or DQ8 restricted. These cells are thought to be crucial in



(A) (B)

Figure 1 Inflammatory response in proximal small intestine. Proximal small intestinal tissue from a normal control (A) and from a celiac patient (B) was stained with hematoxylin and eosin. Shortened and flattened villi, shorter and wider enterocytes, increased numbers of intraepithelial lymphocytes, and a cryptal layer with increased depth are all present in (B) but not in (A).

the actual pathology of the lesion, and clones derived from such cells have been used to characterize the response to gliadin. Lymphocytes in the intraepithelial layer are also increased in number in untreated celiac disease, many of which bear the $\gamma\delta$ TCR. These cells slowly decrease when gluten is removed. An early event in the pathology is an increased expression of class II HLA molecules by enterocytes and antigen-presenting cells such as the macrophages within the lamina propria. These molecules are involved in the presentation of the exogenous and possibly the endogenous antigens that are released in the setting of inflammation.

In addition to the pathological changes in the small bowel mucosa, a potent humoral response occurs in untreated celiac disease. In the intestinal mucosa there are increased numbers of plasma cells secreting IgA, IgG, and IgM directed against gluten peptides and other plasma cells secreting antibodies directed against connective tissue autoantigens, particularly tissue transglutaminase. Those antibodies are found in the intestinal juice and the serum. The dynamics of the humoral response seems to parallel the dynamics of cellular injury, although antibodies may rise prior to mucosal relapse and disappear prior to healing.

There is also increased permeability of the small intestine to macromolecules. It is not clear if this increased permeability precedes the development of gliadin sensitivity. It may persist after healing of the microscopic lesions has occurred. Family members without the disease may have increased permeability. Recently, it has been suggested that gluten itself may rapidly cause an increase in paracellular permeability due to the uncoupling of intercellular tight junctions via the release of zonulin.

Adaptive Immune Response to Gluten

Celiac disease is characterized by an immune response to the storage proteins of wheat, rye, and barley, with wheat as the most immunogenic. Wheat gluten is composed of glutenin and gliadin, and evidence suggests that the gliadin fraction induces disease. Information gathered from T cell clones derived from chronic lesions of the small intestines of celiac patients with established disease demonstrate that gliadin peptides are presented by HLA class II molecules to CD4+ T cells. Several studies have suggested that unaltered native gliadin peptides were antigenic but lacked the negatively charged amino acids needed to bind to the recognition sites of the DQ2 or DQ8 molecules. It has since been recognized that the gliadin peptides are made more antigenic by tissue transglutaminase, and it is these altered (deamidated) peptides that either perpetuate or cause gluten sensitivity in celiac disease.

It is thought that tissue transglutaminase is released by fibroblasts in the setting of intestinal inflammation, since tissue transglutaminase is normally involved in cell to cell signaling and extracellular matrix formation. Released tissue transglutaminase would then bind to its preferred substrate, gliadin, and convert (deamidate) the specific glutamine residues in gliadin to glutamic acid, resulting in improved binding of gliadin peptides to specific pockets in the DQ molecule. Interestingly, this release of tissue transglutaminase into the

inflamed celiac gut also results in a strong autoimmune response to tissue transglutaminase with high levels of circulating anti-tTG IgA present in untreated celiac patients. Thus, the main characteristics of celiac disease are the DQ2/DQ8 restricted responses to gliadin peptides, the strong intestinal T cell response to deamidated gliadin peptides, and the production of circulating autoantibodies against tissue transglutaminase. However, it has been observed that some native gliadin peptides can also induce strong T cell proliferation. The 11-mer native peptide of gliadin, amino acids 206–216, was observed to induce proliferation in DQ8 T cell lines, and when instilled into the jejunum of a patient with celiac disease, induced jejunal inflammation.

Observations in the DQ8 mouse support the notion that both native gliadin peptides and deamidated gliadin peptides can evoke T cell responses in a DQ restricted fashion. Also of interest is that T cell cultures derived from the biopsies of children, who presumably would have had a recent onset of disease, responded to native gluten peptides and a narrower range of gliadin peptides than those derived from adults. Thus, it is possible that naturally occurring (native) epitopes of gliadin may be involved in the initiation of gluten sensitivity, whereas the deamidation process is involved in the perpetuation and amplification of the process.

Innate Immune Response

Many of the studies on gut responses to gluten have been performed in the established chronic lesion. Little is known of innate responses that can elicit effects within minutes to hours of exposure to gluten. *In vitro* studies demonstrated an increase in the expression of HLA antigen on the cells in the surface layers of the intestinal mucosa occurring within 2–4 h after exposure to gluten. Gluten also causes the production of the proinflammatory cytokine IL-15 at the surface epithelium. IL-15 expressed by the surface enterocytes activates NK-like T cells to recognize gluten presented by MHC class 1a molecules in the context of the NKG-2D receptor. The NK-like T cell may be a key player in both the damage to the surface epithelium and be a proinflammatory influence on adaptive response that occurs in the underlying lamina propria. This induction of innate immune responses by gluten may have important consequences. Since the gluten peptides enter into the epithelial compartment and paracellular regions, the Peyer's patch pathway is not the exclusive route for the introduction of the gluten peptides to the immune system. As such, the bypass of the Peyer's patches may lead to a loss of tolerance and even an induction of sensitization, resulting in an uncontrolled immune response in the intestinal mucosa. The stimulation of the surface layer not only explains the substantial changes that occur in the surface but may well condition the inflammatory response beneath. Thus, both arms of the immune system, the innate and the adaptive, play a role in the development of celiac disease, even though most attention has so far been focused on the adaptive arm.

Triggers for Loss of Tolerance

Celiac disease only develops in a minority of DQ2+ individuals. How the consumption of gluten generates an inflammatory state in these individuals can be theorized as follows. First, there may be a trigger of the innate immune response, such as a viral infection or physical injury (surgery) that initiates inflammation and later permeability. Enough triggers repeated over time will alter the immune milieu of the mucosal compartment and perturb gut homeostasis, potentially altering the levels of the regulatory cytokines IL-10 and TGF-β and increasing the levels of inflammatory cytokines like IFN-γ and ILNA.

Determining which factors lead to the loss of tolerance to gluten in DQ2+ individuals who later develop celiac disease will be crucial in understanding the pathogenesis of celiac disease. Possible factors that may lead to the loss of tolerance are recurring gastrointestinal infections, surgery, or pregnancy. The way in which children are first exposed to gluten may also affect whether tolerance to gluten or an inflammatory response develops. Quantity and timing of exposure to gluten during childhood may affect the development of tolerance to gluten. Interestingly, the aging process has also been implicated in the loss of tolerance to gluten. More recently, it has become apparent that most celiac patients initially present with disease as adults, possibly due to an increased propensity to autoimmunity associated with advanced age.

The mechanism of response to gluten in celiac disease is quite different from that of IgE-mediated food allergies. IFN-γ, a potent inflammatory cytokine, is characteristically produced in celiac disease as well as TNF-α. IL10 and TGF-β, which are both counter-inflammatory regulatory cytokines for the intestine, are also expressed in celiac disease, although they are apparently inadequate to prevent the substantial inflammation that occurs.

Overall then, it is thought that environmental triggers, which may be nonspecific, the innate responses to gluten, and finally the adaptive responses to gluten combine to result in the enteropathy that characterizes celiac disease. This process

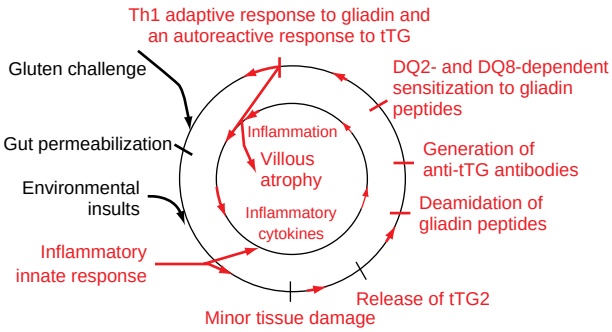


Figure 2 Maelstrom of celiac disease. An initial perturbation of the mucosal immune system would activate the innate immune response, increase epithelial permeability, and inflame the mucosal environment, resulting in overexpression of MHC class II molecules and the presentation of gliadin peptides to CD4+ T cells. tTG produced by fibroblasts in the damaged lining would deamidate gliadin peptides and amplify the CD4+ T cell response to gliadin. The released tTG would subsequently be targeted by the humoral immune response as an autoantigen, further advancing the mucosal immune system down a spiraling pathway towards self-destruction and villous atrophy.

can be best described as a 'maelstrom' of immune activity that leads to celiac disease if unchecked (**Figure 2**). Removal of gluten, the major instigator, will result in the reversal of this process and healing of the intestine. However, reintroduction of gluten will often result in a prompt recurrence of the disease, demonstrating that celiac disease is a permanent intolerance to gluten.

Epidemiology of Celiac Disease

Celiac disease is one of the most common, chronic genetic gastrointestinal conditions affecting just under 1% of Caucasian individuals. Whilst it was initially recognized in Northern Europeans, celiac disease will affect Caucasians wherever they live. It can affect people of mixed ethnic background. It is apparently rare in Southeast Asia and sub-Saharan Africa. While celiac disease was once considered primarily a childhood disease, in many geographic locations, celiac disease is more commonly diagnosed in adulthood. The median age in one study in the US was 50 years of age. Even in childhood, the age at diagnosis is now increased from early infancy to later in childhood or adolescence. The spectrum of disease has also changed with an increasing proportion of patients being diagnosed with mono-symptomatic or less severe celiac disease. Many of these patients would not have been diagnosed in the past, but their symptoms described as functional disorders. It is not uncommon to diagnose celiac disease in those of advanced age. Patients in their 80s may be diagnosed for the first time with celiac disease, some of these having symptoms that persist for many years prior to diagnosis. The delay time to diagnosis may be anywhere between 8 and 11 years from the time of first clinical presentation to the actual diagnosis being made.

Diagnosed celiac disease appears to be more common in women than men. However, in those individuals who are identified by population-based screening, the prevalence of the disease appears to be equal between genders. The explanation for this is unclear, however, presentation of disease may be more common in women because of the nutritional challenges posed by pregnancy and menstruation, especially when producing iron deficiency anemia. The predisposition of women to autoimmune disease may also increase the likelihood of the occurrence of symptomatic celiac disease.

The sister condition of celiac disease is dermatitis herpetiformis, which is the skin manifestation of gluten-sensitive enteropathy. It is an extremely itchy immunobullous disease that affects the extensor surfaces of elbows, knees, buttocks, the hairline, and the torso and is much less common than celiac disease. Probably the ratio between the two in geographic areas where both have been estimated is approximately 10:1. However, in countries where there has traditionally been less celiac disease awareness, such as North America, the ratio may be closer to 3:1.

The incidence of celiac disease and the prevalence of celiac disease have been estimated in a number of geographic locations. These have shown incidence estimates of approximately between 1 and 9 cases per 100 000 person years and prevalence rates of anywhere from 1 in 2000 to 1 in 300. The latter high estimate is based on geographic locations where there have been active case findings such as Tampere in Finland. The lower rates are the prevalence based on clinically diagnosed cases. However, these measures may greatly underestimate the true prevalence of celiac disease in the community. Two specific lines of research would suggest this. One is a systematic follow-up of birth cohorts for the occurrence of celiac disease by using serologic screening. In one geographic location in Denver, Colorado, almost 1% of children became persistently seropositive for markers for celiac disease by 7 years of age. This cumulative prevalence is remarkably similar to population-based studies in both Europe and North America that suggest prevalence in an adult general population of approximately 1 in 133. The estimates of the prevalence of celiac disease in most Caucasian studies have been remarkably similar, varying from 1 in 300 to 1 in 87.

All the studies so far have been carried out in predominantly Caucasian populations. No studies have been carried out that have incorporated substantial numbers of African-Americans, nor have systematic studies been carried out in sub-Saharan Africa. Areas of North Africa are at least as commonly affected as Europe. Individuals from Southeast Asia very rarely develop celiac disease.

The major effect of these screening studies has been to illustrate the wide spectrum of disease that incorporates celiac disease. Many individuals are completely asymptomatic. Others may be severely ill, presenting at an early age. It is possible that some individuals who have developed celiac disease may have no symptoms whatsoever despite the presence of demonstrable serologic and histopathological changes of celiac disease.

Associated Disorders

Dermatitis herpetiformis This is characterized by an extremely pruritic papulovesicular eruption, which usually occurs symmetrically on the elbows, knees, buttocks, and back. About 80% of patients with dermatitis herpetiformis have small intestine histology indistinguishable from celiac sprue. The diagnosis is established by skin biopsy demonstrating granular IgA deposits in areas of normal appearing skin. A majority of patients with the skin lesion who undergo small bowel biopsy have intestinal mucosal changes of celiac disease. The skin lesions, as well as small bowel histology, improve on a gluten-free diet. Dapsone is an effective short-term treatment for dermatitis herpetiformis; however, it does not have any impact on management of small bowel enteropathy. Also, those with dermatitis herpetiformis who are not compliant with the gluten-free diet are at higher risk for malignancy, as are those with celiac disease.

Celiac disease has also been associated with other autoimmune as well as nonautoimmune disorders. It has been reported that the longer there is exposure to gluten in patients with celiac disease, the greater the occurrence of other autoimmune diseases. There is evidence for a strong association between type 1 diabetes and celiac disease. About 8% of patients with type 1 diabetes have the characteristic features of celiac sprue on small bowel biopsy. When the two diseases coexist, 90% have the diagnosis of diabetes before celiac disease. Among the symptoms that may be suggestive of coexisting celiac disease, in addition to those considered classical for celiac disease, are delayed puberty, hypertransaminasemia, anemia, iron deficiency, arthralgias, dental enamel defects, hypoglycemia, and unexplained reduction in insulin requirements. Treatment with a gluten-free diet may improve diabetic control and decrease the occurrence of hypoglycemia episodes.

There is also a strong association between selective IgA deficiency and celiac sprue. Studies including adults and children in Ireland and Italy reported the frequency of selective IgA deficiency in celiac sprue to be about 2%, and 5–11% of IgA-deficient individuals have celiac disease.

There is a strong association between Down's syndrome and celiac disease. Individuals with Down's syndrome and celiac disease more commonly have gastrointestinal manifestations such as intermittent diarrhea, failure to thrive, anemia, and low serum iron and calcium. The prevalence of celiac disease in patients with Down's syndrome varies between 5 and 12%. An increased prevalence of celiac sprue has also been reported in individuals with Turner's syndrome and Williams syndrome.

Clinical Presentation

Celiac disease may present in a wide variety of ways (**Table 1**). In children, the onset of celiac disease is classically described as occurring within the first to seventh year of life with the introduction of cereals to the diet. Symptoms may vary with the age of the child at onset of disease. Young children may develop chronic diarrhea, failure to thrive, muscle wasting, abdominal distension, vomiting, and abdominal pain. Older children may present with anemia, rickets, behavioral disturbances, or poor performance in school. In some children constipation, pseudo-obstruction, and intussusception may be seen. It has been estimated that 2–8% of children with unexplained short stature may have celiac disease. Dental enamel defects involving secondary dentition as well as

Table 1 Presentations of celiac disease

Gastrointestinal presentations
• Classic malabsorption syndrome – diarrhea, steatorrhea, weight loss, bloating, failure to thrive, multiple deficiencies
• Monosymptomatic – anemia, diarrhea, lactose intolerance, constipation
• Acute abdomen – abdominal pain, intussusception, vomiting, obstruction perforation, lymphoma
Nongastrointestinal presentations
• Neurological diseases – migraine, ataxia, peripheral neuropathy, dementia, depression, epilepsy
• Autoimmune diseases – dermatitis herpetiformis, pulmonary hemosiderosis
• Diseases associated with nutritional deficiencies – bruising, epistaxis, chronic fatigue, infertility, bone disease, short stature

neurological syndrome and epilepsy with intracranial calcification have also been reported in children with celiac disease.

In adults, celiac disease may be overt in presentation with classic gastrointestinal symptoms of diarrhea, weight loss, and abdominal pain. The presence of diarrhea and/or steatorrhea, which occurs in about 50% of patients, indicates severe disease and malabsorption. Often celiac disease is diagnosed in adults with nongastrointestinal symptoms including iron deficiency anemia, abnormal liver tests, osteopenic bone disease, neurological symptoms, or menstrual abnormalities. Anemia is common in both children and adults with celiac disease and may be secondary to iron deficiency, folate deficiency, or a combination of the two. Iron deficiency is frequently associated with celiac disease. Six to ten per cent of patients with unexplained iron deficiency anemia when evaluated by upper endoscopy with small bowel biopsy were diagnosed with celiac sprue in the absence of any other features suggestive of the disease.

Unexplained elevated serum transaminases (ALT, AST) should also raise the suspicion of undiagnosed celiac disease. Up to 9% of adults with unexplained elevated serum transaminases have been diagnosed with celiac disease based on serological testing or small bowel biopsy. Liver biopsies in these individuals may show reactive hepatitis. In this setting, adherence to a gluten-free diet results in improvement or normalization of the liver enzyme levels. The prevalence of celiac sprue is higher in adults with autoimmune liver disease than in the general population. Volta *et al.* demonstrated a 4% prevalence of celiac sprue in 181 patients with autoimmune hepatitis. Similarly, a high prevalence of celiac disease has been reported in other autoimmune liver disorders such as primary biliary cirrhosis, autoimmune cholangitis, and primary sclerosing cholangitis.

Patients with untreated celiac disease are at increased risk for the development of osteoporosis and low bone mineral density. Celiac disease can result in malabsorption of calcium and vitamin D. Malabsorption of calcium results from impaired transport by the diseased small bowel as well as precipitation of the ingested calcium with unabsorbed intraluminal fats to form insoluble soaps that are then excreted in the stool. Untreated, patients with celiac disease have been observed to have increased bone turnover and elevated levels of 1,25 dihydroxycholecalciferol because of secondary hyperparathyroidism that helps maintain a positive calcium balance. This results in diminished bone densities associated with increased risk of fractures in patients with classical celiac disease. Untreated patients with celiac sprue are at risk for developing low bone mineral density and osteoporosis reported that 34% of their study population with celiac disease had fractures in the peripheral skeleton. For those with classical celiac symptoms, the odds ratio for fracture was 5.2 compared to those without celiac disease. Although the reduced bone mineral density improves on a gluten-free diet, adults with celiac sprue may be at increased risk for the development of peripheral bone fractures, although this is not universally agreed.

Infertility and recurrent spontaneous abortions have been reported in women with celiac disease. Male infertility has also been observed in patients with untreated celiac disease. Restoration of fertility both in males and females has been observed following treatment with a strict gluten-free diet and may be unexpected in some couples who have been unable to become pregnant before this time.

Patients with celiac disease may in addition present with neurological symptoms such as ataxia, muscle weakness, paresthesias, weight sensory loss, epilepsy, and bilateral parieto-occipital calcification. Symptoms of depression, epilepsy, and migraine have been reported in 30% of patients with celiac disease.

Diagnosis

Small bowel biopsy remains the gold standard for diagnosis of celiac disease. Over the past decade the diagnostic criteria for celiac sprue have changed. Based on the 1990 revised criteria of the European Society of Paediatric Gastroenterology and Nutrition, the diagnosis of celiac sprue can be made with a diagnostic small bowel biopsy in a patient with highly suggestive clinical symptoms, followed by an objective clinical response to a gluten-free diet. Endoscopic biopsies from the distal duodenum are preferable because the presence of Brunner glands in the duodenal bulb and proximal second portion of the duodenum may affect histologic interpretation. The original criteria requiring a series of three biopsies, i.e., first to confirm the diagnosis, second for demonstration of response to a gluten-free diet and the third for deterioration after gluten challenge, are only required in those few patients in which there is still some diagnostic uncertainty.

Endoscopic features observed in patients with celiac disease include scalloped or fissured folds, absence of folds when the duodenum is inflated, and visible submucosal blood vessels; however, these findings are unreliable in diagnosing celiac

disease as only roughly half of the patients will have the findings detected endoscopically. Other causes of atrophy are indistinguishable from celiac disease.

Characteristic histologic changes described are partial or total villous atrophy, elongation of crypts, a decreased villous:crypt ratio, and increased intraepithelial lymphocytes (>30 per 100 enterocytes). Marsh and colleagues proposed a classification for the spectrum of histologic changes ranging from type 0 or preinfiltrative/normal, type1 or infiltrative lesion (increased intraepithelial lymphocytes), type 2 or hyperplastic lesion (presence of crypt hyperplasia), type 3 or destructive lesion (variable degree of villous atrophy), and type 4 or the hypoplastic lesion (total villous atrophy with crypt hypoplasia).

The role of radiological studies in the initial diagnosis of celiac sprue is limited. The findings of flocculation and segmentation of barium representing excessive fluid secretion in the lumen of the small intestine, thickened mucosal folds, and dilation of the small intestine are nonspecific and insensitive for celiac disease. Reversal of the fold patterns between the jejunum and ileum may also be seen. Computerized tomography techniques may be useful in diagnosing the complications of celiac sprue such as development of lymphoma, malignancy, hyposplenism, or cavitating mesenteric lymphadenopathy. CT enterography techniques are currently under investigation and may become an accepted diagnostic test in the future.

Serological Screening Tests

Serological tests are helpful in detecting celiac disease in individuals with nongastrointestinal symptoms, and high-risk groups who may or may not have signs of disease. The clinicians often use the serological results to triage those who need small bowel biopsy. The high-risk groups include first-degree relatives of confirmed cases of celiac disease, those with type 1 diabetes mellitus, Down's syndrome, Turner's syndrome and unexplained dental enamel deficits, and children with unexplained short stature. Serological tests are also used to monitor progress after diagnosis as well as in prevalence studies in unselected populations. The serological tests utilized in current clinical practice include the endomysial antibody, tissue transglutaminase antibody, and the anti-gliadin antibodies (IgA and IgG).

An enzyme-linked immunosorbent assay (ELISA) for both the IgA and IgG subclass of antibodies to gliadin has been used for the diagnosis of celiac disease. Their role in diagnosis is limited because of moderate sensitivity and specificity. The antigliadin antibodies are found in intestinal secretions as well as in serum of patients with untreated celiac disease. However, these antibodies are also found in a variety of autoimmune disorders including rheumatoid arthritis, Sjögren's syndrome, sarcoidosis, inflammatory bowel disease, and cows' milk protein intolerance. IgA antigliadin antibodies have sensitivity of 75–90% and specificity of 82–95%. The IgG antigliadin antibodies range in sensitivity from 69% to 85% and have specificity of 73–90%; they are useful in the diagnosis of celiac patients with IgA deficiency. Other than this use gliadin antibodies have fallen from favor as a screening test for celiac disease (National Institute of Health consensus panel).

The IgA antiendomysial antibody (EMA) assay is directed against the connective tissue protein found in the collagenous matrix of human and monkey tissue. This antibody is found in association with celiac sprue. The test is based on immunofluorescence techniques using monkey esophagus or human umbilical cord as a substrate. Although quite sensitive (85–98%) and specific (97–100%), the test has several limitations including false-negative results in 23% of patients with celiac disease who have selective IgA deficiency. Other factors that have an impact on the sensitivity and specificity of this test include laboratory variations and disease severity. In a study of 101 patients with untreated celiac disease the sensitivity of the endomysial antibody in those with total villous atrophy was excellent (100%), but decreased remarkably (31%) in patients with partial villous atrophy. The endomysial antibody performed by the indirect immuno-flourescent assay (IFA) technique is technically challenging and is being replaced by the tissue transglutaminase antibody test.

Tissue transglutaminase (tTG) is a cytosolic protein released by damaged epithelial cells. This is the autoantigen recognized by the endomysial antibody indirect immunofluorescence assay in patients with celiac disease. The advantages of this test are that it is performed using ELISA techniques, which makes it easier to perform, is widely available, and less costly. It eliminates the use of monkey esophagus as well as the subjective interpretation of immunofluorescence analysis of the endomysial antibody test. Though the tTG test is comparable to EMA in sensitivity, there is loss of specificity in patients with autoimmune disorders, hence it is important to confirm the diagnosis with small intestine biopsy.

In some patients biopsies are taken during an endoscopy that has been performed for another

reason. In these patients serological tests can be useful to help confirm the diagnosis if there is some uncertainty with an equivocal biopsy, while negative serology in this circumstance may indicate another cause. Celiac disease may occur in the absence of antibodies however.

Treatment

Once the presumptive diagnosis of celiac disease is made then treatment may be commenced. It is important that the patient does not start to restrict their diet until each of the steps including the biopsy have been completed. Once confirmed, the responsibility for directing the management of the patient lies with the physician.

The treatment starts with an explanation of the condition and its cause. It is important that the patient understands that this is a chronic inflammatory condition of the gut and not a simple food allergy, that it is permanent eventhough the intestine will heal, and that the central and indeed only treatment at present is a gluten-free diet for life. The clinician should expect shock and even a fully expressed grief reaction on the part of the patient. Disbelief that something as basic to the Western diet as wheat is responsible is common. Some patients are overwhelmed both by the realization of having a chronic illness and others by relief that an explanation for their suffering has been found. The tone that the physician sets is crucial to the patient's success. A positive and upbeat though serious demeanor on the part of the doctor is appropriate, as most patients will do very well so long as they stick to the diet. Probably the most important thing that the doctor can do beyond diagnosis is to refer the patient for professional dietary advice that is up to date on how to achieve a gluten-free life style.

Patients should be encouraged to join both local and national support groups as an essential adjunct to management. The feeling of isolation so common in newly diagnosed patients in the past can be quickly dispelled by participation in an active support group.

It is important to identify and to correct deficiencies with nutritional supplementation. Deficiencies of the fat-soluble vitamins (D, E, A, and K), iron, folate, B_{12}, and even zinc or selenium are common. Baseline bone mineral density should be measured, as osteoporosis and osteomalacia are common.

Occasionally, intensive nutritional support and fluid replacement may be needed in very ill patients. Coexisting malignancy/autoimmune disease should be considered especially in elderly or ill patients. Follow-up of patients to ensure response to gluten-free diet and compliance is crucial to ensure long-term compliance as well as detecting potential complications of the disease. Screening of at-risk family members should be considered.

The Gluten-Free Diet

The term 'gluten' as it is used in the context of celiac disease refers to the storage proteins of wheat (gliadins and glutenin) and of barley (prolamines), and rye (hordeins), and oats (avedins). Gluten is defined in the setting of celiac disease as any protein-containing derivative of the offending grains or their derivatives. Grains that should be avoided are as follows:

1. wheat;
2. barley;
3. rye;
4. spelt;
5. kamut; and
6. triticale.

To achieve healing and maintain health, a well-balanced, interesting dietary life style that avoids gluten should be adopted. However, it is not enough to simply avoid gluten; patients need to be enabled to explore dietary alternatives and strategies that minimize impact on their life style.

The role of oat toxicity in celiac disease is still controversial. Several recent, well-constructed studies have demonstrated no ill effects when a moderate amount of oat products have been included in the diet of either newly diagnosed or already treated celiac patients. These recent studies have clearly demonstrated that oats are nontoxic for most patients with celiac disease, however, the concern is that contamination of oat products by gluten is taking place during the growing, milling, or processing of oat products. When completely gluten-free oats become available then it will probably be safe to recommend them to most celiac patients. A small number of patients may still react to oats, some markedly so. Where a patient intends to include oats in their diet they need to have a careful and informed medical follow-up.

While foods such as bread, cookies, biscuits, and pasta are obvious sources of gluten, many other seemingly 'safe' foods contain hidden gluten. It is important to enquire if a food has any ingredients that are in any way derived from, or processed with, wheat, barley, or rye. This part of the diet is not self evident, and the patient needs both expert counseling from a dietitian as well as being versed with and up to date on the gluten-free diet with ongoing

support from a local or national support group. Nonfood items, such as medications and communion wafers, may also be unappreciated sources of gluten, as are fat substitutes and food contaminants. Ingredients that should be viewed with suspicion include:

1. malt or malt flavoring;
2. hydrolyzed vegetable protein (HVP);
3. modified food starch (and starch in foreign foods);
4. natural flavorings;
5. vegetable gum; and
6. fat substitutes.

It is important that the most up to date instruction manuals are used. Older manuals may contain out of date or even misleading information.

Some ostensibly gluten-free foods may become contaminated with gluten during processing, packaging, transport, handling in the store, or even preparation in the patient's own kitchen.

Home testing kits for gluten are now available but generally are not helpful or practical for most patients. Large listings of the commercially available processed foods that are gluten free are available but must be updated at least yearly and must be tailored to geographic location. Patients should not rely on the self test of reaction to gluten as a means of detecting gluten in foods as symptoms may be delayed.

Bone Metabolism

Osteomalacia is a well-recognized, although uncommon, complication of celiac disease, with bone pain and pseudofractures as features. It is associated with elevated alkaline phosphotase and often normal levels of calcium and phosphate. It usually responds well to a gluten-free diet and calcium and vitamin D supplementation.

Osteoporosis, which is common in adults with celiac disease, affects both men and women, and the exact mechanisms are not clear. The prevalence of osteoporosis is even higher in refractory sprue compared to gluten-free-diet-responsive patients. Diagnosis depends on bone mineral density testing with a T-score less than 2.5 SD below mean peak value in young adults. The primary treatment for the osteoporosis in a celiac is the strict gluten-free diet with adequate calcium ($1500\,mg\,day^{-1}$ and vitamin D). Other measures directed at preserving or building bone density may be necessary if the boss mineral loss has been substantial or does not recover with a gluten-free diet.

Complications

Nonresponsive Celiac Disease

Whilst most patients with celiac disease respond appropriately to a gluten-free diet, usually with responses to symptoms occurring within days to weeks of institution of the diet, a small proportion of patients (approximately 5%) do not have the expected complete response to a gluten-free diet or they have a relapse of symptoms while apparently on a gluten-free diet. This scenario termed 'nonresponsive celiac disease' is multifactorial in nature.

The single most common cause of continued or relapsing symptoms in patients with celiac disease is that of inadvertent gluten ingestion. There are many ways in which gluten can get into the diet, and in one series the most common source was commercial cereal in which minor ingredients were derived from the offending grains. However, other sources such as communion wafers and environmental contamination with flour, particularly of baked goods, are also possible.

In patients whose serologic tests have returned to normal and where a careful dietary review, including a detailed food record, does not reveal any potential source of gluten contamination, the occurrence of a second associated disease or a complication of celiac disease must be considered. A common associated disorder would be microscopic colitis, either lymphocytic or collagenous. Typically, these patients will have watery diarrhea whereas symptoms related to malabsorption such as weight loss, bloating, and steatorrhea will have resolved. The patient will continue to have watery diarrhea or may, indeed, develop watery diarrhea while on a gluten-free diet. The taking of biopsies from the colon can readily identify this condition. Whilst in some patients adhering to a strict gluten-free diet may improve the colitis, in many circumstances, it does not or it has not sufficed. The use of empiric therapy such as Pepto-Bismol®, loperamide, or, in some circumstances, delayed release budesonide may be valuable. Another cause of continued diarrhea is disaccharidase deficiency such as lactose intolerance. In most patients with celiac disease, the lactose intolerance that occurs is secondary to the injury and resolves its treatment. In a few unfortunate patients there may be a genetic predisposition to lactose intolerance. Avoidance of lactose or the use of lactase enzyme supplementation may suffice for correction of symptoms. In patients who have continued steatorrhea but in whom small bowel biopsies are found to have become normal, pancreatic exocrine insufficiency or bacterial overgrowth syndrome might be considered.

Where the small intestine has failed to recover histologically, particularly in patients who have continued symptoms and signs of malabsorption, the

diagnosis of refractory sprue is made. This condition is often associated with severe illness, significant bone disease, and hypoalbuminemia. These patients are particularly prone to ulceration in the proximal small intestine, so-called ulcerative jejunitis. Some have clonal expansion of T cells within their intestine. These patients are probably entering a pre-lymphoma state and the mortality in these circumstances is high with a relatively poor response to immunosuppression; many will progress to lymphoma within 5 years. Other patients appear to have refractory sprue but without clonality, and they tend to respond much better to immunosuppression. This probably represents a now self-perpetuating autoimmune process within the intestine. The rare case of collagenous sprue, which has features similar to celiac disease but is characterized by a thick layer of collagen in the intestine subepithelial layer in the colon, typically responds poorly to all therapies and often require long-term nutritional support.

The approach to diagnosing and treating nonresponsive celiac disease is outlined in **Figure 3**.

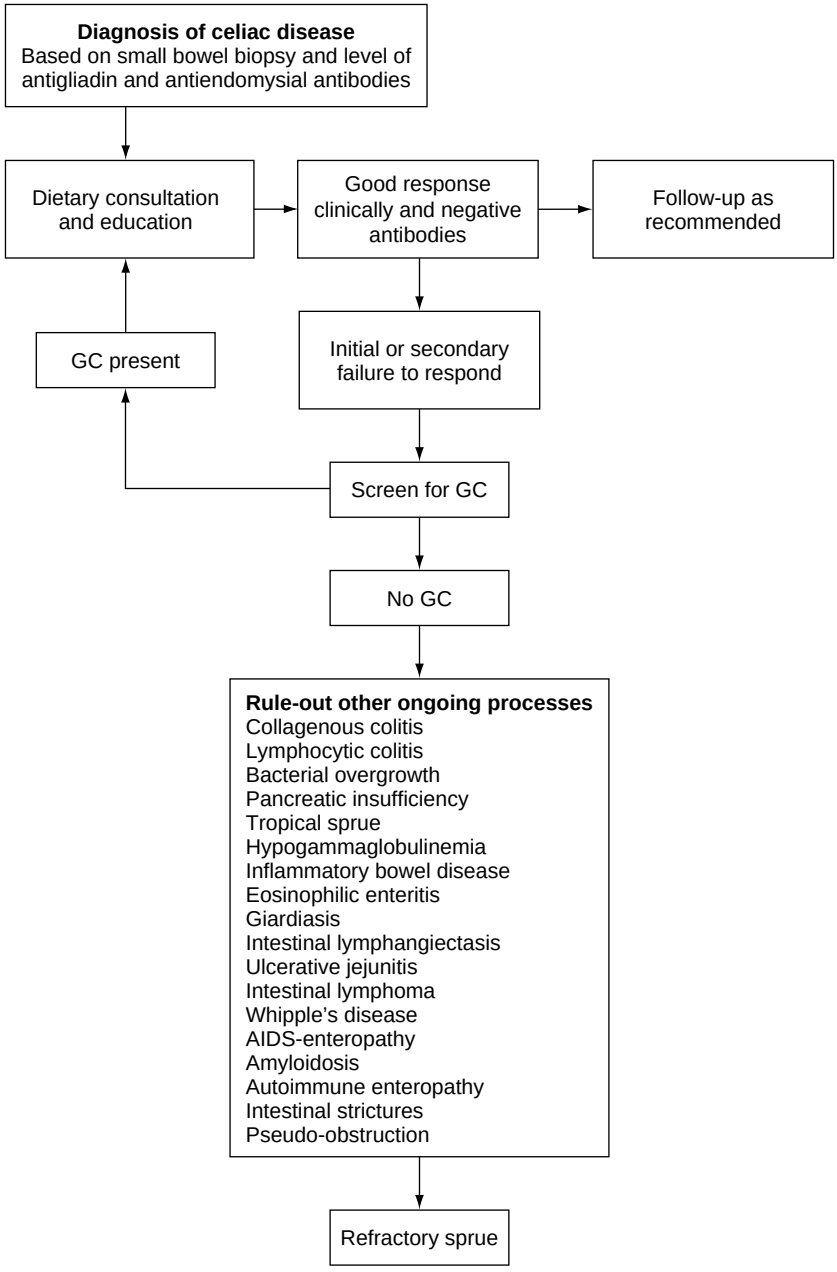


Figure 3 Flow chart for diagnosing and treating celiac patients. The proper procedure for diagnosing a patient who potentially has celiac disease, education and treatment of that patient, followed by the steps that need to be taken in the event that the patient is not responsive to a gluten-free diet are illustrated in a flow chart. GC, gluten challenge.

Malignant Complications of Celiac Disease

The complications of celiac disease can be divided into malignant and nonmalignant complications. In addition, the malignant complications of celiac disease are most commonly that of non-Hodgkin's lymphoma of a T cell variety. This particular tumor occurs in patients who have not been compliant with the diet or within 3 years of diagnosis. The risk of lymphoma or other malignancies appears to drop once a gluten-free diet has been instituted. While the relative risk of malignancy in celiac disease is greatly increased for specific diseases, the actual absolute risk is relatively small. The presentation of lymphomas in the small intestine can be acute with a surgical emergency such as obstruction, perforation, and bleeding, or gradual with insidious return or progress of severe malabsorptive symptoms, often associated with hypoalbuminemia and severe weight loss and malnutrition. The treatment for lymphoma is often unsuccessful. Those patients presenting acutely and managed surgically appear to do better than those who have a slow insidious onset. Occasional cases of response to stem cell transplantation have been reported.

The second most common malignancy occurring in celiac disease is that of adenocarcinoma of the small intestine. This adenocarcinoma seems to occur in the setting of the chronic inflammation of celiac disease. It is associated with defects and mismatch repair and whilst this is an unusual tumor, the survival with aggressive surgical therapy may be better than that for small bowel adenocarcinomas that occur sporadically. Usually, these patients present with iron deficiency anemia, gastrointestinal bleeding, obstruction, or pain. Other malignancies such as esophageal cancer or melanoma are increased in frequency in celiac disease, though again the absolute risk is low. Some recent evidence suggests that risk of breast cancer may be reduced in patients with celiac disease, though this is yet to be confirmed.

Nonmalignant Complications of Celiac Disease

Nonmalignant complications of celiac disease include ulcers and structuring within the intestine that occasionally may present with small bowel obstruction and/or bleeding, and recurrent acute pancreatitis as the result of inflammation, probably of the sphincter of Oddi. Nongastrointestinal complications are usually the consequence of malnutrition or specific deficiencies. However, others such as neurological problems including ataxia, peripheral neuropathy, or dementia are of uncertain mechanism and perhaps autoimmune in nature. Other consequences of celiac disease have been discussed in the section on atypical or nongastrointestinal presentations (**Table 1**). Many, but not all, of these nonmalignant complications of celiac disease will respond to a gluten-free diet.

See also: **Cereal Grains**. **Cytokines**. **Handicap**: Down's Syndrome. **Osteoporosis**. **Vitamin D**: Rickets and Osteomalacia.

Further Reading

Abdo A, Meddings J, and Swain M (2004) Liver abnormalities in celiac disease. *Clinical Gastroenterology and Hepatology* **2**: 107–112.

Abdulkarim AS, Burgart LJ, See J, and Murray JA (2002) Etiology of nonresponsive celiac disease: results of a systematic approach. *American Journal of Gastroenterology* **97**: 2016–2021.

Ackerman Z, Eliakim R, Stalnikowicz R, and Rachmilewitz D (1996) Role of small bowel biopsy in the endoscopic evaluation of adults with iron deficiency anemia. *American Journal of Gastroenterology* **91**: 2099–2102.

Farrell RJ and Kelly CP (2002) Celiac sprue. *New England Journal of Medicine* **346**: 180–188.

Farrell RJ and Kelly CP (2001) Diagnosis of celiac sprue. *American Journal of Gastroenterology* **96**: 3237–3246.

Holmes GK (2002) Screening for coeliac disease in type 1 diabetes. *Archives of Disease in Childhood* **87**: 495–498.

Marsh MN (1992) Gluten, major histocompatibility complex, and the small intestine. A molecular and immunobiologic approach to the spectrum of gluten sensitivity ('celiac sprue'). *Gastroenterology* **102**: 330–354.

Murray JA, Watson T, Clearman B, and Mitros F (2004) Effect of a gluten-free diet on gastrointestinal symptoms in celiac disease. *American Journal of Clinical Nutrition* **79**: 669–673.

Pengiran-Tengah DS, Wills AJ, and Holmes GK (2002) Neurological complications of coeliac disease. *Postgraduate Medical Journal* **78**: 393–398.

Schuppan D (2000) Current concepts of celiac disease pathogenesis. *Gastroenterology* **119**: 234–242.

Sollid LM (2000) Molecular basis of celiac disease. *Annual Review of Immunology* **18**: 53–81.

Sulkanen S, Halttunen T, Laurila K *et al.* (1998) Tissue transglutaminase autoantibody enzyme-linked immunosorbent assay in detecting celiac disease. *Gastroenterology* **115**: 1322–1328.

Anonymous (1990) Working Group of European Society of Paediatric Gastroenterology and Nutrition: Revised criteria for diagnosis of coeliac disease. *Archives of Disease in Childhood* **65**: 909–911.

COFACTORS

Contents
Inorganic
Organic

Inorganic

E D Harris, Texas A&M University, College Station, TX, USA

Introduction

In the next article, a cofactor is defined as any nonenzyme component that promotes the catalytic prowess of an enzyme. The definition emphasizes function rather than structure. Since nearly a third of all enzymes require metal ions for catalytic function, it is apparent that inorganic components make up a substantial number of cofactors. Most of the trace metals have a common denominator in their intimate involvement with enzymes. Many are active site components that bind substrates, accept electrons, stabilize tertiary and quaternary structures, or even regulate the pace of metabolic pathways. In this article, the individual metal ion cofactors are discussed.

History

The nutritional history of the mineral elements, unlike the vitamins, had little early focus on humans. Rather, it was domestic livestock eating forage from mineral-poor soils that exhibited deficiency symptoms (thought at first to be due to toxicity). Typical signs were the crimping of wool in sheep, aortic rupture in pigs and cattle, and loss of myelin in brains of newborn lambs. Symptoms were lessened sharply by supplementing the feed with salts of metal ions such as $CuSO_4$, $Fe(NH_4)_2(SO_4)_2$, and $ZnCl_2$. Reversing symptoms and reestablishing optimal growth to livestock provided the first evidence for essential metals. In time, biochemical studies led to the isolation of enzymes that required metal ions for function, and soon after that specific enzymes could be linked to the deficiency symptoms. Metal ion interactions were viewed as detrimental as well as valuable to the system. An early study by Hart *et al*. showed that copper potentiated the effects of iron in alleviating an anemia condition in laboratory rats fed milk-based diets. That observation was repeated in chicks and pigs and soon attracted the attention of clinicians who adopted a similar bimetal protocol in the treatment of anemic humans. Coupled with the advent of semipurified diets in that same era, the science of nutrition stood on the threshold of major discoveries as to the roles of the essential mineral elements.

General Properties

Mineral cofactors comprise a large group of inorganic substances the bulk of which are the metal ions. The domain of metal ions include macro metals, such as Na^+, K^+, Ca^{2+} and Mg^{2+}, trace metal ions, including Fe^{2+}, Zn^{2+}, Cu^{2+} and Mn^{2+}, and metalloids, such as Se, Si and B (**Table 1**). In seeking a reason for their necessity, one must realize that metal ions are suited to the task of executing dangerous chemical reactions on enzyme surfaces, reactions that would otherwise harm the more sensitive organic side chains of amino acids in an enzyme. For example, redox metals such as iron, manganese, and copper can accept electrons into their structure, hold them temporarily, and then donate them to oxygen, forming water as a way to dispose of the electrons safely. In essence, one should consider a metal cofactor as extending the

Table 1 Inorganic cofactors

Metal	*Common biological form or valence*
Iron	Fe^{2+}, Fe^{3+}
Zinc	Zn^{2+}
Copper	Cu^{+}, Cu^{2+}
Manganese	Mn^{2+}, Mn^{4+}
Cobalt	Co^{+}, Co^{2+}, Co^{3+}
Vanadium	VO^{2+}
Molybdenum	MoO_2^{2+}, MoO_4^{2-}
Nickel	Ni^{2+}
Selenium	Selenocysteine, selenomethionine
Silicon	$Si(OH)_4$, SiO_2
Potassium	K^+
Sodium	Na^+
Calcium	Ca^{2+}
Magnesium	Mg^{2+}
Boron	$B(OH)_3$

repertoire of catalytic functions available to and performed by enzymes.

Metal-Activated Enzymes versus Metalloenzymes

Enzymes that depend on metal ions as cofactors fall into two categories: metal-activated enzymes and metalloenzymes. As the name implies, metal-activated enzymes are prompted to greater catalytic activity by the presence of a mono- or divalent metal ion exterior to the protein (in the assay medium). The metal may activate the substrate (e.g., Mg^{2+} with ATP), engage the enzyme directly, or enter into equilibrium with the enzyme exploiting its ionic charge to render a more favorable substrate binding or catalytic environment. Therefore, metal-activated enzymes require the metal to be present in excess, perhaps 2–10 times more than the enzyme concentration. Because the metal cannot be bound in a more permanent way, metal-activated enzymes typically lose activity during purification. An example is pyruvate kinase, which has a specific requirement for K^+ and is inactivated by dialysis (diffusion through a semiporous membrane). Other examples of metal-activated enzymes are shown in **Table 2**.

Metalloenzymes, in contrast, have a metal cofactor bound firmly to a specific region on the protein surface. Some may even require more than one metal ion and in rare instances could be two different metals as, for example, in Cu_2,Zn_2 superoxide dismutase. With few exceptions, trace metals fit into the picture as cofactors for metalloenzymes. Fe, Zn, Cu, and Mn, referred to as first transition series metals, are the most common. Their counterparts, Mg, K, Ca, and Na, are not considered 'trace' and only in rare instances are these so-called macroelements strongly bound to the surface of enzymes. Tight binding precludes loss of the metal ion by dialysis or loss to weakly dissociating agents. Metalloenzymes, however, can lose their metal cofactor and hence be rendered inactive when treated with metal chelators that have a stronger binding affinity than the enzyme and out compete the enzyme protein for the metal ion. As prosthetic groups, metals in metalloenzymes have a stoichiometric relationship (metal ion–enzyme protein ratio) represented by a whole integer. Metalloenzymes seldom are primed to greater activity by adding its conjugate metal ion to the enzyme. Spatial geometry is also a concern. Metals in the first transition series metals (Mn, Fe, Co, Ni, Cu, Zn) must adhere to strict geometric configurations around the metal-binding site. Examples of the more common geometrical arrangements are shown in **Figure 1**. For metals in the first transition series one takes note of the 3*d* and 4*s* orbitals in assigning valence states and likely geometric shapes. Apart from those with Zn, enzymes with first transition series metals tend to be highly colorful; for example, the beautiful red color of hemoglobin (iron) or the blue color of ceruloplasmin (whose name means heavenly blue) associated with copper. **Table 3** gives some examples of metalloenzymes and the specific metal each requires.

Table 2 Metal-activated enzymes and metalloenzymes

Metal or metal cofactor	*Enzyme*	*Function*
Metal-activated enzymes		
K^+	Pyruvate kinase	Synthesize pyruvate
Mg^{2+}	Hexokinase	Phosphorylate glucose
	DNase	Cleave DNA
	RNase	Cleave RNA
	ATPase	Cleave ATP
Metalloenzymes		
Cu^{2+}, Zn^{2+}	Superoxide dismutase	Destroy superoxide anion
Fe	Catalase	Destroy H_2O_2
Zn	Alcohol dehydrogenase	Metabolize alcohol
	DNA polymerase	Synthesize DNA
Mn	Pyruvate carboxylase	Synthesize oxaloacetate
	Arginase	Synthesize urea
Ca	Alpha amylase	Cleave glycogen, starch

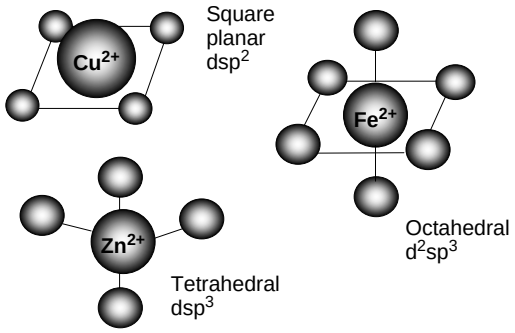


Figure 1 Some common geometries of metal complexes.

Individual Metal Cofactors

Iron

Most iron enzymes engage iron either as heme or as a special arrangement of iron with sulfur groups referred to as iron-sulfur centers (Fe_nS_n). Iron in heme bears a striking resemblance to the magnesium ion in chlorophyll (**Figure 2**). Heme, which is basically a porphyrin ring system with iron positioned in the center, is the most common form of iron in

Table 3 Important iron enzymes

Enzyme	*Source*	*Function*	*Form of iron*
Cytochrome *c* oxidase	Mitochondria	Electron transport	Heme
Aconitase	Mitochondria	Krebs cycle	Fe_4S_4
Succinate dehydrogenase	Mitochondria	Krebs cycle	Fe_4S_4
Catalase	Peroxisomes	H_2O_2 destruction	Heme
Peroxidase	Peroxisomes	Peroxide destruction	Heme
Prolyl hydroxylase	Cytosol	Collagen synthesis	Fe^{2+}
Ribonucleotide reductase	Cytosol	DNA synthesis	Fe-O-Fe
Cytochrome P450	Microsomes	Sterol synthesis	Heme

biological proteins. In cytochrome *c*, a common heme protein in the mitochondria, the axial ligands to the iron are occupied by histidine and methionine from the protein. Heme enzymes include calalase and peroxidase. As components of iron-sulfur centers, iron enters into multiple cluster arrangements with cysteine residues on enzymes that offer a more direct contact with the protein. These centers differ in their complexity from the simple 2Fe-2S to the more elaborate 4Fe-4S (**Figure 3**). Iron in these centers binds substrates as well as transfer electrons and takes part in reactions involving dehydrations and rearrangements. Enzymes with iron-sulfur centers include xanthine oxidase, succinate dehydrogenase, aconitase, and nitrogenase. A third class, represented by ribonucleotide reductase, has a FeO_2 cluster with a dioxygen as a peroxide anion O_2^{2-} straddled between two iron centers (**Figure 4**). This arrangement allows the enzyme to remove a hydrogen atom from a very stable C—H bond. No metal can replace iron in these complexes. Enzymes with a heme group generally are reddish-brown in color (depending on the oxidation state of the iron). The color led to early interest in these proteins and was the motivating factor behind naming heme proteins in the mitochondria 'cytochromes.' Although only a relatively few soluble enzymes have iron as a cofactor, iron is especially prominent in membrane-bound proteins that comprise electron transport pathways. Examples of the latter include the cytochromes in the mitochondria, endoplasmic reticulum, and photosystem I, II in chloroplasts. Perhaps the most unusual iron protein is ferritin, a huge multisubunit iron storage protein that has the capacity to bind more than 2500 iron atoms in its structure.

Reactivity The redox property of iron carries over to much of its chemistry as a cofactor. Iron is nearly always involved with the transfer of electrons and many times donates the electrons to a molecule of oxygen. Two important properties that fit that role are: (1) an iron atom that can readily undergo reversible valence changes from Fe^{2+} to Fe^{3+}, which allows facile exchange of electrons; and (2) the ferrous-ferric ion pair has a relatively low electrochemical potential (−0.1 V), which allows iron to be on the high (reducing) end of an electron transport chain. In cytochrome P450 a single oxygen atom is transferred to the substrate after O_2 binds to Fe(II). In the mechanism the Fe(II)−O_2 complex is converted into FeO, which features an Fe(V) species that attacks the substrate and incorporates the single oxygen atom into its structure. Although higher valence states such as Fe(IV) and Fe(VI) are formed by removing additional 3*d* electrons, only rarely are these higher valences of iron seen in biological

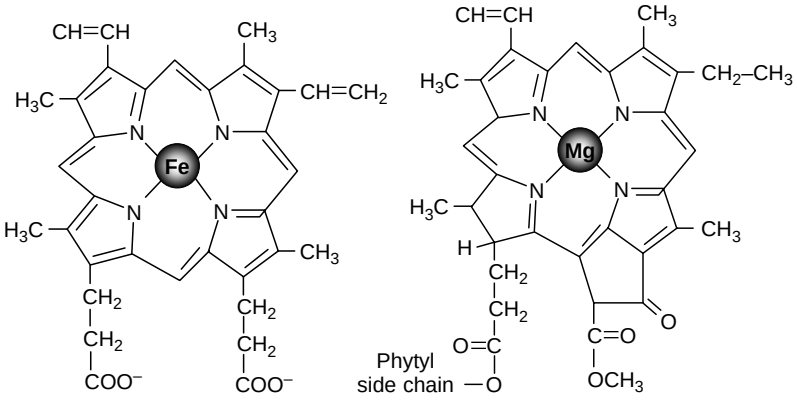


Figure 2 Heme iron in hemoglobin. Heme is a porphyrin ring with iron in the center. Four heme b groups are present in hemoglobin, the iron protein in erythrocytes. A similar structural arrangement is seen with magnesium in chlorophyll a from plants.

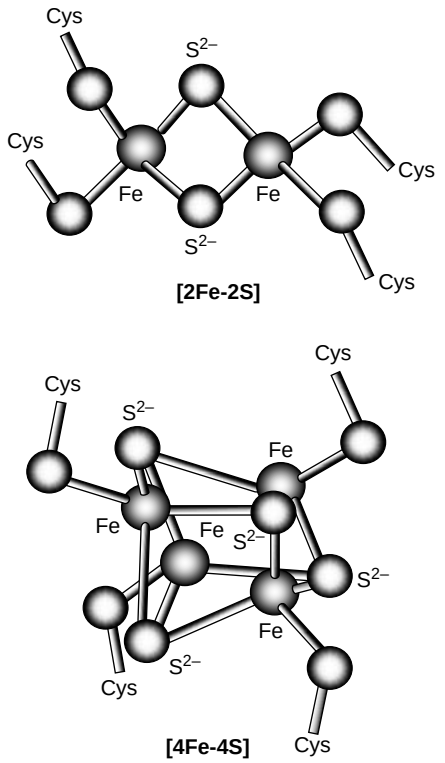


Figure 3 Iron-sulfur clusters. Both the Fe_2S_2 and Fe_4S_4 clusters are bound to the protein via cysteine residues. The iron in these complexes either engages a substrate or holds and passes electrons.

systems. As noted above, catalase and peroxidase, two heme enzymes, use iron to engage dangerous oxidants. Both enzymes are located in the cytosol and in peroxisomes where harmful oxidation reactions occur during the course of normal metabolic events. Perhaps the most familiar iron-containing enzyme is cytochrome *c* oxidase, the terminal electron acceptor in the mitochondrial electron transport chain and the enzyme capable of splitting a molecule of oxygen to form water.

$$4 \text{ cytochrome } c\ (Fe^{2+}) + O_2 + 4H^+ \rightarrow 4 \text{ cytochrome c } (Fe^{3+}) + 2H_2O$$

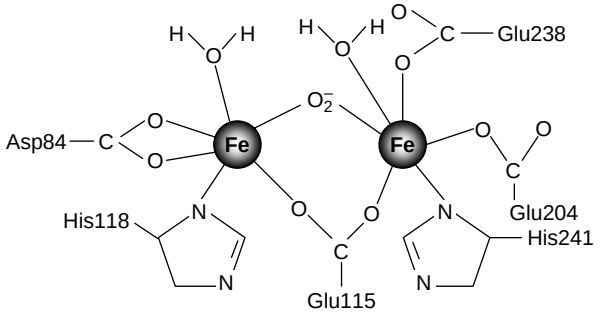


Figure 4 Fe-O-Fe center in ribonucleotide reductase. The two iron atoms are in close juxtaposition to bind dioxygen as the peroxide-anion O_2^{2-}. Side chains of aspartic and glutamic acid residues as well as two histidine residues assist in linking the center to the protein. The center assists in the formation of a free radical that forms on a neighboring tyrosine residue. (After Fraústo da Silva JJR and Williams RJP (1991) *The Biological Chemistry of the Elements. The Inorganic Chemistry of Life.* Oxford: Oxford University Press.)

Zinc

Zinc is perhaps the most ubiquitous and versatile of all metal cofactors. More than 300 enzymes have a zinc cofactor. **Table 4** lists some of the important zinc enzymes. Zinc-binding proteins that engage DNA, the so-called zinc finger proteins, attest to the versatility of zinc in biological systems. Approximately 3% of the genome of mammals codes for zinc finger protein. As a cofactor, zinc can perform both structural and catalytic functions. In carbonic anhydrase, for example, zinc enters into a coordinate bond with the CO_2 substrate (**Figure 5**). In carboxypeptidase, zinc takes an active part in the cleavage of the peptide bond (**Figure 6**). Multisubunit enzymes such as aspartate transcarbamylase use zinc to coordinate the positions of the catalytic and regulatory subunits, a structural role. Cu_2,Zn_2 superoxide dismutase requires zinc to position the copper atom in the channel accessed by the substrate $HO^{\bullet}_2$, another structural role. In zinc finger proteins, Zn^{2+} contributes to the stability of the loop structure that contacts the major and minor grooves of DNA. These examples illustrate why zinc is an important companion to enzymes and proteins.

Reactivity Zinc is considered a bland metal because it behaves as a divalent cation with no special geometric preference. It is perhaps this blandness that allows zinc to adapt to so many different enzyme environments. Zinc exists in one

Table 4 Important zinc enzymes

Enzyme	*Source*	*Function*	*Zn/ protein*
Alcohol dehydrogenase	Liver	Alcohol metabolism	4
Alkaline phosphatase	Placenta	Unknown	4
Carbonic anhydrase	Erythrocyte	CO_2 hydration	1
Carboxypeptidase	Pancreas	Protein catabolism	1
Glutamate dehydrogenase	Liver	Glutamate synthesis	2–6
Leucine aminopeptidase	Intestine	Peptide catabolism	4–6

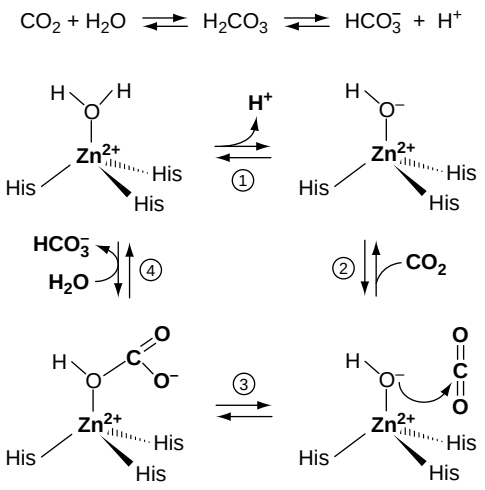


Figure 5 Zinc in carbonic anhydrase. Zinc in the enzyme 'activates' a water molecule (1) creating a better nucleophile to attack the CO_2 (2). Once formed (3) the hydrated CO_2 as HCO_3^- is displaced from the enzyme via a second water molecule (4) regenerating the active enzyme.

valence state, Zn^{2+}, and hence has no redox properties. The Zn^{2+} ion is configured as a $3d^{10,}$ which denotes a filled $3d$ orbital. For that reason, zinc complexes lack color and zinc itself behaves mostly as a cation. Zn^{2+} is a good electron acceptor (Lewis acid) that can enter into a coordinate bonding arrangement that polarizes groups to which it binds. This property allows zinc to increase the susceptibility of a chemical bond to attack. For example, Zn^{2+} polarizes water:

$$Zn^{2+} + H_2O \rightleftarrows Zn^{2+}(OH^-) + H^+$$

This makes the water behave more like a hydroxide ion and be more effective in attacking the CO_2 to form HCO_3^- in the reaction catalyzed by carbonic anhydrase. Another example is the use of zinc to polarize the ester or amide bonds thus promoting nucleophilic attack of water on the bond as in reactions catalyzed by carboxypeptidase and aminopeptidase.

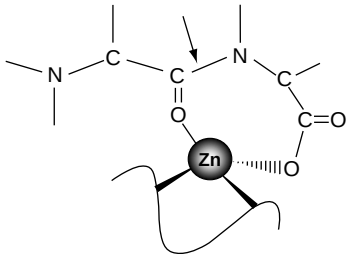


Figure 6 Zinc in carboxypeptidase. In carboxypeptidase, the zinc atom forms a binary complex with groups on the C-terminal end of the protein. Arrow shows bond that will be cleaved with water. Only the C-terminus residue is released from the protein.

Copper

Copper, like iron, is a redox metal. Like iron, copper exists in multiple valence states; Cu^+ and Cu^{2+}(cuprous and cupric) are the most common. Copper enzymes, while not nearly as numerous as zinc, fill important biological functions, mostly within the cytosol. Many fit the category of oxidoreductases, or more specifically 'oxidases,' meaning they catalyze reactions in which electrons from the substrate are transferred to O_2. Copper enzymes can be simple or complex, depending on the number of Cu atoms in the enzyme. Simple enzymes generally contain one Cu per subunit. The more complex enzymes include the multicopper oxidases, which may have as few as four, e.g., laccase, or as many as eight copper atoms per enzyme, e.g., dopamine-β-monooxygenase. Copper in these enzymes exists in three different chemical environments referred to as type 1, type2, and type 3 copper sites. Ceruloplasmin, for example, contains 6–7 Cu atoms in three distinct sites. The type 1 copper site gives a blue color to ceruloplasmin and other blue copper proteins. The copper-binding sites in a multicopper oxidase form a triad consisting of one type two and two type 3 coppers arranged as an isosceles triangle. Oxygen binds to the two type 3 coppers at the base of the triangle. Examples of copper enzymes include cytochrome *c* oxidase, lysyl oxidase, and ascorbate oxidase (**Table 5**).

Reactivity Because it is prone to accept electrons, copper is a powerful oxidant in biological systems. The copper sites in ceruloplasmin have the capacity to oxidize Fe^{2+} to Fe^{3+}, which prepares ferric ions to bind to transferrin and deliver iron to the organs and tissues. This reaction links iron with copper metabolism and could explain how an absence of copper in the diet impairs the transport of iron and causes anemia in humans. In Cu_2, Zn_2 superoxide dismutase, the Cu^{2+} at the active site removes the single nonbonding electron from one superoxide anion (O_2^-) and transfers it to another:

$$O_2^- + O_2^- + 2H^+ \rightarrow H_2O_2 + O_2$$

Seldom is copper destined to perform only a structural role and many enzymes that possess copper as a cofactor use the metal at the active site. More recent studies have linked copper ions with the formation of blood vessels or angiogenesis. One of the more exciting discoveries yet to be fully understood is that depriving an animal or human of copper delays or even inhibits the growth of cancerous tumors. From a nutritional perspective, this could

Table 5 Important copper enzymes

Enzyme	*Source*	*Function*	*Cu/protein*
Ascorbate oxidase	Squash	Ascorbate catabolism	8
Ceruloplasmin	Plasma	Iron oxidation	6–7
Cytochrome *c* oxidase	Mitochondria	Electron transport	2
Dopamine-β-monooxygenase	Adrenal	Noradrenaline synthesis	8
Lysyl oxidase	Aorta	Collagen, elastin synthesis	1
Superoxide dismutase	Erythrocyte	Superoxide radical destruction	2

mean that copper is essential for the development of the microvascular system.

Manganese

Whereas zinc may be the most common transition metal in enzymes, manganese is perhaps the least common. Part of the reason is that complexes of manganese with proteins tend to be weakly stable and dissociate readily. Notable manganese metalloenzymes include pyruvate carboxylase and manganese superoxide dismutase in the mitochondria and arginase in the urea cycle. Manganese can also function as a metal-activating cofactor for many enzymes that require magnesium.

Reactivity Although manganese is not considered a redox metal based on reactivity, it nonetheless can exist in six oxidation states (Mn^{2+} to Mn^{7+}), three of which (Mn^{5+} to Mn^{7+}) are not seen in biological systems. The most common form of manganese is Mn^{2+}. The highest number of multiple valences of manganese occurs in the water splitting enzyme that is found in chloroplasts of plants as part of photosystem II.

Cobalt

The role of cobalt as a cofactor is limited to its presence in vitamin B_{12}. Cobalt can exist in three valence states, Co^{+}, Co^{2+}, and Co^{3+} with Co^{2+} being the most common in 5′ deoxyadenosylcobalamin, the familiar form of vitamin B_{12} coenzyme. Cobalt is bound by a planar ring system analogous to heme but with very special features (see 00059). Cobalt and nickel are ions that may have figured more prominently in primitive systems when the atmosphere contained H_2 and CH_4 as common environmental gases. The argument has been made that as biological systems gradually adapted to O_2 the necessity for these two metals became less.

Reactivity Cobalt in the structure of vitamin B_{12} resembles iron in heme by being bound in a square planar arrangement to a ring (corrin). Unlike heme, however, cobalt has two axial ligands that are free from the protein, which allows nonprotein groups to access the central metal from above and below the plane. In the octahedral complex, one axial position (the fifth coordinate) is normally occupied by a benzimidazole and the other by a methyl group (as in methyl cobalamin). The arrangement is unique and allows cobalt to form carbon–metal bonds with the potential for two different reactivities. The methyl group, for example, may be removed as a carbonium ion retaining both electrons on the cobalt, which then reverts to a less stable Co(I). This is typical of the reaction in which vitamin B_{12} acts as a methyl donor. In positional rearrangements, cobalt retains only one electron and forms a stable Co(II) or d^7 ion with the release of a free radical. Free radicals are highly reactive and overcome energy barriers that would stymie other reactants. Thus, cobalt's chemical properties transfer groups as carbonium ions or highly reactive carbon-centered radicals. Both products are possible and hence explains the necessity for cobalt as a cofactor for a reaction that proceeds via a free radical mechanism. An example of the latter is the intramolecular rearrangement of methylmalonyl-CoA to succinyl-CoA as catalyzed by methylmalonyl-CoA mutase.

Vanadium

Although a defined biochemical function for vanadium in higher animals and humans is yet to be described, recent reports of vanadium in bacteria and algae have provided clues as to the functional necessity of this metal in enzyme catalysis. About 10 years ago, vanadium was found to be essential for the activity of bromoperoxidase, an enzyme found in brown and red algae. Shortly thereafter, a vanadium-dependent iodoperoxidase was characterized. Vanadium was also found in high concentrations in mushrooms and was shown to accumulate in large quantities in ascidians, specifically the blood cells (vanocytes) of these organisms. Speculation as to the function of vanadium in microorganisms ranges from antimicrobial action to electron transfer and the trapping of oxygen. In higher animals, however, vanadium has been shown to have insulin-mimetic properties and to stimulate cell proliferation and

differentiation. It is also believed to regulate phosphorylation and dephosphorylation reactions through control of ATPases, phosphatases, and adenylcyclase, which have widespread effects on cell functions. These are the most plausible theories to date. However, it should be emphasized that vanadium has not been shown to be a specific activator (or inhibitor) of any enzyme in humans.

Reactivity Vanadium is like molybdenum (see below) in being able to form both oxyanions and oxycations, VO_4^{2-} (MoO_4^{2-}), VO_2^+ (MoO^{2+}, MoO_2^+, and MoO_2^{2+}), as well as sulfur centers, e.g., VS_4^{3-} (MoS_4^{2-}). Vanadate differs from molybdate in being a rather strong oxidizing agent ($E \sim +0.5\,V$ at pH 7), which may relate to its electron transfer function in lower life forms but has questionable significance in humans.

Calcium

Calcium is a cofactor for a limited number of important enzymes apart from the more familiar actin–myosin complex in muscle; alpha amylase and thermolysin are two of the most familiar. As a free ion or working through calmodulin, calcium is better understood as an activator of enzymes in hormone-dependent cell signaling pathways. Enzymes that have been referred to as Ca-ATPases and H^+/Ca-ATPases should not be mistaken for calcium-dependent enzymes. This is a misnomer in that the Ca^{2+} is the object of the enzyme's action rather than the cofactor for activity. The ATPases comprise a large group of membrane-bound enzymes that either pump Ca^{2+} from the cytosol into the endoplasmic reticulum or expel calcium from the cell through membrane channels.

Reactivity As a group IIa metal, calcium is limited to a 2+ valence state and serves primarily as a divalent cation in its interactions with enzymes. The role of Ca^{2+} is limited mainly to structure stability.

Molybdenum

Molybdenum is widely distributed in plants and animals. The metal exists in three valence states: Mo^{4+}, Mo^{5+}, and Mo^{6+}. A limited number of redox reactions exploit the multivalence states. Molybdenum-dependent enzymes are found in pathways that metabolize purines, pyrimidines, pterins, aldehydes, and sulfites. A cofactor structure for molybdenum has been proposed (**Figure 7**) and is referred to as molybdopterrin. Enzymes that use the cofactor include xanthine oxidase, sulfite oxidase, and aldehyde oxidase. In microorganisms, molybdenum is a key metal for the fixation of nitrogen. Xanthine oxidase is the enzyme with importance relevance to a mammalian system.

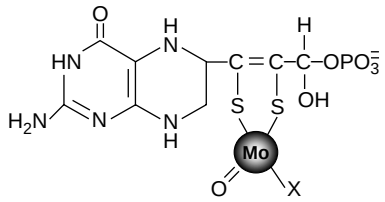


Figure 7 Proposed structure for the molybdenum cofactor in nitrogenase. This center consists of a special pterin cofactor, a relative of tetrahydrofolate. The molybdenum engages two sulfur atoms as a dithiolate complex.

Reactivity A major nutritional concern of molybdenum is its ability to antagonize copper. Indiscriminant spraying of soils with molybdenum has been shown to affect the growth and productivity of ruminants. The effect relates to the formation of thiomolybdates in the rumen. The thiomolybdates interact and bind copper preventing its absorption from the rumen. Thiomolybdates have a very high affinity for copper almost to the exclusion of other metal ions. Lately, thiomolybdates have been used to control copper toxicity in Wilson's disease, a genetic disease of copper poisoning in humans.

Nickel

As a cofactor, nickel occurs infrequently. About the only known occurrence of nickel is in microbial and plant enzymes such as urease from jack bean, soy bean, rice, and tomatoes. There are roughly two gram-atoms of nickel per mole of the 96 000 Da subunits of the enzyme. Other metalloenzymes containing nickel include Factor F430 found in the membrane of methanogenic bacteria, carbon monoxide dehydrogenase, and hydrogenases I and II. Nickel has drawn the attention of nutritionists because of the observation that nickel concentrations in the serum of women rise sharply immediately after parturition.

Reactivity Some consider nickel the 'metal that was.' As biosystems evolved and moved from an atmosphere of no oxygen to one rich in oxygen, where methane and H_2 have tended to be minimized as energy substrates, metals that formed a major cofactor in the anaerobic environment and were used by the more primitive organisms such as archaebacteria have been replaced in favor of a metal or cofactor more suitable to the present day environment. Thus, nickel, like cobalt, may

have had its greatest era in enzymes that catabolized CH_4 or H_2.

Other

The sodium ion is generally not considered a specific cofactor because one has yet to demonstrate an enzyme whose catalysis depends strictly on sodium ions. Sodium-activated enzymes often respond to surrogate metal cofactors such as Li^+ or even divalent cations. The magnesium ion is required by a large number of enzymes referred to as kinases, enzymes that transfer the terminal phosphate group of ATP to a substrate. Kinase enzymes figure prominently in many biochemical pathways such as glycolysis (hexokinase, fructose-6-phosphate kinase, pyruvate kinase), hormone responses mediated by cyclic AMP, cell signaling, and regulation of cell division. The potassium ion is a specific cofactor for pyruvate kinase in the glycolysis pathway. Both potassium and magnesium form no permanent bonds with their respective enzymes and hence act more in the capacity of activators.

Although chromium, tin, arsenic and strontium have been postulated by some investigators to be essential for optimal growth and health of organisms, as well as having a positive influence on biological systems, cofactor functions for their ions have not been assigned because specific enzymes that may require them for activity have not been found.

Nonmetal Mineral Cofactors

Selenium

Selenium belongs to the category of redox nonmetals. Selenium is included in the same class as sulfur (sometimes referred to as metalloids), which implies that selenium should be able to substitute for sulfur in biological complexes. As a congener of sulfur, selenium becomes part of a protein's structure as selenocysteine and selenomethionine, not as a selenium atom ligated directly to the protein as a prosthetic group. The former are the active cofactors in selenium enzymes.

Reactivity Although a selenium ion is clearly capable of redox reactions, there is still little information available as to how selenium functions as a cofactor. Enzymes such as glutathione peroxidase are soluble enzymes that transfer electrons to and from substrates. Replacing the selenium with sulfur in the enzyme negates the activity. With only a few selenoenzymes available, there is little information as to the precise catalytic role of selenium. Glutathione peroxidase in the reduced (resting) form is believed to contain an ionized selenol that can react with either organic peroxides or H_2O_2 according to the reaction (a) below:

$$\text{(a) Enz-Se}^- + \text{ROOH} \rightarrow \text{Enz-SeOH} + \text{ROH}$$

$$\text{(b) Enz-Se}^- + \text{RI}_4 \rightarrow \text{Enz-SeI} + \text{RI}_3$$

A selenol enzyme is also believed to be an intermediate in the reaction (b) catalyzed by 5′ deiodinase, the enzyme that catalyzes the removal of iodine from thyroid hormone. The enzyme–selenenic acid complex (Enz-SeOH) is regenerated by reduced glutathione (GSH), which forms a mixed selenide sulfide intermediate (Enz-Se-S-G). This intermediate then reacts with a second GSH to restore $Enz\text{-}Se^-$ and releases oxidized glutathione (GSSG) as a product. Regeneration of the Enz-SeI of 5′ deiodinase also requires a reducing agent whose identity is still uncertain although dithiothreitol can perform the reduction *in vitro*.

Silicon

There is still some question as to whether silicon is a cofactor. It is included here because of the importance of silicon in a number of biochemical reactions leading to the synthesis of glycoproteins and polysaccharides in the extracellular matrix of connective tissue ground substance. Silicon as $Si(OH)_4$ is very abundant in soils and minerals and is as common in human tissues as magnesium. In plants, especially grasses, silicon is a major component of a mineral skeleton and has a metabolic turnover nearly on a par with carbon. In humans, the highest concentrations of silicon occur in connective tissues such as aorta, trachea, tendon, bone, and skin. Lesser amounts are found in liver, heart, and muscle. The epidermis and hair are significantly high in silicon.

Reactivity Silicon, as silicic acid, has been shown to be required for maximal activity of prolyl hydroxylase, the enzyme that converts proline residues to hydroxyproline in collagen. High levels (0.2–2.0 mM) are needed to stimulate the enzyme, which catalyzes a rate-determining factor in collagen biosynthesis.

Boron

Manipulating the boron content of a diet leads to a wide number of metabolic responses, which is testament to the potential importance of boron in human nutrition. Early studies reported increased levels of steroid hormones, testosterone, and estradiol in animals supplemented with boron. Further studies suggest that boron has a regulatory role in

the metabolism of other minerals such as calcium and may affect bone metabolism. In a comparative way the role of boron is well established in vascular plants, diatoms, and marine algal flagellates. Zebra fish deprived of boron tend to suffer developmental defects. These data have prompted investigations into the biological functions of boron in higher vertebrates. To date, however, few studies have supported boron's essential role in vertebrates. In a comparison to Zebra fish, pregnant rats fed one-fiftieth the level of boron as control rats exhibited no impairment in fetal growth or development. Fewer two-cell embryos from the deficient rats, however, reached the blastocyst stage when cultured *in vitro*, suggesting boron deprivation did have an impact at a very early stage of development. Perhaps the strongest holdup to accepting boron as essential is the failure to define and link a specific organoboron compound with a physiological function. A report of boron associated with a naturally occurring antibiotic is an exception. The data, however, tend to support the notion that boron complexes with biological components are too unstable to be isolated and studied. This clearly has put a damper on the forward thrust of establishing boron's precise metabolic function.

Conclusions

The mineral cofactors described above may be thought of as representing a special subset of the biominerals. Rather than contributing to skeletal mass and fluid homeostasis, however, mineral cofactors are subtler and are devoted specifically to enzymes. The words 'mineral' and 'cofactor' combine to designate an inorganic component required by an enzyme in order to achieve optimum catalytic efficiency. In seeking a reason for mineral confactors, one must consider that to meet its functional obligations, an enzyme faces many challenges. The protein surface can easily be modified chemically through interaction with substrates and the enzyme protein can readily lose its biological form through denaturation. Electrons and groups that are transferred to and from substrates have the potential to permanently modify the enzyme. This happens frequently and instead of undergoing repair, old enzymes are replaced by new ones. The mineral cofactors fit into the daily wear and tear by making the enzyme better able to stand up to the harsh environment of their existence. They also have been shown to be effective binders of substrate and to interact with oxidants and reductants in a facile manner. Some trace metals such as zinc can accept electron pairs in forming a covalent attachment that polarizes and facilitates rupture of the chemical bonds in the substrate. Other metals such as copper and iron can accept electrons from the substrate and pass them to oxygen. Catalysis and structure stability are the two primary functions of metals in enzymes. Many organic factors serve as electron-capturing and group-transferring agents (see 00059). This suggests that metalloenzymes may back up enzymes with organic cofactors. This view is rather narrow and oversimplified since there are many enzyme-catalyzed reactions where only a metal will suffice, such as in the metalloenzymes that catalyze the destruction of oxygen radicals. In biology seldom does one factor become indispensable. What nutritionists refer to as essential metals are on the same level as vitamins in that they are needed in very small quantities to maintain the status quo in a system and, like vitamins, are available strictly through the diet. Therefore, one must conclude that essential minerals and vitamins have common ground in the enzymes, which they literally permit to function.

See also: **Calcium**. **Cofactors**: Organic. **Copper**. **Iron**. **Magnesium**. **Manganese**. **Potassium**. **Selenium**. **Zinc**: Physiology.

Further Reading

Berthon G (1995) *Handbook of Metal Ligand Interactions in Biological Fluids. Bioinorganic Medicine*, vols I and II. New York: Marcel Dekker.

Eichhorn GL (1973) *Inorganic Biochemistry*, vols 1 and 2. Amsterdam: Elsevier Scientific Publishing.

Fraústo da Silva JJR and Williams RJP (1991) *The Biological Chemistry of the Elements. The Inorganic Chemistry of Life*. Oxford: Oxford University Press.

Harris ED (2003) Basic and clinical aspects of copper. *Critical Reviews of Clinical and Laboratory Sciences* **40**: 547–586.

King TE, Mason HS, and Morrison M (1988) *Oxidases and Related Redox Systems: Progress in Clinical and Biological Research*, vol. 274. New York: Alan R. Liss.

Lanoue L, Taubeneck MW, Muniz J, Hanna LA, Strong PL, Murray FJ, Nielsen FH, Hunt CD, and Keen CL (1998) Assessing the effects of low boron diets on embryonic and fetal development in rodents using *in vitro* and *in vivo* model systems. *Biological Trace Element Research* **66**: 271–298.

Mertz W (1987) *Trace Elements in Human and Animal Nutrition*, 5th edn, vols I and II, New York: Academic Press.

Prasad AS (1993) *Essential and Toxic Trace Elements in Human Health and Disease: An Update*. New York: Wiley-Liss.

Stadtman TC (1996) Selenocyteine. *Annual Reviews of Biochemistry* **65**: 83–100.

Stryer L (1995) *Biochemistry*, 4th edn. New York: WH Freeman & Co.

Organic

E D Harris, Texas A&M University, College Station, TX, USA

Introduction

Cofactors are important accessories to biochemical processes. Generally present as small organic compounds or metal ions, cofactors empower enzymes to functional at maximal catalytic effectiveness or endurance. A related term, coenzymes, relates to a subgroup of cofactors whose structure in part is derived from water-soluble B vitamins. Historically, cofactors were often inadvertently removed during purification and had to be added back to restore enzyme activity. Today, we regard a cofactor as an obligatory component of the catalytic mechanism. Compounds meeting the criteria are either: (1) small organic molecules that bind directly to the enzyme surface forming an active site for the substrate to bind or interact, or assist in these events indirectly, or (2) inorganic ions that bind to specific groups on an enzyme surface and aid in substrate binding, catalysis, stabilizing the transition state, or contributing to the overall stability of the enzyme's structure. Practically speaking, any substance in an assay medium that promotes the catalytic activity or stability of an enzyme is a candidate for its cofactor.

As will be illustrated in this and was in the last article, cofactors are indispensable adducts of the catalytic machinery of the body and have provided nutritionists with the strongest insights into the essential role of vitamins and trace elements. It is still fashionable to consider coenzymes as vitamin derivatives that bind loosely to enzymes or serve as transient active sites. Cofactors and coenzymes are terms that are used interchangeably. It is important to note, however, that the prefix 'holo' is used to refer to an enzyme and its coenzyme together as a catalytic unit and 'apo' when the coenzyme is missing. Apoenzymes are functionless and are of no benefit to the organism.

History

Early studies of vitamins found that many, especially the water-soluble B vitamins, formed the nucleus of compounds that partook in enzyme catalysis. The discovery established a bridge between nutrition and the fledgling science of biochemistry. Indeed, many early biochemical investigations were devoted to learning the biological functions of essential nutrients, which included the vitamins. A general principle that emerged at the time was that a vitamin had to be changed to another compound in order to be metabolically functional. With the diet as the only source, it was possible to learn the specific effects of individual vitamins by omission studies. With deeper insights into biological processes, it was soon realized that canceling an enzyme in a critical biochemical pathway was behind many of the vitamin deficiency diseases such as beriberi, pellagra, and pernicious anemia. This put dramatic new emphasis on enzyme functions and the search for enzymes that depended on vitamins for function.

Cofactors in Biochemical Pathways

Table 1 lists vitamins and nonvitamins that are known to give rise to many of the organic cofactors in humans. **Figure 1** provides a glimpse into their importance by showing the location of organic cofactors in the biochemical pathway for oxidizing glucose and other biocompounds to CO_2 and H_2O. That overall reaction for glucose is:

$$C_6H_{12}O_6(\text{glucose}) + 6O_2 \rightarrow 6CO_2 + 6H_2O$$

One sees that at least seven distinct B vitamin-derived coenzymes are needed to complete the transition. Nicotine adenine dinucleotide (NAD^+), derived from niacin, is required for the oxidation of glucose to pyruvate and thiamine pyrophosphate (TPP) derived from the vitamin thiamine (sometimes written as thiamin), flavin adenine dinucleotide (FAD) from riboflavin, panthothenic acid from pantothene, and lipoic acid all take part in the oxidation of pyruvate to acetyl-coenzyme A in the middle stage. In addition, flavinmononucleotide (FMN) also from riboflavin and coenzyme Q from ubiquinone take part in completing the oxidation to CO_2 and H_2O in the oxidative-phosphorylation pathway in the mitochondria. All told, some 20 organic cofactors engage enzymes in the various biochemical pathways of humans. Below is a brief description of each coenzyme/cofactor. **Table 2** summarizes the list of key enzymes known to be associated with each coenzyme.

Specific Vitamins as Cofactors

Thiamine (Vitamin B_1)

Best known as the anti-beriberi factor and called at first simply vitamin B by McCollum, thiamine was shown to be involved in the decarboxylation of pyruvate to acetaldehyde in alcohol fermentation and was named 'cocarboxylase' in 1932.

Table 1 Vitamins and nonvitamin cofactors[a]

Name of vitamin[a]	*Related coenzymes*	*Biochemical function*
Thiamine, thiamin B_1	Thiamine-pyrophosphate	Carbonyl group transfer
Riboflavin B_2	FMN, FAD	Redox reactions
Niacin (nicotinamide) B_3	NAD, NADP	Redox reactions
Pantothenic acid B_5	Coenzyme A	Acyl group transfer
Pyridoxine B_6	Pyridoxal 5′ phosphate	Amine group transfer
Folic acid (folacin) B_9	Tetrahydrofolates	One-carbon transfer
Cobalamin B_{12}	5′ Deoxyadenosyl cobalamin	Methylation, rearrangement reactions
L-Ascorbic acid C	Dihydroascorbate	Collagen, adrenaline synthesis
Calciferol D	None	Calcium absorption
Tocophoral E	None	Antioxidant
Biotin H	Biocytin	CO_2 fixation
Phylloquinone K	None	Prothrombin synthesis
Bioflavonoids P	None	Antioxidant
Nonvitamin cofactors		
p-Aminobenzoate	Tetrahydrofolate	One-carbon transfer
α-Lipoic acid	None	Acetyl group transfer
Betaine	None	Methylating agent
Coenzyme Q	Ubiquinone	Electron transfer
PQQ	None	Oxidation reactions
Topaquinone	None	Oxidation reactions
Carnitine	None	Fatty acid transfer
Inositol	None	Membrane lipids
S-adenosyl methionine	None	Methylation reactions
Glutathione	None	Group transfer, anitoxidant
3′ Phosphoadenosine-5′ phosphosulfate	None	Sulfate esterification

[a]Although codified in vitamin literature at one time, B_4, B_{10}, and B_{11} have since been abandoned.

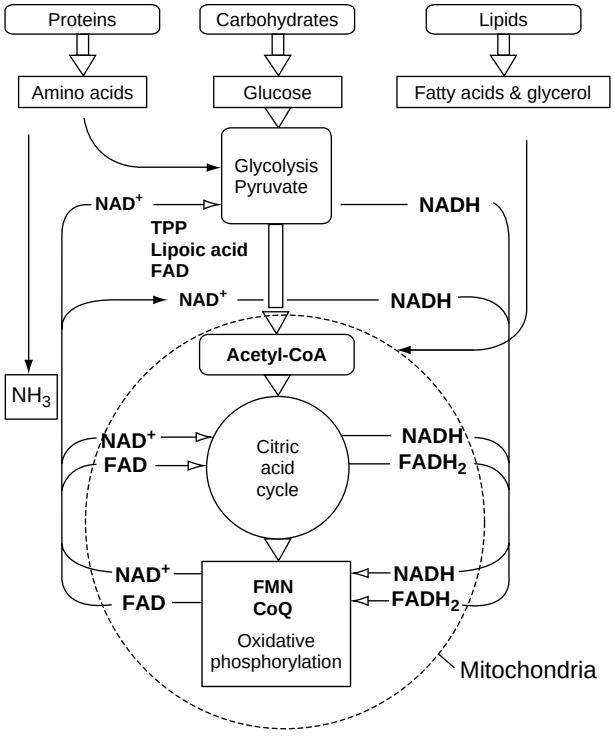


Figure 1 Occurrence of organic cofactors in energy metabolism. Only a few key intermediates in the pathway are shown. Dotted circle shows reactions taking place in the mitochondria. Note how coenzymes NAD^+ and FAD cycle between oxidized and reduced forms.

Confirmation of its structure as TPP came 5 years later. Its name is meant to signify a vitamin containing sulfur (*thios* in Greek).

Reactions

1. Pyruvate dehydrogenase complex in mitochondria.
2. α-Ketoglutarate dehydrogenase complex in mitochondria.
3. Branch chain dehydrogenase.
4. Transketolase reactions in pentose pathway and in reductive pentose pathway of photosynthesis.

Reactivity The structure of thiamine has two rings bridged by a methylene group as seen in **Figure 2A**. The coenzyme (TPP) arises via an ATP-dependent pyrophosphorylation of the primary alcohol group (**Figure 2B**). What may be called the active site of the coenzyme is the carbon in position 2 (C-2) of the smaller five-member thiazolium ring (arrow). A favorable positioning of C-2 between atoms of nitrogen and sulfur causes C-2 hydrogen to exchange protons with water, indicating C-2 can ionize to a carbanion. As a carbanion, C-2 is able to engage positive centers such as carbonyl carbons of α-keto acids and keto sugars. In the reaction with pyruvate, α-ketoglutarate, or branch-chain α-keto acids from valine, leucine, or isoleucine, a carboxyl group is

Table 2 Sample of enzymes associated with each of the coenzymes derived from vitamins

Coenzyme	*Enzyme*
1. Thiamine-pyrophosphate	Pyruvate dehydrogenase complex
	α-Ketoglutarate dehydrogenase complex
	Transketolase
	Branch chain dehydrogenase
2. NAD^+, NADH	Glyceraldehyde-3-PO_4 dehydrogenase
	Pyruvate dehydrogenase complex
	Alcohol dehydrogenase
	Lactate dehydrogenase
3. $NADP^+$, NADPH	Glucose-6-PO_4 dehydrogenase
	Glutamate dehydrogenase
	β-Ketoacyl-ACP synthase
4. FAD, $FADH_2$	Glucose-6-PO_4 dehydrogenase
	Succinate dehydrogenase
	Fatty acyl-CoA dehydrogenase
5. Pyridoxal-5′ phosphate	Aminotransferases
	Glycogen phosphorylase
6. Tetrahydrofolate	Glycine synthase
	Homocysteine methyltransferase
7. Biocytin	Pyruvate carboxylase
	Acetyl-CoA carboxylase
	Propionyl-CoA carboxylase
8. Coenzyme A (pantothenic acid)	Pyruvate dehydrogenase complex
	Acetyl-CoA carboxylase
	Citrate synthase
9. Cobalamin	Homocysteine methyltransferase
	Methylmalonyl-CoA mutase
10. L-Ascorbate	Prolyl and lysyl hydroxylase
	Dopamine-β-monooxygenase

expelled as CO_2 and the electrons remain with the 'active aldehyde' on the C-2 position. Attack on a keto sugar cleaves the first two carbons as a unit, which then attaches to C-2 as an 'active glycoaldehyde' adduct. Yeast disengage active aldehyde as acetaldehyde later to be reduced to ethanol by alcohol dehydrogenase. Bacteria convert 'active aldehyde' to acetyl-phosphate. In the mitochondria of higher organisms, however, active aldehyde is oxidized by an FAD-containing enzyme (part of the pyruvate dehydrogenase complex) and transferred to lipoic acid (see **Figure 10**), which transfers the highly energetic acetyl group to the thiol group of coenzyme A. As a coenzyme for transketolases in the pentose pathway, TPP takes part in the formation of ribose-5-phosphate, glyceraldehyde-3-phosphate, and erythrose-4-phosphate from sedohepulose-7-phosphate, xyulose-5-phosphate, and fructose-6-phosphate, respectively. Each sugar phosphate donates an 'active glycoaldehyde' to an aldose acceptor.

TPP is also the coenzyme for branch chain dehydrogenase, the enzyme that catalyzes the oxidative decarboxylation of α-keto acids derived from leucine, isoleucine, and valine, three essential amino acids. The reaction follows a scheme similar to pyuvate oxidation, only this time the carbon skeleton of the amino acid condenses with coenzyme A (CoA).

Riboflavin (Vitamin B_2)

The first hint that McCollum's vitamin B was in reality a multifactor complex came when yeast void of antineuritis activity still retained growth-stimulating activity. Originally called vitamin G, riboflavin was renamed vitamin B_2 when it was recognized to be part of the yeast B complex. The name riboflavin followed the discovery in 1935 of its association with green fluorescent pigment of whey. Today, we regard riboflavin and niacin as the two principal vitamins that give rise to coenzymes that function with enzymes known as oxidoreductases. Both coenzymes transport electrons to and from substrates and in so doing form oxidized or reduced products. The two are referred to as 'redox' (an abbreviation for oxidation-reduction) coenzymes for that reason. Riboflavin was identified as the biochemical compound that gave the color to Warburg's 'yellow enzyme,' glucose-6-phosphate dehydrogenase (G6PDH). G6PDH was observed to catalyze the transfer of electrons from nicotinamide adenine dinucleotide phosphate (NADPH) to methylene blue, a redox sensitive dye that lost color upon reduction, suggesting that riboflavin probably mediated the electron exchange. G6PDH is a key entrance point for glucose into the pentose pathway and a major contributor of NADPH for the biosynthesis of fatty acids and other fats. The reaction is:

$$\text{Glucose-6-phosphate} + 2\text{NADP}^+ + \text{H}_2\text{O} \rightarrow \text{ribose-5-phosphate} + \text{CO}_2 + 2\text{NADPH} + 2\text{H}^+$$

Riboflavin (**Figure 2C**) is associated with two coenzymes, FMN and FAD. FMN is formed by phosphorylating the primary alcohol on the sugar moiety of riboflavin, an ATP-dependent reaction. FAD results from a further condensation of FMN with the 5′ AMP moiety of ATP (**Figure 2D**). What may be considered the active site is the isoalloxazine ring, which can exist in both oxidized and reduced states depending on whether electron pairs are absent or present, respectively. Enzymes that contain FAD or FMN are referred to as flavoproteins. FMN is limited to the membrane proteins of the mitochondria electron transport system whereas FAD is found in both membrane-bound and soluble

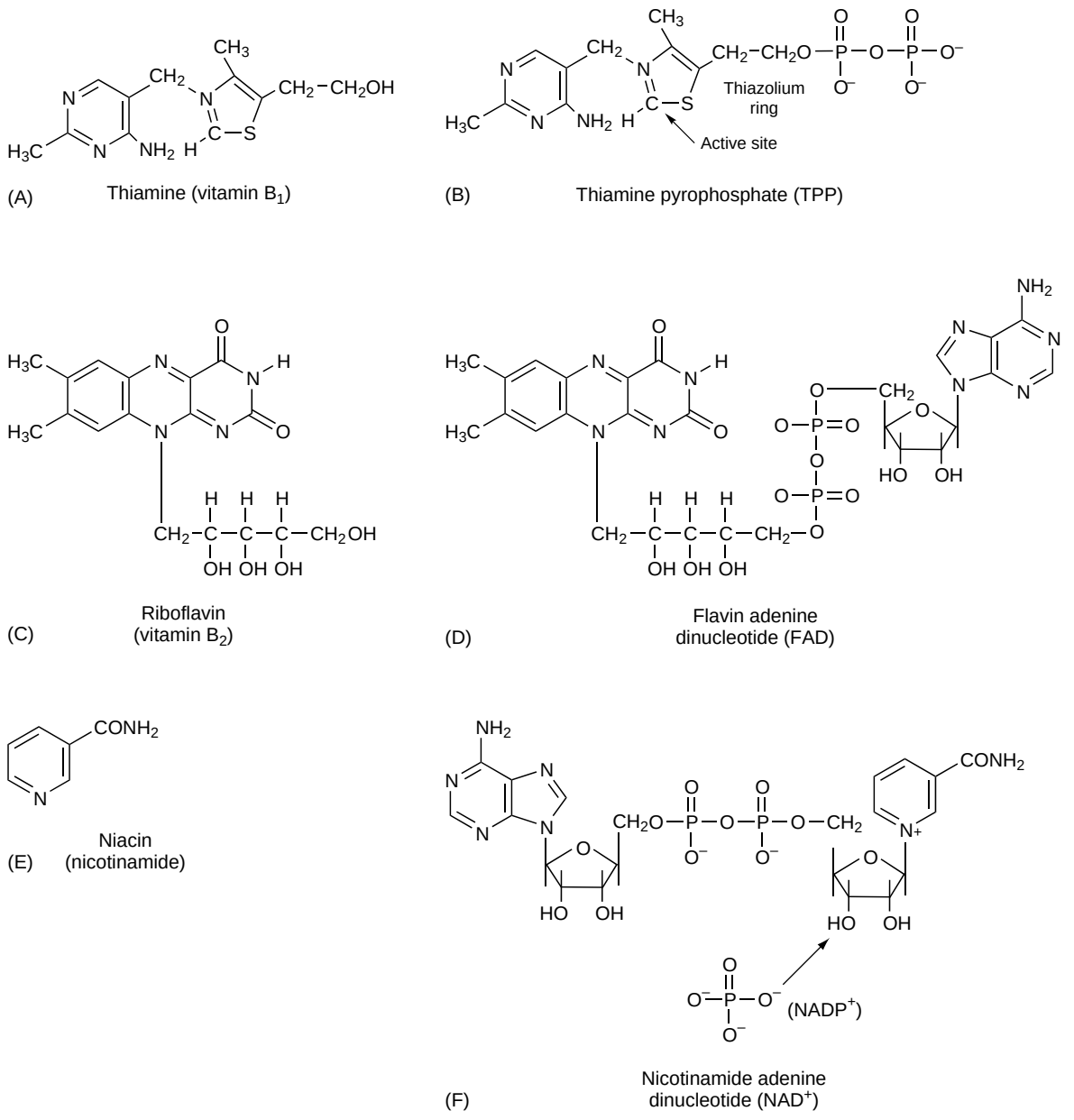


Figure 2 Structural relationship of vitamin-coenzyme for (A) thiamine (B_1), (C) riboflavin (B_2), and (E) niacin. In the left column is the structure of the vitamin, in the right column the coenzyme derived from the vitamin. Note the pyrophosphate group in the coenzyme derived from thiamine (B) and the prevalence of phosphate and adenyl groups in the coenzymes derived from riboflavin (D) and niacin (F), showing the necessity for ATP in their synthesis.

enzymes. The flavin cofactor is bound covalently to the structure preventing disengagement during purification procedures.

Reactions Flavin enzymes are designed to remove (and add) electrons to and from substrates. In general, flavin coenzymes are stronger oxidizing agents than the pyrimidine coenzymes (NAD^+, $NADP^+$) and tend to participate in more complex reactions. Also, flavin coenzymes can accept single electrons from a donor, forming a semiquinone and allowing flavoproteins to take part in reactions that form free radicals. Having a single electron also allows favins to bind molecular oxygen as a hydroperoxyl complex.

Niacin (Nicotinic acid, Nicotinamide)

Niacin presents an unusual twist in that its parent compound, nicotinic acid, had been known for about 70 years (ca. 1867) before its activity as a vitamin first become known (ca. 1937). If thiamine

(vitamin B_1) is the anti-beriberi factor, niacin (vitamin B_3) is the anti-pellagra factor. Pellagra is a disease characterized by a rash or dermatitis on areas of the skin exposed to sunlight as well as swelling in the legs from the knee on down and a painful flush and rash. Niacin (also called nicotinic acid or nicotinamide), its amide derivative (**Figure 2E**), is the active component of the second major redox coenzyme nicotinamide adenine dinucleotide and its phosphate (NAD^+ and $NADP^+$, respectively; **Figure 2F**). As with FMN and FAD, NAD^+ and $NADP^+$ arise by phosphorylation and condensation of the basic vitamin structure with ATP. NAD^+ differs from $NADP^+$ by having a phosphate group on the 3′ position of the pentose nearest to the adenine (**Figure 2F**). NAD^+ and NADH were discovered in dialyzable extracts of yeast, meaning these coenzymes readily disengaged from the enzyme that bound them. The ability to come on and come off an enzyme is fundamental to the electron delivery scheme shown in **Figure 1**.

Reactivity With few exceptions, the redox-related reactions of NAD^+ and $NADP^+$ are with dehydrogenase enzymes, i.e., enzymes that catalyze the removal and addition of electrons (as hydride ions) to substrates. A typical reaction in which NAD^+ is the oxidizing agent is the conversion of L-lactic acid to pyruvate:

$$\text{L-Lactate} + NAD^+ \rightleftarrows \text{pyruvate} + NADH + H^+$$

Therefore, the coenzyme is a major participant in energy-yielding catabolic reactions. $NADP^+$ performs less of a catabolic role, but its reduced form, NADPH, is a major reductant in anabolic reactions, especially the biosynthesis of fatty acids and other lipids.

The active site of both NAD^+ and $NADP^+$ is the nicotinamide ring. The oxidized form of nicotinamide has a quaternary nitrogen (four attaching bonds) that is written as a positive charge (**Figure 2F**). The oxidized ring accepts two electrons and one proton from a substrate (literally a hydride ion, H^-) reducing the ring and abolishing the positive charge on the nitrogen:

$$NAD^+ + 2e^- + 2H^+ \rightarrow NADH + H^+$$

$$NADP^+ + 2e^- + 2H^+ \rightarrow NADPH + H^+$$

More than 200 enzymes are known to catalyze reactions in which NAD^+ or $NADP^+$ accepts a hydride ion from a substrate. Moreover, there is a strong stereospecificity to this reaction. Addition of an H^- to the ring can either be in front (A-type) or in back (B-type), depending on the enzyme. Reducing the ring weakens it bonding to the enzyme and causes the NADH (NADPH) to dissociate and engage other cell components for the purpose of transferring the electrons.

NAD^+ is also a source of 5′ adenosine monophosphate (5′ AMP) in a limited series of activation or inhibition reactions. The transferred 5′ AMP becomes a leaving group for subsequent bond formation. In DNA ligase in bacteria, for example, the 5′ AMP is transferred to a lysine on the enzyme to form an unusual phosphoamide adduct that subsequently is transferred to one of the DNA strands. Attack by the 3′ hydroxyl group on the adjacent DNA strand releases the 5′ AMP concomitant with forming a phosphodiester bond and sealing the two DNA strands together. A related reaction, referred to as an ADP-ribosylation results in the nicotinamide group being split from the NAD^+ and the ribose moiety forming a covalent glycosylamine bond with a protein. ADP-ribosylation of sensitive proteins is one of the deadlier effects of bacteria toxins such as cholera toxin, diphtheria toxin, or pertussis toxin.

Pyridoxine (Vitamin B_6)

Vitamin B_6 was discovered in the 1930s and named pyridoxine because of its structural resemblance to pyridine (**Figure 3**). Pyridoxine's principal involvement is with a family of enzymes known collectively as amino transferases. These enzymes exchange amine groups from amino acids to α-keto acids. Familiar names include serum glutamate-oxalate transaminase (SGOT). The coenzyme, pyridoxal-5′

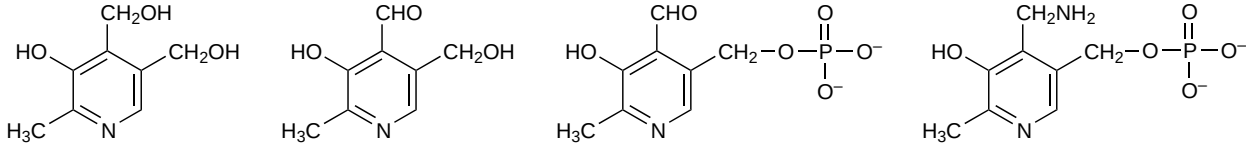


Figure 3 Pyridoxine (vitamin B_6) and its coenzymes. The coenzyme form is capable of receiving (pyridoxal) and donating (pyridoxamine) NH_3 groups.

phosphate (PLP), is the predominant form and is synthesized in a two-step reaction involving the oxidation of the hydroxymethyl group in the para position on the pyridine ring to an aldehyde, and the phosphorylation of the hydroxymethyl group on the 5 position. PLP is also a coenzyme for glycogen phosphorylase and ornithine decarboxylase. With tetrahydrofolate (see below), PLP takes part in serine to glycine interconversion. On the enzyme surface, the reactive species is not an aldehyde, but rather an aldamine formed by a Schiff base bond between the aldehyde and an ε-amino group of lysine in the active site.

Reactivity The reactivity of PLP is due to a number of features. First, the carbonyl (aldehyde) on the ring is positioned to engage the amino groups from amino acids and tether the amino acid to the enzyme. Through a series of electron rearrangements promoted by the PLP, the nitrogen on the amino acid substrate is disengaged and the carbon skeleton (an α-keto acid) is set free retaining the amine group on the coenzyme (pyridoxamine-5′ phosphate). The enzyme then binds a second α-keto acid and transfers the amino group generating a new amino acid and restoring the carbonyl function on the PLP. Other changes can also occur to a tethered structure. An electron rearrangement can result in the loss of a carboxyl group (as CO_2) or a molecule of H_2O. Thus, PLP enzymes also take part in decarboxylation and dehydration reactions. In the glycogen phosphorylase reaction, the phosphate group of the coenzyme acts as a general catalyst, promoting the attack of phosphate on the glycosidic bond of glycogen.

Folic Acid

Folic acid was first recognized as the yeast or liver factor that could cure a severe megoblastic anemia in chicks, monkeys, and humans. Later proof that the active substance was a growth factor for certain bacteria such as *Lactobacillus casei* and *Streptococcus faecalis* provided a rapid bioassay for isolating, identifying, and eventually synthesizing the vitamin and its coenzymes. The name folic acid was given in 1941 in recognition of its abundance in leafy green vegetables or 'foliage' and its structure was confirmed as monopterylglutamic acid in 1946. Today, we recognize folic acid as one of our most complex vitamin coenzymes because of its presence in many biochemical forms. Despite such enormous complexity, however, the biochemical role of folic acid narrows down to a specific set of synthetic reactions whose common denominator is one-carbon units.

The structure of folic acid (*N*-pteroyl-L-glutamic acid) can be pictured as a composite of three covalently linked molecules: a methylated pteridine ring attached to *p*-aminobenzoic acid (PABA), which in turn is linked via the carboxyl group to the α nitrogen of glutamic acid (**Figure 4A**). The coenzyme form is tetrahydrofolate (FH_4) formed in mammals by adding four electrons and four hydrogens to the pteridine ring (**Figure 4B**). The reduction is catalyzed by dihydrofolate reductase with NAPDH as the electron donor. The addition of one or more glutamic acid residues completes the structure. In the reductive step, a new asymmetric center is generated at C-6 and appears to be critical to the biological role since only one stereoisomer of this center

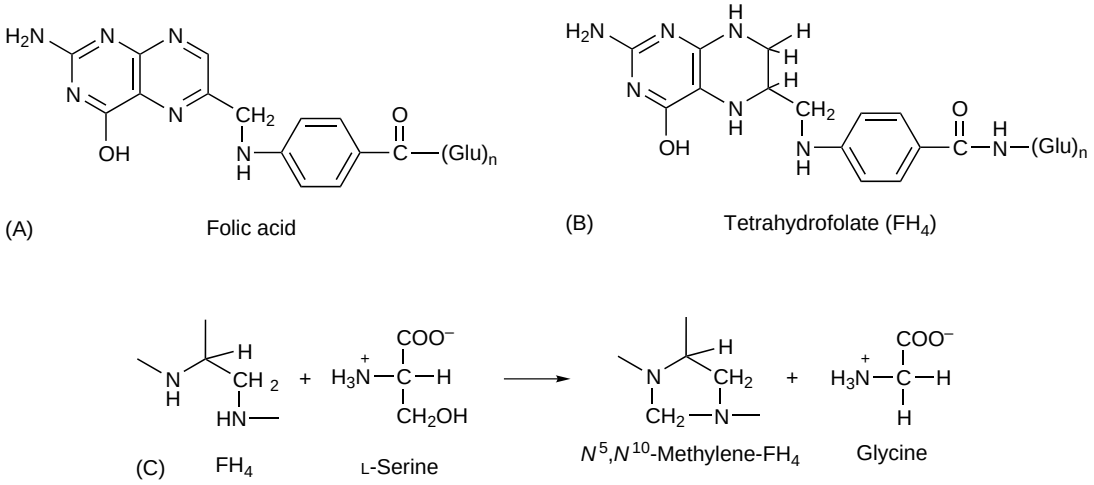


Figure 4 Folic acid (A) and its coenzyme form. Activation requires folic acid to be converted into its tetrahydrofolate derivative (B). (C) Specific function of one of the tetrahydrofolate derivatives, N^5,N^{10}-methylene-FH_4, in the synthesis of glycine from serine. Serine is thus able to donate a carbon to the coenzyme for a subsequent one-carbon transfer reaction.

is active. FH_4 may have up to seven glutamic acid residues and exist in many different chemical forms, most of which are interconvertible.

The basic reactions take part at the N^5 and N^{10} positions on the molecule, which serve as attachment points for one-carbon units in transit (**Figure 4C**). N^{10}-formyl- and N^5,N^{10}-methenyl-FH_4 are two synthetic forms that are biologically active. Complexes of N^5-formyl-FH_4 (folinic acid) transfer formyl groups to specific substrates. Active folic acid derivatives have carbon in the oxidation state of formate as well as formaldehyde (methylene) and a methyl derivative, N^5-methyl-FH_4, is known to take part in the enzymatic conversion of homocysteine to methionine. These observations reveal that the family of folic acid coenzymes is quite complex but all seem to involve the attachment of a single carbon atom to the substrate.

Reactivity Enzymes that require folic acid participate in what is referred to as 'one-carbon metabolism.' This takes the form of group transfers involving methyl groups, formyl groups, formimino groups, and methylene groups. Folic acid does not take part in acetylations or carboxylations. A typical reaction in higher vertebrates is the synthesis of glycine by the enzyme glycine synthase:

$$CO_2 + NH_4^+ + NADH + H^+ + N^5,$$
$$N^{10}\text{-methenyl-}FH_4 \rightarrow \text{glycine} + NAD^+ + FH_4$$

In perhaps its only major requirement as a methyl group donor, N^5-methyl-FH_4 is needed by the enzyme homocysteine methyltransferase to synthesize (regenerate) methionine from homocysteine. This reaction also uses a methylated derivative of vitamin B_{12} (see below) to mediate the group transfer. Another important reaction is the interconversion of serine and glycine. As shown in **Figure 4C**, the reaction requires the N^5,N^{10}-methylene-FH_4 derivative. Today, the list of folate-catalyzed reactions is quite large and includes one-carbon units in the synthesis of a purine ring of nucleic acids, methylation of DNA and RNA, thymidine biosynthesis, choline and *S*-adenosylmethionine biosynthesis, and histidine and tyrosine catabolism.

Biotin

Early interest in biotin involved the so-called egg white injury factor. When it was confirmed that egg white injury was caused by a deficiency and not a toxicity, pursuit of the missing substance led eventually to the discovery of biotin. Research on the vitamin brought a new concept to nutrition, that of the 'antivitamin' or substances capable of negating the action of vitamins before their use as cofactors. In the case of biotin, the 'antivitamin' turned out to be the protein avidin, which bound biotin tenaciously and limited its intestinal absorption.

Reactivity Biotin can be thought of as another one-carbon cofactor, but for biotin this is CO_2. Thus, biotin-requiring enzymes catalyze carboxylation, decarboxylation, or transcarboxylation reactions. The active form of biotin is 'biocytin' (ε-*N*-biotinyl L-lysine), which is formed by the covalent attachment of the biotin side chain to the ε-amino group of a lysine residue on the apoenzyme as catalyzed by a specific synthetase (**Figure 5B**). The condensation requires ATP and proceeds via a biotinyl-AMP intermediate with the apoenzyme catalyzing formation of the amide bond. The resulting unique structure combines the aliphatic chains or biotin and lysine permitting the ring structure of biotin to extend about 14 Å from the enzyme's surface (**Figure 5B**).

The active site on the biotinyl group is one of the *N* in the 5-member ring (**Figure 5C**). An *N*-carboxyl derivative serves as a donor of CO_2 to an appropriate substrate acceptor. The reaction occurs in two steps and requires an ATP-dependent formation of a carboxy biotinyl enzyme. If the enzyme is a carboxylase, there are two main substrate types: (1) acyl-CoA derivatives, which include acetyl-CoA,

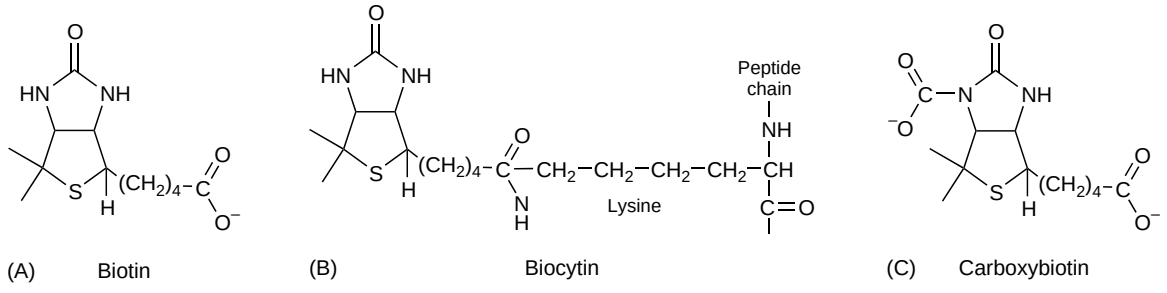


Figure 5 Biotin and its coenzyme. (A) Biotin as a vitamin, (B) biotin attached to the ε-amino group of lysine to form the coenzyme biocytin, and (C) the carboxy derivative of biotin prepared to donate CO_2 to a substrate.

propionyl CoA; and (2) simple α-keto acids such as pyruvate. Each substrate must contain a carbonyl group adjacent to or conjugated with the carbon receiving the carboxyl group from carboxy biocytin. Perhaps the most familiar biotin carboxylase enzymes in mammalian systems are acetyl-CoA carboxylase in fatty acid biosynthesis, propionyl CoA carboxylase in odd-chain fatty acid catabolism, pyruvate carboxylase in gluconeogenesis, and β-methylcrotonyl CoA carboxylase in leucine catabolism.

Pantothenic Acid

Pantothenic acid was named by Roger Williams who recognized its ubiquitous (greek, *pantothene*, from all quarters) occurrence in tissues of all organisms and all food sources. There are two twists to the story of its discovery. First is that the coenzyme form, CoA, was discovered long before the vitamin and, second that investigations of the substructure of CoA led to an understanding of how the coenzyme was synthesized. CoA was known to be a dialyzable cofactor essential for the acetylation of sulfonilamide and choline. Indeed, the 'A' designation in CoA recognized the importance of the acetylation reactions. The first clue to the vitamin's structure came when digests of CoA with intestinal phosphatase and liver extracts were found to contain β-alanine and pantothenic acid as hydrolysis products. Treating CoA with a specific 3′ nucleotidase inactivated the coenzyme and a specific pyrophosphatase cleaved the coenzyme to a panthothene-containing product that could be restored to CoA by adenylation with ATP. These studies showed that pantothenic acid was an essential component of CoA and had been locked into the structure of a rather complex coenzyme (**Figure 6**).

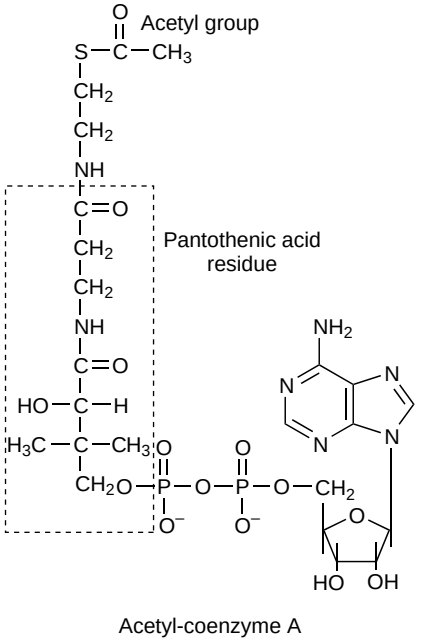


Figure 6 Pantothenic acid in the structure of coenzyme A.

Reactivity As a major component of CoA and its derivatives, pantothenic acid is involved in acetylation reactions, which include the synthesis of acetoacetyl-CoA, a precursor of cholesterol, and the biosynthesis of citrate from acetate. CoA can engage in acyl thiotransfer reactions such as accepting the acetyl groups from lipoic acid and forming acetyl-CoA as well as fatty acyl CoAs. Pantothene is also found as a prosthetic group (4′-pantothene) attached to a serine residue of acyl carrier protein (ACP), which plays a prominent role in the biosynthesis of fatty acids.

Cobalamin (Vitamin B_{12})

Few vitamins have been more challenging to structure–function studies than vitamin B_{12}. Among its many unique features, B_{12} is the only vitamin-coenzyme known to have a transition metal ion (cobalt) coordinated to its structure. The metal allows some usual chemistry (see 00058). The vitamin is present in a variety of foods but is almost totally lacking in plants. Although the vitamin can be synthesized *de novo* by intestinal flora, the absorption site anatomically is prior to the synthesis site in the gut, which means little benefit is derived from endogenous synthesis. Isolating the active form of the vitamin meant developing an *in vitro* assay for 'pernicious anemia,' one of the deficiency symptoms. In 1950, Shive introduced an assay in which homocysteine was converted to methionine in a B_{12} -dependent reaction. A second assay showed that the derivative 5-deoxyadenosyl cobalamin was essential for the interconversion of L-glutamate and β-methyl aspartate. The latter discovery led to the isolation of 5′ adenosylcobalamine, the principal active form of the vitamin.

Reactivity The core of the vitamin consists of a corrin ring with a central cobalt atom. Corrin contains four pyrrole rings linked together, which vaguely resembles structurally the porphyin ring in heme (**Figure 7**). An inactive form of the vitamin contains a displaceable CN group bound to the cobalt; hence the early name cyanocobalamin for one of its more familiar forms (**Figure 7**). The cobalt atom in the ring can have a +1, +2, or +3 oxidation state. The fifth valence (below the ring plane) has a dimethylbenzimidazole attached to the cobalt and the six can be either a methyl group, an –OH

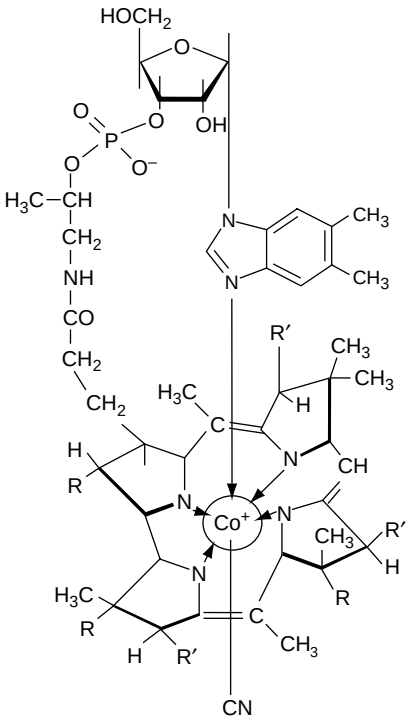


Figure 7 Cyanocobalamin, one of the forms of vitamin B_{12}. Shown are the corrin ring and the attachment of the dimethylbenzimidazole to the cobalt above the ring. The cyano group is shown attached to the cobalt below the plane of the ring.

group, or a 5′ deoxyadenosyl group depending on the reaction or enzyme. As noted, 5′ deoxyadenosylcobalamin is the most common form of the coenzyme. The 5′ deoxyadenosylcobalamine arises by an attack on the 5′ carbon of ATP by Co^+, which displaces the triphosphate group of ATP, a rare action in biochemistry. Known enzymes that require B_{12} fit one of two functional categories: those that transfer methyl groups from the coenzyme to the substrate, and those that take part in positional rearrangements of neighboring groups on the substrate, or group transfer reactions.

As noted above, methylation reactions in mammalian systems that involve B_{12} are limited to the transfer of a methyl group to homocysteine to form methionine. Recall, N^5-methyl-THF is the methyl group donor in the reaction and B_{12} mediates the transfer. Restoring the methyl group on methionine primes the system to further methylation since methionine itself, acting through its active form, *S*-adenosylmethionine (see below), is a primary donor of methyl groups to other substrates.

Of late, there has been considerable interest in vitamin B_{12} reactions that have free radicals as intermediates. This may be one of the principal advantages of the coenzyme, i.e., the ability to form and retain a stable free radical in its structure. The stability of the free radical is due to the unusual chemistry of the cobalt ion.

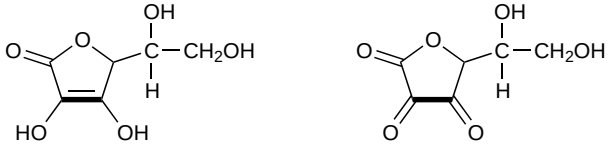
Ascorbic Acid

The vitamin (L-ascorbic acid) linked with scurvy is known to be a cofactor for two enzymes that take part in the biosynthesis of collagen, the major connective tissue protein; the formation of hydroxyproline and hydroxylysine residues as catalyzed by prolyl hydroxylase and lysyl hydroxylase enzymes, respectively (**Figure 8**). Collagen is an essential component of the extracellular matrix. As a cofactor for dopamine-β-monooxygenase, L-ascorbic acid is also required for the synthesis of adrenaline (epinephrine) and noradrenaline (epinephrine) in the adrenal medulla. Another important noncofactor role of L-ascorbate is as an antioxidant in cells and blood.

Vitamin K

The antihemorrhagic role of vitamin K had been established long before it was realized that the vitamin with no structural modification was essential for the biosynthesis of functional prothrombin. The two forms of the vitamin, phylloquinoine (vitamin K_1) and menadione (vitamin K_2), differ only in the structure of their side chains (**Figure 9**).

Reactivity Vitamin K takes part in an extensive series of carboxylation reactions involving all the glutamic acid residues in the first half of prothrombin, a blood-clotting factor. The carboxyglutamic acid residues or 'gla' that occupy this region of the prothrombin molecule are able to bind the calcium ions needed to catalyze formation of thrombin. Carboxylation is dependent on oxygen and uses the hydroquinone form of the vitamin as the driving force. Warfarin, a clotting inhibitor, interferes with the carboxylation reaction, which explains the basic mechanism of this inhibitor. A synthetic substrate, Pro-Leu-Glu-Glu-Val, has been found to substitute for prothrombin in the reaction, opening the way to learning the finer details of the reaction mechanism and the specific role of vitamin K.



L-Ascorbic acid (ascorbate)

Dehydroascorbic acid

Figure 8 Vitamin C in its reduced (dihydro) and oxidized (dehydro) forms.

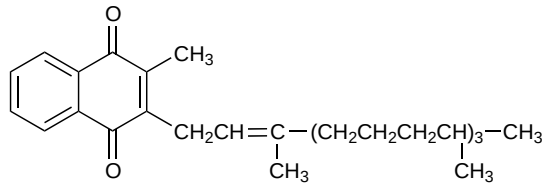


Figure 9 Phylloquinone, one of the active forms of vitamin K (Vitamin K_1).

Nonvitamin Cofactors

In addition to the above vitamin-coenzymes, there is a list of cofactors that do not fit the general category of vitamin-derived cofactors. Nonetheless, these organic cofactors are essential components of the catalytic mechanisms of important enzymes or membrane-bound enzyme systems.

Lipoic Acid

Lipoic acid's role as a growth factor for microorganisms and as a cofactor for biochemical reactions in all organisms is well established. The cofactor was discovered originally in the conversion of pyruvate to acetate and as a factor essential for the oxidation of pyruvate. Lipoic acid is known to occur in α-keto acid dehydrogenases from a variety of organisms. It is normally bound to the ε-amino group of a lysine residue (analogous to biotin) allowing the cofactor to extend out and away from the enzyme surface as a 'swinging arm.'

Reactivity The reactions taking place in the pyruvate dehydrogenase complex best reveal the cofactor function of lipoic acid. As a cofactor, lipoic acid (6,8 dithiooctanoic acid) exists in both an oxidized (disulfide) and reduced form (**Figure 10**). The disulfide form oxidizes active acetaldehye bound to TPP simultaneously with the transfer of the acetate product to one of the -SH groups of the now reduced lipoic acid. As a thioester, the acetate group is subsequently transferred to coenzyme A forming acetyl-CoA and regenerating a free -SH group on lipoic acid. The reduced lipoic acid, which still contains the electrons, is then oxidized by a flavoprotein (FAD) restoring the disulfide group of lipoic acid for another round of catalysis. Eventually, $FADH_2$ passes the electrons to NAD^+, which links to the electron transport chain of the mitochondria. Lipoic acid is thus an oxidizing agent and a carrier of acetate in the reaction. One can picture the long arm of the cofactor swinging between sites on the subunits of the pyruvate dehydrogenase complex in order to perform the multiple reactions in synchrony with the catalytic events taking place.

Carnitine

Carnitine is readily synthesized from lysine. As a cofactor, carnitine takes part in the membrane-bound enzyme system that transports fatty acids into the mitochondria for energy oxidation. Two enzymes, carnitine acyl transferase I and carnitine acyl transferase II, comprise a cycle that delivers the fatty acid as an acyl carnitine derivative to the interior of the mitochondria and returns the carnitine to the cytosolic side for further transport (**Figure 11**). The structure of carnitine with its hydroxyl group on C-3 is ideally suited for forming an acyl bound with a fatty acid.

Coenzyme Q (Ubiquinone)

First detected in lipid extracts of mitochondria and identified as a quinone, coenzyme Q (CoQ) was so named to signify its cofactor role in oxidation reactions. A second group of investigators discovered a cofactor that had ubiquitous occurrence in oxidative processes, which they named ubiquinone. In time, CoQ and ubiquinone were found to be the same compound. Early studies on the electron transport chain of mitochondria showed at least three complexes required CoQ and recognized its essential role in the electron transfer process overall. CoQ can be synthesized in man from tyrosine in a rather complex synthesis.

Reactivity CoQ and its reduced form $CoQH_2$ are designed to handle electron pairs in transit in oxidation-reduction reactions. A third form, semiquinone ($CoQH^{\cdot}$), exists as a stable radical and is capable of a one-electron transfer (**Figure 12**). Because of its ability to deal with electrons on a single or paired base, CoQ takes part in electron transport chains where one- and two-electron transfers are essential. Its lipid nature allows the cofactor to bind firmly to

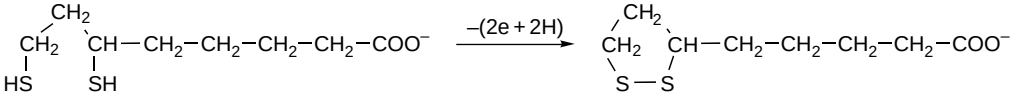


Figure 10 Lipoic acid in its reduced (sulfhydryl) and oxidized (disulfide) forms. The carboxyl group on the end is attached to a lysine group on the enzyme thus forming a swinging arm that is designed to transverse remote donor-acceptor sites across the surface of the enzyme.

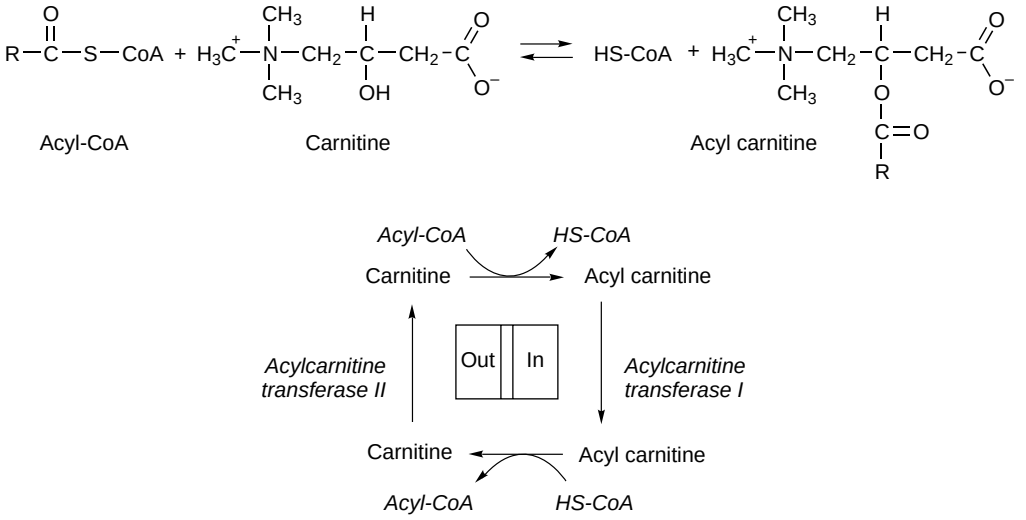


Figure 11 Carnitine-dependent transfer of fatty acyl groups. Two juxtaposed membrane transferase enzymes, acylcarnitine transferase I,II, are designed to transport a fatty acyl carnitine complex into the mitochondria and return the carnitine for additional reactions. Note that the acyl group is transferred to the carnitine from CoA and returned to CoA inside the mitochondria.

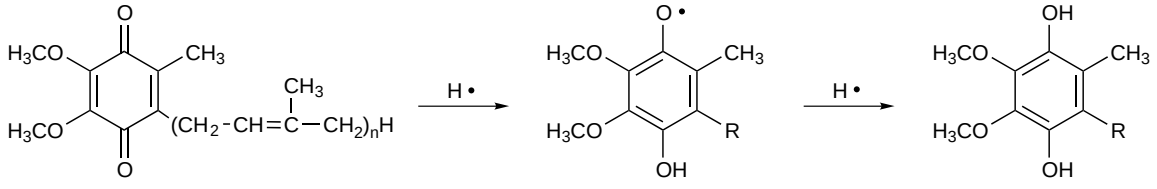


Figure 12 Coenzyme Q (ubiquinone) as an electron carrier. The coenzyme is a mobile electron carrier that moves between protein complexes in the mitochondria membrane. Shown are single (semiquinone) and dual (hydroquinone) forms that permit one- and two-electron transport, respectively.

the mitochondria inner membrane. Besides being a prominent carrier of electrons in the electron transport chain of mitochondria, CoQ is known to be a source and mediator of protons that are pumped across the inner mitochondria membrane to form the high-energy proton gradient associated with oxidative phosphorylation.

Pyrroloquinoline Quinone (PQQ) and 6-Hydroxydopa (topa) Quinone

A cofactor known to be present in methylogenic bacteria and other microorganisms, PQQ was considered at first to be a cofactor for mammalian copper amine oxidases and other copper enzymes. Its essential nature led some workers to consider PQQ an undiscovered vitamin. This, however, turned out not to be the case and PQQ as a cofactor has now been relegated to the world of microorganisms.

What at first was thought to be PQQ in copper oxidases turned out to be a cofactor with quinone properties that was derived by modifying a tyrosine residue in the enzyme (**Figure 13**). The synthesis of 6-hydroxy (topa) quinone, a derivative of tyrosine, requires copper in an apparently autocatalytic reaction. Though rare and limited, this most unusual biochemical reaction opens a new chapter on cofactors by showing that some enzymes have a limited but specific capacity

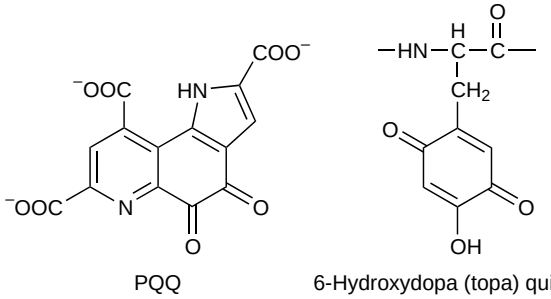


Figure 13 Comparison of structures of pyrroloquinoline quinone (PQQ) with topa quinone. Cofactor on the right has been identified in mammalian systems.

to synthesize cofactors on their surface through modification of existing amino acid side chains.

Other

There are a number of natural organic compounds that do not quite fit the description of cofactors, yet have been identified as being involved with stabilizing enzymes or taking part in a select series of reactions. We include them here with the caveat that some may behave more as substrates than cofactors.

Glutathione

Glutathione is a naturally occurring tripeptide that is known to exist in millimolar quantities in cells. The reduced form (GSH) has exposed -SH groups associated with an internal cysteine residue that has been shown to figure prominently in the stability of enzymes that have -SH groups at the active site. Glutathione partakes in many biological reactions such as amino acid transport, heavy metal transport, and antioxidant activity. Its cofactor role, however, should not be overlooked since GSH has been shown to activate many enzymes or retain their catalytic effectiveness during assays. One suspects, with justification, that this could be one of the functions of GSH *in vivo*.

Betaine

Betaine *(N,N,N*-trimethylglycine) arises from choline by an oxidation reaction. The structure is a trimethyl derivative of glycine. Betaine occurs in very small quantities in cells and has been shown to be a methyl donor for a limited number of reactions, notably synthesis of methionine from homocysteine.

S-Adenosylmethionine

To consider *S*-adenosylmethionine (SAM, also known as adomet) a cofactor is to recognize its role in a multitude of reactions that transfer methyl groups to substrates. Thus, SAM is involved in an extensive series of methylation reactions that surpass methylations of either N^5-methyl-FH_4 or methyl cobalamin combined. The methyl group transferred is the terminal carbon of methionine. To activate the methyl group, methionine reacts with ATP, adding an adenosyl group to the sulfur atom and causing a high-energy methyl-donating species to form (**Figure 14**). Although SAM is perhaps more of a substrate than a cofactor, its inclusion here is to denote the importance of methionine and its reactive

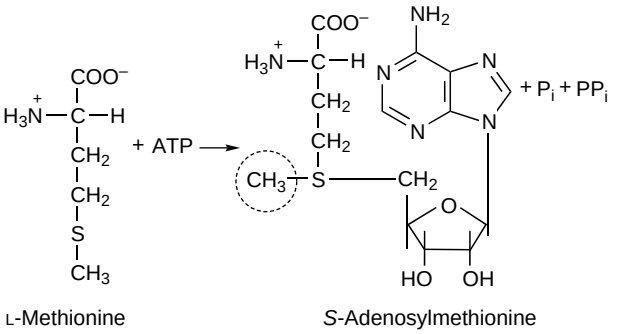


Figure 14 Synthesis of *S*-adenosylmethionine from methionine. Note the favorable positioning of the methyl group of methionine (dotted circle) as a result of the condensation with ATP.

form, SAM, in a series of extremely important biosynthetic reactions.

3′ Phosphoadenosine-5′ Phosphosulfate (PAPS)

PAPS is a cofactor for sulfation reactions, a process confined largely to plants and bacteria, but an important metabolic reaction in humans. PAPS serves as an active agent for sulfate esterification, as in the synthesis of sulfated polysaccharides such as chondroitin sulfate, keratin sulfate, and heparin.

Conclusions

Of the 20 or so organic cofactors that have been discovered over the years, the structures of more than half are derived from the nucleus of vitamins, primarily the water-soluble B vitamins. As companions to enzymes, organic cofactors relate to all forms of life. While we may think of vitamins as being needed by only higher organisms, many organic factors were designed to serve exclusively with enzymes and imperfections in enzyme systems that gave vitamins an essential character were revealed in bacteria and yeast systems long before they became known in humans. The carryover between cofactor-dependent reactions in microorganisms and humans has been remarkable, an illustration of the structure-function principle of biochemistry. Still, we must not overlook the fact that the study of cofactors has brought a sharper focus to the role of dietary components in human health and nutrition. Whereas mutated bacteria may fail to grow for want of a vitamin synthesized *de novo*, a susceptible human will develop a deficiency symptom. Both need the vitamin factor in a failing enzyme system in order for that system to perform at a healthy capacity. Nutritionists are challenged to learn the function of all cofactors

because that information provides fundamental insights into a chemical blueprint that applies to a wide spectrum of different organisms at the molecular level.

See also: **Anemia**: Megaloblastic Anemia. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements; Deficiency States. **Biotin**. **Cobalamins**. **Cofactors**: Inorganic. **Folic Acid**. **Niacin**. **Pantothenic Acid**. **Riboflavin**. **Thiamin**: Physiology; Beriberi. **Vitamin A**: Deficiency and Interventions. **Vitamin B_6**. **Vitamin D**: Rickets and Osteomalacia. **Vitamin K**.

Further Reading

Brown DE, McGuir MA, Dooley DM, Jane SM, Mu D, and Klinman JP (1991) The organic functional group in copper-containing amine oxidases. Resonance Raman spectra are consistent with the presence of TOPA quinone (6-hydroxy quinone) in the active site. *Journal of Biological Chemistry* **266**: 4049–4051.

Carpenter KJ (2003) A short history of nutritional science: Part 3 (1912–1944). *Journal of Nutrition* **133**: 3023–3032.

Freedland RA and Briggs S (1977) *A Biochemical Approach to Nutrition*. London: Chapman and Hall.

Harris ED (1992) The pyrroloquinoline quinone (PQQ) coenzyme: a case of mistaken identity. *Nutrition Reviews* **50**: 263–267.

Jukes TH (1977) Adventures with vitamins. In: Klemm WR (ed.) *Discovery Processes in Modern Biology*, pp. 152–170. New York, NY: Krieger.

Kutsky JK (1973) *Handbook of Vitamins & Hormones*. London: Van Nostrand Reinhold.

Marks J (1975) *A Guide to the Vitamins. Their Role in Health and Disease*. Baltimore: University Park Press.

Mathews CK and van Holde KE (1990) *Biochemistry*. Redwood City, CA: Benjamin Cummings.

Shive W and Lansford EM Jr. (1980) Roles of vitamins as coenzymes. In: Alfin-Slater RB and Kritchevsky D (eds.) *Human Nutrition. A Comprehensive Treatise: Nutrition and the Adult, Micronutrients*, vol. 3B, ch. 1, pp. 1–71. New York: Plenum Press.

Coffee *see* **Caffeine**

COLON

Contents

Structure and Function

A Maqbool, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

The colon is a dynamic organ involved in the absorption of salts, fluids, and nutrients, and it has a primary role in defecatory function. Additionally, the colon is an immunologically active tubular cavity playing an important part in host immune responses and defense from pathogens.

Gross Morphology

The colon is a continuous structure originating at the ileocecal valve and extending to the anus. The cecum is the first part of the colon, which lies in a posterior position at the right iliac fossa and has an ovoid-like shape. This cavity is more generous in proportion than other compartments of the colon. The appendix (a blind-ending out pouching) originates in the cecum and its opening is usually visible during colonoscopy.

The ascending colon runs cephalad and anteriorly from the cecum to just inferior to the liver, to the

hepatic flexure, emerging into the peritoneum. The transverse colon continues from the hepatic flexure to the splenic flexure, from which it travels distally and once again posteriorly to the sigmoid colon, an S-shaped, tortuous, narrow peritoneal structure. At the peritoneal reflection, the rectum arises and, closely following the sacral curve, leads to the anal canal. The rectum is a vault-like structure that can distend in order to accommodate fecal load. The anal canal bears two sphincters, an internal and an external anal sphincter. The internal sphincter is composed of inner circular smooth muscle fibers and a distal external fiber on the other side of a muscular pelvic diaphragm. The fibers of the external sphincter are intertwined with those of the levator ani, tethered anteriorly and posteriorly to the perineal body and the coccyx, respectively.

With respect to colonic mobility within the abdominal, peritoneal, and pelvic cavities, the cecum and flexures are less mobile, with the sigmoid colon being the most mobile. The transverse colon supports the greater omentum and has a variable degree of mobility.

Cross-sectionally, the colon has an external longitudinal muscle and an inner layer of circular musculature, the former of which has coalescence of fibers forming band-like structures known as taeniaecoli. These taeniae are particular to the large intestine, are located at one-third of the circumference from each other, and run continuously from one end of the colon to the other. Haustra are hemilunar-like outpouchings present between taeniae. The more proximal rectal taenial fibers surround the rectum; the inner fibers form the internal anal sphincter. The external fibers are intertwined with those of the levator ani and sandwiched between fibers running anterior to posterior from the peroneal body to the coccyx, forming the external sphincter.

Vasculature

The ascending colon and portions of the transverse colon are perfused by branches of the superior mesenteric artery, with the remainder of the colon receiving arterial blood from tributaries of the inferior mesenteric artery. Distal iliac arterial branches perfuse the anal canal. Venous drainage is achieved via the superior and inferior mesenteric veins laying in close proximity to their arterial counterparts and subsequently dumping into the portal vein.

Additional gross morphologic structures include lymphatic vessels, in close approximation to the vasculature, leading to lymph nodes in the celiac, superior, and inferior preaortic regions. Perianal drainage is via the inguinal lymph nodes.

Innervation

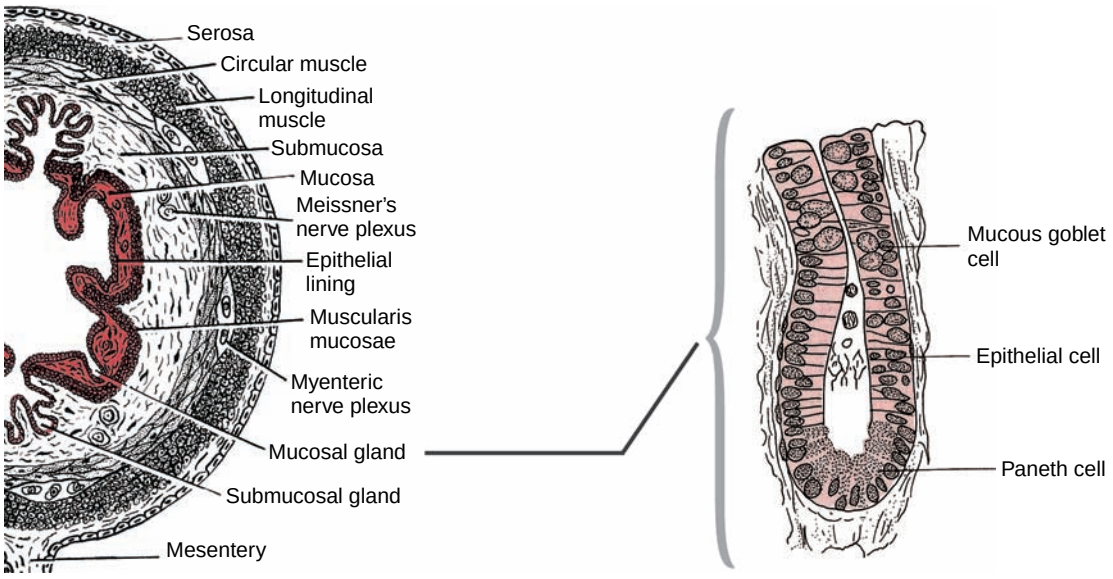
Parasympathetic innervation to the proximal colon is provided via the vagus nerve; the distal colon and rectum are innervated via pelvic parasympathetic fibers. The sympathetic nervous system innervates the proximal colon via lower thoracic fibers and the distal colon and rectum via lumbar fibers. Prevertebral sympathetic vertebrae receive fibers from neurons projecting out of the gut.

Histology

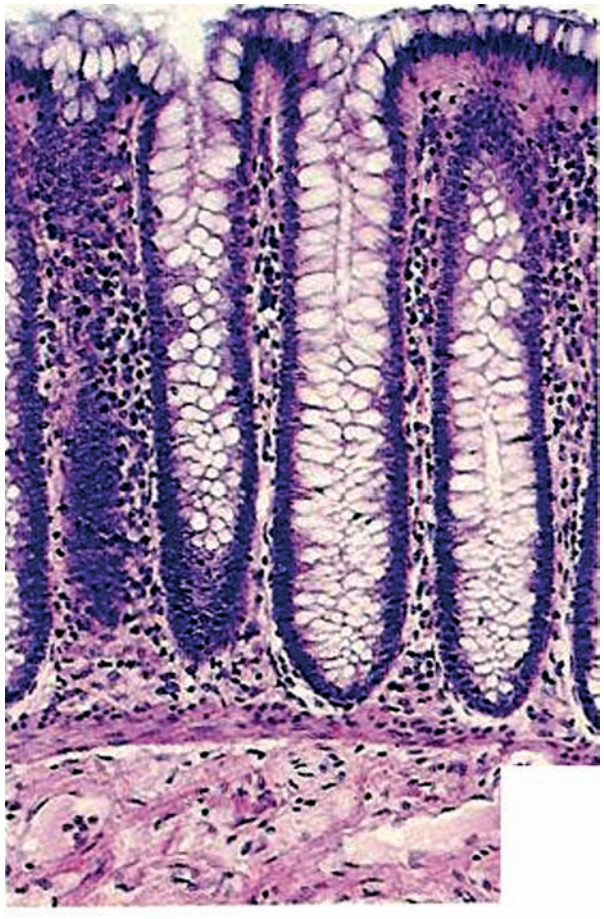
Cross-sectionally, the intestinal wall is divided into four layers, with the serosa, a monolayer of mesothelial cells comprising the outermost layer, followed by the muscularis externa. These muscle layers comprise an external longitudinal layer and an internal circular layer. Sandwiched between these two layers lies Auerbach's (myenteric) plexus. The submucosa is the next more medial layer. A rich admixture of cells, including structural elements such as fibroblasts and dense connective tissue, immunologically important cells (plasma cells, lymphocytes, macrophages, eosinophils, and mast cells), and vascular tissue and innervation to Meissner's plexus (ganglion cells) and lymphatics comprise this layer. The muscularis mucosa, a thin sheet of smooth muscle, separates the deeper submucosa from the mucosa. The lamina propria runs interior to this layer, is composed of connective tissue, and is lined by the luminal epithelium (**Figure 1A**).

The intestinal epithelium is a tight monolayer of cells that function to absorb nutrients, electrolytes, and liquids as well as to secrete mucous and fluids. The epithelial surface is punctuated by numerous tightly packed crypts, which contain epithelial precursor cells, enterendocrine cells, other undifferentiated cells, and Paneth cells. Goblet cells, which secrete mucin, are also located in the crypt (**Figures 1B** and **1C**, **Table 1**). As undifferentiated and precursor cells mature, they migrate superiorly to the surface to the monolayer of absorptive cells present in crypts. The average life span of a colonocyte is 3–6 days.

The absorptive colonocyte develops short microvilli while in the colonic crypt, which elongate during its migration to the surface. The hydrophobic lipid bilayer of the colonocyte epithelium prevents passive transport of charged particles. The epithelial membrane contains specific protein transporters, carrier proteins, and channels allowing electrolyte transport. The electrochemical gradient formed by active transport facilitates passive flow across cell membranes.



(A) (B)



(C)

Figure 1 (A) Cross section of the gut and (B) a colonic crypt. (Reproduced with permission from Guyton (1991) *Guyton's Textbook of Medical Physiology*, 8th edn. Philadelphia: WB Saunders.) (C) Histology H&E stain of a colonic crypt, with prominence of mucin-containing goblet cells. (Reproduced with permission from Burkett HG, Young B, and Heath JW (1993) *Wheater's Functional Histology*, 3rd edn. London: Churchill Livingstone.)

Table 1 Colonic cell types

Cell type	*Location*	*Function(s)*
Stem cells	Crypt (base) Nonmigratory until differentiated	Pluripotent
Undifferentiated crypt cell	Crypt	Secrete water and chloride into intestinal lumen
Paneth cells	Crypt base Nonmigratory Basophilic cytoplasm Proximal one-third of colon only	Growth factor secretion, digestive enzyme synthesis Antimicrobial peptide synthesis and release
Goblet cells	Colonic crypt Most common cell type in the colon	Mucin release
Enteroendocrine cells	Mostly in small intestine Basolateral membrane	Receptor-mediated epithelial cell function modulators
Enterocytes	Predominantly small intestinal; present in the colon	Digestive enzyme synthesis (small intestine) Ion transporters and channels involved in fluid and electrolyte transport
M cells	Small and large intestines Overlying lymphoid follicles	Bind, process, and present antigens to components of the mucosal lymphoid immune system
Intraepithelial lymphocytes	Small and large intestines Basolateral membranes	Memory T cells Mucosal immune defense

Electrolyte Transport: Ion Channels

Fluids and electrolytes are absorbed via either the transcellular or the paracellular pathway. Active and passive transport systems exist via both of these pathways.

There is a clear polarity to the distribution of protein transporters, channels, and pumps distinguishing the apical from the basolateral membrane. Active transport utilizes a transcellular, energy-driven protein pump or channel to facilitate passage of an electrolyte from an area of low concentration to one of high concentration (electrochemical gradient). A prime example of this is the NA-K ATPase pump, the principal pump present along the basolateral membrane. The net effect of the three Na ions expelled for every two K ions accepted into the cell is a lowered intracellular Na content and resultant net negative charge. The negative charge formed by this active transport creates an electrochemical gradient facilitative to the passive flow for other ions across the cell membrane—a process known as secondary active transport (**Figure 2**).

Ion transporters may be subclassified into symporters, in which ions move in the same direction, or antiporters, in which ions move in opposite directions across the cell membrane (**Figure 3**). Cotransport of ions occurs with other molecules, such as Na and glucose. The intracellular concentration of glucose is regulated both by uptake at the apical surface and by exit through the basolateral membrane, allowing for conditions favorable to uptake from the lumen. The Na-glucose transporter system allows for therapeutic interventions, such as the use of oral rehydration solution in cases of severe diarrhea related to cholera or other processes. Similar cotransporters are linked to the transport of bile salts and amino acids (**Table 2**).

Whereas sodium is the primary cation involved in ion transport, short-chain fatty acids constitute the primary anion in the colon and the primary metabolic fuel for colonocytes. Their transport is postulated to be linked to Na-H transporters and pH, specific bicarbonate-linked transporters, and the concentration gradient across cell membranes. Chloride transport occurs via both active and passive processes, and it is the major intestinal anion involved in intestinal secretion of fluids.

Colonic smooth channels also possess ion channels and are involved in active and secondary ion

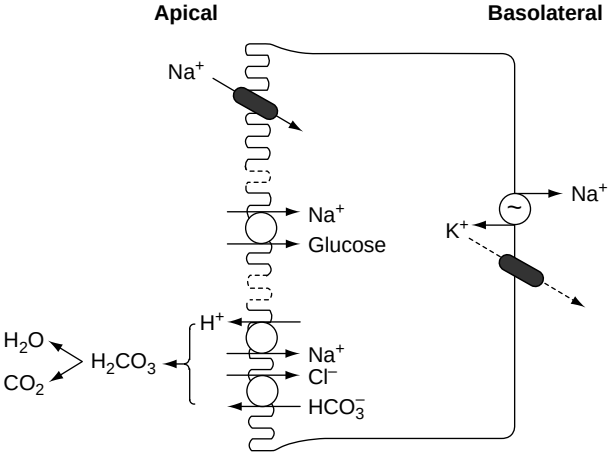
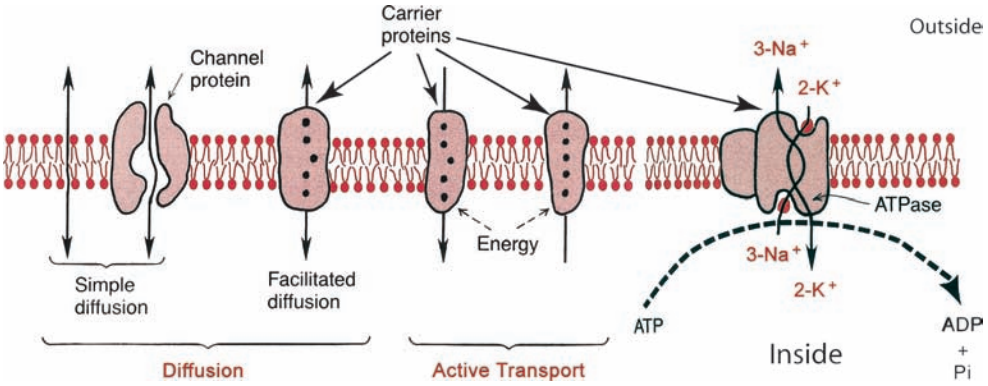
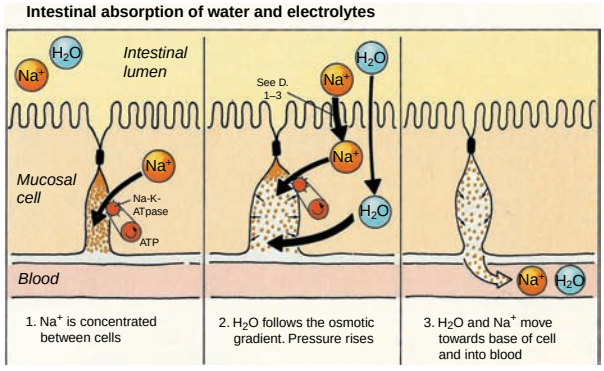


Figure 2 Electrolyte transport at the colonocyte level. (From Despopoulos A and Silbernagl S, *Color Atlas of Physiology*. New York: Thieme; 2000. Reprinted with permission.)



(A)



(B)

Figure 3 Electrolyte transport across the cell membrane and the different types of tranporters. (Reproduced with permission from Guyton (1991) *Guyton's Textbook of Medical Physiology*, 8th edn. Philadelphia: WB Saunders.)

transport processes involving calcium. The electrochemical gradient formed by the activity of these ion channels facilitates the function of smooth muscle action potential generation upon depolarization. With the generation of smooth muscle action potentials attaining threshold voltage, contractility of the smooth muscle is possible. The efflux of calcium into these active transport channels activates the process of contraction. Interaction with the enteric nervous system stimulates the release of calcium ions in intracellular stores. The function of ion channels can be modified by calcium channel-blocking drugs. This contractile activity, when it occurs in a coordinated fashion and is modulated by neurotransmission, effects peristalsis and colonic motility.

Fluid Transport

There is heterogeneity to the mucosal epithelium in several aspects dependent on the location in the alimentary canal. The type, variety, and number of ion transporters, channels, and carrier proteins vary from region to region (e.g., from jejunum to colon). Additionally, the nature of interepithelial cell junctions varies from the proximal to distal intestinal tract, influencing the 'leakiness' of the respective regions. Finally, a clear gradient in cell composition and function between colonic crypt cells and those on the surface exists. Physiologic heterogeneity follows the aforementioned patterns, defining tissue function in these respective areas. For example, the colonic crypts serve more of a secretory function, whereas the villus structures seen most notably in the jejunum exhibit greater absorptive function. This heterogeneity is key in understanding changes in intraluminal osmolality and fluid shifts that occur in the intestine.

Approximately 98% (9 l per day) of the daily fluid load handled by the intestine is reabsorbed. Of this, the jejunum absorbs 85%, and the colon absorbs approximately 13% (1.5 l).

Passive reabsorption of water occurs in the intestines, regulated primarily by electrolyte transport (i.e., following an osmotic gradient). Na-driven or -related transport mechanisms are the primary driving force

Table 2 Electrolyte transport

Ion	*Transporter*	*Location*	*Type*	*Function(s)*
Na	Na-K-ATPase	Basolateral membrane	Active; antiport	Principal ion involved in water absorption
	Na-H exchangers	Apical and basolateral membrane	Secondary; antiport	
Na and Cl	NaCl	Apical	Antiport; passive; electrochemically neutral	
Cl	Protein channel	Apical	Diffusion; passive (secretion) and some active transport proteins at the apical surface (absorption and secretion), including CFTR	Principal ion involved in water secretion
				Basal rate of secretion influenced by several mediators (endocrine, paracrine, neural, luminal, etc.)
Cl	Protein channel			
K	Protein channel		Antiport; active transport (basolateral membrane) Active secretion at the apical membrane; linked to Cl transport function Absorptive active apical K-ATPase pumps in distal colon	
HCO_3		Apical and basolateral channels	Alkaline phosphatase linked Passive transport mechanisms Na-HCO_3 cotransporter postulated CFTR-synchronized apical channel and Cl-HCO_3 exchanger postulated	
Short-chain fatty acids	Apical		Postulated link to NA-H ion transport	Principal anion of the colon

allowing water absorption. This osmotic gradient facilitates water absorption via both transcellular and paracellular pathways.

Transcellular water transport mechanisms such as aquaporins, or water channels, have been described. The paracellular pathway of water transport has been studied extensively, a process often described as 'solvent drag' (**Figure 4A**).

The leakiness of paracellular pathways, which varies by location in the lower alimentary tract (more prominent in the jejunum, with subsequent decrease distally), and the magnitude of the osmotic gradient (also affected by dietary Na content) are important factors affecting solvent drag. The nature of the intercellular junctions in a particular region of the colon determines the permeability or leakiness of that particular epithelial area. Several intercellular structures have been described, including the zona accludens (tight junction), desmosomes (connections between cells), and the zona adherens, which functions in cell adhesion and contributes to maintaining cellular polarity across the membrane. Zona occludens are more apical in location and form junctional complexes between cells. It has been postulated that these junctional complexes may be more dynamic than previously believed, responding to signaling mechanisms and subject to regulation, thereby influencing their function and resultant permeability characteristics (**Figure 4B**).

The Enteric Nervous System and Gastrointestinal Motility

The enteric nervous system (ENS) operates both in conjunction with and independent of the peripheral nervous system. As discussed previously, nerve plexi exist within the bowel wall, with Auerbach's plexus sandwiched between longitudinal and circular muscle layers, and Meissner's plexus located more medially in the submucosa. The ENS is the largest component of the autonomic nervous system, based on nerve cell number.

Interstitial cells of Cajal, a cell type unique to the alimentary tract, are present medial to the inner smooth muscle layer. These specialized cells interact with myenteric neurons and are thought to exhibit independent electrical activity, generating and transmitting slow waves to smooth muscle, functioning as pacemakers for colonic motility. The ENS

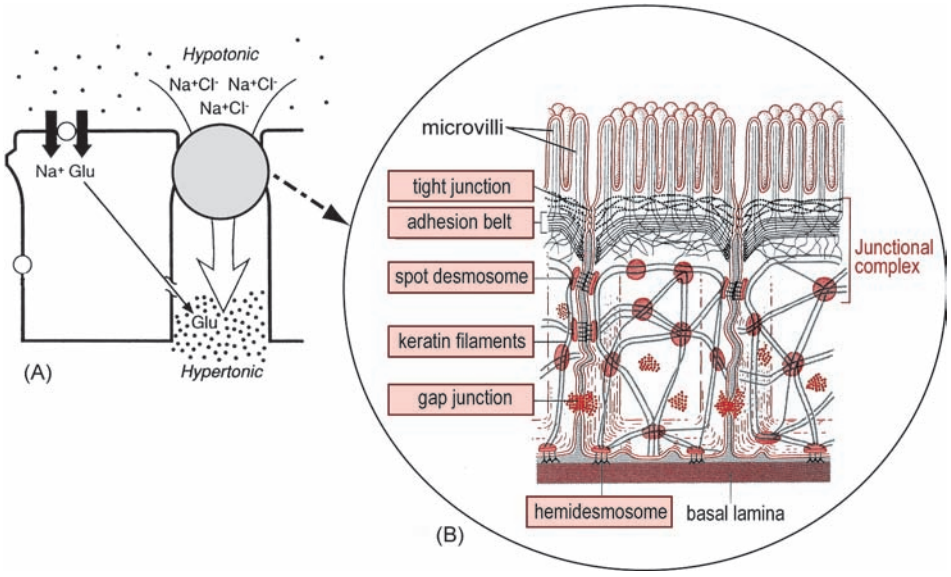


Figure 4 (A) Fluid transport across the cell membrane. (Reproduced with permission from Shils ME, Olson JA, Shike M, and Ross AC (1999) Modern Nutrition in Health and Disease, 9th edn. Baltimore: Lippincott Williams & Wilkins.) (B) Intercellular junctions. (Reproduced with permission from Alberts B, Bray D, Lewis J, Raff M, Roberts K, and Watson JD (1989) *Molecular Biology of the Cell*, 2nd edn. Garland Publishing, NY.

functions independent of the central nervous system, with reflex activity in response to luminal stimuli, including muscle contraction and coordination (i.e., motility, blood flow, and glandular secretion). Modulation of the ENS is via the sympathetic and parasympathetic nervous system.

Colonic Motility

The colon functions to delay passage of luminal contents to allow for water absorption and for mixing of luminal contents with the mucosa, to store fecal matter prior to defecation, and, at the time of defecation, to propel contents forward.

The frequency and duration of propagative, high-pressure waves in the colon in part are determined by pressure exerted by intraluminal contents (mechanical) and the degree of stretch stimulation, the (chemical) composition of the contents, and other stimuli interacting with the colon.

The gastrocolic reflex, an anterograde postperistaltic process, occurs following a meal, originating proximally and propagating anterograde. Both the caloric content and the fat composition of the meal influence colonic peristalsis. Gastric distention by food contents, water, or gas also has a stimulatory effect. Gastrointestinal hormones secreted in response to a meal, such as cholecystokinin, are thought to mediate peristaltic responses to a (fatty) meal. Irritant laxatives also stimulate peristalsis, even when administered rectally. Opiates are known to inhibit the ENS and, as a consequence, retard peristalsis. Colonic motility diminishes significantly during sleep, resuming upon awakening.

Motor activity varies by colonic region; in degree, frequency, amplitude, and velocity; in being propagative versus nonpropagative (the latter is more common in the distal colon); in relative distance of propagation; and in direction of propagation (anterograde versus retrograde, the latter most commonly seen in the proximal colon). Approximately one-third of these colonic peristaltic waves are propulsive, and those associated with propulsion of stool tend to be slower but greater in amplitude.

Defecation involves the integration of peristaltic activity in the majority of colonic regions, not exclusively to that solely in the anorectal region. In the predefecatory phase, approximately 1 h prior to actual defecation, the majority of the colon exhibits an increase in propulsive peristaltic waves, first in the proximal colon and then advancing distally.

The sensation of defecatory urge is not evident until approximately 15 minutes prior to defecation. At that time, there is a marked increase in propagative peristaltic activity, originating more distally in the colon. Each of these late propagative waves successively originates proximate to the preceding one with greater amplitude and presents over a greater distance of colonic length.

Stool contact with the receptors in the upper anal canal can effect relaxation of the inner anal sphincter. In addition, stretch receptor stimulation of the

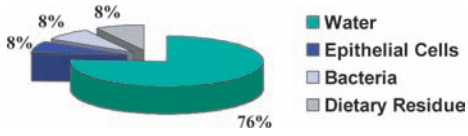


Figure 5 Composition of feces (60–80 g/day).

rectal vault walls results in the urge to defecate. Failure of relaxation of the external anal sphincter (which is under voluntary control) results in retrograde passage of stool into the rectum, with subsequent diminishing of more proximal peristaltic propagative waves, thereby maintaining continence when immediate defecation is not desirable or convenient.

Evacuation of the rectum and defecation require correcting the angle of the anal canal in the anterior–posterior plane, which is accomplished by assuming a squatting position. Contraction of the abdominal musculature and of the diaphragm, with a relaxed pelvic floor, facilitates defecation, even in the absence of colonic peristalsis.

Stool size and consistency vary based on diet, water intake, and transit time, as well as bacterial content (a major component of stool). Higher water content tends to result in larger, softer stools. The more fusiform-shaped the stool is, the less likely that its passage is associated with straining. The transit time through the colon is inversely related to the stool's water content and, hence, its consistency (**Figure 5**).

Colonic Immune Function and Colonic Bacterial Flora

The immune system of the gastrointestinal tract defends against infection (bacterial, viral, and parasitic) and luminal antigens ingested/formed by bacteria. Nonspecific and specific mechanisms exist.

The mucin secreted by colonic goblet cells serves a barrier function for the mucosal surface. Mucosal integrity is an important barrier to luminal pathogens. Interepithelial cell junctions function both to control permeability as pertains to fluid and electrolyte absorption and to prevent pathogen access beyond this layer.

The enteric immune system is vast and complex; it interacts with the rest of the immune system as well as with luminal contents. Gut-associated lymphoid tissue consists of both discretely organized tissue, such as Peyer's patches (lymphoid follicles with proliferative potential in response to antigen presentation) containing M cells, and the more diffuse lymphocytes and macrophages distributed among the submucosa, mucosa, and lamina propria (**Figure 6**). M cells function in antigen sampling of

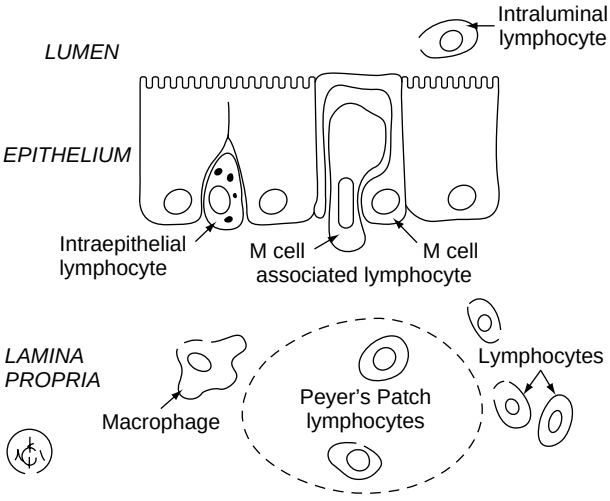


Figure 6 Mucosal immunology. (Reproduced with permission from Shils ME, Olson JA, Shike M, and Ross AC (1999) Modern Nutrition in Health and Disease, 9th edn. Baltimore: Lippincott Williams & Wilkins.)

intraluminal contents by binding antigens, endocytosis, antigen processing, and subsequent interaction with lymphocytes and macrophages within Peyer's patches, eliciting host responses. The lymphocyte complement of Peyer's patches originate in either the bone marrow or the thymus, enter the systemic circulation to migrate to Peyer's patches, interact, return to the intestinal mucosa or, via mesenteric lymph nodes, re-enter the systemic circulation to other organs.

The gastrointestinal tract houses up to 80% of the body's immunoglobulin-producing cells. Intraepithelial T lymphocytes, plasma cells, macrophages, dendritic cells, eosinophils, and mast cells also function in a specialized manner (**Figure 7**).

Secretory IgA is an important host immune defense mechanism. Unlike the monomeric, systemic form of IgA, intestinal secretory IgA is polymeric (specifically, dimeric) in nature. This dimeric immunoglobulin is secreted by B lymphocytes situated in the lamina propria, and it contains a unique 'J' chain instrumental in polymer formation. This IgA binds to the Ig receptor of the epithelial cell on the basolateral membrane and, following endocytosis and transport across the cell, is secreted from the apical side.

Secretory IgA binds to intraluminal antigens, including dietary ones, and functions in preventing their absorption. Additionally, secretory IgA has the ability to bind to microorganisms, thus preventing adherence, colonization, and invasion. Secretory IgA is secreted in breast milk, and in breast-fed neonates and infants it confers a degree of passive immunity to infection by limiting luminal contents from interacting with, or directly binding to or invading, the mucosa.

Interaction of intraluminal bacteria with the immune system may affect intestinal permeability, and it may modulate the intestinal immune system. Certain bacterial species are believed to interact with other enteric flora as well as with the host immune system to effect a healthier gastrointestinal tract and enhance nutrient digestion. Organisms studied include Lactobacillus, *Vibrio* species, and sacromyces. These findings, among others beneficial to the host, have prompted investigation into oral supplementation of single and multiple species of these probiotics for prevention and treatment of antibiotic-associated diarrhea, bacterial overgrowth in short bowel syndrome, and as an adjuvant therapy for inflammatory bowel disease, as well as for the treatment and prevention of recurrent clostridium difficile colitis.

Regulation of the quantity of bacteria, in addition to the specific profile of bacterial species present, is dependent on a host of factors, including gastric acid output, gastrointestinal motility, luminal contents, and the milieu created therein. Additionally, the intraluminal environmental milieu is affected by the specific properties of different species of bacteria and their interactions with other luminal species and with the host.

The colon accommodates the largest number of enteric flora, on the order of 10^{10}–10^{12}—more than 100 000 flora and more than 100-fold greater diversity of species than in any other location in the alimentary canal. Efflux of bacteria into the ileum is hindered by the ileocecal valve, which functions to restrict several of these bacterial species to the large intestine. The majority of these colonic bacteria are anaerobic in nature (**Table 3**).

Table 3 Colonic enteric flora

Bacterial genus	*Prevalence (%)*	*Total count (CFU/g or ml)*
Anaerobes		**10^9–10^{12}**
Bacteroides	100	
Porphyromonas	100	
Bifidobacterium	30–70	
Lactobacillus	20–60	
Clostridium	25–35	
Peptostreptococcus	—	
Peptococcus	—	
Methanogens	—	
Facultative aerobes		**10^2–10^9**
Enterococcus	100	
Escherichia coli	100	
Staphylococcus	30–50	
Other Enterobacteriaceae	40–80	

Table 4 Examples of biochemical reactions by intestinal flora

Reaction type	*Reaction*	*Example substrate*
Hydrolysis	Amides	Methotrexate
	Glucuronides	Estradiol-3-glucuronide
Dehydroxylation	Decarboxylation	Amino acids
	Deamination	Amino acids
	Dehydrogenase	Bile acids, cholesterol
Reduction	Double bonds	Unsaturated fatty acids
	Acetylation	Histamine

From Klein S, Cohn SM and Alpers DH (1999) The alimentary tract in nutrition. In: *Modern Nutrition in Health & Disease*, 9th edn, pp. 605–631. Baltimore: Williams & Wilkins.

The enteric flora plays several important roles. It interacts with the enteric immune system, effecting cellular immune activity; it is associated with the size and number of Peyer's patches present, influencing intestinal motility; and it has nutritively important functions, including bile salt deconjugation (facilitating enterohepatic circulation of bile salts), bilirubin metabolism (deconjugation and urobilin formation, allowing excretion), mucin degradation, and lipid metabolism (generation of short-chain fatty acids). Androgens and estrogens are hydrolyzed, facilitating resorption and conservation of these sterols, whereas cholesterol is processed into coprostanol, a nonabsorbed sterol. Ammoniagenesis via protein and urea degradation may play a role in hepatic encephalopathy (**Table 4**).

Consumption of lipids, carbohydrates, and protein also occurs by colonic bacteria, in addition to that of vitamins (vitamin B_{12} and folic acid are consumed; vitamin K and biotin are produced by these bacteria).

Dietary Fiber and the Colon

Nondigestable carbohydrates, traditionally defined as deriving from plant sources (but recently encompassing some non-plant-derived polysaccharides), that escape digestion and reach the colon nearly 100% intact compromise dietary fiber.

The metabolic fate of this fiber is influenced primarily by the colonic bacterial complement, which, depending on its structure, may render it susceptible to fermentation (such as pectin and oat bran). Common by-products of colonic fermentation include carbon dioxide, methane (in addition to other gases), oligofructases (among the classes of compounds of oligosaccharides termed prebiotics because of their nutritive support for the sustainment of certain colonic bacteria thought to be beneficial to human health—the so-called probiotics), and short-chain fatty acids.

The common short-chain fatty acids produced by fermentation include acetate, butyrate, and proprionate. The pattern of short-chain fatty acid production is dependent on several dynamic factors, including the type of fiber or oligosaccharide present in the diet, the transit time and exposure to bacteria, and the bacteria flora to which the substrate is being exposed. Short-chain fatty acids influence colonic physiology by stimulating colonic blood flow as well as fluid and electrolyte uptake. Butyrate in particular is thought to be preferred fuel for the colonocyte. This short-chain fatty acid is thought to have a role in maintaining the normal phenotype in these cells (i.e., in decreasing the risk of dyplasia by promoting differentiation and apoptosis of these cells). Because there is a lack of agreement on the *in vivo* and *in vitro* effects of the latter chemopreventive role of butyrate, it is commonly referred to the 'butyrate paradox.' Comparisons between these two models may have to do with the assumptions made with respect to environment and conditions. These include duration and patterns of exposure of the colonic to butyrate, interactions with other dietary components (such as the omega-3 dietary fats), and the confounding effect of dietary fiber that is resistant to digestion as well as to fermentation by the colon.

Nondigestible dietary fiber is also thought to play an important role in chemoprevention by diluting toxins, carcinogens, and tumor promoters; by decreasing transit time, thereby decreasing colonic mucosal exposure; and by promoting their expulsion in the fecal stream.

Dietary fiber resistant to colonic degradation may also play a role in maintaining and promoting stool bulk and in the regulation of intraluminal pressure/colonic wall resistance, disordered colonic motility, or both. These consequences of lack of dietary fiber have been linked to constipation as well as to the development of diverticulosis. Epidemiologically, the observed pattern of diverticulosis is striking: A dichotomy between the industrialized world and developing countries is clearly noted, and it is believed to be very closely linked to dietary fiber content. The nutrition transition and changes in lifestyle may impact dietary fiber intake and, as a consequence, may eventually be reflected by changing patterns of incidence and prevalence of processes linked to low dietary fiber intake, as mentioned previously.

See also: **Colon**: Disorders; Nutritional Management of Disorders. **Diarrheal Diseases**. **Dietary Fiber**: Physiological Effects and Effects on Absorption; Potential Role in Etiology of Disease; Role in Nutritional Management of Disease. **Electrolytes**: Water–Electrolyte Balance. **Fatty Acids**: Metabolism. **Microbiota of the Intestine**: Prebiotics; Probiotics. **Potassium**. **Small Intestine**: Structure and Function. **Sodium**: Physiology.

Further Reading

Feldman (2002) *Sleisenger & Fordtran's Gastrointestinal & Liver Disease*, 7th edn. Philadelphia: WB Saunders.

Food and Nutrition Board (2001) *Dietary Reference Intakes: Proposed Definition of Dietary Fiber*, A report of the Panel on the Definition of Dietary Fiber and the Standing Committee on the Scientific Evaluation of Dietary Reference Intakes. Institute of Medicine. Washington DC: National Academy Press.

Klein S, Cohn SM, and Alpers DH (1999) The alimentary tract in nutrition. In: *Modern Nutrition in Health & Disease*, 9th edn, pp. 605–631. Baltimore, MD: Williams & Wilkins.

Netter FH (1999) *Netter's Interactive Atlas of Human Anatomy*, version 2.0 East Hanover, NJ: Novartis Pharmaceuticals Medical Education Division.

Walker WA, Durie PR, Hamilton JR, Walker Smith JA, and Watkins JB (eds.) (2000) *Pediatric Gastrointestinal Disease: Pathology, Diagnosis, Management*, 3rd edn. Hamilton, Ontario, Canada: BC Decker.

Willie R and Hyams J (1999) *Pediatric Gastrointestinal Disease*, 2nd edn. Philadelphia: WB Saunders.

Disorders

A Maqbool, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

Diarrhea

Diarrhea is defined as a decrease in stool consistency and/or an increase in stool frequency and volume. It results from a complex interplay between colonic epithelial cell function, luminal factors, intestinal motility, and other factors.

Stool consistency and volume are determined partly by dietary factors (e.g., fiber intake) and fluid and electrolyte transport. Electrolyte transport mechanisms and diffusion processes in the small intestine render the fluid milieu isotonic. Active (primary and secondary) electrolyte transport mechanisms create an electrochemical gradient by which means cotransport of additional electrolytes can occur. Sodium is the major cation involved in the process of fluid absorption. Chloride constitutes the major anion that plays a significant role in fluid transport, and its active export into the intestinal

lumen is an important mechanism of intestinal fluid secretion. Potassium and bicarbonate also play a role in intestinal absorption and secretion mechanisms. Water transport is facilitated by this osmotic gradient. It is then absorbed by processes of transcellular passage facilitated by aquaporins as well as by solvent drag via paracellular pathway; paracellular permeability is regulated by junctional complexes. Glucose transport is linked to sodium transport, as is the case for certain amino acids. Electrolyte transporter function can be influenced by glucocorticoids and mineralocorticoids.

Intestinal motility also influences stool volume and consistency. The enteric nervous system, with some modulation by the autonomic nervous system, is the primary regulator of gastrointestinal motility. Neuropeptides, gastrointestinal hormones, and luminal stimuli, such as dietary factors and interactions with bacteria, influence colonic motility.

Mechanisms of diarrhea can also be viewed from the perspective of absorptive capacity of the small intestine and colon. Of the 8–10 l of fluid processed by the small and large intestines daily (composed of intake as well as gastrointestinal secretions), the smaller intestine absorbs 80–90% of the net load. The normal adult colon absorbs approximately 1 l of fluid per day but has a capacity to absorb 3 or 4 l per day; diarrhea results when this threshold is exceeded.

From a pathophysiological perspective, four mechanisms of diarrhea are traditionally described: osmotic, secretory, motility, and inflammatory. A degree of overlap occurs between these different types of diarrhea.

Osmotic diarrhea occurs when the failure to absorb a solute (usually a carbohydrate) in the proximal small intestine occurs, thus rendering the fluid hypertonic rather than isotonic, as would regularly occur. Whereas electrolytes may be reabsorbed, the carbohydrate is not; rather, a portion of it is metabolized by enteric flora to short-chain fatty acids, carbon dioxide, hydrogen, and methane. With sodium and other electrolytes absorbed readily by the colon, and resultant low-sodium concentration in the lumen, compounded by the presence of nonabsorbed carbohydrate, the high osmotic gradient draws fluid into the lumen and results in diarrhea. This type of diarrhea is characterized by a significant osmotic gap that can be calculated; an additional clinically significant feature of this type of diarrhea is that it diminishes upon cessation of enteral intake. Malabsorbed carbohydrate and its metabolites effect a lowering of the pH of the stool as well. Lactose deficiency is a good example of osmotic diarrhea in both children and adults. Ingestion of nonabsorbable sugars, such as sorbitol, can also lead to osmotic diarrhea. In children, excess intake of fruit beverages or of carbohydrates when recovering from a bout of acute gastroenteritis can occur, which resolves upon cessation of consumption of the carbohydrate.

Secretory diarrhea occurs when the net secretion of fluids and electrolytes from the colon exceeds their absorption. This type of diarrhea exists independent of eating and is not influenced by fasting or bowel rest. The prototypical example of pure secretory diarrhea (i.e., in the absence of inflammation or blood present in the stool) is of congenital chloride transport defects and of gastrointestinal hormonal disorders, such as in Zollinger–Ellison syndrome and disorders of vasoactive intestinal peptide or in other neuroendocrine tumors (**Figure 1**).

Cholera occurs when the toxin interacts with the colonocyte stimulating chloride, potassium, and bicarbonate secretion via toxin A stimulation of cyclic adenosine monophopshate; some degree of inflammation may accompany this. Oral rehydration solution, which contributes fluid, sodium, and glucose, relies on cellular mechanisms to effect rehydration and is the mainstay of therapy.

Motility disorders influence intestinal function as pertains to absorption; whereas decreased transit enhances absorption of nutrients, significant decreases in motility can result in stasis. Deconjugation of bile acids by enteric flora can result in malabsorption and inflammation. Increases in motility can occur in the clinical picture of an inflamed colon, such as can occur in infants and adults. Acute hormonal influences are more common in the adult population, such as those seen with thyrotoxicosis

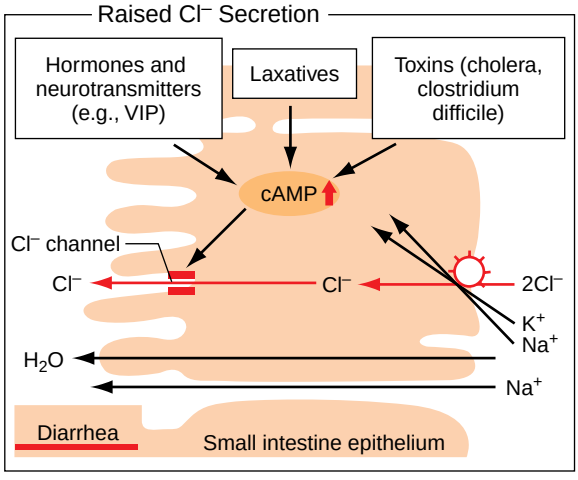


Figure 1 Chloride mechanisms of secretory diarrhea. (Reproduced with permission from Silbernagl S and Lang F (2000) p. 151 *Color Atlas of Pathophysiology*, p. 333. New York: Thieme.)

and carcinoid syndrome. Pharmacological agents or substance abuse can also influence motility.

Inflammatory diarrhea results in secretion of mucus, typically with the presence of blood in the lumen, which is also a cathartic agent. The integrity of the epithelial barrier is often compromised, with resultant exudation of water and proteins. Bacterial invasion of the mucosa may occur and is one example of inflammatory diarrhea. Additional disorders that may cause inflammatory diarrhea include allergic colitis and inflammatory bowel disease (IBD).

Lastly, diarrhea can be categorized clinically into acute and chronic forms, with the latter being defined as persistence of symptoms for more than 3 weeks. Each type of diarrhea can be further clinically divided based on age with respect to likelihood of cause.

Constipation

Constipation is defined as the infrequent, painful passage of large, hard stools; the frequency of defecation may be variable. Breast-fed infants may stool more than five times per day or as little as once every 3 days. Children age 2 years average two stools per day. Adults stool between three times a day and three times per week on average. The colon functions to store stool as well as to absorb fluids.

Eating triggers the gastrocolic reflex, mediated by the enteric nervous system and gastrointestinal hormones to stimulate propogative peristalsis. Stool traverses the colon and accumulates, filling the rectum. The stretch upon the rectal walls stimulates reflexive relaxation of the internal anal sphincter, which is sensed by the individual. If defecation is desired, relaxation of the external anal sphincter and correction of the angle of the rectum with respect to the anus (achieved by relaxation of the puborectalis muscle, the Valsalva maneuver, and proper posturing) allow for passage of stool. Abnormalities in sensation, conditions in which paradoxical contraction of the puborectalis muscle occurs instead of relaxation (anismus), pelvic floor muscle dyssenergia, and congenital absence of innervation of the rectal vault (Hirschprung's disease) also result in constipation. Anatomical structural defects, such as an anteriorly displaced anus, or an imperforate anus present as constipation in childhood.

Stool frequency and consistency are influenced by dietary factors. Fats decrease intestinal motility and facilitate greater absorption of nutrients. Dietary fiber acts as a stool bulking agent and is a major determinant of stool weight, size, and transit time; larger, fiber-laden stools are defecated more frequently than smaller ones. Additional factors, including physical activity level, stress, and functional changes in the environment, can influence stooling frequency. Medical conditions, such as spinal cord injuries, abdominal surgery, and hypothyroidism, as well as the use of pharmacological agents such as opioid derivatives, can decrease colonic motility and result in constipation. Transit time for stool through the colon varies greatly by age, with 8 or 9 h being the average for infants and 1.5 to 2 days for adults.

Encopresis is chronic fecal incontinence with frequent fecal soiling seen in children with otherwise normal colonic anatomy and physiology, usually triggered by a social stressor as opposed to an event such as painful passage of a large stool. The passage of this large, painful stool may result in rectal tearing, and, as a consequence, behaviors regarding defecation and its avoidance may occur, leading to stool withholding. The rectal vault is capable of accommodating stool, and, with time, the stretch receptors will diminish nerve stimulation. The stool present undergoes desiccation as more stool is propagated to the rectum, distending the rectal vault further. Liquid stool may pass the perimeter of stool and continue caudally, resulting in anal leakage. Rectal vault distension may reach a threshold at which defecatory urge may be diminished or lost.

Diagnosis of constipation and encopresis is based on a thorough history and physical examination, with additional laboratory and radiological testing as indicated. Correction of the underlying condition, and correction of dietary composition, and pharmacological therapy usually improve constipation; an initial disimpaction and colonic clean-out may also be warranted. For encopresis, additional behavioral therapy is often required. Conditions such as anismus may require additional modalities, such as anorectal manometry to diagnose and biofeedback training to aid in correcting.

Inflammatory Disorders of the Colon

Infections and Enteric Parasites

For viral and bacterial agents to cause inflammatory disease involving the gastrointestinal tract, nonspecific host defense factors of gastric acidity, gastrointestinal motility, enteric flora, barrier functions of mucus secretion and mucosal integrity (in some cases), and specific enteric mucosal immunity and systemic immune mechanisms have to be overcome.

These infections can results in vomiting, diarrhea, and abdominal pain, in addition to systemic effects such as fever. Clinical symptoms vary according to pathogen.

Bacterial virulence is facilitated by enterotoxin secretion (which may be site specific in it action, secreted prior to introduction or while within in the lumen), adherence and invasion of the mucosa, and cytotoxin production, which function to disrupt mucosal and cellular function.

Bacteria can be classified based on their pathological mechanism (**Table 1**) as well as by their site of activity and the nature of clinical signs and symptoms manifest. Signs and symptoms vary significantly by pathogen and age at presentation, with some forms presenting as crampy abdominal pain with watery diarrhea of relatively short duration, bloody diarrhea, systemic signs and symptoms of inflammation with frank sepsis, and shock. Common bacterial, viral, and parasitic infections involving the colon are outlined in **Tables 2** and **3**.

Polyps

Intestinal polyps are intraluminal protuberant tumors characterized by their gross morphological appearance, location(s), number, size, and presence (pedunculated) or absence (sessile) of a stalk. Additional salient features include specific histological features used to discriminate between types and to aid in predicting malignant potential. Extraintestinal manifestations are also associated with specific polyposis syndromes.

Age of occurrence is important with respect to clinical significance and malignant potential. Family history of polyps or polyposis syndromes can also be predictive of disease evolution and aid in screening and surveillance of family members.

In children, juvenile polyps account for approximately 90% of colonic polyps. They can be classified as hamartomatous or inflammatory and are also commonly referred to as retention polyps. Grossly, they are large (up to 3 cm), mostly pedunculated, erythematous, and friable. Fluid-filled cysts on the surface can be appreciated endoscopically; microscopically, this corresponds to dilated, mucous-laden cystic glands and some inflammatory cells (**Figure 2**).

This type of polyp is not in itself neoplastic; however, focal areas of adenomatous and epithelial changes indicate a risk of carcinoma. The presence of these findings in an index case also confers increased risk of carcinoma in first-degree relatives. The age of the subject, family history and inheritance patterns, number and location of polyps, and histology guide the frequency of surveillance colonoscopy. Symptoms of rectal bleeding usually bring these children to the attention of a physician. Polyps can cause clinically significant, but often painless, bleeding so as to cause anemia, and they can be linked to abdominal pain, rectal prolapse, or lead points associated with intussusceptions.

Syndromes associated with juvenile polyps are summarized in **Table 4**. Other intestinal polyposis syndromes are outlined in **Table 5**.

Adenomas are composed by immature cells, with the growth rate exceeding regenerative/replacement needs of the colonic crypt. Three features of these polyps aid in detecting malignant potential. Regarding size, adenomatous polyps less than 1 cm have a 2% incidence of being malignant; larger ones have a risk more than 2.5 times that of colorectal carcinoma for the general population. Regarding histology, the degree of cellular atypia defines the degree of irregularity; this can vary from region to region even within a single polyp. Villous structures are associated with the highest risk group, followed by tubulovillous types (**Figure 3**), with simple tubular types conferring the lowest risk of dysplasia. Smaller polyps tend to display a tubular nature.

Adenomatous polyps are much more common in adults than in children and bear significant malignant potential. They are usually inherited in an autosomally dominant manner. When they occur in children, it is usually in the context of familial adenomatous polyposis–neoplastic polyps that appear gradually, rarely in the first decade of life. Adenomatous polyps may be singular and isolated, multiple, limited to the colon or spread throughout the gastrointestinal tract, or occur in a syndromic manner. Gardner's syndrome, Turcot's syndrome, and Cronkhite–Canada syndrome are all adenomatous in nature.

Lymphoid hyperplasia may be mistaken for polyps and consists of sessile projections. Histologically,

Table 1 Bacterial pathogens grouped by pathogenic mechanism

Adherent	*Invasive*	*Toxigenic*	*Cytotoxic*
Enteropathogenic *Escherichia coli*	Shigella	Shigella	Shigella
Enterohemorrhagic *E. coli*	Salmonella	Enterotoxigenic *E. coli*	Enteropathogenic *E. coli*
Enteroaggregative *E. coli*	*Yersinia enterocolitica*	*Yersinia enterocolitica*	Enterohemorrhagic *E. coli*
Diffuse-adherent *E. coli*	*Campylobacter jejuni*	Aeromonas	*Clostridium difficile*
	Vibrio parahemolyticus	*Vibrio cholerae*	

Table 2 Bacterial enteric infections

Name	*Epidemiology and pathogenesis*	*Clinical features*	*Diagnosis and treatment*
Shigella *S. dysentriae* *S. sonnei* *S. flexneri* *S. boydii*	Acute infection Highly contagious; low infective dose (10–100 organisms)	Bacterial dysentery Crampy abdominal pain and watery stools Progressive to bloody, mucoid, pus-laden stools Tenesmus Fever Meningismus Febrile seizures in younger patients Hemolytic uremic syndrome (HUS) possible	Diagnosis Crypt abscesses Lymphatic hypertrophy Necrosis Elevated WBC count Stool culture Treatment Fluid and electrolyte replacement Hand washing to prevent transmission Limited role of antibiotics
Salmonella *S. typhi* *S. paratyphi* *S. enteritidis*	Infective load: 10^3–10^5 organisms Reservoirs: poultry and eggs, lizards, amphibians Raw/undercooked foods Mucosal invasion (jejunum and colon Inflammatory response with active secretion	Five clinical syndromes Acute gastroenteritis (12- to 72-h incubation) Focal, nonintestinal infection Bacteremia Asymptomatic carrier state Enteric fever Abdominal cramping, nausea Vomiting Bloody, mucoid stools Rose spots on the trunk Leucopenia Prolonged excretion possible, variable by age Carrier state not uncommon	Diagnosis Stool culture Treatment Supportive Very limited role for antibiotics
Campylobacter	Transmission by contaminated foods (poultry, eggs, milk; water; domestic animals) Initial site(s): jejunum colon	Incubation period of 2–11 days Fever prodrome Severe diarrhea Tenesmus Abdominal pain	Diagnosis Incubation for culture Treatment Supportive Role for antibiotics for limiting excretion period and duration of illness
Clostridium *C. difficile*	Enteric flora Toxin A (enterotoxin: alters permeability; inflammation mediator) Toxin B (cytotoxin) Pseudomembranes	Antibiotic-associated diarrhea May be restricted to the right colon	Diagnosis Stool *C. difficile* toxins A and B ELISA Endoscopy and histology Pseudomembranes (mucin, fibrin, polymorphonuclear lymphocytes, necrotic debris) Erythema Edema Friability Apthous ulcers Treatment Supportive therapy Cessation of offending antibiotic Metronidazole as first-line agent Vancomycin secondary agent Probiotics useful in relapse prevention
Yersinia *Y. enterolytica*	Transmission Contaminated pork, Enterotoxin elaboration		Diagnosis Cultures not very accurate

Continued

Table 2 Continued

Name	*Epidemiology and pathogenesis*	*Clinical features*	*Diagnosis and treatment*
			Endoscopy Mucosal ulcerations, friability throughout colon, and terminal ileum possible Histology Lamina propria infiltration of inflammatory cells Ulcerative and necrotic areas Dilated crypts Treatment Antibiotics
Aeromonas *A. hydrophila*	Water contaminants	Three syndromes Mild watery diarrhea Bloody diarrhea Persistent diarrhea	Diagnosis Stool culture Treatment Antibiotics
Escherichia coli			
Enteropathogenic *E. coli* (EPEC)	Localized adherence to enterocytes Signal transduction Intimate adherence and effacement	Diarrhea Vomiting Malaise Fever Mucoid, nonbloody stools Two-week duration	Diagnosis Presence of adherent organisms on small intestinal/rectal biopsy Treatment Antibiotics
Enterotoxigenic *E. coli* (ETEC)	Enterotoxin elaboration Heat labile (LT) toxin Heat stabile (ST) toxin Fimbriae-based attachment Stimulate adenylate cyclase (LT) and guanylate cyclase (ST) to secrete fluid	Nausea Abdominal pain Watery diarrhea Traveler's diarrhea	Diagnosis Bioassays Immunoassays Gene probes for ST or LT Treatment Supportive Antibiotics decrease duration of excretion; not recommended for children
Enteroinvasive *E. coli* (EIEC)	Colonize colon Invade tissue Replicate within cells Secretory enterotoxins	Shigella-like Watery diarrhea Then, bloody mucoid, pus-laden diarrhea Tenesmus and fever possible	Diagnosis Bioassays Serotyping ELISA Treatment Supportive Limited antibiotic role
Enterohemorrhagic *E. coli* (EHEC)	Part of normal enteric flora in healthy animals Cytotoxin similar to Shiga toxin Adherence O157:H7 prototypical Transmission Contaminated, undercooked meat Unpasteurized apple cider Children and the elderly more prone to HUS	Hemorrhagic colitis Crampy abdominal pain Watery diarrhea progressing to bloody stools Absence of fever HUS	Diagnosis Serotyping Serum antibody tests Cytotoxin bioassays DNA hybridization PCR-based tests ELISA Treatment No effective therapy Supportive care Dehydration correction Management of electrolyte abnormalities Blood transfusions as necessary
Enteroaggregative *E. coli* (EAEC)	Localized adherence likely (HEp-2 or HeLa cells) Enterotoxin Increased intestinal mucus secretion	Diarrhea Watery Mucoid Persistent	Diagnosis DNA probes
Diffuse adherent *E. coli* (DAEC)	Diffuse adherence likely (HEp-2 or HeLa cells)	Diarrhea	Diagnosis DNA probes

Table 3 Additional colonic pathogens

Name	*Pathogenesis*	*Clinical symptoms*	*Diagnosis and treatment*
Amoeba *Entamoeba histolytica*	Travel to endemic areas a risk factor Large intestinal commensal organism Transmission Person-to-person contact Contaminated food/water (cysts) Cysts transform into trophozoytes at the terminal ileum Invade mucosa and submucosa	Acute onset Fulminant colitis Bloody, mucoid diarrhea Abdominal distention Abdominal pain Perforation possible Hepatic abscesses possible	Diagnosis Histopathology: Hyperemia and edema Acute inflammation Microulceration Flask ulcer formation (Fresh) stool examination for cysts or trophozoites Treatment Iodoquinol Metronidazole
Helminths *Trichuris trichura* (whipworm)	Primarily colonic	Heavy infestations associated with (bloody) diarrhea Rectal prolapse	Diagnosis Stool assays Treatment Thiabendazole Mebendazole
Schistosomiasis *S. mansoni*	Snail as pathogen Contaminates fresh water	Dysenteric-like illness Bloody diarrhea Perianal fistulas	Diagnosis Endoscopic Focal and diffuse fibrosis Intraluminal Granulamatous masses (bilharziomas) Stool exam for viable eggs Treatment Praziquantal

hyperplastic lymphoid follicles are present; it is not uncommon for ulcerations to overlie these areas.

Inflammatory polyps (also referred to as pseudopolyps) can be seen during the recovery phase from inflammation or in inflammatory diseases, and they are often seen in the context of IBD. They can be associated with phases of regeneration and are pleimorphic in nature.

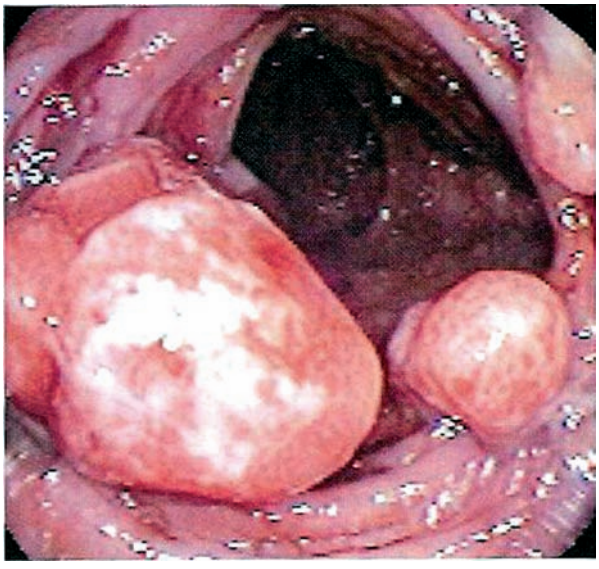


Figure 2 Endoscopic view of colonic polyps in a patient with juvenile polyposis col. (Reproduced with permission from Kleinman RE, Gilger MA, Braverman RM, Finegold MS, Hawkins EP, and Klish WJ (eds.) (1998) *Atlas of Pediatric Gastrointestinal Disease*. Hamilton Ontario: Decker.)

Inflammatory Bowel Disease

The term inflammatory bowel disease encompasses ulcerative colitis and Crohn's disease. Indeterminate colitis is a diagnosis attributed to a condition in which clear distinction cannot be made between the two aforementioned forms of IBD, as opposed to a heterogeneous group of diseases that present over a wide clinical and histological spectrum.

Epidemiology

IBD presents in a bimodal manner as pertains to age, first in late adolescence or early adulthood and a smaller peak in the fifth decade of life. The sexes are equally affected by ulcerative colitis; in adults, the incidence of Crohn's disease is 20–30% higher in women.

In terms of trends in disease over time, the incidence of ulcerative colitis remained stable during the second half of the twentieth century; Crohn's disease has demonstrated a marked increase across all age groups since 1950. Although IBD can affect all races, Caucasians are affected significantly more than Africans or people of African origin.

Table 4 Hamartomatous intestinal polyps

Syndrome	*Location of polyps*	*Pathology*	*Extraintestinal abnormalities*	*Cancer risk*
Juvenile polyposis	Colon; some small intestinal	Up to 3 cm Mucus retention and inflammatory cells in the lamina propria cysts Mostly pedunculated		Colonic; low risk
Peutz–Jeghers	Mostly small intestinal; some gastric and colonic	1–3 cm Either sessile or pedunculated Glandular epithelium and smooth muscle branching	Macular pigmentation on hands, lips, and mouth	Up to 18 times versus the general population; lower than other polyposis syndromes
Cowden's syndrome	Colon and stomach	Multiple polyps Hamartomatous	Lipomas Papillomas Orocutaneous hamartomas	Fibrocystic or fibroademomatous, ductal breast cancer Nodular thyroid hyperplasia or follicular adenoma

Table 5 Polyposis syndromes

Type/syndrome	*Location(s)*	*Histology*	*Clinical features*	*Cancer risk*
Familial polyposis coli	Colonic; fundic gland hyperplasia (stomach)	Thousands of adenomas Elevated ornithine decarboxylase levels APC gene	Apparent after puberty Diarrhea most common symptom Abdominal pain Hypertrophic retinal lesions	Thyroid cancer Pancreatic cancer Risk of colon cancer 100% by 55 years of age
Gardner's syndrome	Colon, stomach, duodenum, small intestine	2–5 mm Sessile mostly Adenomas in the antrum and periampular regions More than 1000 over time	Triad of: Polyps Osteomas Soft tissue tumors Also dental abnormalities	Duodenal tumors at highest risk Associated risk of Pancreatic carcinoma Ampullary cancer Hepatoblastoma
Turcot's syndrome	Colonic	Adenomatous polyps	Presents in adolescents with cancer; family history Autosomal recessive	Associated neural tumors Medulloblastmas Gliomas
Cronkhite–Canada syndrome	Throughout gastrointestinal tract	Adenomatous lesions within adenomatous polyps	Alopecia Nail dystrophy Brown macular skin lesions Edema related to protein-losing enteropathy	5% of cases evolve into gastrointestinal carcinomas
Inflammatory polyposis	Colonic; pseudopolyps	Pleiomorphic Regenerative tissue	Systemic signs and symptoms of inflammation	Colonic; risk of cancer from inflammatory bowel disease (Crohn's disease and ulcerative colitis)

Ashkenazi Jews have a markedly increased risk of IBD compared to other Jewish groups. The incidence in the Ashkenazi Jewish population roughly parallels that of the respective geographical community in which they reside, albeit at a level that can be three or four times that of the general population, suggesting a genetic predisposition. The majority of individuals affected by these disorders reside in North America and northern Europe. The remainder of Europe, Latin America, and Australia have lower incidence rates, and rare cases occur in Africa and Asia.

Etiology

The exact etiology of IBD is unclear and an area of active research. A multifactorial interaction between genetic predisposition, environmental stimuli, endogenous triggers, immunological dysregulation, and modifying factors is postulated.

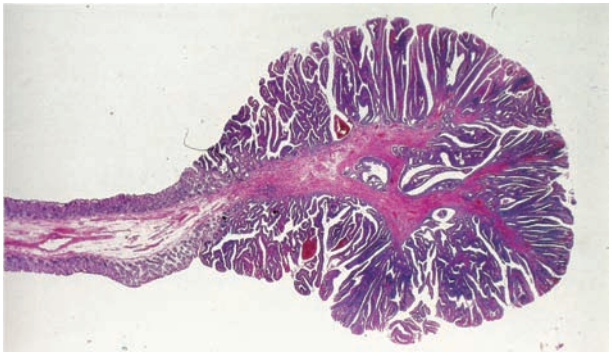


Figure 3 Polyp histology: tubullovillous adenoma. (Reproduced with permission from Wheater PR, Burkitt HG, Stevens A, and Lowe JS (1991) *Basic Histopathology*, 2nd edn. Churchill Livingston, UK.)

Genetics

A positive family history confers significant risk (10–20%) of disease occurrence of either disease type in a first-degree relative. The roles of race and ethnicity were discussed previously, with northern European and North American populations, particularly the Ashkenazi Jewish population globally, having the highest risk of disease.

A high rate of concordance among Swedish monozygotic twins versus dizygotic twins has been reported for Crohn's disease (44 vs 3.8%). In the same study, the incidence rate observed in monozygotic twins for ulcerative colitis was 6.3%. These data, although supportive of a genetic role, show less than 100% penetrence, suggesting that although genetics are more important in Crohn's disease than in ulcerative colitis, environmental influences play a significant role. Simple Mendelian models of inheritance are inadequate to address the complex inheritance patterns of IBD. Candidate gene studies have suggested modest HLA associations that differ in different populations. Systemic genome searches performed on families with several members with IBD have employed linkage analyses. Evidence that the *NOD-2* gene on chromosome 16 is involved in Crohn's disease has led to it being labeled the *IBD1* gene locus. This gene is involved with the encoding of a protein associated with monocytic nuclear factor-κB; this protein and pathway are involved in the interaction of monocytes with bacterial peptidoglycans. Note that only approximately 30% of individuals with Crohn's disease are positive for this particular gene mutation.

Environmental Influences

Because of the rapid increase in Crohn's disease during the past 50 years, increasing trends in immigrant populations, as well as incomplete genotype–phenotype associations, attention has focused on environmental factors. In particular, the search to identify an antigenic trigger for the enteric immune system has been pursued by several investigators. Postulated microbial intraluminal triggers include mycobacterium and viruses. Dietary antigens or toxins have not been identified; Westernized diet has been explored and remains an active area of research. Exposures early in the life cycle (birth environment) and nutritive factors (breast vs formula feeing; the former confers protective effects) have also been considered. Additional modulating factors include smoking and the use of oral contraceptives.

Pathogenesis

The interactions between the enteric immune system and the intestinal lumen are dynamic; some degree of inflammation in response is always present in the normal mucosal lamina propria of the colon and small intestine, which handle a very large antigenic load daily. An intact mucosal barrier, in addition to normally functioning immunoregulatory mechanisms, prevents this interaction from progressing to the level at which tissue injury occurs.

Current chronic, inflammatory relapsing disease processes may represent an inappropriate persistent immune response to a luminal antigen/stimulus versus an appropriate immune response to a persistent, abnormal stimulus or perhaps a prolonged immune response to a ubiquitous stimulus.

Enteric flora may play a role in this process, although no evidence strongly indicates a single pathogen. Defective mucosal barrier function and increased intestinal permeability may also be involved, with the latter being documented in patients with IBD and in up to 10% of nonaffected first-degree relatives.

The immune response is primarily T cell mediated and of a Th-1 nature—interleukin-12, interferon-gamma, and tumor necrosis factor-alpha (TNF-α). White blood cells respond to these inflammatory mediators and proliferate the immune response. These recruited cells synthesize agents such as arachidonic acid metabolites, platelet activating factor, proteases, and free radicals such as reactive oxygen species—all of which cause direct injury to cells and the mucosa.

Pathology

Pathology differs between these two disorders in terms of anatomical distribution and tissue involvement. Ulcerative colitis is limited to the colon and rectum, usually beginning distally in the rectum and extending to varying lengths proximally in a continuous manner (**Figure 4**). Usually, a clear distinction

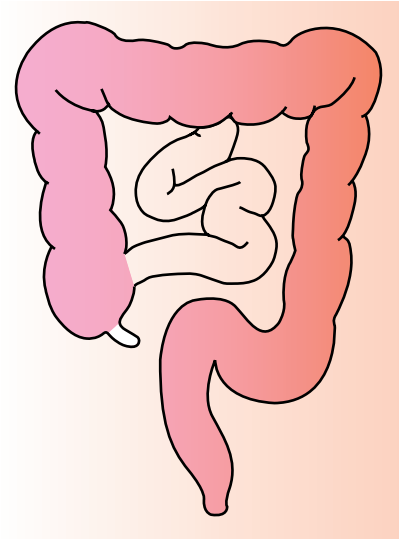


Figure 4 Continuous distribution of ulcerative colitis.

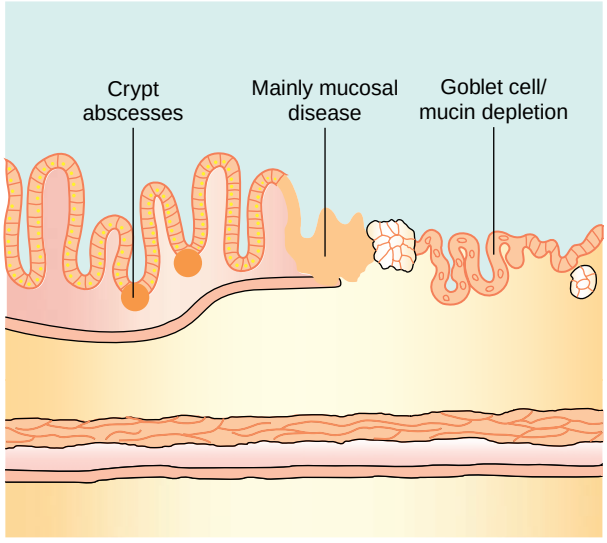


Figure 5 Illustration of ulcerative colitis. (Reproduced with permission from Kelly DA and Booth IW (1996) *Pediatric Gastroenterology & Hepatology*. London: Mosby–Wolfe.)

can be made where disease ends and normal mucosa can be appreciated grossly or endoscopically. The gross appearance of the mucosa is dependent on the severity of the disease process. Mild disease presents with a diffuse erythema and loss of the characteristic appearance of the vasculature. Numerous small, superficial ulcerations, exudates, and bleeding are seen in moderate disease; larger, deeper ulcerations, increased exudates, and the development of pseudopolyps are seen in severe disease, with loss of normal gross architectural landmarks such as the folds. Microscopically, ulcerative colitis is limited to the mucosa; with more severe disease, deeper layers may show a degree of involvement, with inflammatory cell infiltrates, shortening, branching, and decreases in the number of crypts as well as crypt abscesses (**Figure 5**).

Crohn's disease may involve any part of the alimentary tract from the mouth to the anus, and it frequently does so in a discontinuous manner, leaving 'skip areas'—regions that are grossly and histologically normal; in the colon, this lends a cobblestone appearance. Macroscopically, wall thickening is evident in longstanding disease. By definition, this disease is a transmural process (**Figure 6**). With chronic disease, fibrostenosis occurs, narrowing the intestinal lumen. Stricturizing disease may follow fibrosis of superficial and deeper layers of the intestinal wall, evident on radiographic studies (**Figure 7**).

The mesentery may also demonstrate inflammation, with resutant adhesion and fixation of the colon. Adjacent loops of bowel may become matted together. As luminal diameter narrows, intraluminal pressure may increase; in the case of nonabating inflammation, this transmural process may lead to fistula formation. Enteroenteric fistulas are limited to the bowel; enterovaginal, enterovesicular, and enterocutaneous fistulization may occur. Inflammatory intraabdominal masses called phlegmons may also form by this fistulization process.

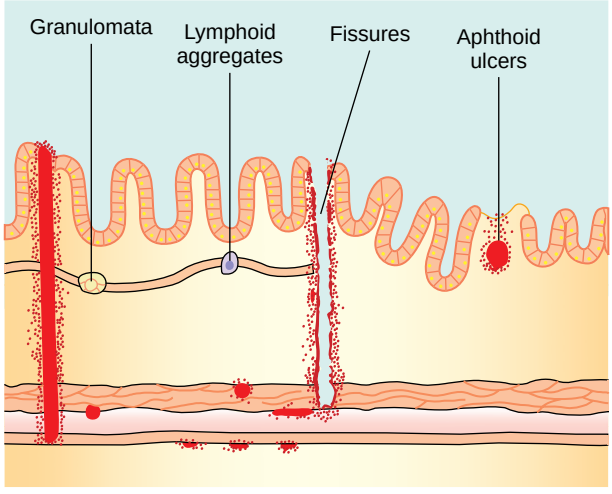


Figure 6 Illustration of Crohn's disease. (Reproduced with permission From Kelly DA and Booth IW (1996) *Pediatric Gastroenterology & Hepatology*. London: Mosby–Wolfe.)

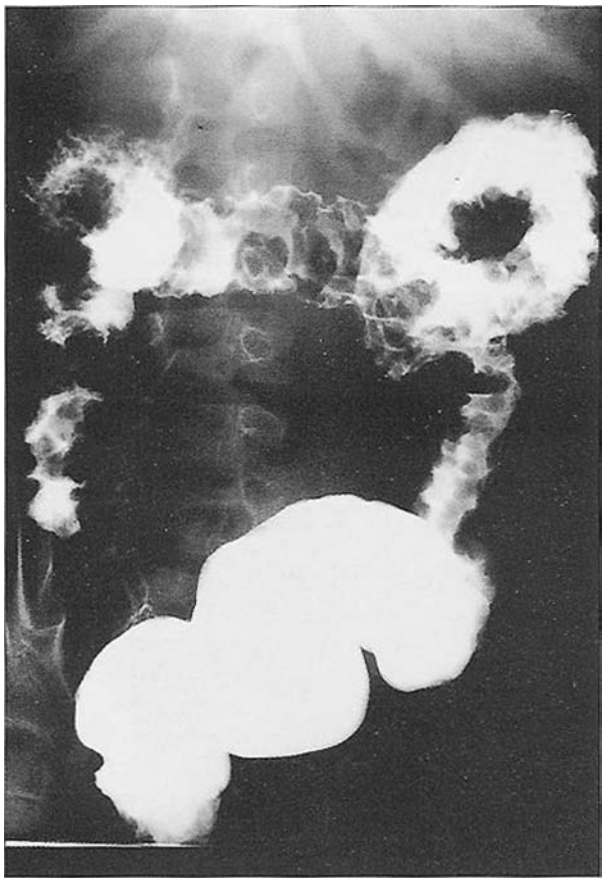


Figure 7 Barium enema in a patient with colonic Crohn's disease. Note the bowel wall ulcerations, structuring, and asymmetric bowel wall edema. (Reproduced with permission from Kelly DA and Booth IW (1996) *Pediatric Gastroenterology & Hepatology*. London: Mosby–Wolfe.)

The endoscopic appearance of Crohn's disease varies by location and time relative to disease evolution. Intestinal Crohn's disease may initially present with apthous ulceration overlying Peyer's patches in the colon. Ulcerations eventually grow, with frankly friable, exudative lesions.

Histological findings of affected areas include intense inflammatory cell infiltrates extending into the crypts, with shortening and forking of these structures, and associated abscesses. The inflammation is transmural; fibrosis and histiocyte proliferation are also seen. Noncaseating granulomatous submucosal and mucosal lesions, which are a hallmark of this disease, are not found in a majority of biopsy specimens. Granulomas can also be seen in intestinal infections, such as in intestinal tuberculosis and sacoidosis. Even macroscopically appearing normal tissue may yield histological findings of inflammation compatible with Crohn's disease, thus indicating that examination of the entire alimentary canal and surveillance biopsies are required prior to arriving at a diagnosis.

IBD and the Terminal Ileum

The terminal ileum is one of the most commonly involved sites in the intestine in Crohn's disease, often originating at the lymphoid follicle; strictures may form. A phenomenon of ileal involvement has been postulated in cases of apparent ulcerative colitis that involve the ileum, in which cecal inflammation is postulated to 'backwash' into the ileum; this finding being consistent with ulcerative colitis is controversial.

Extraintestinal manifestations are common in both Crohn's disease and ulcerative colitis, including ophthalmologic manifestations (uveitis), joint involvement (arthralgias and arthritis of the large joints), and manifestations of the skin, hepatobiliary system, pancreas, renal system, and vascular system. Anemia and weight loss are common at the time of presentation. Growth and pubertal delay are very common at the time of presentation in children; short stature occurs in up to 50% of children. Some of these findings relate to the inflammatory process; others are linked to malnutrition associated with IBD. Perianal disease with fistulization and/or skin tags is perhaps the most common extraintestinal abnormality associated with Crohn's disease.

Nutritional Consequences of IBD

Malnutrition includes weight loss acutely, partly attributable to anorexia associated with inflammation and partly to the disease process (i.e., inadequate intake as well as excessive (malabsorptive) losses). These deficiencies can evolve from frank losses as well as be malabsorptive in nature. An example delineating all of these mechanisms is anemia, which can result from frank blood loss from associated gastrointestinal bleeding, anemia of chronic disease mediated by the inflammatory mediators, anorexia with decreased dietary iron intake, and, as in the case of duodenal and jejunal disease activity (as can occur in Crohn's disease), anorexia with decreased absorption.

Intestinal disease can result in both decreased nutrient absorption and disruption of the mucosal barrier, resulting in exudation of proteins, a process known as protein-losing enteropathy (PLE). The latter can result in hypoalbuminemia; third spacing of fluids as a result of decreased intravascular oncotic pressure can occur. Increased energy expenditure as

a consequence of inflammation is noted, particularly in the febrile state or with sepsis. Inflammation and discomfort also contribute to decreased enteral intake—factors contributory to a catabolic state.

In addition to iron, other mineral and trace element deficiencies are noted in IBD. Iron deficiency was discussed previously. Zinc is closely associated with gut mucosa and is susceptible to deficiency; low albumin levels resulting from PLE and increased intestinal epithelial cell turnover probably represent a significant source of zinc depletion. Vitamin B_{12} and folic acid deficiency has also been documented among the water-soluble vitamins, particularly when the terminal ileal disease is noted. Vitamin D deficiency is the most common among the fat-soluble vitamins.

Treatment of IBD

Several antiinflammatory treatment modalities have been employed in the treatment of IBD. Their use is dictated by disease type, location, extent, and severity. Steroids provide the cornerstone of initial therapy for acute inflammation. Five aminosalicylate derivatives, antimetabolites such as azothioprine and 6-mercaptopurine methotrexate, and newer biological agents including anti-TNF-α are currently employed.

Nutritional therapies including semielemental enteral feedings or parenteral therapy have a role in the management of IBD. Although enteral therapy is not considered a first-line therapy for ulcerative colitis in the United States, its use as such is popular in Europe and Canada and allows for steroid sparing/avoidance. The time to onset of remission using enteral therapy in Crohn's disease is much less with steroids than with enteral therapy, however, with the former occurring typically within 2 weeks and the latter taking usually 6–8 weeks to achieve similar clinical remission. Smaller studies have been conducted employing low-fat diets and less processed sugar foods. The latter has shown promise in studies of Crohn's disease, thought to be secondary to the immunomodulatory effects of these fatty acids. Another approach under investigation is enteral therapy containing TGF-β, which is thought to modulate the enteric immune response. Also under investigation are dietary fiber, short-chain fatty acids, and influencing the colonic flora with the use of prebiotics and probiotics.

Surgical treatment is indicated in ulcerative colitis when acute, fulminant disease does not respond to medical therapy or when persistent chronic disease is refractory to medical (steroid) therapy and the diagnosis has been confirmed (i.e., Crohn's disease has been ruled out). Colectomy is curative in this instance.

Crohn's disease is more complex, and surgical intervention, limited to involved segments only, is not curative. Failure of medical therapy to reduce inflammation, critical stenosis of the involved segments with fibrosis leading to obstruction, perforation, fistulization, and/or abscess formation not amenable to medical therapy, and frank gastrointestinal hemorrhage are indications for surgical intervention. Reactivation of disease can occur postoperatively at the site of anastomoses or elsewhere.

The natural history of IBD is such that long-standing disease increases the risk of colonic dysplasia, particularly in the case of ulcerative colitis, even though it is a curative intervention, making colectomy more attractive in the older patient. Ileoanal continuity can be achieved by means of surgical anastomoses. Pouchitis secondary to bacterial overgrowth, smoldering pockets of disease activity that may not have been resected or become evident after resection, and loss of continence are common complications of these procedures.

See also: **Colon**: Structure and Function; Nutritional Management of Disorders. **Diarrheal Diseases**. **Dietary Fiber**: Physiological Effects and Effects on Absorption; Potential Role in Etiology of Disease; Role in Nutritional Management of Disease. **Small Intestine**: Structure and Function; Disorders.

Further Reading

Balfour R (2002) Mucosal immunology & mechanisms of gastrointestinal inflammation. In: Feldman M, Friedman LS, and Sleisenger MH (eds.) *Sleisenger & Fordtran's Gastrointestinal and Liver Disease: Pathophysiology, Diagnosis Management*, 7th edn, pp. 21–54. Philadelphia: WB Saunders.

Bayless TM and Hanauer S (eds.) (2001) *Advanced Therapy of Inflammatory Bowel Disease*. Hamilton, Ontario, Canada: BC Decker.

Griffiths AM and Bueller HB (2000) Inflammatory bowel disease. In: Waker WA, Durie P, Hamilton R, Watkins J, and Walker-Smith J (eds.) *Pediatric Gastroenterology: Pathophysiology, Diagnosis, Management*, 3rd edn, pp. 28–38. Hamilton, Ontario, Canada: BC Decker.

Guandalini S (2000) Acute diarrhea. In: Waker WA, Durie P, Hamilton R, Watkins J, and Walker-Smith J (eds.) *Pediatric Gastroenterology: Pathophysiology, Diagnosis, Management*, 3rd edn. Hamilton, Ontario, Canada: BC Decker.

Homer DH and Gorbach SL (2002) Infectious diarrhea and bacterial food poisoning. In: Feldman M, Friedman LS, and Sleisenger MH (eds.) *Sleisenger & Fordtran's Gastrointestinal and Liver Disease: Pathophysiology, Diagnosis Management*, 7th edn, pp. 1864–1932. Philadelphia: Saunders.

Pickering LK and Cleary TG (1998) Approach to patients with gastrointestinal tract infections and food poisoning. In: Feigin RD and Cherry JD (eds.) *Textbook of Pediatric Infectious Diseases*, 4th edn, pp. 567–600. Philadelphia: WB Saunders.

Shashidhar S and Mobassaleh M (1998) Bacterial infections. In: Altshuler S and Liacouras C (eds.) *Clinical Pediatric Gastroenterology*, pp. 131–142. Philadelphia: Churchill Livingstone.

Steffen R and Loering-Burke V (1999) Constipation and encopresis. In: Willie R and Hyams J (eds.) *Pediatric Gastroenterology*, 2nd edn, pp. 43–50. Philadelphia: WB Saunders.

Yamada F, Alpers DH, Laine L, Owyang C, and Powell DW (eds.) (1999) *Textbook of Gastroenterology*, 3rd edn. Philadelphia: Lippincott, Williams & Wilkins.

Nutritional Management of Disorders

D M Klurfeld, US Department of Agriculture, Beltville, MD, USA

The primary functions of the colon are to absorb water and to form and store feces for excretion. The length of the large intestine in an adult is approximately 1.5 m; several divisions and landmarks of the colon are shown in **Figure 1**. Disturbances in colonic function are symptoms of diseases or disorders, including constipation, diarrhea, diverticular disease, irritable bowel syndrome, and inflammatory bowel diseases; due to surgical treatment of inflammatory bowel diseases, stomas are often created. Symptoms of these conditions range from mild discomfort to life-threatening emergencies, although most are chronic and can benefit from nutritional management. Although pharmacological therapy in some of these conditions may often be indicated, nutritional management is recognized as the first choice and is particularly effective in preventing and treating some of these disorders. As in many issues related to nutrition, there is a tremendous amount of misinformation believed by many. For example, there is no basis for the claims that meat remains undigested in the colon for years or that colonic cleansing with enemas or herbal preparations is of any value.

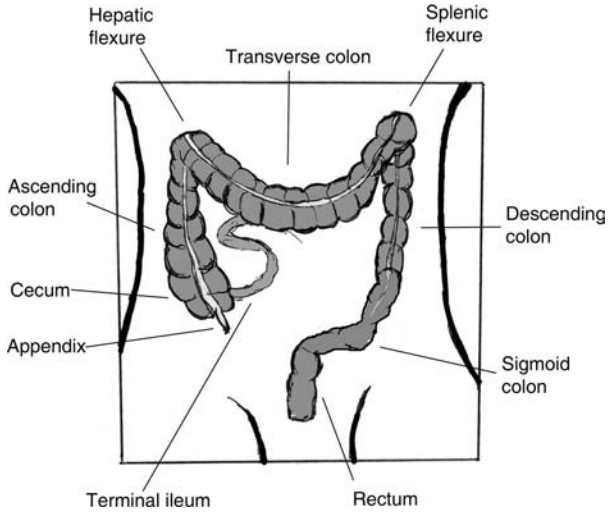


Figure 1 Diagram of the colon showing anatomical divisions and landmarks. Intestinal chyme enters via the ileum, ferments in the proximal ascending portion of the colon, and becomes feces, which is stored in the transverse and distal descending portions for elimination.

Constipation

Constipation can be defined as the slow movement of feces through the large intestine that results in the passage of dry, hard stool. Another acceptable definition is the infrequent passage of small, dry, hard feces accompanied by discomfort or pain. Many individuals self-diagnose the condition based on perceived deviations from 'normal' bowel habits; this often leads to unnecessary use of chemical laxatives that can irritate the colon and eventually lead to dependence on such preparations for evacuation. Normal bowel habits are generally deemed as at least three stools per week to no more than three per day. Constipation is often accompanied by symptoms of distension and flatulence. Diverticular disease is characterized by thinning and outpouching of the colonic wall. The diverticula are generally asymptomatic but can become infected with the potential for rupture. The major complications of this condition are bleeding and bacterial infection; the latter may result in abscess formation or perforation of an existing diverticulum with subsequent peritonitis. Although past practices were to prescribe low-residue diets to rest the bowel, it is now known that high-fiber diets are effective in the treatment and prevention of diverticular disease as well as for reducing the complication rate.

In the early 1970s, it was proposed that reduced fiber consumption in Western countries resulted in an incidence of diverticular disease that approximated 50% beyond the age of 70 years, whereas this condition was almost nonexistent in sub-Saharan Africa. Since that time, there have been contradictory studies on this point, and it seems that dietary fiber may not be the only factor that influences the development of diverticular disease. However, several explanations for some of the contrary findings in this area are evident. First, many factors in the diet are correlated; that is, a diet low in fiber tends to be high in meat and fat. Second, the measurement of dietary fiber in many studies has not included total dietary fiber, thereby suffering

from measurement bias. Third, not all sources of dietary fiber have a therapeutic effect and the benefit varies according to the specific type of food or fiber supplement along with the amount of water available.

Despite past controversy, it now seems clear that diets high in dietary fiber will prevent the development of diverticular disease, can be used successfully for symptomatic treatment, and reduce the risk of infection of the diverticula. It must be understood that once formed, diverticula do not spontaneously resolve, and surgery is the only means of removing them. Recommended dietary modifications are increases in water and dietary fiber, particularly wheat bran or psyllium. Although fruits and vegetables also contribute to the prevention or reduction of symptoms, there is controversy about including those that have seeds. Seeds often pass through the gastrointestinal tract undigested; these have been found in infected diverticula, and it was assumed that they were the nidus for infection. Therefore, many practitioners have prohibited patients from consuming foods with small seeds, such as raspberries, cranberries, and blueberries, or larger seeds, such as tomatoes, peppers, and cucumbers. In addition, seeds added to foods, such as caraway, sesame, and poppy, have been proscribed. Some do not prohibit consumption of these seed-containing foods but there is little evidence to demonstrate the safety of abandoning this advice.

Diarrhea

Diarrhea is generally defined as loose, watery stools occurring more than three times a day and is a symptom of an underlying condition. It is estimated that the average adult has four episodes of diarrhea each year. Most events are self-limited and resolve within 24–48 h; it is presumed that the majority of these are viral in etiology and numerous agents have been implicated. Bacterial causes include *Campylobacter*, *Escherichia coli*, and, less commonly, *Salmonella* or *Shigella*. The latter organism is prevalent in tropical areas of the world and one species of *Shigella* is responsible for dysentery, which is characterized by profuse, watery diarrhea particularly in children and the elderly. Parasitic infections with *Giardia lamblia*, *Entamoeba histolytica*, or *Cryptosporidium* are often linked to chronic diarrhea. Alterations of normal colonic flora secondary to antibiotic therapy may result in diarrhea; the best studied organisms that overgrow the normal bacteria are *Clostridium* species. Noninfectious causes of diarrhea may be lactose intolerance, excess consumption of sugar alcohols, irritable bowel syndrome, inflammatory bowel disease, or celiac disease.

Clearly, the underlying cause of chronic diarrhea needs to be established and proper therapy instituted. Although most episodes of diarrhea will resolve spontaneously without specific therapy, nutritional management of diarrhea is primarily concerned with replacement of lost fluid. Copious or chronic diarrhea increases the need for electrolytes. Any episode of diarrhea in young children or the elderly may require electrolytes in addition to fluid replacement; this is easily obtained from oral rehydration therapy solutions made for this purpose if diarrhea is severe or protracted. Such solutions contain starch, proteins, and electrolytes and have been shown to reduce stool volume significantly. In less severe cases, maintenance of, or return to, the usual diet after 24 h is recommended. Four foods traditionally recommended for children—bananas, rice, applesauce, and toast—and referred to as the BRAT diet were thought useful because they do not irritate the colon since they are low in fiber and residue. However, this diet is no longer recommended by some pediatrics organizations because it is low in energy density, protein, and fat. During bouts of diarrhea, it is generally recommended that individuals avoid spicy foods, fatty foods, high-sugar foods, or high-fiber foods. Milk is sometimes proscribed, but evidence suggests that 80% of children with diarrhea can tolerate full-strength milk so most do not need to avoid this food, which provides more energy and protein than alternative fluid sources. Clear broth or soup is often recommended. Although clear fruit juices or soft drinks are sometimes recommended, these should be not be used or they should be diluted to avoid an osmotic effect of the sugars drawing more fluid into the intestine. There is considerable controversy regarding the consumption of specific foods during or just after a bout of diarrhea. Some studies justify use of complex carbohydrates (rice, wheat, potatoes, bread, and cereals), lean meats, yogurt, fruits, and vegetables because they are well tolerated even during active diarrhea. Many health professionals recommend a more limited diet of toast, rice, bananas, cooked carrots, and skinless chicken until symptoms abate. The choice of specific foods will depend on the tolerance of an individual, keeping in mind the fluid and energy needs of that person. Research suggests that repopulating the colonic bacteria through consumption of yogurt may provide more healthful organisms (Lactobacilli, Bifidobacteria, and *Streptococcus thermophilus*) and a quicker return of the total flora to normalcy, particularly in antibiotic-induced diarrhea.

Irritable Bowel Syndrome

Irritable bowel syndrome (IBS) is also called spastic colon or mucus colitis, and it is one of the most common causes for referral to gastroenterologists. Symptoms include abdominal pain, bloating, constipation and/or diarrhea, heartburn, belching, and mucus in the stool. IBS is diagnosed by eliminating other diagnoses or organic causes and, therefore, treatment is aimed at alleviating symptoms. Women are affected twice as frequently as men, and it is estimated that as much as 8% of the US population is afflicted with this condition. Although stress and other psychological factors seem to play a significant role in IBS, a number of dietary interventions have been suggested. Although dietary fiber has been advocated for alleviation of the symptoms of IBS, it is clear that some patients will benefit but others will get worse, so individual trial and error may be the logical therapeutic plan. Patients with abdominal distension or excessive flatulence should reduce consumption of gas-provoking foods, such as beans, lentils, cabbage, broccoli, onions, garlic, raw fruits and juices, bananas, and nuts. Caffeinated, alcoholic, and carbonated beverages cause exacerbations in some patients. In addition, high-fat foods, such as deep-fried foods, processed meats, gravies, and chocolate, as well very spicy or pickled foods, may increase symptoms of IBS. Since many IBS patients are also lactose intolerant, reduction of milk products is often recommended empirically but should probably be done based on a patient's response to these food products. Sweeteners such as sorbitol and fructose are often associated with increased diarrhea, so products containing these should be tested for effects on symptoms. In patients with constipation, the addition of fiber to the diet along with increased fluid intake will help to alleviate this symptom. The most commonly recommended types of fiber are wheat bran and psyllium. In patients with diarrhea, some have benefited from the addition of wheat bran, pectin, or kaolin to the diet, but other patients seem to do better with a low-fiber diet. Individualized trials of dietary intervention seem to be indicated for IBS. In addition to nutritional therapy, drug treatment to reduce intestinal transit and emotional or psychological support are major parts of therapy for IBS. Aerobic exercise, consumption of smaller, more frequent meals, relaxation techniques, cessation of cigarette smoking, and a variety of other environmental changes have been tried with varying degrees of efficacy.

Inflammatory Bowel Diseases

Inflammatory bowel disease (IBD) refers to a group of conditions in which inflammation involves portions of the small or large intestines. The ileum and colon are most commonly affected. The types of IBD are Crohn's disease and ulcerative colitis (UC). Although the two conditions have many common features, they can be distinguished based on clinical, x-ray, and pathological findings. Symptoms include chronic abdominal pain, cramps, rectal bleeding, or bloody stools; diarrhea and rectal bleeding are more common in UC. The diagnosis is usually made in adults younger than age 30 years and there is a preponderance of cases among whites, especially Jews. There is familial clustering of these conditions, indicating that genetic predisposition is important; approximately 20% of patients have a close relative with the same diagnosis. The onset of either disease can be acute or insidious, and the course is protracted. Both are characterized by exacerbations and remissions, sometimes of long duration. The international epidemiology of Crohn's disease suggests that it is uncommon in developing countries and has become more common in Western countries, where fiber intakes are lower and consumption of refined carbohydrates is high, but there is no direct evidence that diet plays a role in its etiology. Much research has focused on the etiology, but no clear-cut factors have been identified. Infectious and immunological mechanisms have been most thoroughly investigated with no conclusive evidence. Cigarette smoking has been strongly linked to Crohn's disease but not to UC.

As a result of chronic abdominal pain and diarrhea, many patients lose large amounts of weight. Since oral intake is often limited, there are frequently protein and multiple micronutrient deficits. It is important to replenish these nutrients, but often the early nutritional intervention for IBD relies on bowel rest. Total parenteral nutrition (TPN) accomplishes this best, but enteral liquid formulae can be used in some cases. In fact, TPN often results in short-term remission of symptoms in Crohn's disease. Unfortunately, long-term remissions are not maintained in most patients by this intervention.

Crohn's disease may involve any portion of the gastrointestinal tract but is most commonly found in the terminal ileum, often with extension to the proximal colon. In the majority of patients, multiple areas of the intestine are involved, usually separated by areas of normal intestine. The inflammatory changes are nonspecific but tend to be granulomatous and involve all layers of the intestinal wall. Because inflammation involves the entire thickness of the intestinal wall, there is a high propensity for development of fistulae into adjacent structures that are often infected by the bacteria from the

colon. Bowel obstruction from strictures or adhesions tends to occur more frequently in Crohn's disease than in UC.

Ulcerative colitis often begins in the distal colon or rectum and progresses to involve most or all of the colon. Approximately one-fourth of cases have involvement of the terminal ileum, usually in continuity with the colonic manifestations. Ulcerative colitis involves primarily the mucosal layer. The chronic inflammation often leads to shortening and narrowing of the muscle layer in which the colon has been likened to resembling a garden hose. The ulcers tend to undermine portions of the mucosa and this frequently gives rise to pseudopolyps. The incidence of colon cancer is greatly increased in patients with chronic UC; tumors are often multiple and tend to have a poorer prognosis than in sporadic colon cancer. Thus, prophylactic subtotal colectomy is required in approximately 30% of patients who have persistent signs of colonic dysplasia. One of the surgical developments in the therapy of UC is the production of an ileoanal reservoir or 'J pouch,' which becomes a functional rectum, allowing a patient to defecate through the rectum rather than via a colostomy. Some patients with this type of surgery continue to have many bowel movements per day, so limitation of dietary fiber or fruits and vegetables is often required; gas-inducing foods should be avoided. A complication of this surgery is infection of the reservoir or pouchitis. Although no foods or food groups have been identified as contributing to this condition, it is recommended that patients pay attention to foods that may be associated with episodes of pouchitis.

In acute phases of IBD, bland low-residue or elemental diets are recommended; sometimes, total bowel rest is required and TPN is prescribed. Malabsorption of many nutrients has been documented in patients with IBD; the estimation of macronutrient needs is not difficult, but the determination of vitamin and mineral losses is problematic. In UC, elimination of milk products is often advised to reduce the amount of fermentable carbohydrate (lactose) entering the colon, which contributes to bloating, cramps, and diarrhea. In Crohn's disease, intestinal strictures are often found; these are contraindications to high-fiber diets because of the possibility of intestinal obstruction, which is a fairly common complication of the condition. The current consensus is to use low-fiber, low-milk, low-fat diets in patients with IBD; however, controlled studies have found little significant benefit of dietary intervention. This is because some dietary components are beneficial in certain patients but have no, or detrimental, effects in others. It is important for the patient to receive nutritional counseling to avoid deficiencies of calories and most nutrients. However, there is little specific nutritional therapy available for these conditions. Some patients find that spicy foods or alcoholic beverages exacerbate symptoms; in others, wheat bran or raw fruits have the same effect. The only hard and fast rule is to avoid foods that provoke symptoms while maintaining as nutritious and balanced a diet as possible.

Research studies in which UC patients were treated with short-chain fatty acid (SCFA) enemas generally showed considerable improvement in symptoms. Because dietary fiber is fermented to SCFA, this suggests that high-fiber diets may be beneficial. If this turns out to be the case, it is likely that specific types of dietary fiber will be recommended to achieve defined concentrations of one or more of the SCFAs.

Surgery and drug treatments are the mainstays of therapy, but there is no cure. In fact, surgical removal of an affected portion of the intestine must be weighed quite seriously because the disease may recur in previously normal tissue. Immunosuppressive medications are the standard therapy for these conditions. Because fish oils, rich in *n*-3 fatty acids, have immunosuppressive properties, these products have been studied and found to have benefit in treating Crohn's disease that is approximately equal in effectiveness to immunosuppressive drugs, but side effects of indigestion and bad breath were major limiting factors in acceptance of the fish oil. The best study done in this area was conducted in Italy, and it is unknown if the fat types and amounts in the Italian diet play some interacting role with the fish oil treatment. In addition, because Crohn's disease is heterogeneous, it is not clear if all patients will benefit from this treatment.

Sublingual vitamin B_{12} has been recommended by some for patients with Crohn's disease because absorption of this vitamin occurs in the ileum, which is the segment of gut frequently affected. However, the absence of the specific transport system for absorption of the vitamin and intrinsic factor, produced in the stomach, call into question the benefit of the route of administration. The efficacy of sublingual vitamin B_{12} is primarily a result of swallowing the vitamin, with subsequent intestinal absorption. Patients should have regular blood counts and intramuscular injection of vitamin B_{12} may be necessary if macrocytic anemia develops.

One of the complications in patients with Crohn's disease who have had ileal resections is increased formation of calcium oxalate kidney stones. This is

due to enhanced absorption of oxalate. Since limiting dietary sources of oxalate is considered too restrictive, calcium supplementation and increased fluid intake are recommended. The calcium will decrease oxalate absorption if taken with meals, and the increased fluids will dilute the urine. Since ascorbic acid increases urinary oxalate, supplements and dietary sources rich in vitamin C should be used judiciously, if at all.

Stomas

Although the creation of a surgical stoma is associated with some dietary restrictions, many patients find that there are fewer prohibited foods following the creation of the stoma than prior to surgical resection. Jejunostomies are usually indicated for treatment of Crohn's disease. Ileostomies are created when the colon is removed typically due to a disorder such as ulcerative colitis or bypassed due to trauma (usually short term and reversible). Continent ileostomies are made by creating an internal reservoir that is drained by the patient several times a day via a tube inserted through the stoma; this avoids the necessity of wearing an external appliance. Colostomies are made by attaching a portion of the colon to the abdominal wall, often as a treatment for cancer, IBD, or trauma. The site of the colostomy often determines the nutritional modifications a patient requires; a stoma in the proximal colon (**Figure 1**) produces more liquid feces, whereas one placed distally results in more solid fecal material. Intestinal excreta are collected into a plastic bag attached to a device around the stoma. Postsurgical diets reflect a transition from clear liquids through low-fiber diets to a relatively unrestricted diet. A number of concerns regarding patients with stomas are reflected in nutritional management.

One of the more important complications of bowel resection and stoma placement is the development of short bowel syndrome. Patients with this condition have reduced absorption of most nutrients, which is accompanied by diarrhea. The degree of symptoms depends on the portion and length of intestine resected. Postsurgical nutritional management usually consists of TPN to reduce an osmotic effect of food in the remaining gut. There is concern about long-term bowel rest inducing intestinal atrophy and allowing bacterial translocation; this is based almost exclusively on studies in animals but there are few data from humans. Formula diets given enterally are indicated if the length of the remaining small bowel is insufficient for adequate digestion and absorption of a normal diet. Combinations of enteral, parenteral, and normal feeding may be required. Specific advice depends on the length of remaining intestine and which portions were resected.

Since one of the primary physiological functions of the colon is reabsorption of water and electrolytes, surgical removal of this organ can result in excessive losses. Many patients voluntarily restrict their fluid intake in the mistaken belief that this will reduce effluent volume; however, excess water intake is eliminated primarily via the kidneys. In addition, the typical patient will need to be reminded to consume adequate sodium and potassium. When stomal losses exceed oral intake, parenteral fluids and minerals are required. Absorption of most trace elements and vitamins can also be reduced and, when needed, these are also supplied parenterally. A number of foods have been found to influence the amount of effluent. Those that increase volume are beans, cabbage, broccoli, spinach, raw fruits and juices, many spicy foods, red wine, and beer. Foods that are associated with decreased effluent volume include rice, bananas, applesauce, cooked milk, and peanut butter. Limited research indicates that pectin supplements can control excessive ileal effluent. Often, there are differences in the way individuals react to specific foods or combinations of them, so a food diary is highly recommended to relate changes in effluent to diet.

Because the diameter of a surgical stoma can be less than that of the intestine, certain foods can cause problems if chewed insufficiently. Therefore, chewing food thoroughly is an important part of the nutritional advice for a patient with a stoma. Some high-fiber foods are also relatively undigested and may need to be reduced or eliminated to control fecal volume and viscosity. These include corn, popcorn, nuts, coconut, celery, and raw fruits, particularly the skins and seeds.

Foods that are associated with gas production should be avoided if the patient is concerned with odor or flatulence from the stoma, which are common problems with this type of surgery. Such foods include the cabbage family, onions, beans, nuts, peppers, chocolate, carbonated beverages, eggs, and alcoholic beverages. Habits that encourage swallowing of air should also be minimized, such as gum chewing or drinking through a straw.

Weight control is important for patients with stomas to avoid some of the complications involving the skin surrounding the stoma. There is also a tendency for patients to gain weight once their primary gastrointestinal disease is treated successfully. There are a number of support groups for patients

that provide nutrition advice, but it is important to distinguish between claims based on hype and soundly conducted studies. Nutritionists with proper academic credentials are generally the best source of accurate information.

See also: **Colon**: Structure and Function; Disorders. **Diarrheal Diseases**. **Nutritional Support**: Adults, Enteral; Adults, Parenteral; Infants and Children, Parenteral.

Further Reading

Beers MH and Berkow R (eds.) (1999) *The Merck Manual of Diagnosis and Therapy*, 17th edn. Whitehouse Station, NJ: Merck.

Coulston AM, Rock CL, and Monsen ER (eds.) (2001) *Nutrition in the Prevention and Treatment of Disease*. San Diego: Academic Press.

Haubrich WS, Shaffner F, and Berk JE (1995) *Bockus Gastroenterology*, 5th edn. Philadelphia: WB Saunders.

Kinney JM, Jeejeebhoy KN, Hill GL *et al.* (eds.) (1988) *Nutrition and Metabolism in Patient Care*. Philadelphia: WB Saunders.

Lewis JD and Fisher RL (1994) Nutrition support in inflammatory bowel disease. *Medical Clinics of North America* **78**: 1443–1456.

Nightingale JM (1995) The short-bowel syndrome. *European Journal of Gastroenterology and Hepatology* **7**: 514–520.

Ozick LA, Salazar CO, and Donelson SS (1994) Pathogenesis, diagnosis, and treatment of diverticular disease of the colon. *Gastroenterologist* **2**: 299–310.

Sax WC and Souba WW (1993) Enteral and parenteral feedings. Guidelines and recommendations. *Medical Clinics of North America* **77**: 863–880.

Shils ME, Olson JA, Shike M *et al.* (eds.) (1999) *Modern Nutrition in Health and Disease*, 9th edn. Baltimore: Williams & Wilkins.

Spíller GA (ed.) (1993) *CRC Handbook of Dietary Fiber in Human Nutrition*, 2nd edn. Boca Raton, FL: CRC Press.

Zeman FJ and Ney DM (1996) *Applications in Medical Nutrition Therapy*, 2nd edn. Englewood Cliffs, NJ: Prentice Hall.

COMPLEMENTARY FEEDING

K G Dewey, University of California—Davis, Davis, CA, USA

Introduction

Complementary feeding has been defined as

> . . .the process starting when breast milk alone is no longer sufficient to meet the nutritional requirements of infants, and therefore other foods and liquids are needed, along with breast milk.
>
> (PAHO/WHO, 2003, p. 8)

In the past, such foods were often called 'weaning foods.' However, the term 'complementary foods' is preferred because weaning implies the cessation of breastfeeding, whereas the goal is that such foods should complement breast milk, not replace it. Breast milk alone is generally sufficient to meet the nutrient needs of infants during the first 6 months of life, but after that time infants need additional sources of nutrients. The World Health Organization (WHO) recommends exclusive breastfeeding for 6 months and continued breastfeeding thereafter (along with the provision of safe and appropriate complementary foods) until 2 years of age or beyond. Therefore, the period of complementary feeding usually refers to the age range of 6–24 months. This is a critical time because it is the peak age for growth faltering, deficiencies of certain micronutrients, and common childhood illnesses such as diarrhea.

In 1998, WHO and UNICEF jointly published a document entitled *Complementary Feeding of Young Children in Developing Countries: A Review of Current Scientific Knowledge*, which provided background information to assist in the development of scientifically based feeding recommendations and intervention programs. In 2003, an update to that document was published (see Further Reading), and a separate document entitled *Guiding Principles for Complementary Feeding of the Breastfed Child* was published by the Pan American Health Organization and WHO (see **Table 1**). This chapter will summarize the information presented in these three documents.

Age of Introduction of Complementary Foods

In 2001, the WHO Expert Consultation on the Optimal Duration of Exclusive Breastfeeding reviewed the evidence regarding the age of introduction of complementary foods and concluded that exclusive breastfeeding for 6 months is beneficial for both the infant and the mother. Risks arising from the early introduction of complementary foods include reduced breast-milk intake and a higher incidence of infant gastrointestinal infections. On a population basis, there is no adverse effect on infant growth of waiting to

Table 1 Guiding principles for complementary feeding of the breastfed child

1. *Duration of exclusive breastfeeding and age of introduction of complementary foods*
   Practice exclusive breastfeeding from birth to 6 months of age, and introduce complementary foods at 6 months of age (180 days) while continuing to breastfeed.
2. *Maintenance of breastfeeding*
   Continue frequent on-demand breastfeeding until 2 years of age or beyond.
3. *Responsive feeding*
   Practice responsive feeding, applying the principles of psychosocial care. Specifically: (a) feed infants directly and assist older children when they feed themselves, being sensitive to their hunger and satiety cues; (b) feed slowly and patiently, and encourage children to eat, but do not force them; (c) if children refuse many foods, experiment with different food combinations, tastes, textures, and methods of encouragement; (d) minimize distractions during meals if the child loses interest easily; (e) remember that feeding times are periods of learning and love – talk to children during feeding, with eye-to-eye contact.
4. *Safe preparation and storage of complementary foods*
   Practice good hygiene and proper food handling by (a) washing caregivers' and children's hands before food preparation and eating, (b) storing foods safely and serving foods immediately after preparation, (c) using clean utensils to prepare and serve food, (d) using clean cups and bowls when feeding children, and (e) avoiding the use of feeding bottles, which are difficult to keep clean.
5. *Amount of complementary food needed*
   Start at 6 months of age with small amounts of food and increase the quantity as the child gets older, while maintaining frequent breastfeeding. The energy needs from complementary foods for infants with 'average' breast-milk intake in developing countries are approximately 200 kcal day^{-1} at 6–8 months of age, 300 kcal day^{-1} at 9–11 months of age, and 550 kcal day^{-1} at 12–23 months of age. In industrialized countries these estimates differ somewhat (130, 310, and 580 kcal day^{-1} at 6–8, 9–11, and 12–23 months, respectively) because of differences in average breast-milk intake.
6. *Food consistency*
   Gradually increase food consistency and variety as the infant gets older, adapting to the infant's requirements and abilities. Infants can eat puréed, mashed, and semi-solid foods from the age of 6 months. By 8 months most infants can also eat 'finger foods' (snacks that can be eaten by children alone). By 12 months, most children can eat the types of foods consumed by the rest of the family (keeping in mind the need for nutrient-dense foods, as explained in #8 below). Avoid foods that may cause choking (i.e., items that have a shape and/or consistency that may cause them to become lodged in the trachea, such as nuts, grapes, and raw carrots).
7. *Meal frequency and energy density*
   Increase the number of times that the child is fed complementary foods as he/she gets older. The appropriate number of feedings depends on the energy density of the local foods and the usual amounts consumed at each feeding. For the average healthy breastfed infant, meals of complementary foods should be provided two or three times per day at 6–8 months of age and three or four times per day at 9–11 and 12–24 months of age. Additional nutritious snacks (such as a piece of fruit or bread or chapatti with nut paste) may be offered once or twice per day, as desired. Snacks are defined as foods eaten between meals – usually self-fed, convenient, and easy to prepare. If energy density or amount of food per meal is low, or the child is no longer breastfed, more frequent meals may be required.
8. *Nutrient content of complementary foods*
   Feed a variety of foods to ensure that nutrient needs are met. Meat, poultry, fish, or eggs should be eaten daily, or as often as possible. Vegetarian diets cannot meet nutrient needs at this age unless nutrient supplements or fortified products are used (see #9 below). Fruits and vegetables rich in vitamin A should be eaten daily. Provide diets with adequate fat content. Avoid giving drinks with low nutrient value, such as tea, coffee, and sugary drinks such as soda. Limit the amount of juice offered to avoid displacing more nutrient-rich foods.
9. *Use of vitamin–mineral supplements or fortified products for infant and mother*
   Use fortified complementary foods or vitamin–mineral supplements for the infant, as needed. In some populations, breastfeeding mothers may also need vitamin–mineral supplements or fortified products, both for their own health and to ensure normal concentrations of certain nutrients (particularly vitamins) in their breast milk. (Such products may also be beneficial for pre-pregnant and pregnant women.)
10. *Feeding during and after illness*
    Increase fluid intake during illness, including more frequent breastfeeding, and encourage the child to eat soft, varied, appetizing, favorite foods. After illness, give food more often than usual and encourage the child to eat more.

Reproduced from Pan American Health Organization/World Health Organization (2003) *Guiding Principles for Complementary Feeding of the Breastfed Child*. Washington, DC: Pan American Health Organization.

introduce complementary foods until 6 months, and the risk of micronutrient deficiencies is very low among full-term infants of normal birth weight whose mothers are well-nourished. Although iron deficiency may occur prior to the age of 6 months in infants whose iron reserves at birth are low (e.g., low-birth-weight infants and those whose mothers were iron deficient during pregnancy), the recommended approach is to provide medicinal iron drops from 2–3 months of age to such infants, rather than introducing complementary foods. Zinc status may also be marginal in low-birth-weight infants, and deficiency can be prevented by using zinc supplements. When the mother's diet is poor, the concentrations of certain vitamins (e.g., vitamin A and many of the B vitamins) and trace elements (e.g., iodine and

selenium) in breast milk may be lower than desirable. In such situations, improving the mother's diet or giving her micronutrient supplements are the preferred approaches, rather than providing complementary foods to the infant before the age of 6 months.

At 6 months of age, infants are developmentally ready to consume complementary foods and can benefit from the additional nutrients that they provide. Continued breastfeeding is recommended because breast milk remains an important source of energy, fat, protein, and several micronutrients. In addition, continued frequent breastfeeding beyond 6 months protects child health by reducing the risk of diarrhea and other infections in disadvantaged populations and by delaying maternal fertility (thereby increasing the interval before the next pregnancy among women who do not use other forms of contraception).

Nutrient Needs from Complementary Foods

The amounts of nutrients provided by breast milk can be estimated by multiplying average breast-milk intake by the concentration of each nutrient in human milk. By subtracting these values from the total recommended nutrient intakes (RNIs) one can derive estimates of the amounts of nutrients needed from complementary foods after 6 months of age. Using this approach, **Table 2** lists these estimates for three age ranges: 6–8 months, 9–11 months, and 12–23 months. In this table, the RNIs for energy and protein are taken from the update report on complementary feeding published in 2003, and the RNIs for micronutrients are taken from the most recent FAO/WHO estimates or the US dietary reference intakes. The estimated amount of each nutrient provided by breast milk is based on the average milk intake during each of the age intervals, calculated separately for infants in developing countries and in industrialized countries, using data from the studies compiled in the 1998 WHO/UNICEF document. Because of differences between developing and industrialized countries in average milk intake and in the assumed breast-milk concentration of vitamin A, the estimated amount of each nutrient provided by breast milk may vary. Within each column of **Table 2**, the first value listed refers to developing countries, and the second value refers to industrialized countries.

The first row of **Table 2** shows the total energy requirements and the estimated amounts of energy obtained from breast milk and required from complementary foods at each age. In developing countries, the average expected energy intake from complementary foods is approximately 200 kcal (837 kJ) at 6–8 months, 300 kcal (1256 kJ) at 9–11 months, and 550 kcal (2302 kJ) at 12–23 months. These values represent 33%, 45%, and 61% of total energy needs, respectively. In industrialized countries, the corresponding values are approximately 130 kcal (544 kJ) at 6–8 months, 310 kcal (1298 kJ) at 9–11 months, and 580 kcal (2428 kJ) at 12–23 months (21%, 45%, and 65% of total energy needs, respectively). Of course, these values will differ if the child is consuming more or less breast milk than the average.

The second row of **Table 2** shows the same estimates for protein. Assuming average breast-milk intake, the amount of protein needed from complementary foods increases from about $2\,g\,day^{-1}$ at 6–8 months to $5–6\,g\,day^{-1}$ at 12–23 months, with the percentage from complementary foods increasing from 21% to about 50%. The remaining rows show the estimates for the key vitamins and minerals. For vitamin B_{12} and selenium, the amounts needed from complementary foods prior to 12 months are zero because human milk contains generous amounts of these nutrients if the mother is adequately nourished. For the other micronutrients, the percentage of the RNI needed from complementary foods varies widely. At 6–8 months, for example, complementary foods need to provide less than 30% of the RNI for vitamin A, folate, vitamin C, copper, and iodine but more than 70% of the RNI for niacin, vitamin B_6, vitamin D, iron, and zinc. The values for the amount of niacin needed from complementary foods are high in all age intervals (75–88% of the RNI), but, because niacin needs can also be met by the contribution of tryptophan in the diet, niacin is not likely to be a limiting nutrient among infants who receive adequate protein. Similarly, the percentage of vitamin D needed from other sources is very high (more than 92%) because there is relatively little vitamin D in human milk; however, it should be noted that adequate exposure to sunlight can meet the child's needs for vitamin D even if complementary foods are not rich in this nutrient.

Complementary foods need to provide relatively large amounts (at least 80% of the RNI in all age intervals) of iron, zinc, and vitamin B_6. Because the amount of iron in human milk is very low (even though what is present is well absorbed), it is likely to be one of the first limiting nutrients in the diets of infants who rely predominantly on breast milk.

Fat is not listed in **Table 2** because there is uncertainty about the optimal intake of fat during the first 2 years of life. Dietary lipids are important not only as a source of essential fatty acids but also because they influence dietary energy density and sensory qualities. Breast milk is generally rich in fat (approximately 30–50% of energy) relative to most complementary foods, so as breast-milk intake

Table 2 Recommended nutrient intakes, average amount provided by breast milk, and amount needed from complementary foods at 6–8 months, 9–11 months, and 12–23 months

	6–8 months				*9–11 months*				*12–23 months*			
	RNI[a]	*Amount from breast milk*[b]	*Amount needed from CF*[c]	*% from CF*[c]	*RNI*[a]	*Amount from breast milk*[b]	*Amount needed from CF*[c]	*% from CF*[c]	*RNI*[a]	*Amount from breast milk*[b]	*Amount needed from CF*[c]	*% from CF*[c]
Energy (kcal d^{-1})	615	413; 486	202; 129	33; 21	686	379; 375	307; 311	45	894	346; 313	548; 581	61; 65
Protein (g d^{-1})	9.1	7.2	1.9	21	9.6	6.5; 5.6	3.1; 4.0	32; 42	10.9	5.8; 4.7	5.1; 6.2	47; 57
Vitamin A (μg RE d^{-1})	400	337; 461	63; 0	16; 0	400	308; 354	92; 46	23; 12	400	275; 300	125; 100	31; 25
Folate (μg d^{-1})	80	58	22	28	80	52; 45	28; 35	35; 44	160	47; 38	113; 122	71; 76
Niacin (mg d^{-1})	4	1	3.0	75	4	0.9; 0.8	3.1; 3.2	78; 80	6	0.8; 0.7	5.2; 5.3	87; 88
Riboflavin (mg d^{-1})	0.40	0.24	0.16	40	0.40	0.22; 0.19	0.18; 0.21	45; 53	0.50	0.19; 0.16	0.31; 0.34	62; 68
Thiamin (mg d^{-1})	0.30	0.14	0.16	53	0.30	0.12	0.18	60	0.50	0.11	0.39	78
Vitamin B_6 (mg d^{-1})	0.30	0.06	0.24	80	0.30	0.06	0.24	80	0.50	0.05	0.45	90
Vitamin B_{12} (μg d^{-1})	0.50	0.66	0	0	0.50	0.60; 0.51	0	0	0.90	0.53; 0.47	0.37; 0.43	41; 48
Vitamin C (mg d^{-1})	30	28	2	7	30	25; 21	5; 9	17; 30	30	22; 18	8; 12	27; 40
Vitamin D (μg d^{-1})	5	0.4	4.6	92	5	0.3	4.7	94	5	0.3; 0.2	4.7; 4.8	94; 96
Calcium (mg d^{-1})	270	191	79	29	270	172; 148	172; 148	36; 45	500	154; 125	346; 375	69; 75
Copper (mg d^{-1})	0.20	0.17	0.03	15	0.20	0.14	0.06	30	0.30	0.14; 0.11	0.16; 0.19	53; 63
Iodine (μg d^{-1})	90	75	15	17	90	68; 58	22; 32	24; 36	90	60; 49	30; 41	33; 46
Iron[d] (mg d^{-1})	9.3	0.2	9.1	98	9.3	0.2	9.1	98	5.8	0.2; 0.1	5.6; 5.7	97; 98
Magnesium (mg d^{-1})	54	24	30	56	54	22; 19	32	59	60	19; 16	41; 44	68; 73
Phosphorus (mg d^{-1})	275	95	180	65	275	86; 74	189; 201	69; 73	460	77; 63	383; 397	83; 86
Selenium (μg d^{-1})	10	14	0	0	10	12	0	0	17	11	6	35
Zinc (mg d^{-1})	3	0.6	2.4	80	3	0.5; 0.4	2.5; 2.6	83; 87	3	0.4; 0.3	2.6; 2.7	87; 90

[a]Recommended nutrient intakes, from FAO/WHO (2002) except for energy and protein (from Dewey and Brown, 2003) and calcium, copper, phosphorus, and zinc (from the US-Canada Dietary Reference Intakes).

[b]Based on average milk volumes of 674, 616, and 549 ml d^{-1} in developing countries and 688, 529, and 448 ml d^{-1} in industrialized countries for 6–8, 9–11, and 12–23 months, respectively (WHO 1998) and milk nutrient concentrations from the Institute of Medicine (IOM 1991, *Nutrition During Lactation*, Washington, DC: National Academy Press), except for vitamin A in milk of women from developing countries (WHO 1998) and zinc (Krebs NF *et al. Am J Clin Nutr* 1995; **61**: 1030–1036). For each nutrient, the first value refers to developing countries and the second value (after the semicolon) refers to industrialized countries, whenever there is a difference between the two.

[c]CF, complementary foods. For each nutrient, the first value refers to developing countries and the second value (after the semi-colon) refers to industrialized countries, whenever there is a difference between the two.

[d]Assuming medium bioavailability of iron.

declines with age, total fat intake is also likely to decline. If one assumes that the percentage of energy from fat in the total diet should be at least 30% and that the concentration of fat in breast milk averages $38\,g\,l^{-1}$, the amount of fat needed from complementary foods (assuming average breast-milk intake) is zero at 6–8 months, approximately $3\,g\,d^{-1}$ at 9–11 months, and $9–13\,g\,d^{-1}$ at 12–23 months, or 0%, 5–8%, and 15–20% of the energy from complementary foods, respectively. As infants decrease their intake of breast milk they also need other good sources of essential fatty acids, such as fish, egg, liver, nut pastes, and most vegetable oils.

Meal Frequency, Energy Density, and Consistency of Complementary Foods

The frequency with which complementary foods need to be fed (i.e., the number of meals per day) depends on the total amount of food required and the amount of food that a child can consume at a single meal (gastric capacity). The total amount of food required is a function of the amount of energy needed from complementary foods (which varies with age and breast-milk intake) and the energy density of the foods (i.e., kcal g^{-1}). The functional gastric capacity of infants and young children is assumed to be $30\,g\,kg^{-1}$ reference body weight. Thus, for a given age interval and level of breast-milk intake, calculating the recommended number of meals requires information about the energy density of the foods. To cover the needs of nearly all children, these calculations use as a starting point the average total energy requirement plus two standard deviations (25%). For children with average breast-milk intake who consume complementary foods with an energy density of at least $0.8\,kcal\,g^{-1}$, the number of meals required is two at 6–8 months and three thereafter. Children who consume less breast milk or who consume complementary foods with a lower energy density would need a greater number of meals. These calculations assume that children are fed to their gastric capacity at each meal, which may not be the case. For this reason, the guidelines recommend that additional nutritious snacks be offered once or twice per day, as desired (see guiding principle 7, **Table 1**).

A meal frequency greater than necessary may lead to excessive displacement of breast milk and may also require more time and effort by caregivers. Thus, it is useful to adapt meal-frequency guidelines to the characteristics of the target population.

The consistency of complementary foods needs to be appropriate for the child's stage of neuromuscular development. Semi-solid or puréed foods are needed at first because young infants do not have the ability to chew and swallow food of thick or solid consistency. By the age of 8 months most infants can also eat 'finger foods,' and by 12 months they can generally consume 'family foods' of a solid consistency. Thus, there should be a gradual change in the consistency of foods offered between 6 and 12 months, to match the infant's developmental progression (see guiding principle 6, **Table 1**).

Meeting Nutrient Needs During the Period of Complementary Feeding

As described above and shown in **Table 2**, breastfed infants need considerable amounts of certain nutrients from complementary foods after 6 months of age. It is a challenge to meet nutrient needs at this age because the amount of food consumed is relatively small, yet nutrient requirements during infancy (per unit of body weight) are very high because of the rapid rate of growth. Thus, nutrient-dense complementary foods are needed. The desired nutrient density (e.g., the amount of nutrient per 100 kcal of food) can be calculated by dividing the amount of each nutrient needed from complementary foods by the amount of energy expected to come from complementary foods (as shown in **Table 2**). When the desired nutrient densities are compared with the actual nutrient densities of the typical complementary foods consumed in various populations, protein density is generally seen to be adequate but several micronutrients are 'problem nutrients.' In most developing countries, iron, zinc, and vitamin B_6 are problem nutrients, and even in industrialized countries these nutrients may be limiting. Intake of iron is likely to be marginal in all populations unless iron-fortified products or substantial amounts of meat are consumed. Riboflavin, niacin, thiamin, folate, calcium, vitamin A, and vitamin C may also be problem nutrients, depending on the local mix of complementary foods. At present there is insufficient information to determine the extent to which some of the other micronutrients, such as vitamin E, iodine, and selenium, may be problem nutrients. Guiding principles 8 and 9 (**Table 1**) provide general recommendations to help ensure that the nutrient density of complementary foods will be adequate. It is difficult to develop more specific dietary 'prescriptions' to be used globally because of the great variability across populations in the types of complementary foods available. However, it is clear that predominantly vegetarian diets cannot meet the nutrient needs of breastfed infants unless nutrient supplements or fortified

products are used. Part of the reason for this is that plant-based diets are often high in phytate, which greatly reduces the bioavailability of iron and zinc. Therefore, it is recommended that meat, fish, poultry, or egg be offered daily, if possible. Iron-fortified infant cereals are a good source of iron, but meats can also provide adequate iron if consumed in large enough quantities, and they have the added advantage of being rich in zinc. When the amount of animal-source food available locally is limited, the amounts of iron and zinc absorbed from the diet can be enhanced by, first, reducing the phytate concentrations of the staple complementary food through germination, fermentation, and/or soaking, second, reducing the intake of polyphenols (e.g., from coffee and tea), which are known to inhibit iron absorption, and, third, increasing the intake of enhancers of iron and zinc absorption, such as vitamin C (for iron) and other organic acids (for iron and zinc). Adequate calcium can be obtained from cheese, yoghurt, and other dairy products, but feeding fresh unheated cow's milk is not recommended before 12 months because it is associated with fecal blood loss and lower iron status. Some vegetables can also provide modest amounts of calcium, but the bioavailability of calcium from foods with high amounts of oxalate (such as spinach) is very low. Fruits and vegetables rich in vitamin A are recommended daily because of the importance of preventing vitamin A deficiency, which has been linked with excess child mortality and other adverse outcomes. The bioavailability of pro-vitamin A carotenoids can be enhanced by finely chopping or puréeing the food and serving it with a source of fat to facilitate absorption. Beverages with low nutrient density (e.g., sugary drinks) should be avoided and the amount of juice should be limited because such beverages can displace more nutrient-dense foods and potentially contribute to child obesity.

In most populations, designing a diet that satisfies the requirements for all the 'problem nutrients' without the use of fortified foods is difficult, because, even when animal-source foods are available, infants typically eat very small quantities. One option that is currently being investigated is the addition of micronutrients, in the form of 'sprinkles' or fat-based products, to home-prepared foods.

Preparation and Feeding of Complementary Foods

Complementary feeding involves not only what to feed but also how to feed infants and young children. The first issue is the safe preparation and storage of complementary foods (see guiding principle 4, **Table 1**). Attention to hygienic practices during food preparation and feeding is essential for preventing gastrointestinal illness. In developing countries, the age range of 6–24 months is when diarrhoea is most prevalent, largely because of microbial contamination of complementary foods. Feeding bottles can be a major source of contamination because they are difficult to keep clean, and thus in resource-poor populations it is recommended that they be avoided. The other recommendations in guiding principle 4 stress the importance of washing hands and feeding utensils carefully, serving food immediately, and storing leftovers safely.

The second issue involves the interaction between caregiver and infant during feeding (see guiding principle 3, **Table 1**). Appropriate feeding behavior is termed responsive feeding and is more sensitive to the child's hunger and satiety cues than either a laissez-faire style of feeding (the caregiver rarely encourages the child to eat) or, at the opposite extreme, a controlling style of feeding (the caregiver determines when and how much the child will eat, even to the point of force feeding). Responsive feeding also involves feeding slowly and patiently, minimizing distractions during meals, and positive interactions with the child while feeding. Although there is little evidence to date regarding the impact of promoting responsive feeding, it is hypothesized that it will enhance dietary intake, child growth, and possibly behavioral development.

The third issue regarding behavioral aspects of complementary feeding involves feeding during and after illness (see guiding principle 10, **Table 1**). The need for fluids is often greater during illness, and for this reason it is recommended that breastfeeding frequency be increased and other fluids be offered as needed. During illness, children prefer breast milk over other foods, so frequent breastfeeding is critical for maintaining nutrient intake. Although appetite may be reduced, continued consumption of complementary foods is recommended to enhance recovery. After illness, the child needs more food than usual to make up for nutrient losses during illness and to allow for catch-up growth.

See also: **Breast Feeding**. **Calcium**. **Diarrheal Diseases**. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Fertility**. **Fish**. **Hunger**. **Infants**: Nutritional Requirements; Feeding Problems. **Iron**. **Obesity**: Prevention. **Protein**: Requirements and Role in Diet. **United Nations Children's Fund**. **Vitamin D**: Rickets and Osteomalacia. **World Health Organization**. **Zinc**: Physiology.

Further Reading

Caulfield LE, Huffman SL, and Piwoz EG (1999) Interventions to improve intake of complementary foods by infants 6 to 12 months of age in developing countries: impact on growth and on the prevalence of malnutrition and potential contribution to child survival. *Food and Nutrition Bulletin* **20**: 183–200.

Dewey KG (2001) Nutrition, growth and complementary feeding of the breastfed infant. *Pediatric Clinics of North America* **48**: 87–104.

Dewey KG (2002) Success of intervention programs to promote complementary feeding. In: Black R and Michaelsen KF (eds.) *Public Health Issues in Infant and Child Nutrition*, pp. 199–212. Nestle Nutrition Workshop Series, Pediatric Program, vol. 48, Nestec Ltd. Philadelphia: Vevey/Lippincott Williams & Wilkins.

Dewey KG and Brown KH (2003) Update on technical issues concerning complementary feeding of young children in developing countries and implications for intervention programs. *Food and Nutrition Bulletin* **24**: 5–28.

FAO/WHO Joint Expert Consultation (2002) *Vitamin and Mineral Requirements in Human Nutrition*. Geneva: World Health Organization.

Gibson RS, Ferguson EL, and Lehrfeld J (1998) Complementary foods for infant feeding in developing countries: their nutrient adequacy and improvement. *European Journal of Clinical Nutrition* **52**: 764–770.

Lutter CL and Dewey KG (2003) Proposed nutrient composition for fortified complementary foods. *Journal of Nutrition* **133**: 3011S–3020S.

Pan American Health Organization/World Health Organization (2003) *Guiding Principles for Complementary Feeding of the Breastfed Child*. Washington, DC: Pan American Health Organization.

World Health Organization (2001) *The Optimal Duration of Exclusive Breastfeeding: A Systematic Review*. WHO/NHD/01.08; WHO/FCH/CAH/01.23. Geneva: World Health Organization.

World Health Organization/London School of Hygiene and Tropical Medicine (2000) *Complementary Feeding: Family Foods for Breastfed Children*. WHO/NHD/00.1; WHO/FCH/CAH/00.6. Geneva: World Health Organization.

World Health Organization/UNICEF (1998) *Complementary Feeding of Young Children in Developing Countries: A Review of Current Scientific Knowledge*. WHO/NUT/98.1. Geneva: World Health Organization.

COPPER

X Xu, S Pin, J Shedlock and Z L Harris, Johns Hopkins Hospital and School of Medicine, Baltimore, MD, USA

Introduction

Transition metals occupy a special niche in aerobic physiology: as facile electron donors and acceptors, they are essential participants in oxidation/reduction reactions throughout the cell. These unique properties of transition metals are largely dependent on the electronic configuration of the electrons in the outer shell and in the penultimate outer shell. These metals can exist in different oxidation states, which is critical for their usefulness as catalysts. However, it is during these same committed reactions essential for aerobic metabolism that toxic reactive oxygen species can be generated. As such the transition metals are chaperoned as they traffic through the body and are regulated tightly. Subtle disruptions of metal homeostasis culminate in disease and death. Iron, copper, and zinc are the most abundant and well-studied transition metals. Copper is the oldest metal in use: copper artifacts dating back to 8700BC have been found. The physiology, requirements, and dietary sources of copper are described here with an emphasis on the role of copper in human health and disease.

Copper, as a trace metal, can be found in all living cells in either the oxidized Cu(II) or reduced Cu(I) state. Copper is an essential cofactor for many enzymes critical for cellular oxidation. These include: cytochrome *c*-oxidase, which is essential for mitochondrial respiration as the terminal enzyme in the electron transport chain; superoxide dismutase, a potent antioxidant defense mechanism; tyrosinase, which is critical for melanin production; dopamine B-hydroxylase, a prerequisite for catecholamine production; lysyl oxidase, which is responsible for collagen and elastin cross-linking; ceruloplasmin, a ferroxidase/metallo-oxidase; hephaestin, a ferroxidase/metallo-oxidase; and peptidylglycine α-amidating monooxygenase, a peptide processor (**Table 1**). Mice that lack the copper transport protein Ctr1 are embryonic lethal, which confirms the importance of copper in enzyme function and normal cellular homeostasis.

Copper Homeostasis

Dietary intake of copper is approximately $5\,\mathrm{mg\,day^{-1}}$ with an equivalent amount being excreted by bile in stool. Approximately $2\,\mathrm{mg\,day^{-1}}$ are directly absorbed across the gastrointestinal tract daily and incorporated into blood, serum, liver,

Table 1 Mammalian copper enzymes

Enzyme	*Function*
Cytochrome *c*-oxidase	Mitochondrial respiration
cu,zn-Superoxide dismutase	Antioxidant defense
Tyrosinase	Melanin production
Dopamine B-hydroxylase	Catecholamine production
Lysyl oxidase	Collagen and elastin cross-linking
Ceruloplasmin	Ferroxidase/metallo-oxidase
Hephaestin	Ferroxidase/metallo-oxidase
PAM	Peptide processing

brain, muscle, and kidney. An equal amount is excreted and maintains the sensitive copper balance (**Figure 1**). The main sources of copper are seeds, grains, nuts, beans, shellfish, and liver (**Table 2**). Drinking water no longer contributes significantly. When copper pipes were commonly used for plumbing, copper toxicity was a more recognized phenomenon.

It is difficult to define specific dietary copper requirements because of the lack of suitable indices to assess copper status. As such, knowledge of factors affecting the bioavailability of dietary copper is limited. Ceruloplasmin contains 95% of the copper found in serum and is frequently used as a marker of copper status. However, ceruloplasmin levels vary with pregnancy and inflammation and ceruloplasmin mRNA is regulated by estrogen, infection, and hypoxia among other factors. Currently, investigators are searching for genetic biomarkers in intestinal, liver, and lymphocyte cells that respond to copper levels and may serve as better markers of copper status. Whole-body copper metabolism is difficult to study in human subjects. However, isotopic tracers and kinetic modeling have added a dimension to what can be learned in humans by direct measurement. These studies suggest that the efficiency of copper absorption varies greatly, depending on dietary intake. Mechanisms regulating total body copper seem to be strong, given the relatively small and constant body pool, but they are not yet well understood. Changes in efficiency of absorption help to regulate the amount of copper retained by the body. In addition, endogenous excretion of copper into the gastrointestinal tract depends heavily on the amount of copper absorbed. When dietary copper is high and an excess is absorbed, endogenous excretion increases, protecting against toxic accumulation of copper in the body. When intake is low, little endogenous copper is excreted, protecting against copper depletion. Regulation is not sufficient with very low amounts of dietary copper ($0.38\,mg\,day^{-1}$) and appears to be delayed when copper intake is high.

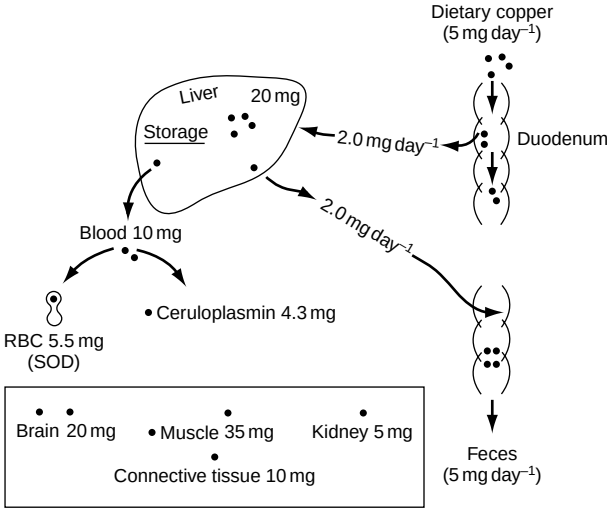


Figure 1 Mammalian copper metabolism: daily copper cycle including oral absorption, tissue distribution, and excretion. Values are for adult men ($mg\,day^{-1}$). An equal amount of copper is absorbed and excreted to maintain copper balance.

Table 2 Copper content of various foods

Food	*Copper concentration (μg wet wt)*	*Size of typical serving (g)*	*Copper/ serving ($mg\,g^{-1}$)*
Fish	0.61	120	0.070
Turkey	0.71	120	0.090
Chicken	0.34	120	0.040
Hamburger	0.95	120	0.110
Roast beef	0.82	120	0.100
Steak	1.2	120	0.140
Sheep liver	157.05	120	18.850
Pork liver	141.14	120	16.940
Egg	0.8	40	0.030
Single sliced cheese	0.43	120	0.050
Whole wheat	1.07	30	0.030
Scallops	6.08	120	0.030
Clams	7.39	120	0.730
Crab	1.75	120	0.890
Shrimp	1.75	120	0.210
Oysters	2.89	120	0.350
Smoked oysters	15	120	1.800
Mussels	4.75	120	0.570
Lobster	36.6	120	4.390
Candy bar	1.18	15	0.020
Milk	0.33	120	0.040
Peas	2.38	120	0.290
Soy beans	109	120	0.130
Applesauce (can)	0.2	120	0.020
Avocado	1.68	120	0.200
Raisins	1.68	30	0.050
Peanut butter	8.53	30	0.260

Recommended Intakes

The Tolerable Upper Intake Limit (UL) for adults is 10 mg daily, based on degree of liver damage associated with intake. UL for children vary with age: 1–3 years/1 mg daily, 4–8 years/3 mg daily, 9–13 years/5 mg daily, 14–18 years/8 mg daily (irrespective of pregnancy or lactation status). UL for children under the age of 1 year are not possible to establish. There are no official recommended daily allowances (RDAs) for copper in children. The RDA for adult males and females is a daily intake of 0.9 mg. Measurements of the dietary requirements for copper in adult men have shown the requirement to range from about 1.0 to 1.6 mg daily. A review of nutrient intakes in the US from 1909 to 1994 confirms that intake varied between 1.5 mg day^{-1} (1965) to 2.1 mg day^{-1} (1909). These trends reflect a diet higher in copper-rich potatoes and grain predominating in 1909 versus a decline in potato popularity in 1965. Daily intake recommendations for children vary with age (see **Table 3**). Persons who consume diets high in zinc and low in protein are at risk of copper deficiency. High intakes of dietary fiber apparently increase the dietary requirement for copper. Diets in Western countries provide copper below or in the low range of the estimated safe and adequate daily dietary intake. Copper deficiency is usually a consequence of low copper stores at birth, inadequate dietary copper intake, poor absorption, elevated requirements induced by rapid growth, or increased copper losses.

Bioavailability

The issue of bioavailability from food sources and the interactions between food groups and copper availability remains a critical question. Lonnerdal *et al.* demonstrated that heat treatment of cows' milk formula decreases the copper bioavailability. Transitional complexes form in the milk upon heating that have a similar configuration to copper and thereby directly inhibit copper absorption. High doses of zinc also reduce copper bioavailability, as does combined iron and zinc supplementation. The dilemma is how to prepare an infant formula containing adequate copper, iron, and zinc that will meet the RDA for copper. Other nutrients dramatically affect copper absorption from foods. Soy protein-based diets promote less copper retention in tissues than lactalbumin-based diets. However, it is unclear if this effect is solely due to the soy protein composition or to the higher zinc in these soy-based formulas. In animals, phytate causes a drop in serum copper but human stable isotope studies reveal no effect on copper absorption in adult men. Patients with low copper indices need to be evaluated for the copper content of their diets, other foods ingested at the same time, and other mineral supplements that may be given.

Table 3 Recommended dietary allowances for copper (mg day^{-1})

Age	*RDA (daily)*
Infants	
<6 months	0.2 (30 mcg/kg)
6–12 months	0.2–0.3 (24 mcg/kg)
Children	
1–3 years	0.34
4–8 years	0.44
9–13 years	0.7
14–18 years	0.89
Adult	
19+ years	0.9
Pregnant women	1
Nursing women	1.3

Absorption and Excretion

Dietary copper is absorbed across the small intestine. It diffuses through the mucous layer that covers the wall of the bowel via the divalent metal transporter DMT1. Copper is thus released into the serum and presumably is transported bound to either albumin or histidine to the multiple sites that require copper or to storage tissues. The liver is the primary storage organ for copper followed by muscle and bone. Not all of the copper ingested is absorbed and gastrointestinal cells that hold on to the excess copper are 'sloughed' when the lining of the gut is turned over every 24–48 hs. Copper bound to albumin or histidine enters the hepatocyte via the high-affinity mammalian copper transporter, hCtr1. Initially identified in yeast by functional complementation studies, this protein has subsequently been cloned in mice and humans. Human Ctr1 has a high homology to the yeast proteins Ctr1 and Ctr3 involved in high-affinity copper uptake. The N-terminus of the protein is rich in histidine and methionine residues, which presumably bind the copper and move it into the cell. Characterization of hCtr1 confirms its localization on the plasma membrane consistent with its role as a copper transporter. *In vitro* work has also identified a vesicular perinuclear distribution for hCtr1 that is copper concentration dependent. Redistribution of the hCtr1 suggests that under different copper states, copper moves through the membrane transporter and into a vesicular compartment for further 'assignment' within the cell.

hCtr2, a low-affinity copper uptake transporter, has also been identified. This low-affinity copper transporter is unable to complement the respiratory defect seen in yeast strains lacking copper transport capabilities. Once inside the cell, copper has one of four fates: (1) bind to and be stored within a glutathione/metallothionein pool; (2) bind to CCS, the copper chaperone for Cu, Zn - SOD; (3) bind to cox 17 for delivery to mitochondrial cytochrome *c*-oxidase; or (4) bind to HAH1 (human Atox1 homolog) for subsequent copper delivery to either the Wilson disease P-type ATPase or the Menkes' P-type ATPase. Copper from HAH1 is incorporated into ceruloplasmin, the most abundant serum cuproprotein, within the trans golgi network (TGN). How the protein unfolds within the TGN to accept copper and how the copper is incorporated into ceruloplasmin is still under study. Ceruloplasmin is then secreted into the serum, and any excess copper not incorporated into ceruloplasmin is recycled in vesicles containing either the Wilson disease P-type ATPase or Menkes' P-type ATPase, and excreted into bile or stored in the liver. Recent characterization of a new protein, Murr1, suggests that this protein regulates copper excretion into bile such that mutations in the Murr1 gene are associated with normal copper uptake but severe defects in exporting copper from hepatocytes.

Approximately 15% of the total copper absorbed is actually transported to tissues while the remaining 85% is excreted. Of that copper pool, 98% is excreted in bile with the remaining 2% eliminated in the urine. The liver is the predominant organ responsible for regulating copper homeostasis at the level of excretion. Whereas copper import is highly conserved between yeast and humans, copper export in vertebrates involves a complex vesicular system that culminates in a lysosomal excretion pathway 'dumping' copper into the bile for elimination. At steady state, the amount of copper excreted into the biliary system is directly proportional to the hepatic copper load. In response to an increasing copper concentration within the hepatocyte, biliary copper excretion increases. There is no enterohepatic recirculation of copper and once the unabsorbable copper complex is in bile it is excreted in stool. Localization studies reveal redistribution of the ATP7b from the TGN to a vesicular compartment that migrates out to the biliary epithelium in response to increasing copper concentrations. Alternatively, under conditions of copper deficiency, the ATP7b remains tightly incorporated with the TGN for maximal copper incorporation into ceruloplasmin.

The highly homologous Wilson disease P-type ATPase (ATP7b) and the Menkes's P-type ATPase (ATP7a) differ only in their tissue expression and both function to move copper from one intracellular compartment to another. The ATP7a is predominantly located in the placenta, blood–brain barrier, and gastrointestinal tract and hence any mutation in the Menkes's P-type ATPase results in a copper deficiency in the fetus, brain, and tissues. In contrast, the Wilson's disease P-type ATPase is expressed in the liver and mutations in this culminate in profound copper overload of the liver because of the inability to shuttle copper into the trans golgi network for incorporation into ceruloplasmin. The excess copper is stored in the liver and eventually leaks out in the serum where it is deposited within sensitive tissues: the eye and brain. The psychiatric illnesses ascribed to Wilson's disease are a result of hepatocyte-derived copper 'leaking' out of the liver and accumulating within the basal ganglia. Similarly, Kayser-Fleischer rings arise from copper deposition in the cornea. The toxic copper in the liver eventually results in cirrhosis and hepatic fibrosis as a result of oxyradical damage. Menkes's syndrome has an incidence of 1:300 000 while Wilson's disease has an incidence of 1:30 000. Expression of these diseases may differ considerably among affected family members.

The recognition of a novel disorder of iron metabolism associated with mutations in the copper-containing protein ceruloplasmin revealed an essential role for ceruloplasmin as a ferroxidase and regulator of iron homeostasis. Patients and mice lacking the serum protein ceruloplasmin have normal copper kinetics: normal absorption, distribution, and copper-dependent activity. These data suggest that although under experimental conditions ceruloplasmin may donate copper, ceruloplasmin is not a copper transport protein. The six atoms of copper are incorporated into three type 1 coppers, one type 2 copper, and a type 3 copper. The type 1 coppers provide the electron shuttle necessary for the concomitant reduction of oxygen to water that occurs within the trinuclear copper cluster comprised of the type 2 and type 3 copper. This reaction is coupled with the oxidation of a variety of substrates: amines, peroxidases, iron, NO, and possibly copper. The recent observation that Fet3, the yeast ceruloplasmin homolog, also has critical cuprous oxidase activity in addition to ferroxidase activity has prompted renaming some of the multicopper oxidases (ceruloplasmin, Fet3, hephaestin) as 'metallo-oxidases' rather than ferroxidases.

Copper Deficiency

Reports of human copper deficiency are limited and suggest that severe nutrient deficiency coupled with

malabsorption is required for this disease state to occur. Infants fed an exclusive cows' milk diet are at risk for copper deficiency. Cows' milk not only has substantially less copper than human milk but the bioavailability is also reduced. High oral intake of iron or zinc decrease copper absorption and may predispose an individual to copper deficiency. Other infants at risk include those with: (1) prematurity secondary to a lack of hepatic copper stores; (2) prolonged diarrhea; and (3) intestinal malabsorption syndromes. Even the premature liver is capable of impressive copper storage. By 26 weeks' gestational age the liver already has 3 mg of copper stored. By 40 weeks' gestational age, the hepatic liver has 10–12 mg copper stored with the majority being deposited in the third trimester. Iron and zinc have been shown to interfere with copper absorption and further complicate the picture of copper deficiency. The most frequent clinical manifestations of copper deficiency are anemia refractory to iron treatment, neutropenia, and bone demineralization presenting as fractures.

The anemia is characterized as hypochromic and normocytic with a reduced reticulocyte count, hypoferremia, and thrombocytopenia. Bone marrow aspirate reveals megaloblastic changes and vacuolization of both erythroid and myeloid progenitor lineages. It is believed that a profound copper deficiency results in a multicopper oxidase deficient state and as such bone marrow demands are unmet by the lack of ferroxidase activity. Bone abnormalities are common and manifest as osteoporosis, fractures, and epiphyseal separation. Other manifestations of copper deficiency include hypopigmentation, hypotonia, growth arrest, abnormal cholesterol and glucose metabolism, and increased rate of infections.

Multiple factors associated with copper deficiency are responsible for the increased rate of infection seen. Most copper-deficient patients are malnourished and suffer from impaired weight gain. The immune system requires copper to perform several functions. Recent research showed that interleukin 2 is reduced in copper deficiency and is probably the mechanism by which T-cell proliferation is reduced. These results were extended to show that even in marginal deficiency, when common indexes of copper are not affected by the diet, the proliferative response and interleukin concentrations are reduced. The number of neutrophils in human peripheral blood is reduced in cases of severe copper deficiency. Not only are they reduced in number, but their ability to generate superoxide anion and kill ingested microorganisms is also reduced in both overt and marginal copper deficiency. This mechanism is not yet understood.

Copper Excess

Excess copper is the result of either excessive copper absorption or ineffective copper excretion. The most common diseases associated with copper excess are: (1) Wilson's disease, a genetic disease resulting in mutations in the Wilson's disease P-type ATPase and excessive hepatocyte copper accumulation; (2) renal disease, in patients on hemodialysis due to kidney failure when dialysate solutions become contaminated with excess copper; and (3) biliary obstruction. Excessive use of copper supplements may also contribute to copper toxicity and is clinically manifested by severe anemia, nausea and vomiting, abdominal pain, and diarrhea. Copper toxicosis can rapidly progress to coma and death if not recognized. Current management of most diseases associated with copper toxicity includes a low-copper diet, a high-zinc diet (competitively interferes with copper absorption), and use of copper chelators such as penicillamine and trientine. Affected individuals should have their tap water analyzed for copper content and drink demineralized water if their water contains more than 100 μg/liter. Given that the liver is the most significant copper storage organ, any activity that can affect hepatic cellular metabolism needs to be monitored. Hence, alcohol consumption is strongly discouraged.

There are reports of chronic copper exposure resulting in toxic accumulation. Fortunately, these events appear to be geographically restricted. Indian childhood cirrhosis (ICC), also known as Indian infantile cirrhosis or idiopathic copper toxicosis, has been associated with increased copper intake from contaminated pots used to heat up infant milk. The milk is stored and warmed in brass (a copper alloy) or copper containers. It is interesting to note that the increased copper absorption alone is not critical for disease formation but rather this occurs in infants that already have prenatal liver copper stores in excess of adult values. How the neonatal liver is able to compartmentalize this toxic metal so effectively is unknown. Perhaps in ICC, this delicate balance is disrupted. Tyrolean liver disease, occurring in the Austrian Tyrol, despite having a Mendelian pattern of inheritance suggestive of an autosomal recessive trait, appears related to use of copper cooking utensils. However, recent reports describe how a persistent percentage of the German population remains susceptible to copper toxicosis despite adjustments in cooking utensils. Perhaps a genetic susceptibility exists in this population that has yet to be determined.

Conclusion

Adult copper homeostasis rests on the foundation of an adequate copper balance in early life. Copper deficiency, either due to inadequate intake or abnormal absorption, may result. While the clinical stigmata of severe copper deficiency are easy to identify, the subtle changes in neurobehavioral development associated with mild copper deficiency are unknown. Given the high copper concentration in the brain, one could postulate that critical copper deficiency during development could lead to significant central nervous system deficits. Recent evidence suggesting that copper metabolism may be involved as an epigenetic factor in the development of Alzheimer's disease (AD) highlights the importance of balance. In this scenario, elevated central nervous system copper, as seen in AD, may initiate increased oxyradical formation and hasten damage. In fact, some are advocating that serum copper might be a good biomarker for AD. Copper is an essential trace metal critical for normal development. The goal of future studies will be to develop sensitive biomarkers for copper status. Only with these tools can we adequately assess copper status and treat copper-deficient and copper excess states appropriately.

Further Reading

Araya M, Koletzko B, and Uauy R (2003) Copper deficiency and excess in infancy: developing a research agenda. *Journal of Pediatric Gastroenterology and Nutrition* **37**: 422–429.

Bush Al and Strozyk D (2004) Serum copper: A biomarker for Alzheimer disease. *Archives of Neurology* **61**: 631–632.

Gitlin JD (2003) Wilson disease. *Gastroenterology* **125**: 1868–1877.

Klein CJ (2002) Nutrient requirements for preterm infant formulas. *Journal of Nutrition* **132**: 1395S–1577S.

Lutter CK and Dewey KG (2003) Proposed nutrient composition for fortified complimentary foods. *Journal of Nutrition* **133**: 3011S–3020S.

Prohaska JR and Gybina AA (2004) Intracellular copper transport in mammals. *Journal of Nutrition* **134**: 1003–1006.

Rees EM and Thiele DJ (2004) From aging to virulence: forging connections through the study of copper homeostasis in eukaryotic microorganisms. *Current Opinion in Microbiology* **7**: 175–184.

Schulpis KH, Karakonstantakis T, Gavrili S *et al.* (2004) Maternal-neonatal serum selenium and copper levels in Greeks and Albanians. *European Journal of Clinical Nutrition* **1**: 1–5.

Shim H and Harris ZL (2003) Genetic defects in copper metabolism. *Journal of Nutrition* **133**: 1527S–1531S.

Tapiero H, Townsend DM, and Tew KD (2003) Trace elements in human physiology. Copper. *Biomedicine & Pharmacotherapy* **57**: 386–398.

Uauy R, Olivares M, and Gonzalez M (1998) Essentiality of copper in humans. *American Journal of Clinical Nutrition* **67**(S): 952S–959S.

Wijmenga C and Klomp LWJ (2004) Molecular regulation of copper excretion in the liver. *Proceedings of the Nutrition Society* **63**: 31–39.

See also: **Bioavailability**. **Zinc**: Physiology.

CORONARY HEART DISEASE

Contents

Hemostatic Factors

W Gilmore, University of Ulster, Coleraine, UK

Introduction

The two major processes that contribute to the pathogenesis of coronary heart disease (CHD) and stroke are atherosclerosis and thrombosis. These, in turn, involve inflammation and hemostasis; two pathways that are linked both at the molecular and cellular levels. In addition to stimulating the inflammatory response, the proinflammatory cytokines, chiefly interleukin-1beta (IL-1β) and tumor necrosis factor alpha (TNFα), may promote the initiation of hemostasis by upregulating tissue factor (TF) on endothelial cells and the blood monocytes. The explosion of interest in the role of cholesterol in vascular diseases in the latter half of the twentieth century led to a decrease in interest in the role of hemostatic factors in this group of diseases. However, the efficacy of thrombolytic drugs in the treatment of acute myocardial

infarction together with the demonstration that clots were involved in sudden ischemic death renewed interest in the role of the blood clotting pathways in these disorders. Hemostasis is part of the body's normal defense system and response to injury. The advancement, by Russell Ross, of the response to injury hypothesis of vascular disease has directed attention towards hemostasis again. Hemostasis involves the interaction between cells of the immune system, blood platelets, smooth muscle cells, endothelial cells, and the blood clotting proteins. The blood clot is eventually broken down by the fibrinolytic mechanism and the normal tissue repair processes promote wound healing. A balance between the processes of hemostasis and fibrinolysis has always been thought necessary to prevent blood loss, on the one hand, and thrombosis, on the other. Many of the components of the hemostatic mechanism have been identified as risk factors for vascular disease and these include: increased plasma levels of the blood clotting proteins fibrinogen and factor VII, increased platelet aggregation, and elevated plasma levels of the inhibitor of plasmin activation, plasminogen activator inhibitor-1 (PAI-1).

Nutrition plays a central role in the activity of the blood coagulation factors with the requirement of vitamin K in the posttranslational carboxylation of glutamic acid residues on key blood clotting proteins and some of their physiological inhibitors. In addition to this involvement of vitamin K in the biosynthesis of biologically active blood clotting proteins other nutrients may modify components of the blood coagulation pathway. However, results from many intervention studies still often yield conflicting results. This article will concentrate on a description of hemostasis, as we currently understand it, rather than an exhaustive account of the studies on dietary factors and hemostasis. This may enable the reader to discern the most appropriate aspects of hemostasis for study by nutrition scientists.

Hemostasis

Platelets

When trauma occurs platelets can form a small primary hemostatic plug that is sufficient to stop bleeding from a small nick in the skin. Platelets are very easily activated and, therefore, difficult to study. They are formed from megakaryocytes in the bone marrow and, to a lesser extent, in the peripheral blood and lungs. Platelets are small (0.2–3.5 μm in diameter) buds off these large megakaryocytes. They have a volume of about 10 fl and number $100-400 \times 10^9\,l^{-1}$ in peripheral blood. When viewed under light microscopy, they have little structure; however, electron microscopy reveals abundant subcellular organelles. Platelets contain large quantities of lipids; their plasma membrane is highly involuted and their cytoplasm contains numerous membrane-bound granules. These granules are of two types: electron-dense granules that contain, *inter alia*, ADP, Ca^{2+} and serotonin, and the so-called specific alpha-granules that contain some blood coagulation proteins and growth factors such as platelet-derived growth factor (PDGF) and transforming growth factor-beta (TGF-β), which promote wound healing.

When activated, platelets express the adhesion molecule P-selectin (CD62P), which, in unactivated platelets, is present in the alpha-granules. Activated platelets stick to damaged endothelium, a process known as platelet adhesion, and then stick to each other, a process known as platelet aggregation. Adhesion molecules form important molecular components of this process. Platelet aggregation can be measured in the laboratory by simple photometric techniques. This *in vitro* aggregation is stimulated by adrenaline, collagen, ADP, and the toxic antibiotic risocetin and has been widely applied in studies of the effect of nutrients on platelet function. However, adhesion and aggregation may only be minor biological functions of platelets. The main function of these tiny fragments seems to be in their ability to release a vast array of biologically active substances. In particular, they release phospholipid from their membranes to participate in the coagulation cascade. Platelets release serotonin, a powerful vasoconstrictant. They also release PDGF and TGF-β both of which are growth factors intimately involved in wound healing and tissue regeneration. Low platelet numbers will lead to an increased bleeding tendency and, interestingly, to a decrease in the integrity of the blood vessel walls. Therefore, platelets have two main functions in the body: they participate in hemostasis, and are responsible for the repair and maintenance of the blood vessel walls.

The products of cyclooxygenases play a crucial role in the action of blood platelets. The substrates of these enzymes are the *n*-6 and *n*-3 long-chain polyunsaturated fatty acids. Platelets cannot synthesize these from the short-chain precursors and so must rely on direct incorporation of the long-chain polyunsaturated fatty acids into their cellular membranes. In platelets, cyclooxygenases and other enzymes convert arachidonic acid (C20:4*n*-6) to thromboxane A_2, which potentiates platelet aggregation, whereas the less potent thromboxane A_3 is

generated when eicosapentaenoic acid (C20:5*n*-3) is the substrate. The endothelial cells constitutively synthesize the arachidonic acid (C20:4*n*-6)-derived eicosinoid, prostaglandin I_2 (PGI_2) which is antiaggregatory in its action and thus the normal endothelium provides a surface that prevents blood clotting.

The Coagulation Cascade

Blood coagulation factors are proteins, most of which were discovered when a genetically inherited bleeding disorder was discovered in patients. A newly discovered coagulation factor was often named after the patient in whom it was first recognized, but nowadays the Roman numerals I through to XIII are designated and accepted names of these proteins. A suffix, 'a' indicates an activated coagulation factor. The components of the blood coagulation pathways are mostly serine proteases that sequentially activate each other in turn, giving rise to a fibrin clot. Some, however, act as cofactors in some of the reactions, e.g., factors Va and VIIIa. There are two different mechanisms whereby blood clotting may be initiated (**Figure 1**). The principal mechanism, known as the tissue factor (TF) pathway or the extrinsic coagulation cascade, involves the expression of TF (thromboplastin or factor III) on the cells of the intima of the damaged blood vessel. TF is only expressed on the cells that are not in contact with the peripheral blood. The exposure of blood to TF, in turn, leads to the sequential activation of the blood clotting factors, and the generation of the fibrin clot. The only cells in day-to-day contact with blood that can be stimulated to produce TF are endothelial cells and monocytes when the proinflammatory cytokines and bacterial endotoxins stimulate these cells to express this cell membrane-bound protein. The presence of a NFκB binding site in the promoter region of the tissue factor gene on chromosome 1 further indicates involvement of hemostasis in inflammatory processes.

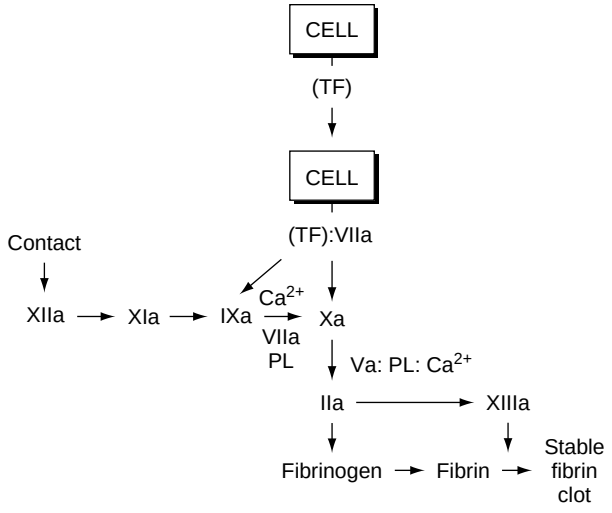


Figure 1 The tissue factor pathway and contact activation pathways of blood coagulation. PL, platelet phospholipids; TF, tissue factor.

TF is an integral membrane protein of 263 amino acids and molecular weight 47 000. It has been assigned CD142 by the VIth International Workshop and Conference on Human Leukocyte Differentiation Antigens held in Kobe, Japan, in 1996. The ligand for this cell surface protein is factor VII or factor VIIa. There are relatively small amounts of factor VIIa constantly present in normal blood. However, only when factor VIIa binds to TF does it become significantly potent as an activator of blood coagulation. A TF:Ca^{2+}:factor VIIa complex is formed and this activates both factor X to factor Xa and factor IX to factor IXa, which forms a complex with factor VIIa, Ca^{2+}, and platelet phospholipids to activate further molecules of factor X. The activated factor X forms a factor Xa:Ca^{2+}:factor Va complex, which converts prothrombin to thrombin. Thrombin acts on fibrinogen molecules to convert them to fibrin monomers. These monomers form an instantaneous clot by associating via noncovalent bonds. The clot is then stabilized in a reaction catalyzed by factor XIIIa by crosslinking the fibrin molecule by covalent bonds between glutamic acid and lysine residues of adjacent fibrin monomers. Thrombin plays a central role in hemostasis in that it not only converts fibrinogen to fibrin but also activates other key players in the pathways; in particular, factor VII, factor XI, and the copper-dependent factors V and VIII. The major physiological activator of factor VII remains unidentified but, in addition to thrombin; factor Xa, factor XIa, and factor XIIa are capable of converting factor VII to factor VIIa.

TF is inhibited by a specific inhibitor of the TF:Ca^{2+}:factor VIIa complex (**Figure 3**). This inhibitor, designated tissue factor pathway inhibitor (TFPI), is a 276 amino acid polypeptide that has three Kunitz-like regions. Therefore, this belongs to the Kunitz-type protease inhibitors whilst most other inhibitors of blood coagulation are serpins. In peripheral blood this inhibitor is associated with the lipid fractions and is, like tissue factor, synthesized by activated endothelial cells and monocytes. The TFPI remains inactive until sufficient amounts of activated factor X (i.e., factor Xa) can bind. The TFPI:factor Xa complex then inactivates the TF:Ca^{2+}:factor VIIa complex.

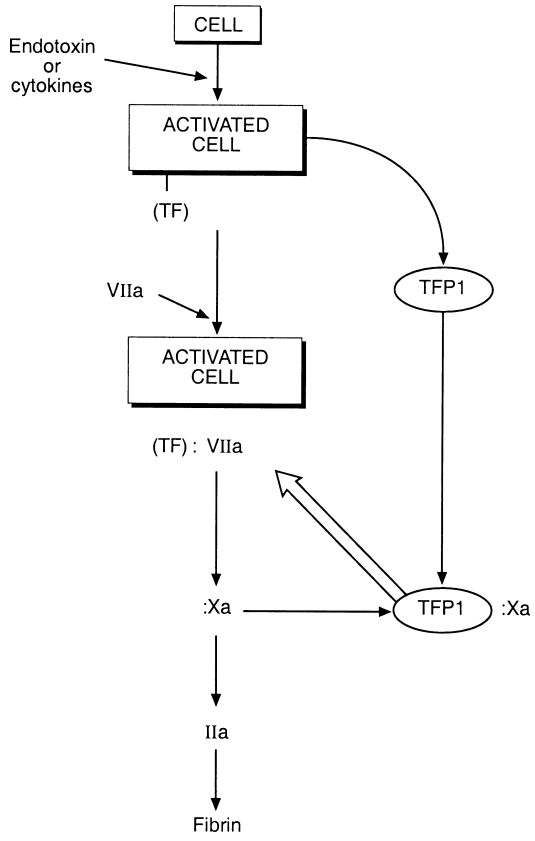


Figure 2 The action of tissue factor pathway inhibitor (TFPI). TF, tissue factor.

In the second mechanism of initiation, the contact activation pathway (**Figure 1** and **Figure 2**), blood coagulation may be triggered by activation upon contact with collagen and other highly charged surfaces, including glass or plastic test tubes. Contact with collagen leads to activation of factor XII and, in turn, factor XIIa activates factor XI. Factor XIa activates factor IX, which together with Ca^{2+}, platelet phospholipid, and factor VIIIa activates factor X to factor Xa. The contact activation pathway merges with the tissue factor pathway at this point and, as before, the factor Xa, together with Ca^{2+}, platelet phospholipid, and factor Va activates prothrombin to thrombin, which converts fibrinogen to fibrin. The existence of this second pathway means that fibrin production remains switched on as long as the blood remains in contact with external surfaces despite the fact that the tissue factor pathway may, at this point, be closed down by the action of TFPI and factor Xa. The formation of factor XIIa has important consequences for other processes involved in the response of the body to injury. For example; high-molecular-weight kininogen and prekallikrein are activated to the kinins and to kallikrein, mediators of inflammation that can activate the contact pathway (**Figure 3**). In addition, fibrinolysis and the complement pathway are both activated by factor XIIa. It is probable that the physiological role of factor XIIa is to act as a mediator in many of the processes involved in the defense of the body during trauma and its role in the activation of blood coagulation may be of minor importance.

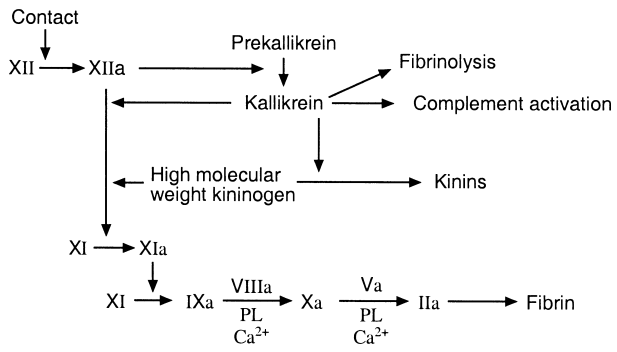


Figure 3 Contact activation pathway. PL, platelet phospholipids.

The coagulation process is inhibited by nonspecific mechanisms such as blood flow and by the presence in plasma of general serine protease inhibitors such as alpha2-macroglobulin. In addition, unique molecular systems will specifically inhibit blood coagulation and one of these involves heparin, a sulfated glycosaminoglycan. Heparin combines with antithrombin III in a one to one molar ratio and the resultant complex inhibits factor Xa and thrombin. Another specific control mechanism involves the vitamin K-dependent proteins, protein S and protein C, which combine with each other in molar ratios to form a complex with cellular membranes that inhibits the activity of factors Va and VIIIa. The complex also inhibits the action of PAI-1 thereby promoting fibrinolysis. Interestingly, a normal factor V molecule is essential for the inhibitory properties of this complex since mutation of factor V will cause a malfunction of the complex and result in a tendency to form clots (thrombophilia). This is the basis of the action of factor V Leiden, a mutant of factor V that causes familial thrombophilia.

Fibrinolysis, Fibrinogen, and the Acute Phase Response

The breakdown of the fibrin clot is initiated by activators of plasminogen, mainly tissue plasminogen activator (tPA) (**Figure 4**). This protein is produced by endothelial cells and activates plasminogen by converting it to plasmin. The plasmin then acts on fibrin to form the fibrin split

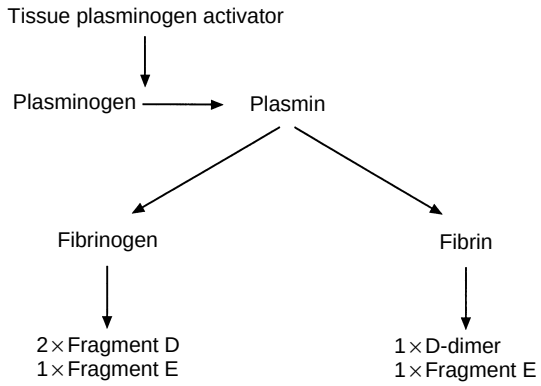


Figure 4 Fibrinolysis and fibrinogenolysis.

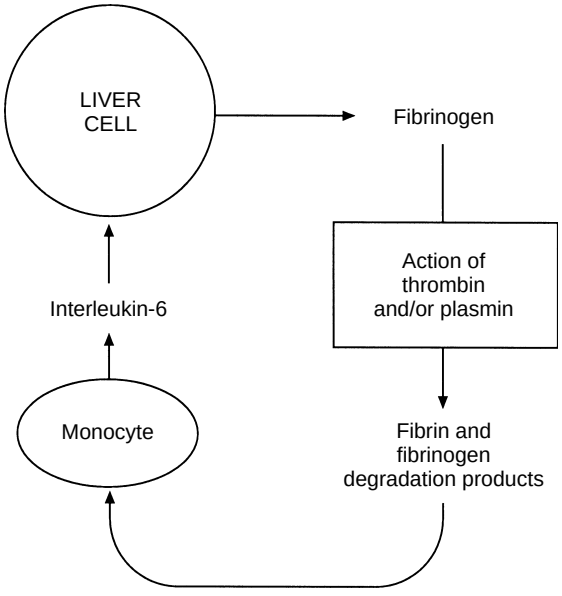


Figure 5 The role of interleukin-6 in fibrinogen synthesis.

products, a D-dimer and a fragment E from each fibrin monomer, and on fibrinogen to form the fibrinogen degradation products, two molecules of fragment D and one of fragment E. These breakdown products act as inhibitors of thrombin. They also stimulate the release of interleukin-6 (IL-6) from blood monocytes; the IL-6 then acts on the liver parenchymal cell to stimulate the synthesis of fibrinogen. The presence of this regulatory loop demonstrates the direct link between a proinflammatory cytokine, interleukin-6, and the production of a protein (fibrinogen) of the so-called acute phase response (**Figure 5**). Fibrinolysis is inhibited by PAI-1 and alpha2-antiplasmin.

Thrombosis

Thrombosis is a major cause of mortality and morbidity in Western societies. Deep venous thrombosis, myocardial infarction, pulmonary embolism, and acute thromboembolic stroke are, probably, all consequences of inappropriate blood coagulation or fibrinolysis. Fibrin formation is instantaneous and therefore clots can quickly occlude an artery and thereby precipitate an acute cardiac or cerebral event. This usually, but not exclusively, occurs in atherosclerotic blood vessels where plaques may rupture and cause thrombus formation.

The expression of tissue factor on cells and their subsequent exposure to blood is probably the major initiator of thrombus formation *in vivo*. It has been demonstrated that the blood monocytes from unstable angina and myocardial infarction patients express increased amounts of tissue factor when compared to healthy controls. Furthermore, it has been shown that monocytes from atherosclerotic plaques from unstable angina patients express increased amounts of tissue factor when compared to controls, thereby further implicating plaque rupture in the pathology of heart disease.

Hemostatic Factors and Risk of Coronary Heart Disease

Elevated plasma coagulation factor levels are risk factors for coronary heart disease. Fibrinogen synthesis in the liver is stimulated by the proinflammatory cytokine interleukin-6 and, therefore, elevated levels are found during the acute phase response. It has been argued that it may be difficult to assign elevated plasma fibrinogen as a definitive risk factor since the pathology of CHD involves inflammation and the acute phase response, which will lead to increased fibrinogen anyway. The same argument has been used in the case of elevated white cell counts, which is also a risk factor for coronary events. However, it has been demonstrated that if fibrinogen and/or white blood cell count remain high after a vascular event then there is greater risk of subsequent events. Therefore, increased plasma fibrinogen and elevated white cell count are now considered a major risk factor for CHD.

Further studies have indicated that other blood clotting factors may act as risk factors for CHD. For example, a prospective study, The Northwick Park Heart Study, identified factor VII as a risk factor for CHD and showed that plasma levels of factor VII were predictive of CHD in a dose-dependent manner. Another study has shown

that factor VIII may be a risk factor for cardiovascular disease. Increased levels of PAI-1 and decreased plasma levels of plasminogen activators have also been identified as risk factors for coronary heart disease.

Dietary Effects on Hemostatic Function

Dietary vitamin K profoundly affects the activity of the blood clotting factors II, VII, IX, and X and the inhibitory molecules, protein C and protein S. These proteins are synthesized in the liver and contain a region or module of the protein that contains modified glutamic acid residues (gamma carboxy glutamic acid) (Gla). Gla formation requires vitamin K and so a deficiency state will cause abnormal bleeding. A dietary deficiency of vitamin K is rare and nearly always occurs in the neonate leading to hemorrhagic disease of the newborn. However, the oral anticoagulant drugs based on warfarin are vitamin K analogs and thus prevent the synthesis of a biologically active coagulation protein.

Fibrinogen levels are elevated in obese individuals although many studies have failed to demonstrate any relationship between plasma fibrinogen levels and dietary fat. Active men have lower fibrinogen levels and a number of studies have demonstrated an inverse relationship between alcohol consumption and plasma fibrinogen. A high-fiber diet was also shown to correlate negatively with plasma fibrinogen. A few studies on the effects of antioxidant vitamins show that they may lower fibrinogen. For example, the Swedish MONItoring CArdiovascular disease (MONICA) study found that high levels of plasma retinol were associated with lower fibrinogen levels. However, plasma tPA levels were also lowered indicating that the fibrinolytic pathway was compromised in these subjects. Fish oil was shown to lower fibrinogen levels when the oil contained vitamin E.

Dietary fat profoundly affects the activity of factor VII. A high fat intake is associated with increased amounts of factor VIIa. Indeed, the postprandial levels of this active blood coagulation factor have been shown to be elevated after a high-fat meal. If exposed to tissue factor, these increased levels of factor VIIa would have major consequences for thrombogenesis. Platelet aggregation is often affected by diet and fish oils can reduce platelet aggregation.

The influence of diet on the components of the coagulation pathway is well recognized by the requirement of vitamin K. However, the role that diet plays in modulating the levels of coagulation factors that are, or may be, risk factors for coronary heart disease is unclear and the results are often conflicting. For example, in intervention studies examining the effects of fish oils on monocyte TF production, one study reported a decrease in expression whilst other studies showed no effect. The study where the positive effects were reported was carried out in Italy and it has been suggested that the Mediterranean diet of the Italians was a contributory factor in these findings. Obviously, more work is required to establish the effects of fish oils on TF expression.

Hemostasis is an exciting and important area of medicine and our knowledge of the molecular and cellular interactions that bring about clot formation is now at an advanced stage. Many of the studies on diet and hemostasis may have used inappropriate laboratory techniques. For example, blood platelets are easily activated making aggregation data difficult to interpret. The advent of automated coagulometers means that the clotting factors may be easily measured. There are commercially available immunoassays for many of the components of hemostasis described in this article. Using flow cytometry, TF can be measured on monocytes and P-selectin may be detected on activated platelets. Since thrombosis is a major cause of acute coronary events the rewards of studies on diet and hemostasis may be great.

See also: **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Coronary Heart Disease**: Lipid Theory; Prevention. **Vitamin K**.

Further Reading

Hutton RA, Laffan MA, and Tuddenham EGD (1999) Normal haemostasis. In: Hoffbrand AV, Lewis SM, and Tuddenham EGD (eds.) *Postgraduate Haematology*, pp. 550–580. Oxford: Butterworth Heinemann.

Laffin MA and Manning RA (2001) Investigation of haemostasis. In: Lewis SM, Bain BJ, and Bates I (eds.) *Practical Haematology*, pp. 339–413. London: Churchill Livingstone.

Ross R (1993) The pathogenesis of atherosclerosis: a perspective for the 1990's. *Nature* **362**: 801–809.

Lipid Theory

D Kritchevsky, Wistar Institute, Philadelphia, PA, USA

Introduction

Arteriosclerosis is a group of conditions characterized by thickening and stiffening of the arterial wall. Atherosclerosis is characterized by the formation of atheromas (lipid-laden plaques) in medium to large arteries. These are associated with calcifications of the arterial wall along with other changes. Eventually, the arterial lumen is reduced and the restricted blood flow due to these changes leads to clinical symptoms. Over the years there have been varying theories about the development of arterial lesions and these theories become more complex as our biochemical and molecular biological skills and knowledge increase.

Arterial fatty streaks are ubiquitous in humans and appear early in life. The fatty streak is comprised of lipid-rich macrophages and smooth muscle cells. Macrophages that accumulate lipid and are transformed into foam cells may be involved in the transformation of the fatty streak to an atherosclerotic lesion. In susceptible persons the fatty streaks may progress to fibrous plaques. Fibrous plaques, at their core, consist of a mixture of cholesterol-rich smooth muscle and foam cells. This core may contain cellular debris, cholesteryl esters, cholesterol crystals, and calcium. The fibrous cap consists of smooth muscle and foam cells, collagen, and lipid. The final stage in this process is the complicated plaque, which can obstruct the arterial lumen. Rupture of the cap may lead to clot formation and occlusion of the artery.

There are several theories of atherogenesis and these may eventually be shown to be interactive. The lipid hypothesis suggests that persistent hyperlipidemia leads to cholesterol accumulation in the arterial endothelium. Hypercholesterolemia may activate protein growth factors, which stimulate smooth muscle cell proliferation.

The lipid infiltration hypothesis proposes that elevated LDL levels increase LDL infiltration which, in turn, increases uptake of epithelial cells, smooth muscle cells, and macrophages. This cascade leads to cholesterol accumulation and, eventually, atheroma formation. The endothelial injury may arise from the action of oxidized lipid.

The endothelial injury hypothesis may help to explain the focal distribution of atheromas, which is not adequately accounted for by the lipid hypothesis. The endothelial injury hypothesis asserts that plaque formation begins when the endothelial cells that cover fatty streaks separate thus exposing the underlying lesion to the circulation. This may lead to smooth muscle proliferation, stimulated by circulating mitogens, or may cause platelet aggregation leading to mural thrombosis.

Another hypothesis relating to atherogenesis is the response-to-injury hypothesis. In this hypothesis the injury may be due to mechanical factors, chronic hypercholesterolemia, toxins, viruses, or immune reactions: these increase endothelial permeability, and lead to monocyte adherence to the epithelium or infiltration and platelet aggregation or adherence at the site of the injury. Injury releases growth factors that stimulate proliferation of fibrous elements in the intima. These growth factors may arise from the endothelial cell, monocyte, macrophages, platelet, smooth muscle cell, and T cell. They include epidermal growth factor, insulin-like growth factors, interleukins 1 and 2, platelet-derived growth factors, transforming growth factors α and β, and tumor necrosis factors α and β, among others. Monocytes and smooth muscle cells carry the 'scavenger' receptor, which binds oxidized but not native low-density lipoprotein (LDL) in a nonsaturable fashion. Uptake of oxidized LDL converts macrophages and smooth muscle cells into foam cells. Another theory of atherogenesis suggests that it begins as an immunological disease, which starts by an autoimmune reaction against the heat stress protein, hsp60. There have been suggestions that oxidized LDL may be an underlying cause of arterial injury.

The term 'atherosclerosis' is derived from the Greek words *athere*, meaning gruel, and *skleros*, meaning hardening. The term was coined by Marchand in 1904 to describe the ongoing process beginning with the early lipid deposits in the arteries to the eventual hardening. The World Health Organization (WHO) definition describes atherosclerosis as a 'variable combination of changes in the intima of the arteries involving focal accumulation of lipids and complex carbohydrates with blood and its constituents accompanied by fibrous tissue formation, calcification, and associated changes in the media' – a decidedly more complex concept than attributing it all to the dietary cholesterol.

Discussions of the etiology of heart disease always describe it as a life-style disease and list a number of risk factors, which include family history, hypercholesterolemia, hypertension, obesity, and cigarette smoking. Having listed these factors, discussion generally reverts to blood cholesterol and its control.

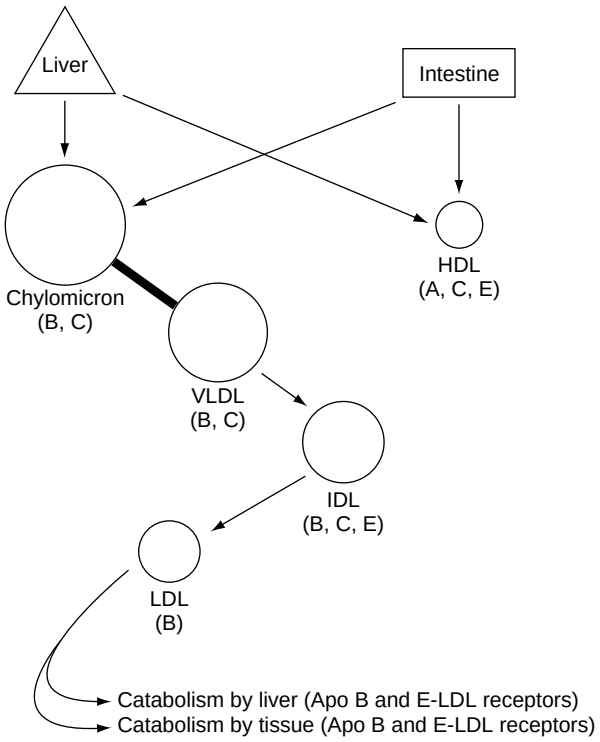


Figure 1 Outline of lipid metabolism. Letters in parentheses refer to apolipoproteins (apo). HDL, high-density lipoprotein; VLDL, very-low-density lipoprotein; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein.

The fasting blood plasma of a healthy individual is a clear, straw-colored liquid, which may contain 400–800 mg of lipids per 100 ml. This clear solution, which is high in lipids, is made possible by the water-soluble complex of lipids with protein, the lipoproteins. A generalized view of lipoprotein metabolism is provided in **Figure 1**. The existence of soluble lipid–protein complexes in serum was suggested about a century ago. Precipitation of a lipoprotein from horse serum was achieved in 1929 and classes of lipoproteins were adduced from studies using moving boundary electrophoresis. The critical experiments were carried out by Gofman and his group in the 1950s. They demonstrated that classes of lipoprotein complexes could be identified by their flotation characteristics in the analytical ultracentrifuge. These complexes were separable because they possessed different hydrated densities and they were defined initially by Svedberg units of flotation (S_f). The lipoproteins vary in chemical composition and although it is common to provide tables describing lipoprotein composition, the values are generally average values. This is so since the lipoproteins exist in a dynamic state exchanging their lipid components with those of tissues or other lipoproteins. Since identification is made according to a physical property, i.e., hydrated density, it is evident that different agglomerates of lipid and protein may have similar hydrated densities. In general, the lipoproteins are a series of macromolecules that, as they progress from low to high density, display decreasing triacylglycerol content and increasing cholesteryl ester, phospholipid, and protein.

Table 1 describes the major lipoproteins. Their chemical composition is described in **Table 2**.

As research continues and as analytical methodology becomes more precise we find a higher resolution of some lipoprotein classes and better definition of their roles. One example is lipoprotein (a) (lp(a)), first described in 1963. Lipoprotein (a) is an LDL whose normal apoprotein (apo B) is linked to an additional protein, apoprotein a, via a disulfide bridge. Lipoprotein (a) interferes with normal fibrinolysis leading to an increased prevalence of blood clots, and is thought to present an especially high risk for myocardial infarction. Characteristics and functions of lipoproteins are described in **Table 3**.

Molecular size influences the ease with which LDL particles can enter the arterial wall. Diabetic rabbits have greatly elevated plasma lipid levels but display surprisingly little atherosclerosis. The reason

Table 1 Major plasma lipoproteins

Lipoprotein class	*Size (nm)*	*Mol. wt*	*Density (g ml^{-1})*	*Electrophoretic mobility*	*Origin*	*Major apoproteins*
Chylomicron	100–400	10^6–10^7	<0.95	Origin	Intestine	A-I, B-48, C-II, C-III, E
VLDL	40–70	5×10^3	0.95–1.006	Prebeta	Liver	B-100, C-II, C-III, E
IDL	30–40	4.5×10^3	1.006–1.019	Between prebeta and beta	Catabolism of VLDL	B-100, C-II, C-III, E
LDL	22.5–27.5	2×10^3	1.019–1.063	Beta	Catabolism of VLDL and IDL	B-100
HDL	7.5–10	0.4×10^3	1.063–1.210	Alpha	Liver, intestine	A-I, A-II, C-II, C-III, E

VLDL, very-low-density lipoprotein; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; HDL, high-density lipoprotein.

Table 2 Plasma lipoprotein composition

Lipoprotein	*Composition (wt%)*				
	FC	*CE*	*TAG*	*PL*	*PROT*
Chylomicron	1	3	90	4	2
VLDL	7	14	55	16	8
IDL	6	22	30	24	18
LDL	7	48	5	20	20
HDL	4	15	4	27	50

FC, free cholesterol; CE, cholesteryl ester; TAG, triacylglycerol; PL, phospholipid; PROT, protein; VLDL, very-low-density lipoprotein; IDL, intermediate-density lipoprotein; LDL, low-density lipoprotein; HDL, high-density lipoprotein.

for this apparent discrepancy is that the lipoproteins of diabetic rabbits are rather large in size and do not penetrate the artery. Since 1982 we have known of an array of LDL particles ranging from small and dense to large and comparatively light. An LDL pattern characterized by an excess of small, dense particles is associated with a threefold increased risk of myocardial infarction, independent of age, sex, or body weight. Commonly, LDL is known as the 'bad' cholesterol and high-density lipoprotein (HDL) as the 'good' cholesterol. These recent findings indicate the presence of 'good, bad' cholesterol and 'bad, good' cholesterol.

Among the apolipoproteins, polymorphism of apoprotein E apparently dictates a subject's chances for successful treatment of lipidemia. The apoE alleles are designated as E2, E3, and E4. The most common pattern (55%) is homozygosity for E3, which gives rise to the E3/E3 phenotype. The next most common phenotype is E3/E4 (26%). The least frequently observed phenotype is E2/E (1%), which is often associated with type III hyperlipoproteinemia. There is some evidence suggesting that subjects bearing the E4 allele have higher levels of LDL than those with the E3/E3 pattern; they may also be more prone to Alzheimer's disease. **Tables 4** and **5** list primary and secondary dyslipoproteinemias.

Cholesterol and Cholesterolemia

In 1913 Anitschkow showed that it was possible to establish atherosclerosis in rabbits by feeding cholesterol. Since then virtually all research on atherosclerosis has centered on cholesterol – circulating cholesterol and dietary cholesterol. The epidemiological data suggest a role for

Table 3 Characteristics and functions of major apolipoproteins

Apolipoprotein	*Lipoprotein*	*(Approximate molecular weight (kD))*	*Source*	*(Average plasma) concentration (mg dL^{-1})*	*(Physiologic) function*
A-1	HDL, chylomicrons	28	Liver, intestine	100–120	Structural apoprotein of HDL, cofactor for LCAT
A-II	HDL, chylomicrons	17	Intestine, liver	35–45	Structural apoprotein of HDL, cofactor for hepatic lipase
A-IV	HDL, chylomicrons	46	Liver, intestine	10–20	Unknown
Apo (a)	Lp(a)	600	Liver	1–10	Unknown
B-48	Chylomicrons	264	Intestine	Trace	Major structural apoprotein, secretion and clearance of chlylomicrons
B-100	VLDL, LDL	550	Liver	100–125	Ligand for LDL receptor, structural apoprotein of VLDL and LDL
C-I	Chylomicrons, VLDL, HDL	5.80	Liver	6–8	Cofactor for LCAT
C-II	Chylomicrons, VLDL, HDL	9.10	Liver	3–5	Cofactor for LCAT
C-III	Chylomicrons, VLDL, HDL	8.75	Liver	12–15	Inhibitor of LPL, involved in lipoprotein remnant uptake
E-2	Chylomicrons, VLDL, HDL	35	Liver, peripheral tissues	4–5	Ligand for cell receptor
E-3	Chylomicrons, VLDL, HDL	35	Liver, peripheral tissues	4–5	Ligand for cell receptor
E-4	Chylomicrons, VLDL, HDL	35	Liver, peripheral tissues	4–5	Ligand for cell receptor

HDL, LDL, VLDL, high-, low-, and very-low-density lipoprotein; LCAT, lecithin-cholesterol acyltransferase; LPL, lipoprotein lipase.

Table 4 The primary dyslipoproteinemias

Type	*Changes in plasma*		*Apparent genetic disorder*	*Biochemical defect*
	Lipids	*Lipoproteins*		
I	TAG ↑	CM ↑	Familial LPL deficiency	Loss of LPL activity
II-a	C ↑	LDL ↑	Familial hypercholesterolemia	Deficiency of LDL receptor and activity
II-b	C ↑, TAG ↑	LDL, VLDL ↑	Familial combined hyperlipidemia	Unknown
III	C ↑, TAG ↑	β-VLDL ↑	Familial type III hyperlipidemia	Defect in TAG-rich remnant clearance
IV	TAG ↑	VLDL ↑	Familial hypertriacylglycerolemia	VLDL synthesis ↑, catabolism ↓
V	TAG ↑, C ↑	VLDL ↑, CM ↑	Familial type V hyperlipoproteinemia	Lipolysis of TGA-rich LP ↓, Production of VLDL TAG ↑
Hyper Lp(a)	C ↑	Lp(a) ↑	Familial hyper apo(a) lipoproteinemia	Inhibits fibrinolysis
Hyperapobeta-lipoproteinemia	TAG ↑	VLDL, LDL ↑	Familial type V hyperlipoproteinemia	CETP deficiency
Familial hypobeta-lipoproteinemia	C ↓, TAG ↓	CM ↓, VLDL ↑, LDL ↓	?	Inability to synthesize apo B-48 and apo B-100
A-beta-lipoproteinemia	C ↓, TAG ↓	CM ↓, VLDL ↓, LDL ↓	?	Apo B-48 and apo B-100 not secreted into plasma
Hypo-alphalipoproteinemia	C ↓, TAG ↓	HDL ↓	?	LCAT deficiency
Tangier disease				Apo A-I ↓, apo C-III ↓
Fish eye disease				Abnormal apo A-I, and apo A-II metabolism

C, cholesterol; CM, chylomicrons; CETP, cholesteryl ester transfer protein; HDL, LDL, VLDL, high-, low-, and very-low-density lipoprotein; LCAT, lecithin-cholesterol acyltransferase; LPL, lipoprotein lipase; TAG, triacylglycerol.

dietary fat, and hypercholesterolemia has been established as a principal risk factor for atherosclerosis. The lipid hypothesis was developed from the data obtained in the Framingham study, which suggested a curvilinear relationship between risk of atherosclerosis and plasma or serum cholesterol levels. However, studies of actual cholesterol intake as it affects cholesterol levels have yielded equivocal results.

Several studies have shown that the addition of one or two eggs to their daily diet did not influence serum cholesterol levels of free-living subjects. Data from the Framingham study show no correlation between cholesterol intake and cholesterol level. So we are left with the anomalous situation that blood cholesterol is an indicator of susceptibility to coronary disease but it is relatively unaffected by dietary cholesterol. It is of interest to point out that we are also seeing a correlation between low plasma or serum cholesterol levels and noncoronary death.

The type of fat in the diet has a strong influence on serum or plasma cholesterol levels. Rabbits fed saturated fat develop more severe atherosclerosis

Table 5 Secondary dyslipoproteinemias

Type	*Associated disease*	*Lipoproteins elevated*	*Apparent underlying defect*
I	Lupus erythematosis	Chylomicrons	Circulating LPL inhibitor
II	Nephrotic syndrome, Cushing's syndrome	VLDL and LDL	Overproduction of VLDL particles, defective lipolysis of VLDL triglycerides
III	Hypothyroidism, dysglobulinemia	VLDL and LDL	Suppression of LDL receptor activity, overproduction of VLDL triglycerides
IV	Renal failure, diabetes mellitus, acute hepatitis	VLDL	Defective lipolysis of triglyceride-rich VLDL due to inhibition of LPL and HL
V	Noninsulin dependent diabetes	VLDL	Overproduction and defective lipolysis of VLDL triglycerides

HDL, LDL, VLDL, high-, low, and very-low-density lipoprotein; HL, hepatic lipase; LPL, lipoprotein lipase.

than do rabbits fed unsaturated fat. In 1965 the groups of Keys and Hegsted independently developed formulae for predicting changes in cholesterol levels based on changes in the diet. Their formulae were based upon changes in quantity of saturated and unsaturated fat and in dietary cholesterol, but the last value makes a very small contribution to the overall number. The Keys formula is:

$$\Delta C = 1.35(2\Delta S - \Delta P) + 1.5\Delta Z$$

where ΔC represents the change in cholesterol level, ΔS and ΔP represent changes in levels of saturated and unsaturated fat, and Z is the square root of dietary cholesterol in mg per 1000 kcal of diet. The Hegsted formula is:

$$\Delta C_P = 2.16\Delta S - 1.65\Delta P + 0.168\Delta C_D + 85$$

where ΔC_P is change in plasma cholesterol and ΔC_D is change in dietary cholesterol in mg per 1000 kcal.

Both studies found that changes in dietary stearic acid did not fit the formula. Since those formulae were introduced a number of newer formulae have appeared, which provide a coefficient for every individual fatty acid, but the original formulae are still used most frequently. Under metabolic ward conditions it has been shown that lauric ($C_{12:0}$), myristic ($C_{14:0}$), and palmitic ($C_{16:0}$) acids raise both LDL and HDL cholesterol levels, and that oleic ($C_{18:1}$) and linoleic ($C_{18:2}$) acids raise HDL and lower LDL levels slightly. Thus, the type of fat is the determining factor in considering dietary fat effects on serum cholesterol. Experiments in which subjects were fed low or high levels of cholesterol in diets containing high or low ratios of saturated to polyunsaturated fat have been reported. When the fat was homologous, changing from low to high dietary cholesterol raised serum cholesterol concentration by 2%. However, even under conditions in which low levels of cholesterol were fed, changing from saturated to unsaturated fat raised serum cholesterol levels by 10% or more.

In nature most, but not all, unsaturated fatty acids are in the *cis* configuration. The major source of fats containing *trans* unsaturated fatty acids (*trans* fats) in the diet of developed nations is hydrogenated fat, such as is present in commercial margarines and cooking fats. Interest in *trans* fat effects on atherosclerosis and cholesterolemia was first evinced in the 1960s. In general, *trans* fats behave like saturated fats and raise serum cholesterol levels, but have not been found to be more atherogenic than saturated fats in studies carried out in rabbits, monkeys, and swine. Studies have also shown that *trans* fat effects may be relatively small if the diet contains sufficient quantities of essential fatty acids.

Studies, clinical and epidemiological, on the influence of *trans* unsaturated fats on the risk of coronary heart disease have continued. The evidence is that *trans* fats may influence the chemical indicators of heart disease risk but final proof must rest on verification by clinical trial. The concerns relative to *trans* fat effects have led to recommendations that the levels of *trans* fats present in the diet be reduced as much as possible. The availability of *trans*-free margarines and other fats may render the entire argument obsolete.

Protein

The type of protein in the diet also influences cholesterolemia and atherosclerosis. In animal studies in which the sole source of protein is of animal or plant origin, the former is more cholesterolemic than atherogenic. However, a 1:1 mix of animal and plant protein provides the higher-grade protein of animal protein and the normocholesterolemic effects of plant protein. The results underline the need for a balanced diet.

Fiber

Dietary fiber may influence lipidemia and atherosclerosis. Substances designated as insoluble fibers (wheat bran, for instance) possess laxative properties but have little effect on serum lipid levels. Soluble fibers (gel-forming fibers such as pectin or guar gum) influence lipidemia and glycemia. Oat bran, which contains β-glucans, which are soluble fibers, will lower cholesterol levels despite its designation.

Variations in Cholesterol Levels

Ignoring the differences of technique involved in cholesterol measurement in the laboratory – variations that are amenable to resolution – there are physiological considerations that should be recognized. Age, gender, genetics, adiposity, and personality traits can affect cholesterol levels, as can diseases unrelated to coronary disease. Stress (job stress, deadlines, examinations) can lead to increased cholesterol levels.

A definite seasonal variation in cholesterol levels (usually higher in winter months) has been seen in a number of studies. Scientists from the National Institutes of Health in the US carried out one of

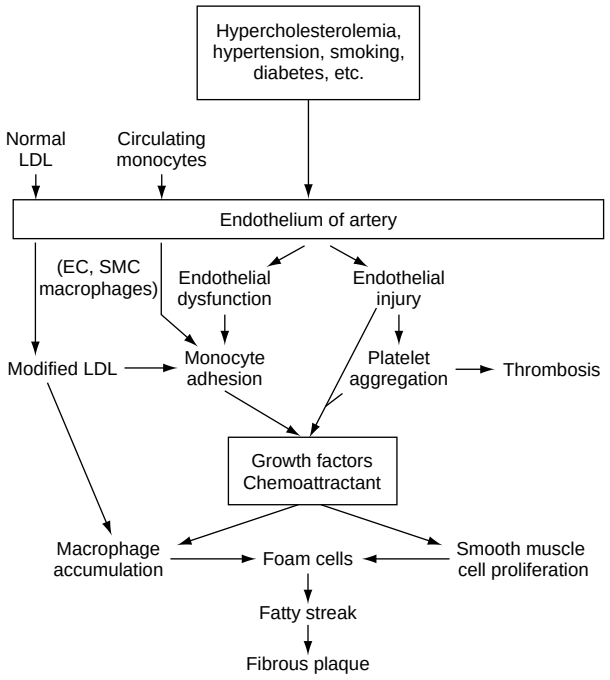


Figure 2 Factors involved in formation of the atherosclerotic plaque.

the finest studies in this area. They examined carefully the data from the 10 American Lipid Research Clinics. They observed that the etiology of their findings was unknown but they found the total and LDL cholesterol levels varied inversely with length of day. The level of HDL cholesterol varied much less, but its variation was correlated directly with ambient temperature. The foregoing does not reduce the importance of measuring cholesterol levels but makes it important to take into consideration the subjects' physical and mental state as well as time of year.

Figure 2 attempts to summarize the many factors now considered to play a role in the formation of the atherosclerotic plaque.

See also: **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Coronary Heart Disease**: Hemostatic Factors; Prevention. **Diabetes Mellitus**: Etiology and Epidemiology. **Fats and Oils**. **Fatty Acids**: *Trans* Fatty Acids. **Hyperlipidemia**: Overview; Nutritional Management. **Lipids**: Chemistry and Classification. **Lipoproteins**.

Further Reading

Ginsburg HN (1998) Lipoprotein physiology. *Endocrinology Metabolism Clinics of North America* **27**: 503–519.

Gold P, Grover S, and Roncari DAK (eds.) (1992) *Cholesterol and Coronary Heart Disease – The Great Debate*. Park Ridge, NJ: CRC Press.

Libby P (2002) Inflammation in atherosclerosis. *Nature* **420**: 868–874.

Lusis AJ (2000) Atherosclerosis. *Nature* **407**: 233–241.

McNamara (2000) Dietary cholesterol and atherosclerosis. *Biochimica Biophysica Acta* **1529**: 310–320.

Nicolosi RJ, Kritchevsky D, and Wilson TA (1999) Pathobiology of hypercholesterolemia and atherosclerosis. In: Rippe JM (ed.) *Lifestyle Management and Prevention of Cardiovascular Disease*, pp. 25–39. London: Blackwell Science Press.

Ross R (1999) Atherosclerosis – an inflammatory disease. *New England Journal of Medicine* **340**: 115–126.

Tabas T (2002) Cholesterol in health and disease. *Journal of Clinical Investigation* **110**: 583–589.

Velican C and Velican D (1989) *Natural History of Coronary Atherosclerosis*. Boca Raton, FL: CRC Press.

White RA (1989) *Atherosclerosis and Arteriosclerosis*. Boca Raton, FL: CRC Press.

Prevention

K Srinath Reddy, All India Institute of Medical Sciences, New Delhi, India

Introduction

Coronary heart disease (CHD) is the leading cause of death in the world. While it is well established as the foremost contributor to mortality in most developed countries, it is also a major and rapidly rising cause of death in many developing countries. Global health transitions, which have seen substantial changes in age-specific coronary mortality rates across the world, in the past half a century, have also been associated with changes in nutrition, which explain a large part of the rise or fall of CHD-related death rates.

Diet and nutrition have been extensively investigated as risk factors for CHD. Many dietary factors have been linked directly to an increased or decreased risk of CHD or to major established risk factors of CHD like high blood pressure, disordered blood fats (dyslipidemia), diabetes and metabolic syndrome, overweight and obesity, and also to emerging risk factors like inflammatory markers and homocysteine. Nutrition influences atherogenesis, thrombosis, and inflammation – all of which are interconnected pathways that lead to CHD.

Observational epidemiological studies and clinical trials have contributed to a wide body of knowledge of the role that some nutrients (like saturated and *trans* fats, salt, and refined carbohydrates) play in

increasing the risk of CHD and of the protective effect of other nutrients (such as fruit and vegetables, polyunsaturated fats, nuts, and fish) against CHD. This knowledge has been successfully applied both in public health and in clinical practice to reduce the risk of CHD in populations as well as in individuals. The present state of that knowledge, as relevant to prevention of CHD, is summarized below.

Global Trends in CHD as a Reflection of Nutrition Transition

Coronary heart disease accounted for 7.2 million deaths in 2002, which forms a large fraction of not only the total number of deaths worldwide due to cardiovascular diseases (16.6 million) but also of the global total number of deaths from any cause (57 million). While age-specific coronary mortality rates have declined in the industrial countries over the past three decades, the absolute burdens of CHD continue to be high. CHD death rates are rising in the developing countries, where about half of these deaths occur below the age of 70 years. In Eastern and Central Europe CHD mortality rates rose sharply in the 1980s and 1990s and have only recently shown signs of stabilization, albeit at high levels.

These changes in CHD mortality rates have accompanied well-documented or clearly discernible shifts in the nutritional state of the populations. The decline of CHD mortality in Western and Northern Europe was linked to a reduction in the consumption of unhealthy fats (saturated fats and *trans* fats) and salt as well as an increased consumption of fruits and vegetables. This is best documented in The Netherlands and Finland. Similarly, the recent decline of CHD mortality in Poland was explained by the increase in fruit and vegetable consumption and growing substitution of vegetable fats for animal fats. Similar evidence of a favorable nutrition transition preceding the decline in CHD mortality rates is available from other developed countries like the US, Canada, Australia, and New Zealand.

The developing countries have, however, witnessed a recent transition in the opposite direction. China, for example, has experienced a large increase in fat consumption over the past two decades, accompanied by a progressive rise in the mean plasma cholesterol levels of the population as well as in the CHD mortality rates. Other developing countries are also increasingly adopting unhealthy dietary patterns that augment the risk of CHD.

Understanding the Links between Nutrition and CHD

The pathogenesis of CHD is mediated through the interconnected pathways of atherogenesis (fat deposition in the walls of the coronary arteries to form plaques), thrombosis (blood clotting over disrupted plaques) and inflammation (which initially damages the blood vessel walls and continues to destabilize the plaques). Nutrition has a major role in influencing each of these pathways and often provides the connecting link between them.

Major coronary risk factors include an abnormal blood lipid profile (especially plasma cholesterol and its subfractions), high blood pressure, and diabetes. Overweight and obesity (both the general and central patterns) are also associated with an increased risk of CHD. Nutrition has a powerful influence on all of these risk factors, with an unhealthy diet pattern tending to elevate them and a healthy diet pattern reducing the levels of risk. Diet becomes especially important in the context of the metabolic syndrome (a complex of central obesity, high blood pressure, dyslipidemia, and glucose intolerance), an entity which is being increasingly identified as a major risk factor for CHD. Nutrition is also linked to the propensity to develop cardiac arrhythmias, in the setting of CHD, and is an important predictor of sudden cardiac death. These links between dietary patterns and several specific nutrients not only manifest as fat deposition in the arteries, plaque growth, plaque instability, and thrombosis but are evident much earlier in the natural history of CHD, as endothelial dysfunction (inability of the arteries to dilate normally), elevated levels of inflammatory markers (such as C reactive protein), and increased intimal medial thickness of arterial walls. These precede and predict the clinical manifestation of CHD.

Nutrients and CHD

Dietary Fats: Cholesterol

The relationship between dietary fats and cardiovascular disease (CVD), especially CHD, has been extensively investigated, with strong and consistent associations emerging from a wide body of evidence accrued from animal experiments, as well as observational studies, clinical trials, and metabolic studies conducted in diverse human populations. This relationship was initially considered to be mediated mainly through the atherogenic effects of plasma lipids (total cholesterol, lipoprotein fractions, and triglycerides). The effects of dietary fats on

thrombosis and endothelial function as well as the relationship of plasma and tissue lipids to the pathways of inflammation have been more recently understood. Similarly, the effects of dietary fats on blood pressure have also become more evident through observational and experimental research.

Cholesterol in the blood and tissues is derived from two sources: diet and endogenous synthesis. Dairy fat and meat are major dietary sources. Dietary cholesterol raises plasma cholesterol levels. Although both high-density lipoprotein (HDL) and low-density lipoprotein (LDL) fractions increase, the effect on the total/HDL ratio is still unfavorable, but small. The upper limit for dietary cholesterol intake has been prescribed, in most guidelines, to be $300\,mg\,day^{-1}$. However, as endogenous synthesis is sufficient to meet the physiological needs, there is no requirement for dietary cholesterol and it is advisable to keep the intake as low as possible. If intake of dairy fat and meat are controlled, then there is no need for severe restriction of egg yolk intake, although some limitation remains prudent.

Saturated Fatty Acids (SFAs)

The relationship of dietary saturated fat to plasma cholesterol levels and to CHD was graphically demonstrated by the Seven Countries Study involving 16 cohorts, in which saturated fat intake explained up to 73% of the total variance in CHD across these cohorts. In the Nurses' Health Study, the effect of saturated fatty acids was much more modest, especially if saturates were replaced by carbohydrates. The most effective replacement for saturated fatty acids in terms of CHD prevention is by polyunsaturated fatty acids (PUFAs). This agrees with the outcome of large randomized clinical trials, in which replacement of saturated and *trans* fats by polyunsaturated vegetable oils effectively lowered CHD risk.

Trans-Fatty Acids (*t*-FAs)

t-FAs (*Trans*-Fatty Acids) are geometrical isomers of unsaturated fatty acids that assume a saturated fatty acid-like configuration. Partial hydrogenation, the process used to create *t*-FAs, also removes essential fatty acids such as LA (Linoleic Acid) and ALNA (Alpha Linolenic Acid). Metabolic studies have demonstrated that *t*-FAs render the plasma lipid profile even more atherogenic than SFAs, by not only elevating LDL cholesterol to similar levels but also decreasing HDL cholesterol. As a result, the ratio of LDL cholesterol to HDL cholesterol is significantly higher with a *t*-FA diet (2.58) than with a SFA diet (2.34) or an oleic acid diet (2.02). This greatly enhances the risk of CHD. Evidence that intake of *t*-FAs increases the risk of CHD initially became available from large population-based cohort studies in the US and in an elderly Dutch population. Eliminating *t*-FAs from the diet would be an important public health strategy to prevent CHD. Since these are commercially introduced agents into the diet, policy measures related to the food industry practices would be required along with public education. *t*-FAs have been eliminated from retail fats and spreads in many parts of the world, but deep-fat fried fast foods and baked goods are a major and increasing source.

Monounsaturated Fatty Acids (MUFAs)

The only nutritionally important MUFA is oleic acid, which is abundant in olive and canola oils and also in nuts. The epidemiological evidence related to MUFAs and CHD is derived from studies on the Mediterranean diet (see below), as well as from the Nurses' Health Study and other similar studies in the US.

Polyunsaturated Fatty Acids (PUFAs)

PUFAs are categorized as *n*-6 PUFAs (mainly derived from linoleic acid) and *n*-3 PUFAs (mainly present in fatty fish and also derived from alpha-linoleic acid). Clinical trials, in which *n*-6 PUFAs (containing linoleic acid) were substituted for SFAs showed a greater impact on reduction of both plasma cholesterol and CHD risk, in contrast to trials where low-fat diets were employed.

Much of the epidemiological evidence related to *n*-3 PUFAs is derived from the study of fish consumption in populations or interventions involving fish diets in clinical trials. Fish oils were, however, used in a large clinical trial of 11 300 survivors of myocardial infarction. After 3.5 years of follow-up, the fish oil group ($1\,g\,day^{-1}$) had a statistically significant 20% reduction in total mortality, 30% reduction in cardiovascular death, and 45% decrease in sudden death.

The Lyon Heart Study in France incorporated an *n*-3 fatty acid (alpha-linolenic acid) into a diet that was altered to develop a 'Mediterranean diet' intervention. In the experimental group, plasma ALNA and EPA (Eicosapentenoic Acid) increased significantly and the trial reported a 70% reduction in cardiovascular mortality at 5 years. Total and LDL cholesterol were identical in the experimental and control groups, suggesting that thrombotic and perhaps arrhythmic events may have been favorably influenced by *n*-3 PUFAs. Since the diet altered many other variables, such as fiber and antioxidants

(by increasing fruit and vegetable consumption), direct attribution of benefits to *n*-3 PUFAs becomes difficult to establish.

The proportions of SFAs, MUFAs, and PUFAs as constituents of total fat intake and total energy consumption have engaged active attention, in view of the strong relationship of these fatty acids to the risk of CHD. The reduction of SFAs in the diet has been widely recommended, but its replacement has been an area of debate, as to whether the place of reduced SFAs should be taken by MUFAs, PUFAs, or carbohydrate. Both MUFAs and PUFAs improve the lipoprotein profile, although PUFAs are somewhat more effective. In view of this, several recent dietary recommendations suggested that SFAs should be kept below 10% of daily energy intake (preferably reduced to 7–8%), MUFAs should be increased to 13–15%, and PUFAs raised to 7–10% of daily energy, with the total fat contributing to less than 30% of all calories consumed. These may need to be adjusted for populations who consume less quantities of total fat, so as to ensure an adequate intake of MUFAs and PUFAs even under those circumstances. The emphasis is now shifting from the quantity of fat to the quality of fat, with growing evidence that even diets with 30–35% fat intake may be protective if the type of fats consumed are mostly from the MUFA and PUFA categories. Enhancing the nutritional quality of dietary fat consumption, to provide greater cardiovascular protection, may be attempted by decreasing the sources of saturated fats and eliminating *t*-FAs in the diet, increasing the consumption of foods containing unsaturated fatty acids (both MUFAs and PUFAs), and decreasing dietary cholesterol consumption.

Carbohydrates

Diets which are high in refined carbohydrates appear to reduce HDL cholesterol levels and increase the fraction of small dense LDL, both of which may impact adversely on vascular disease. This dyslipidemic pattern is consistent with the elevation of plasma triglycerides and is typical of the 'metabolic syndrome.' Carbohydrate diets with high glycemic index might adversely impact on glucose control, with associated changes in plasma lipids, and have been linked to an increased risk of CHD.

Fiber

Most soluble fibers reduce plasma total and LDL cholesterol concentrations, as reported by several trials. Fiber consumption strongly predicts insulin levels, weight gain, and cardiovascular risk factors like blood pressure, plasma triglycerides, LDL and HDL cholesterol, and fibrinogen. Several large cohort studies in the US, Finland, and Norway have reported that subjects consuming relatively large amounts of whole-grain cereals have significantly lower rates of CHD.

Antioxidants

Though several cohort studies showed significant reductions in the incidence of cardiac events in men and women taking high-dose vitamin E supplements, large clinical trials failed to demonstrate a cardioprotective effect of vitamin E supplements. Beta-carotene supplements also did not provide protection against CHD and, in some trials, appeared to increase the risk.

Folate

The relationship of folate to CVD has been mostly explored through its effect on homocysteine, which has been put forward as an independent risk factor for CHD. Reduced plasma folate has been strongly associated with elevated plasma homocysteine levels and folate supplementation has been demonstrated to decrease those levels. Data from the Nurses' Health Study in the US showed that folate and vitamin B_6, from diet and supplements, conferred protection against CHD (fatal and nonfatal events combined) and suggested a role for their increased intake as an intervention for primary prevention of CHD. Recommendations related to folate supplementation must, however, await the results of ongoing clinical trials. Dietary intake of folate through natural food sources may be encouraged in the meanwhile, especially in individuals at a high risk of arterial or venous thrombosis and elevated plasma homocysteine levels.

Flavonoids and Other Phytochemicals

Flavonoids are polyphenolic antioxidants, which occur in a variety of foods of vegetable origin, such as tea, onions, and apples. Data from several prospective studies indicate an inverse association of dietary flavonoids with CHD. The role of these and other phytochemicals (such as plant stanols and sterols) in relation to CHD needs to be elucidated further.

Sodium and Potassium

High blood pressure (HBP) is a major risk factor for CHD. The relative risk of CHD, for both systolic and diastolic blood pressures, operates in a continuum of increasing risk for rising pressure but the absolute risk of CHD is considerably modified by coexisting risk factors (such as blood lipids and diabetes), many of which are also influenced by diet. A cohort study in Finland observed a 51%

greater risk of CHD mortality with a 100 mmol increase in 24-h urinary sodium excretion. Several clinical trials have convincingly demonstrated the ability of reduced sodium diets to lower blood pressure. A meta-analysis of long-term trials suggests that reducing daily salt intake from 12 g day^{-1} to 3 g day^{-1} is likely to reduce CHD by 25% (and strokes by 33%). Even more modest reductions would have substantial benefits (10% lower CHD for a 3-g salt reduction). The benefits of dietary potassium in lowering blood pressure have been well demonstrated but specific effects on CHD risk have not been well studied. Keeping the dietary sodium:potassium ratio at a low level is essential to avoid hypertension.

Food Items

Fruits and Vegetables

A systematic review reported that nine of ten ecological studies, two of three case–control studies, and six of sixteen cohort studies found a significant protective association for CHD with consumption of fruits and vegetables or surrogate nutrients. In a 12-year follow-up of 15 220 male physicians in the US, men who consumed at least 2.5 servings of vegetables per day were observed to have a 33% lower risk for CHD, compared with men in the lowest category (<1 serving per day). A follow-up study of NHANES (National Health and Nutrition Examination Survey), a large national survey in the US, also reported a coronary protective effect of regular fruit and vegetable intake. Persons who consumed fruits and vegetables 3 or more times a day were at 24% lower risk than those who consumed less than one portion a day. A global study of risk factors of CHD in 52 countries (INTERHEART) also reported low consumption of fruit and vegetables to be a major risk factor, across all regions.

Fish

In the UK diet and reinfarction trial, 2-year mortality was reduced by 29% in survivors of a first myocardial infarction in those receiving advice to consume fatty fish at least twice a week. A meta-analysis of 13 large cohort studies suggests a protective effect of fish intake against CHD. Compared with those who never consumed fish or did so less than once a month, persons who ate fish had a lower risk of CHD (38% lower for 5 or more times a week, 23% lower for 2–4 times a week, 15% lower for once a week, and 11% lower for 1–3 times a month). Each 20 g day^{-1} increase in fish consumption was related to a 7% lower risk of CHD.

Nuts

Several large epidemiological studies, the best known among them being the Adventist Health Study, demonstrated that frequent consumption of nuts was associated with decreased risk of CHD. The extent of risk reduction ranged from 18% to 57% for subjects who consumed nuts more than 5 times a week compared to those who never consumed nuts. An inverse dose–response relationship was demonstrated between the frequency of nut consumption and the risk of CHD, in men as well as in women. Most of these studies considered nuts as a group, combining many types of nuts (walnuts, almonds, pistachio, pecans, macadamia nuts, and legume peanuts).

Soy

Soy is rich in isoflavones, compounds that are structurally and functionally similar to estrogen. Several animal experiments suggest that intake of these isoflavones may provide protection against CHD, but human data on efficacy and safety are still awaited. Naturally occurring isoflavones, isolated with soy protein, reduced the plasma concentrations of total and LDL cholesterol without affecting the concentrations of triglycerides or HDL cholesterol in hypercholesterolemic individuals.

Dairy Products

Dairy consumption has been correlated positively, in ecological studies, with blood cholesterol as well as coronary mortality. Milk consumption correlated positively with coronary mortality rates in 43 countries and with myocardial infarction in 19 regions of Europe.

Alcohol

The relationship of alcohol to overall mortality and cardiovascular mortality has generally been J-shaped, when studied in Western populations in whom the rates of atherothrombotic vascular disorders are high. The protective effect of moderate ethanol consumption is primarily mediated through its effect on the risk of CHD, as supported by more than 60 prospective studies. A consistent coronary protective effect has been observed for consumption of 1–2 drinks per day of an alcohol-containing beverage but heavy drinkers have higher total mortality than moderate drinkers or abstainers, as do binge drinkers.

Composite Diets and CHD

The Mediterranean diet

The traditional Mediterranean diet has been described to have eight components:

1. high monounsaturated-to-saturated fat ratio;
2. moderate ethanol consumption;
3. high consumption of legumes;
4. high consumption of cereals (including bread);
5. high consumption of fruits;
6. high consumption of vegetables;
7. low consumption of meat and meat products; and
8. moderate consumption of milk and dairy products.

Most of these features are found in many diets in that region. The characteristic component is olive oil, and many equate a Mediterranean diet with consumption of olive oil.

A secondary prevention trial of dietary intervention in survivors of a first recent myocardial infarction (the Lyon Heart study), which aimed to study the cardioprotective effects of a 'Mediterranean type' of diet, actually left out its most characteristic component, olive oil. The main fat source was rapeseed oil. Vegetables and fruits were also increased in the diet. On a 4-year follow-up, the study reported a 72% reduction in cardiac death and nonfatal myocardial infarction. The risk of overall mortality was lowered by 56%. Large cohort studies in Greece and in several elderly European population groups have also recently reported a protective effect against CHD and better over all survival in persons consuming a Mediterranean type of diet. The protection was afforded by the composite diet rather than by any single component. Improvement in metabolic syndrome and reduction of inflammatory markers has also been observed with this diet, which may explain part of the protection against CHD.

DASH Diets

A composite diet, employed in the Dietary Approaches to Stop Hypertension (DASH) trials, has been found to be very effective in reducing blood pressure in persons with clinical hypertension as well as in people with blood pressure levels below that threshold. This diet combines fruits and vegetables with food products that are low in saturated fats. The blood pressure lowering effect is even greater when the DASH diet is modified to reduce the sodium content. Though the effects on CHD prevention have not been directly studied, the blood pressure and lipid-lowering effects of the low salt-DASH diet are likely to have a substantial impact on CHD risk.

Vegetarian Diets

A reduced risk of CVD has been reported in populations of vegetarians living in affluent countries and in case–control comparisons in developing countries. Reduced consumption of animal fat and increased consumption of fruit, vegetables, nuts, and cereals may underlie such a protective effect. However, 'vegetarian diets' *per se* need not be healthful. If not well planned, they can contain a large amount of refined carbohydrates and t-FAs, while being deficient in the levels of vegetable and fruit consumption. The composition of the vegetarian diet should, therefore, be defined in terms of its cardioprotective constituents.

Prudent versus Western Patterns

In the Health professionals follow-up study in the US, a prudent diet pattern was characterized by higher intake of vegetables, fruits, legumes, whole grains, fish, and poultry, whereas the Western pattern was defined by higher intake of red meat, processed meat, refined grains, sweets and dessert, French fries, and high-fat dairy products. After adjustment for age and other coronary risk factors, relative risks, from the lowest to the highest quintiles of the prudent pattern score, were 1.0, 0.87, 0.79, 0.75, and 0.70, indicating a high level of protection. In contrast, the relative risks, across increasing quintiles of the western pattern, were 1.0, 1.21, 1.36, 1.40, and 1.64, indicating a mounting level of excess risk. These associations persisted in subgroup analyses according to cigarette smoking, body mass index, and parental history of myocardial infarction.

Japanese Diet

The traditional Japanese diet has attracted much attention because of the high life expectancy and low CHD mortality rates among the Japanese. This diet is low in fat and sugar and includes soy, seaweeds, raw fish, and a predominant use of rice. It has been high in salt, but salt consumption has recently been declining in response to Japanese Health Ministry guidelines.

Prevention Pathways

The powerful relationship of specific nutrients, food items and dietary patterns to CHD has been persuasively demonstrated by observational epidemiological studies (which indicate the potential for primary

prevention in populations) and by clinical trials (which demonstrate the impact on secondary prevention in individuals).

Atherosclerotic vascular diseases (especially CHD) are multifactorial in origin. Each of the risk factors operates in a continuous manner, rather than across an arbitrary threshold. When multiple risk factors coexist, the overall risk becomes multiplicative. As a result of these two phenomena, the majority of CHD events occurring in any population arise from any individuals with modest elevations of multiple risk factors rather than from the few individuals with marked elevation of a single risk factor.

These phenomena have two major implications for CHD prevention. First, it must be recognized that a successful prevention strategy must combine population-wide interventions (through policy measures and public education) with individual risk reduction approaches (usually involving counseling and clinical interventions). Second, diet is a major pathway for CHD prevention, as it influences many of the risk factors for CHD, and can have a widespread impact on populations and substantially reduce the risk in high-risk individuals. Even small changes in blood pressure, blood lipids, body weight, central obesity, blood sugar, inflammatory markers, etc., can significantly alter the CHD rates, if the changes are widespread across the population. Modest population-wide dietary changes can accomplish this, as demonstrated in Finland and Poland. At the same time, diet remains a powerful intervention to substantially reduce the risk of a CHD-related event in individuals who are at high risk due to multiple risk factors, prior vascular disease, or diabetes.

A diet that is protective against CHD should integrate: plenty of fruits and vegetables (400–600 g day^{-1}); a moderate amount of fish (2–3 times a week); a small quantity of nuts; adequate amounts of PUFAs and MUFAs (together constituting about 75% of the daily fat intake); low levels of SFAs (less than 25% of the daily fat intake); limited salt intake (preferably less than 5 day^{-1}); and restricted use of sugar. Such diets should be culturally appropriate, economically affordable, and based on locally available foods.

National policies and international trade practices must be shaped to facilitate the wide availability and uptake of such diets. Nutrition counseling of individuals at high risk must also adopt these principles while customizing dietary advice to specific needs of the person. CHD is eminently preventable, as evident from research and demonstrated in practice across the world. Appropriate nutrition is a major pathway for CHD prevention and must be used more widely to make CHD prevention even more effective at the global level.

See also: **Alcohol**: Absorption, Metabolism and Physiological Effects; Disease Risk and Beneficial Effects; Effects of Consumption on Diet and Nutritional Status. **Antioxidants**: Diet and Antioxidant Defense; Observational Studies; Intervention Studies. **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Coronary Heart Disease**: Hemostatic Factors; Lipid Theory. **Dietary Fiber**: Role in Nutritional Management of Disease. **Fatty Acids**: Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated. **Fish**. **Folic Acid**. **Fruits and Vegetables**. **Nuts and Seeds**. **Potassium**. **Sodium**: Physiology; Salt Intake and Health. **Vegetarian Diets**.

Further Reading

Appel LJ, Moore TJ, Obarzanek E *et al.* (1997) A clinical trial of the effects of dietary patterns on blood pressure. DASH Collaborative Research Group. *New England Journal of Medicine* **336**: 1117–1124.

De Lorgeril M, Salen P, Martin JL *et al.* (1999) Mediterranean diet, traditional risk factors, and the rate of cardiovascular complications after myocardial infarction: final report of Lyon Diet Heart Study. *Circulation* **99**: 779–785.

He FJ and MacGregor GA (2003) How far should salt intake be reduced? *Hypertension* **42**: 1093–1099

He K, Song Y, Daviglus ML *et al.* (2004) Accumulated evidence on fish consumption and coronary heart disease mortality: a meta-analysis of cohort studies. *Circulation* **109**: 2705–2711.

INTERSALT Cooperative Research Group (1988) INTERSALT: an international study of electrolyte excretion and blood pressure. Results for 24 hr urinary sodium and potassium excretion. *British Medical Journal* **297**: 319–328.

Kris-Etherton P, Daniels SR, Eckel RH *et al.* (2001) Summary of the scientific conference on dietary fatty acids and cardiovascular health: conference summary from the nutrition committee of the American Heart Association. *Circulation* **103**: 1034–1039.

Ness AR and Powles JW (1997) Fruit and vegetables, and cardiovascular disease: a review. *International Journal of Epidemiology* **26**: 1–13.

Reddy KS and Katan MB (2004) Diet, nutrition and the prevention of hypertension and cardiovascular diseases. *Public Health and Nutrition* **7**: 167–186.

Sacks FM, Svetkey LP, Vollmer WM *et al.* (2001) Effects on blood pressure of reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. DASH-Sodium Collaborative Research Group. *New England Journal of Medicine* **344**: 3–10.

Seely S (1981) Diet and coronary disease. A survey of mortality rates and food consumption statistics of 24 countries. *Medical Hypotheses* **7**: 907–918.

Trichopoulou A, Costacou T, Bamia C, and Trichopoulos D (2003) Adherence to a Mediterranean diet and survival in a Greek population. *New England Journal of Medicine* **348**: 2599–2608.

Verschuren WMM, Jacobs DR, Bloemberg BP *et al.* (1995) Serum total cholesterol and long-term coronary heart disease mortality in different cultures. Twenty-five year follow-up of the Seven Countries Study. *JAMA* **274**: 131–136.

World Health Organization (2003) Diet, nutrition and the prevention of chronic diseases. *Technical Report Series* **916**: 1–149.

World Health Organization (2002) *The World Health Report 2002. Reducing Risks, Promoting Healthy Life*. Geneva: WHO.

Yusuf S, Hawken S, Ounpuu S *et al.* (2004) INTERHEART study Investigators. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet* **364**: 937–952.

CYSTIC FIBROSIS

J Dowsett and O Tully, St Vincent's University Hospital, Dublin, Ireland

Definition and Etiology

Cystic fibrosis (CF) is a multisystem autosomal recessive disorder caused by the mutation of a single gene on the long arm of chromosome 7 that codes for the cystic fibrosis transmembrane regulator (CFTR). This protein regulates the passage of chloride through the membrane of secretory epithelia; the dysfunction of which results in an altered composition of epithelial secretions. Clinically, CF is characterized by chronic pulmonary infection with periods of acute exacerbation, pancreatic insufficiency and excessive losses of sweat electrolytes. The latter forms the basis for the diagnostic test. The mutated gene was identified in 1989 and since then over 800 CFTR mutations have been reported, the most common of these being △ F508.

Prevalence

Approximately 5% of the Caucasian North European and North American populations are carriers of the gene defect causing CF, leading to an approximate incidence of 1 in 2500 live births. This inheritance is illustrated in **Figure 1**. The incidence of CF in non-Caucasians is much lower and estimated to be around 1 in 100 000 in Oriental populations.

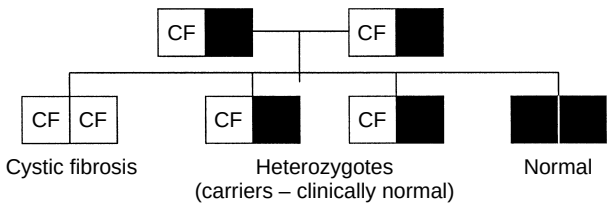


Figure 1 Mode of inheritance of CF: a Mendelian inherited recessive characteristic.

Prognosis

The median age of survival has dramatically risen from approximately 2 years in the 1940s to around 30 years in the 1990s. A current survival estimation following diagnosis is approximately 40 years. This improved prognosis can be attributed to a combination of factors including aggressive management of infections, effective antibiotics, improved nutritional management, modern physiotherapy techniques, and the centralization of treatment in specialist centers. The survival age for females with CF would appear to be less than that for males. This may be related to poorer nutritional status amongst female CF patients. Expert management started immediately after an early diagnosis of CF by neonatal screening results in an important beneficial effect on outcome and may be critical to the clinical course of the condition and long-term prognosis. Even though optimized nutrition, antibiotics, and chest physiotherapy remain the mainstay of CF management, new approaches to treatment are being developed that may add to the traditional medical therapy for CF. As prognosis and survival improves nutritional related issues become more prevalent including the effective management of pregnancy, diabetes, osteoporosis, and transplantation.

Clinical Features

The clinical features of CF are listed in **Table 1**.

Pathogenesis of Lung Disease

Pulmonary disease can be demonstrated within the first few months of life. Bacterial infection is characterized by high levels of neutrophils and mediators of infection in the form of interleukin 1, 8 and elastases. Mucous glands become dilated leading to obstruction, secondary infection, and progressive lung damage. Frequent periods of respiratory infection and exacerbation are common in CF with increased cough, increased sputum production, and shortness of breath. The immune response appears

Table 1 Clinical features of CF

Respiratory features of cystic fibrosis	
Atelectasis	Incomplete expansion of a lung or part of a lung due to airlessness or collapse
Bronchiectasis	Chronic dilatation of the bronchi associated with coughing and expectoration of purulent mucus
Bronchitis	Inflammation of one or more bronchi
Pneumonia	Inflammation of the lungs with air spaces becoming filled with exudates
Pneumothorax	An accumulation of air in the pleural space
Gastrointestinal features of cystic fibrosis	
Cholelithiasis	The presence or formation of gallstones
Cirrhosis	Liver disease characterized by loss of normal liver tissue and fibrosis
Distal intestinal obstruction syndrome	Blockage of the bowel with feces, mucus, and undigested food
Gastroparesis	Paralysis of the stomach or delayed gastric emptying
Malabsorption	Impaired intestinal absorption of nutrients
Maldigestion	Impaired intestinal digestion of nutrients
Meconium ileus	Blockage of the bowel with meconium
Osteoporosis/ Osteopenia	Reduction in bone mass
Pancreatic insufficiency	Reduction of enzyme production from the pancreas
Portal hypertension	High pressure in the portahepatic artery
Rectal prolapse	Protrusion of the rectal mucous membrane through the anus
Splenomegaly	Enlargement of the spleen

to be of great significance. Chronic inflammation has been cited as the cause of so much of the lung damage seen in CF. Steroidal anti-inflammatory drugs have been shown to be beneficial but have nutritional side effects such as hyperglycemia and osteoporosis. Nonsteroidal anti-inflammatory drugs such as ibuprofen have been used in some centers with positive results but their long-term effect on renal function is not yet known. The impact of malnutrition on lung disease and respiratory muscle function has been extensively studied in patients with CF. Malnutrition and deterioration of lung function are interdependent. Prevention of malnutrition from the time of diagnosis is associated with better lung function and improved survival.

Gastrointestinal Complications

Individuals with CF can develop a variety of gastrointestinal (GI) disorders related to the pathophysiological changes associated with CF. Pancreatic insufficiency, which is present in the majority of CF patients leads to many of the GI manifestations of CF including steatorrhea, abdominal pain, distal intestinal obstruction syndrome (DIOS), and rectal prolapse. Gastroesophageal reflux (GOR) occurs frequently in CF due to decreased lower esophageal sphincter pressure and is usually treated by proton pump inhibitors. In patients with advanced lung disease vomiting is common after strenuous bouts of coughing and this over time may lead to decline in nutritional status. Peptic ulcer disease, pancreatitis, and intussusception also occur to varying degrees in patients with CF. Crohn's disease and celiac disease occur more frequently in the CF population than in controls and gastrointestinal tumors, although rare, have an increased incidence in CF.

Meconium ileus is the presenting complaint in up to 15% of infants with CF. This is a condition in which the small intestine is blocked with tenacious meconium and surgical intervention is required to correct it. Excessive mucus in the small bowel of patients with CF can provide a physical barrier to the absorptive surface. Undigested or unabsorbed food in association with this mucus, and possibly a reduced gut motility, can lead to a partial or complete obstruction of the GI tract in older children and adults known as meconium ileus equivalent, or more accurately distal intestinal obstruction syndrome (DIOS). This is a condition specific to CF. The usual clinical presentation is one of abdominal pain, abdominal distension, and constipation. It can be precipitated by dehydration, change in eating habits, change in enzyme brand or dose, or immobility. DIOS is treated with a laxative regime and should have a diet and enzyme review.

CF-Related Diabetes Mellitus (CFRD)

Diabetes requiring insulin is the most common comorbidity in CF. The islets of Langerhan are the last cells to be damaged in the process of fibrosis of the pancreas. The incidence of diabetes in CF has been reported to be 8–15% but this may be underestimated due to lack of screening. It is estimated that 50% of patients over 30 years will have some degree of glucose intolerance. The primary cause of CFRD is insulin deficiency secondary to pancreatic fibrosis. Diagnostic criteria for CFRD are the same as for non-CF-related diabetes. Glucose metabolism is also affected by many factors including infection, malabsorption, abnormal intestinal transit time, and steroid use, all features of CF. While CFRD shares many of the characteristics of both type 1 and type 2

diabetes, it is itself a distinct clinical condition. Hyperglycemia may adversely influence weight and pulmonary function and as the age of survival increases may lead to the development of microvascular complications. Retrospective studies have shown in those presenting with overt diabetes mellitus, deterioration in weight and respiratory status for 2 years before diagnosis are reversed once insulin therapy is instituted. A program of multiple daily insulin injections and self-monitoring of blood glucose with the aim of normoglycemia is the preferred treatment with regular follow-up with the Endocrinology team. All patients with CF should be screened annually for CFRD using the oral glucose tolerance test. Minimal dietary restrictions are imposed on this group of patients in an attempt to maximize nutritional intake. See section on dietary management of CF.

Liver Disease

Another complication associated with increased longevity in CF is liver disease, which affects between 2 and 37% of adults with CF. The development of liver disease in CF has been attributed to the blockage of small bile ductules with thick secretions, and the subsequent development of progressive cholestasis, biliary fibrosis, and eventually biliary cirrhosis and portal hypertension. The persisting acidic conditions in the upper small bowel lead to bile salt precipitation and defective lipid emulsification. Unhydrolyzed fat and other products of maldigestion may interfere with bile acid reabsorption in the terminal ileum, thereby reducing the total bile salt pool. Fecal losses of primary and secondary bile acids leads to an imbalance of bile salts, which further increases the viscosity of the already tenacious bile. Treatment with ursodeoxycholic acid has led to an improvement in bile excretion and liver function tests. Complications of liver disease including ascites, gastro and esophageal varices may further exacerbate a patient's nutritional status. In a small number of patients liver failure may require liver transplantation. See section on dietary management of CF.

Nutritional Management

Aggressive nutritional management of patients with CF is key in their overall management. Nutritional management of CF involves maximizing dietary intake, minimizing malabsorption and maldigestion, monitoring vitamin intakes and serum levels, and adapting eating patterns in the event of diabetes, osteoporosis, DIOS, or liver disease. Nutritional support in the form of nocturnal gastrostomy feeding may be necessary if nutritional failure persists ($BMI < 18.5\,kg/m^2$). It is well recognized that the malnutrition seen in CF is due to an energy imbalance caused by three main factors: decreased dietary intake, increased energy requirements, and increased energy losses. There appears to be a direct association between the degree of malnutrition and the severity of pulmonary disease, affecting overall prognosis. Many patients are capable of balancing these factors effectively and have a normal growth velocity and good nutritional status. However, as lung function deteriorates, energy requirement increases and appetite decreases leading to a loss of energy stores and lean tissue further contributing to progressive deterioration of lung function (see **Figure 2**).

Decreased Dietary Intake

People with CF are advised to consume a diet high in energy with no fat restriction. Prior to the development of enteric-coated enzymes in the mid 1980s, patients with CF were advised to follow a low-fat diet in an attempt to minimize fat malabsorption and steatorrhea. Unfortunately, older patients continue this practice as they have developed an aversion to fatty foods after many years of avoiding them. Decreased dietary intake secondary to anorexia is common in CF and can become more of a problem during recurrent chest infections. There have also been an increased number of reports of eating disorders and abnormal eating behavior in the CF population. In addition polypharmacy, repeated exacerbations of CF, organomegaly, gastrointestinal problems, food intolerance, and poor social circumstances can reduce oral intake.

Increased Energy Requirements

Energy requirements are increased during periods of infection by catabolism and fever and continue to increase with advanced pulmonary disease. It has been estimated that CF patients require 120–150% of the estimated average requirement for energy. As pulmonary function deteriorates, mobility also decreases and overall energy expenditure is reduced as a result. Owing to the heterogenicity of CF the energy requirements of individuals will vary and should be assessed on an individual basis. Energy losses through sputum may also be significant in a patient with a marginal energy intake. Salbutamol, often used as a bronchodilator in CF, can increase basal metabolic rate.

Increased Energy Losses

Pancreatic changes are caused by the obstruction of small ducts with thick secretions and cell debris.

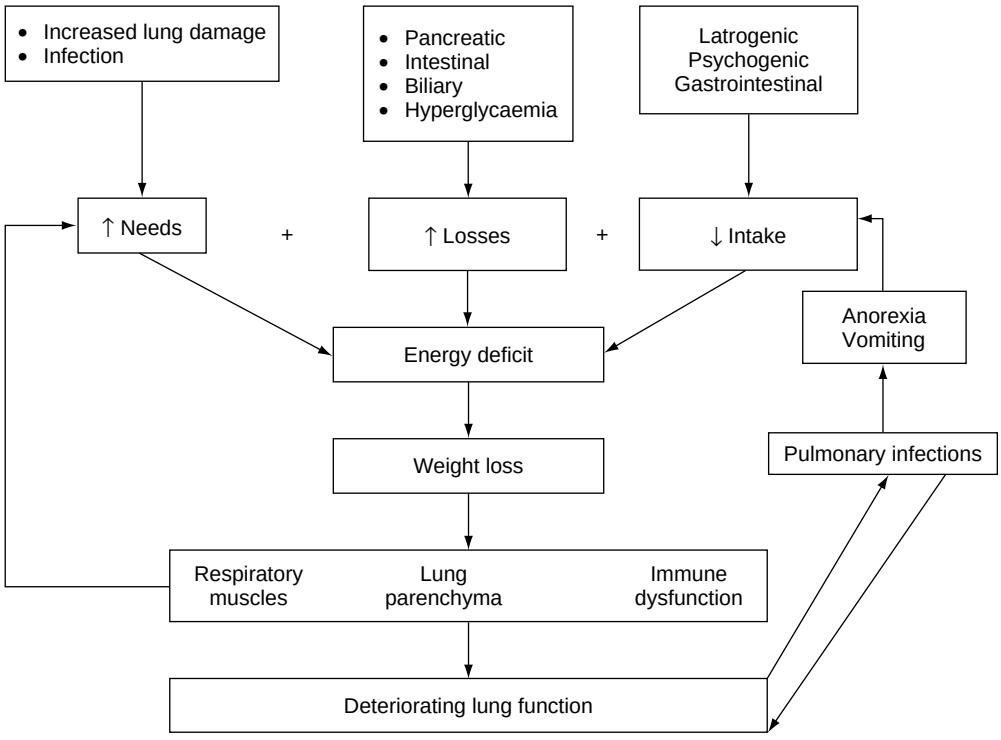


Figure 2 Interdependent factors that may give rise to progressive energy deficit as lung function deteriorates.

Functional tissue becomes replaced with fibrotic tissue leading to pancreatic exocrine insufficiency when more than 90% of the normal structure of the pancreas is lost. Pancreatic insufficiency is the most common gastrointestinal manifestation in CF, occurring in at least 95% of patients. The production of pancreatic secretions including enzymes and bicarbonate is reduced, necessitating pancreatic enzyme replacement therapy (PERT). PERT is supplied in the form of gelatin capsules containing microspheres, which are swallowed whole with food. The capsule dissolves within the stomach and releases the microspheres, which are protected from the gastric acid by an enteric coating. Enzymes should be taken immediately before or during a meal to maximize their efficacy. The microspheres mix with the stomach contents and pass through the pylorus into the duodenum where they become activated. Microspheres should be less than 1.5 mm in diameter to ensure they leave the stomach with food. Fibrosis of the pancreas tends to be a progressive process so increasing amounts of oral enzyme supplements are often required as patients get older. All people with CF have some level of pancreatic dysfunction but requirements of enzymes are variable and must be assessed individually. Clinically, the aim of PERT is to correct symptomatic steatorrhea, relieve any abdominal pain, reduce the mass and frequency of stool passed, and achieve weight gain within normal limits.

The enteric coating on enzyme supplements is designed to dissolve at a pH of 6, the optimal pH for pancreatic enzymatic action. Owing to the reduced production of bicarbonate and the resulting lower pH of the duodenum in patients with CF, the enteric coating of the enzyme may fail to dissolve so that the enzyme does not become activated at the absorptive surface of the small bowel. Increasing the duodenal pH by taking proton pump inhibitors may improve absorption. Changing the brand of enzyme may also improve absorption as dissolution characteristics of the enteric coating and proportions of enzymes contained within the microspheres vary. Patients should be dissuaded from chewing enzymes as this breaks the enteric coating and leads to deactivation in the acid medium of the stomach. Even with maximal PERT it has been estimated that between 10 and 20% of ingested fat will be malabsorbed. Colonic strictures known as fibrosing colonopathy (FC) in CF populations receiving high-potency enzymes with a more concentrated dose of lipase and protease per capsule have been reported The etiology of this FC remains unclear. Recently, it has been suggested that FC may be related to the presence of methacrylic acid copolymer (MAC) coating present in some preparations rather than

actual enzyme strength. Some adult patients continue to take high-dose enzymes and are advised to do so within recommended levels. The working group on PERT use recommends that no more than 10 000 units of lipase per kilogram body weight be taken per day.

Dietary Management of CF

Patients with CF are encouraged to consume a diet providing 150% of the recommended intake for age and sex. However, this is only a guideline, since in practice the energy requirement for a patient with CF is that which maintains their ideal body weight when malabsorption has been controlled. Maximizing energy intake from everyday foods should be the initial step in the promotion of a high-energy diet. As fat is the most concentrated source of energy in the diet, liberal use of fat should be encouraged; this can best be achieved by recommending frequent consumption of high-fat meals and snacks including confectionery, desserts, and cakes. PERT should be dosed accordingly.

Dietary Supplements

The energy intake of many patients with CF is commonly suboptimal. Many patients find it difficult to eat sufficient food daily to attain or maintain their ideal body weight. During a respiratory exacerbation of CF, energy requirements are at a maximum, but appetite is often reduced. Dietary supplements in the form of sip feeds can be a useful adjunct to a high-energy diet. Care should be taken to ensure that supplements are used in addition to a diet and not as a substitute for normal foods.

Enteral Feeding

When diet and oral dietary supplements are undesirable or ineffective and nutritional failure persists, i.e., BMI $<18.5\ kg/m^2$, enteral feeding should be considered. Research has demonstrated a sustained weight gain and a slowing decline in respiratory function associated with supplemental enteral feeding. Artificial nutritional support can be provided via nasogastric or gastrostomy tube depending on patient preference. Gastrostomy feeding is becoming more popular, whether passed endoscopically or under fluoroscopic guidance. The introduction of low-profile gastrostomy feeding tubes or 'button' tubes have made this method of nutritional support more acceptable of patients. The type of feed used and the PERT given with it, varies between centers. Feeds are usually administered overnight in an attempt to provide 30–50% of energy requirements and to allow for maximal oral intake during the day. Gastrostomy feeds can be used over longer periods during periods of acute pulmonary infection, loss of appetite, or in the severely malnourished patient. Patients with a previous poor intake should be monitored for refeeding syndrome.

Specific Dietary Considerations

There are some medical complications of CF that warrant particular nutritional attention.

Liver Disease

Patients with liver disease as a complication of their CF may have ascites, gastric, or esophageal varices, all of which may affect nutritional status and options for nutritional support. Dietary management of the patient with CF and liver disease centers on maximizing energy intake and is best achieved by encouraging small, frequent, energy-dense meals, snacks, and drinks. Suboptimal oral intake can arise in patients with hepatomegaly or splenomegaly, who often have a feeling of fullness after eating referred to as the 'small stomach syndrome.' The benefits of gastrostomy insertion should be carefully weighed in the patient with gastric varices or splenomegaly due to risk of bleeding. A moderate sodium restriction may alleviate ascites. If coagulation is impaired, supplementation with vitamin K may be indicated.

Treatment of liver disease in CF is with ursodeoxycholic acid, which has a positive effect on liver enzymes. Whether this improvement is associated with improvement in nutritional status is unknown.

Cystic Fibrosis-Related Diabetes (CFRD)

The dietary treatment of CF-related diabetes varies from standard diabetic dietary advice. The principle of the diet centers on maintaining caloric intake whilst ensuring glycemic control. The treatment of CFRD should enhance rather than impair a patient's nutritional status. This is done by encouraging a high-fat diet and confining the intake of refined carbohydrate to mealtimes. Insulin doses should be increased so as to maximize the flexibility of the diet, particularly in those patients who are already nutritionally compromised Patients taking oral nutritional supplements and/or overnight gastrostomy feeds need to have their insulin doses carefully monitored and adjusted accordingly.

Bone Disease in CF

Osteopenia and osteoporosis are now widely recognized in the CF population. There are a number of contributing factors to this early development of bone disease including steroid usage, malabsorption of calcium and, more importantly, vitamin D, poor nutritional status, decreased levels of physical activity, and a reduced peak bone mass in CF patients compared to healthy individuals. Assessment of bone health is by dual energy X-ray absorptiometry (DXA) scanning and there are a variety of treatment options available depending on the severity of disease ranging from dietary calcium and vitamin D supplementation to the use of bisphosphonate drugs, which aim to halt the progression of bone loss and promote bone formation.

Fertility Issues

As the number of people with CF of a reproductive age increases, so does the incidence of pregnancy in this group. Although almost all males with CF are infertile owing to the absence of the vas deferens, most females are fertile. Pregnancy in women with CF requires special nutritional attention with regular monitoring, particularly with respect to adequate weight gain, and vitamin and mineral status.

Body Composition Studies in CF

Studies of body composition in CF patients have shown deficits in total body mass, lean body mass, and body fat, which affect body density. As skinfold thickness percentiles are derived from body density, it has been suggested that the assessment of the body fat content of children with CF using, or derived from, body density such as skinfold thickness is invalid. Muscle function indices have been shown to respond to refeeding in malnourished patients with CF before body composition or biochemical indices of protein status improved, and so appear to be sensitive markers of nutritional status.

Assessment of Nutritional Status

Malnutrition in CF remains a major clinical problem. Growth and nutritional status should be monitored at each clinic visit to ensure early detection of any deterioration, and to prompt appropriate nutritional intervention. The many factors that complicate nutritional status in CF are shown in **Table 2**.

When weight falls to a BMI of less than 18.5 kg/m^2 nocturnal enteral feeding should be considered. At diagnosis and when the patient shows clinical deterioration the following should be determined:

Table 2 Factors affecting nutritional status

- Variation in gene mutation
- Frequency of pulmonary exacerbations
- Gastroesophageal reflux
- Distal intestinal obstruction syndrome
- Pancreatitis
- Liver disease
- Diabetes mellitus
- Drug therapy
- Dietary dislikes and misconceptions
- Psychological problems/eating disorders
- Pregnancy
- Transplantation

electrolytes, serum albumin and other liver function tests, oral glucose tolerance test, full blood count, serum retinol, and alpha tocopherol. If there is any evidence of iron deficiency, iron status should be assessed. Other medical disorders should be considered in the evaluation of nutritional failure. These include diabetes mellitus, liver disease, Crohn's disease, celiac disease, chronic abdominal pain, DIOS, and esophagitis.

Vitamin Status in CF

At least 85% of CF patients have some level of pancreatic insufficiency leading to a degree of fat malabsorption. For this reason, unless supplemented, most patients are at risk of developing either clinical or subclinical deficiencies of the fat-soluble vitamins, vitamin A, D, E, and K. Those most at risk appear to be individuals with poorly controlled malabsorption, poor adherence to treatment, liver disease, bowel resection, or following a late diagnosis.

Vitamin A

Vitamin A should be supplemented at a dose of 4000–10 000 IU per day. However, low serum levels of retinol have been noted even at this dose. If retinol levels are persistently low despite adequate supplementation, an assessment of compliance, retinol-binding protein (RBP), and zinc levels should be checked. Special care should be given to vitamin A supplementation during pregnancy as high levels are reported to be teratogenic.

It is important to consider hepatotoxicity with large supplemental doses of vitamin A in a patient who may store vitamin A in the liver, yet shows low serum levels of retinol, and who may display ocular signs of deficiency. The free alcohol retinol is almost entirely attached to RBP, which is synthesized in the liver. Decreased levels of RBP, which may occur in up to 25% of patients with CF, may be due to an

abnormality in its production by the liver, zinc deficiency, or protein energy malnutrition. Even with adequate vitamin supplementation and pancreatic enzyme replacement treatment, up to 20% of patients may have ocular signs of deficiency of retinol. Xerosis may improve by increasing the dose of vitamin A alone, or combined with zinc. It has been suggested that there may exist a specific defect in the handling of retinol in the GI tract of people with CF unrelated to the level of fat malabsorption. A correlation has been demonstrated between low levels of vitamin A and poor lung function.

Beta-Carotene

Beta-carotene is one of the carotinoids present in plasma and a precursor of vitamin A. It is effective as an antioxidant at lower oxygen saturation states than vitamin E. It has a biological role as a lipid-soluble chain-breaking antioxidant in biomembranes. Routine supplementation with beta-carotene could diminish lipid peroxidation and improve essential fatty acid status.

Vitamin D

Vitamin D deficiency may be caused by malabsorption, underexposure to sunlight or defects in metabolism due to liver disease. Even though skin exposure to sunlight is the major source of vitamin D, serum concentrations will vary between individuals depending on endogenous production in the skin. Rickets as a result of vitamin D deficiency is rare but has been described in CF. Osteopenia and retarded bone maturation have been reported in a number of CF patients, even with supplementation to recommended levels. Bone density has been shown to be significantly decreased in all sites compared with that of normal young adults. Other variables such as activity levels and nutritional status have not been adequately researched, although the incidence of osteoporosis was found to be higher in those patients with severe respiratory disease. To attain and maintain normal serum levels a daily dose of 400–2000 IU is generally required in adults.

Vitamin E

Cholestasis and a reduced enterohepatic circulation of bile acids contribute to the malabsorption of fat-soluble vitamins from the small intestine. Vitamin E is highly lipophilic and deficiency correlates with degree of fat malabsorption. Subclinical neuroelectrophysiological abnormalities are already present in about 40% of patients by 2 months of age. Neurological signs of vitamin E deficiency are responsive to supplementation if initiated early but are irreversible if treatment starts after the neurological lesions are present. As circulating alpha tocopherol is transported in the blood attached to lipid it is should be expressed as a ratio to total lipid to be correctly interpreted. Current recommendations are to monitor serum vitamin E levels annually and adjust supplementation accordingly. A daily dose of 400 IU per day should achieve normal serum levels in adults.

Vitamin K

A review of the literature provides conflicting opinions in the area of routine supplementation of vitamin K as the prevalence of vitamin K deficiency has not been established. Theoretically, the risk factors for patients developing vitamin K deficiency are pancreatic insufficiency, severe liver disease, extensive small bowel resection, and chronic broad-spectrum antibiotic use. Monitoring the coagulation system is advised, as vitamin K estimations are not generally routinely available. It seems prudent to prescribe vitamin K supplements to patients with the above risk factors. Vitamin K has recently been shown to play an important role in bone health. There are no specific guidelines on supplementation, but doses of 5–10 mg appear to be a prudent guide. Annual monitoring of fat-soluble vitamin levels should be carried out and doses of vitamins altered as appropriate.

Water-Soluble Vitamins

Supplementation with water-soluble vitamins is, in general, thought to be unnecessary in CF. In cases where dietary intake is poor or unbalanced, supplementation of vitamin C is advised. Supplementation with other water soluble vitamins is not routinely recommended.

Mineral Status in CF

Fat malabsorption can lead to the formation of insoluble fatty acid complexes with minerals in the gut, leading to a reduction in their absorption. CF may also be associated with intestinal mucosal defects, which may further retard the absorption of nutrients. Suboptimal levels of zinc, selenium, manganese, and iron have all been described in CF. Routine iron supplementation is not recommended as it has been suggested that *Pseudomonas aeruginosa* grows in tissues with a high concentration of iron. In addition, levels of iron may be suppressed as a normal body response in times of infection, and attempting to correct this is potentially harmful. Sodium and chloride do not need to be

supplemented unless in very hot climates or during excessive exercise.

The Oxidant/Antioxidant Imbalance in CF

Patients with CF frequently exhibit increased oxygen free radical generation from activated neutrophils due to chronic lung inflammation. This, coupled with antioxidant deficiencies due to exocrine pancreatic insufficiency, results in an oxidant/antioxidant imbalance. Consequently, free radical attack on unsaturated fatty acids of lipid structures occurs leading to lipid peroxidation. An efficient antioxidant supply is suggested to control tissue damage by restoring the oxidant/antioxidant balance.

Conclusions

There is a complex relationship between physiological, environmental, and genetic variables leading to a great variability in energy requirements among individuals with CF. Despite advances in the treatment of CF the need for good nutritional strategies in CF will continue. Individually tailored nutritional advice for each patient with CF by a dietitian experienced in the area of CF is essential.

See also: **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Eating Disorders**: Anorexia Nervosa; Bulimia Nervosa. **Liver Disorders**. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **Nutritional Assessment**: Anthropometry; Biochemical Indices; Clinical Examination. **Nutritional Support**: Adults, Enteral; Adults, Parenteral; Infants and Children, Parenteral. **Vitamin A**: Physiology; Biochemistry and Physiological Role; Deficiency and Interventions. **Vitamin D**: Physiology, Dietary Sources and Requirements; Rickets and Osteomalacia. **Vitamin K**.

Further Reading

Borowitz DS, Grand RJ, and Durie PR, and the Consensus Committee (1995) Use of pancreatic enzyme supplements for patients with cystic fibrosis in the context of fibrosing colonopathy. *Journal of Paediatrics* **127**: 681–684.

Dodge JA (1992) Nutrition in cystic fibrosis: a historical overview. *Proceedings of the Nutrition Society* **51**: 225–235.

McDonald A, Holden C, and Harris G (1991) Nutritional strategies in cystic fibrosis: current issues. *Journal of the Royal Society of Medicine* **84**(supplement 18): 28–35.

Moran A, Hardin D, Rodman D, and Allen HF, and the consensus committee (1999) Diagnosis screening and management of cystic fibrosis related diabetes mellitus: A consensus conference report. *Diabetes Research and Clinical Practice* **45**: 57–68.

Ramsey BW, Farrell PM, and Pencharz P, and the Consensus Committee (1992) Nutritional assessment and management in cystic fibrosis: a consensus report *American Journal of Clinical Nutrition* **55**: 108–116

Rosenstein *et al.* (1998) The diagnosis of cystic fibrosis: A consensus statement. *Journal of Pediatrics* **132**(4): 589–595.

Sinaasappel M, Stern M, Litttlewood J, Wolfe S, Steinkamp G, Harry, Heijerman HGM, Robberecht E, and Döring G (2002) Nutrition in patients with cystic fibrosis: a European Consensus. *Journal of Cystic Fibrosis* **1**: 51–75.

Warner J (ed.) (1992) Cystic fibrosis. *British Medical Bulletin* **48**(4): 717–978.

Zentler-Munro PI (1987) Cystic fibrosis: a gastroenterological cornucopia. *Gut* **28**: 1531–1547.

CYTOKINES

R F Grimble, University of Southampton, Southampton, UK

Chemistry and Classification

Cytokines comprise a wide range of proteins that are released mainly from cells of the immune system in response to invasion of animals by pathogens or severe injury. Cytokines induce a state of inflammation in the body and modulation in the activity of the immune system. Research shows that cytokine production is not restricted to cells in the immune system, but that fibroblasts, endothelial cells, adipocytes, and specialized tissues, such as the ovary, produce cytokines. Although largely influencing immune function, a number of cytokines act as growth factors and lead to the proliferation and differentiation of a wide range of cell populations in the body. Cytokines are proteins of low molecular weight. They act generally in an autocrine or paracrine fashion and are active in the subnanomolar range. Cytokines are subclassified as interleukins (ILs), tumor necrosis factors (TNFs), interferons, and colony-stimulating factors. Examples from the family of cytokines are detailed in **Table 1**. All influence cells of the immune system; however,

Table 1 Main properties of the pro-inflammatory cytokines

Cytokine	*Mol. wt*	*Cell sources*	*Main cell targets*	*Main actions*
Interleukin-1α Interleukin-1β	33 000 17 500	Monocytes, macrophages, astrocytes, epithelial cells, endothelium, fibroblasts, dendritic cells	Thymocytes, neutrophils, T and B cells, skeletal muscle, hepatocytes	Immunoregulation, inflammation, fever, anorexia, acute-phase protein synthesis, muscle proteolysis, enhanced gluconeogenesis
Interleukin-6	20 000	Macrophages, T cells, fibroblasts, some B cells	T and B cells, thymocytes, hepatocytes	Acute-phase protein synthesis, immune cell differentiation
Tumor necrosis factor-α	50 000 (trimer)	Macrophages, lymphocytes	Fibroblasts, endothelium, skeletal muscle hepatocytes	As for IL-1

only three exert metabolic effects upon the host. These are denoted as pro-inflammatory cytokines IL-1, IL-6, and TNF-α. A summary of the cell sources, main cell targets, and actions of the proinflammatory cytokines is shown in **Table 1**.

Metabolism and Metabolic Functions

Widespread metabolic changes occur as a result of cytokine production (**Figure 1**). These responses are powerful, focused, and dangerous to both host and pathogens. A hostile environment for pathogens is created within the body by the release of oxidant molecules (superoxide, hydrogen peroxide, perchlorous acid, and nitric oxide) from phagocytes.

Nutrients are provided for the immune system as a result of wasting of peripheral tissues. Amino acids released as a consequence of increased proteolysis in muscle, skin, and bone provide substrate for the synthesis of cells in the system. Glutamine, released from muscle, and glucose, derived from increased hepatic gluconeogenesis of amino acids, are major sources of nutrition for the immune system. Likewise, increased lipolysis in adipose tissue provides fatty acids as metabolic fuel for the body. Zinc, an important cofactor in DNA synthesis, is released from peripheral tissues, incorporated into the zinc transporting protein metallothionein in liver and kidney, and subsequently utilised by the immune system. A loss of appetite often occurs. This may be purposeful in permitting a situation in which substrate is more closely tailored to the requirements of the immune system than would occur from the vagaries of habitual dietary intake. This concept, however, is a matter of debate. Nonetheless, it is important that the immune system receives a guaranteed source of nutrition immediately after the body is infected or damaged because bacterial cells

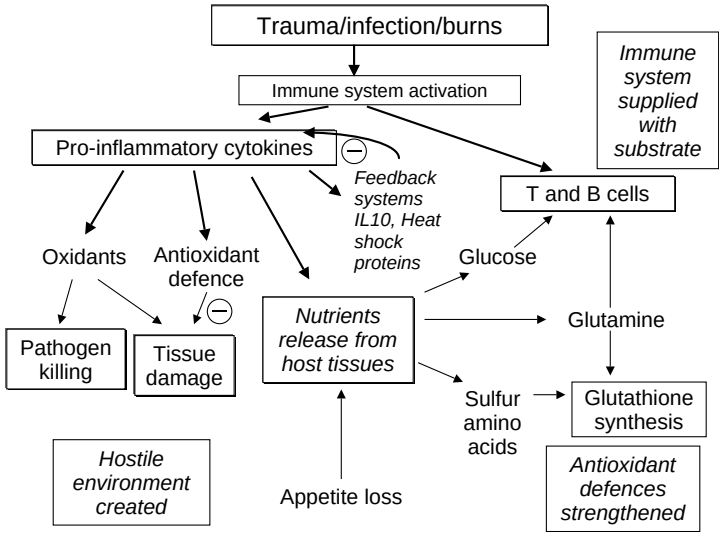


Figure 1 The coordinated actions of cytokines upon metabolism designed to defeat pathogens and protect the host. The resultant effects are shown in italics.

multiply at least 50 times more rapidly than T cells under favorable conditions. Under the actions of cytokines, the metabolic activity of the liver is greatly enhanced and modified. Large increases in the rates of gluconeogenesis, glycogen breakdown, and urea and fat synthesis occur. Blood glucose, urea, and triacylglycerol concentrations may rise. The increase in triacylglycerol levels may have functional importance due to the ability of these molecules to bind and neutralize endotoxin, thereby reducing the impact of this toxic bacterial product upon the host. Paradoxically, however, metabolism of xenobiotics is decreased due to a reduction in the activity of cytochrome P450. The profile of export proteins synthesized by the liver is changed, synthesis of albumin is reduced, and the synthesis of a group of proteins closely associated with inflammation (acute-phase proteins) is increased. Acute-phase proteins are multifunctional and include caeruloplasmin (an antioxidant and copper transport protein), C-reactive protein (to improve macrophage activity), fibrinogen (for blood clotting), complement proteins (for enhanced phagocytosis and pathogen destruction), and metallothionein (a zinc transport protein).

The antioxidant defences of the body are strengthened by increases in the activities of superoxide dismutase, catalase, glutathione peroxidase, and reductase and by increases in the hepatic synthesis of the reduced form of glutathione (GSH). The liver thus becomes the main focus for the synthesis of molecules for the nutrition, support, and direction of the immune system and for the protection of the body from the adverse effects of cytokine action. Indeed, when the ability of the liver of patients with sepsis (a severe clinical form of inflammation induced by infection) to extract amino acids from the circulation was assessed, it was found that the livers of patients who subsequently died had only half of that of livers of patients who survived.

A number of molecules synthesized in enhanced amounts when cytokines are produced are part of complex feedback systems that limit cytokine production and effects (**Figure 2**). These include GSH and some acute-phase proteins that suppress cytokine production and also cytokine receptor antagonist molecules for IL-1 and TNF. The first two types of molecule are derived from liver and the last two from lymphocytes and the cellular targets for TNF, respectively. Other molecules also moderate cytokine actions. Anti-inflammatory cytokines such as IL-10 and heat shock proteins exert an anti-inflammatory influence in the latter stages of the inflammatory response. This downregulation of inflammation, once the infectious agent has been defeated, is important for survival since the inflammatory process has a large capacity to deplete the body. The balance between the pro- and anti-inflammatory process is of key importance for survival since excessive production of IL-10 has been associated with increased mortality.

Role in Disease and Disease Processes

Despite the importance of cytokines in protecting the host from pathogens, the molecules may have damaging and even lethal effects on the host. Thus, the response of the host to a pathogen may play as significant a part in the demise of the host as the effects of the pathogen. Cytokines may also play a major role in tissue damage in chronic inflammatory disease, in which no infective agent is operating. Excessive or inappropriate cytokine production has been associated with increased morbidity and mortality in a wide range of diseases and conditions in which inflammation plays a role. These include

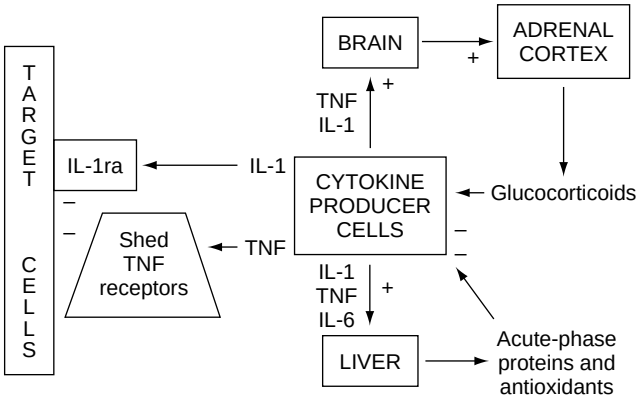


Figure 2 Innate systems for controlling the production and actions of pro-inflammatory cytokines. Stimulatory actions are indicated by plus signs and inhibitory actions by minus signs.

diseases where the immune system is clearly interacting with invading pathogens, such as malaria, meningitis, sepsis, and AIDS, and conditions such as asthma, inflammatory bowel disease, rheumatoid arthritis, and cancer, in which inflammatory disease develops without obvious involvement of pathogens. Furthermore, pro-inflammatory cytokines may be involved in the progression of disease processes such as plaque development in atherosclerosis and demylination in multiple sclerosis and Alzheimer's disease (**Figure 3**).

Damage may also be exerted on the host by release of free radicals and other oxidant molecules that are also released from phagocytic cells in response to the inflammatory stimulus and IL-1 and TNF. Furthermore, oxidant molecules upregulate production of IL-1, TNF, and IL-8 by activation of the transcription factor, nuclear factor-kappa B (NF-κB). The factor is normally held quiescent in the cytoplasm due to attachment of an inhibitory component (IκB) to it. In the presence of oxidants, IκB dissociates from NF-κB, migrates to the nucleus, and brings about the transcription of a large range of genes associated with the inflammatory process (**Figure 4**). Unfortunately, the human immunodeficiency virus (HIV) has an NF-κB response element on its genome. Thus, a by-product of the inflammatory response is increased replication of this virus and progression toward AIDS.

The fact that insulin resistance and disordering of lipid metabolism occur in obesity, diabetes mellitus, and during the inflammatory response has led to investigation of the possibility that obesity exerts an inflammatory influence on individuals. Large population studies show a strong association between indices of inflammation, abnormal lipid and carbohydrate metabolism, obesity, and atherosclerosis. This association is particularly strong in populations with a high incidence of obesity, diabetes, and cardiovascular disease (e.g., Pima Indians and Southeast Asians). TNF-α is produced not only by cells of the immune system but also by adipocytes and may provide the link between inflammation, insulin sensitivity, and the diseases associated therewith (**Figure 5**). TNF-α results in insulin insensitivity indirectly by stimulating stress hormone production and directly by down-regulating insulin receptor substrate-1 and by negative regulation of PPAR-γ, an important insulin-sensitising nuclear receptor. Adipose tissue produces both TNF-α and leptin. Production of the latter relates positively to adipose tissue mass and through its actions on immune function exerts a pro-inflammatory influence. It is unclear whether chronic inflammation is a trigger for chronic insulin insensitivity and conditions associated therewith or whether the reverse is the case. Evidence favours the former interpretation of the data.

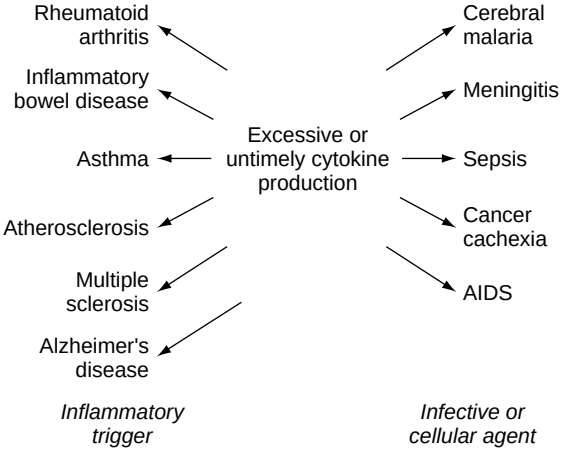


Figure 3 Diseases and conditions in which cytokines play a role.

Genetic factors play a role in the propensity of individuals to produce damaging or life-threatening amounts of cytokines during inflammation. Males and postmenopausal females possess a genetically determined propensity to produce high, medium, or low levels of cytokines in response to stimuli. Single base changes in the promoter regions of cytokine genes (single nucleotide polymorphisms (SNP)) result in these different levels of production. In the case of TNF-α, production of the cytokine is influenced by SNP in the TNF-α (*TNF2*) and TNF-β (*TNFB2*) genes. Individuals with the *TNF2* or *TNFB2* alleles produce higher amounts of TNF. In premenopausal women, the capacity to produce cytokines is influenced by the hormones of the oestrous cycle. Although the capacity for genetically determined levels of cytokines produces no apparent harm in healthy subjects, in disease, genetics has an impact on mortality. Studies in The Gambia showed that subjects who were homozygous for *TNF2* had a seven times higher rate of death or serious neural symptoms than subjects with one or no copies of the allele. Likewise, in patients with severe sepsis, possession of a *TNFB22* genotype resulted in 72% mortality in men compared with 42% mortality in men with a *TNFB11* genotype. Women were less severely affected by this genotypic influence. There is controversy about the reason for the retention of this lethal characteristic within the gene pool of the population. It is possible that in heterozygotes the presence of the genetic characteristic gives an immunological advantage. Homozygotes, who are less numerous than

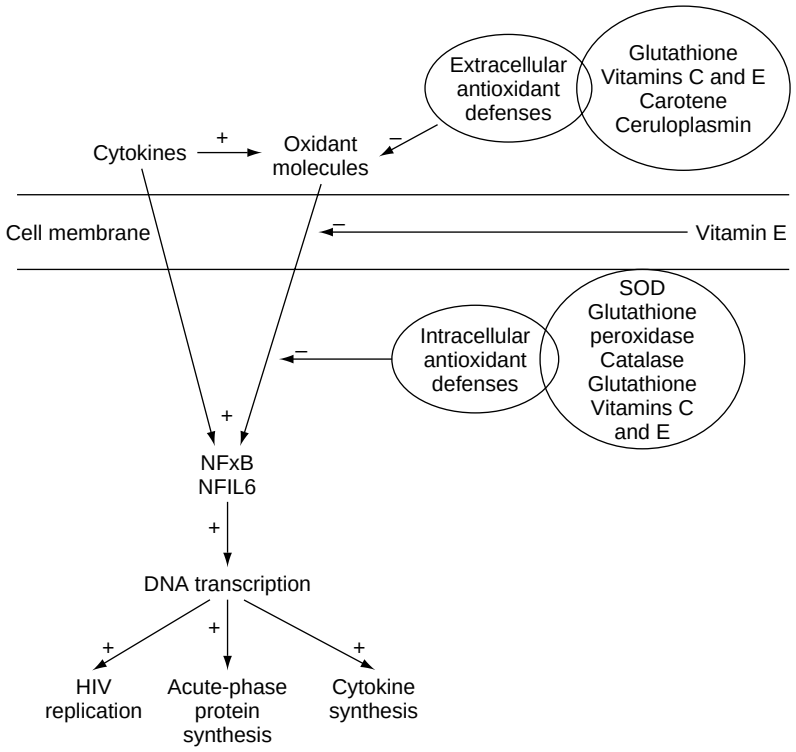


Figure 4 The interactions among oxidants, cytokines, and antioxidant defenses during inflammation. Minus signs indicate an inhibitory effect, and plus signs indicate a stimulatory effect. NF, nuclear factor; SOD, superoxide dismutase.

heterozygotes might pay the price for the advantageous retention of the genetic characteristic within the population. It is also interesting to note that in longevity studies a SNP that results in raised IL-10 production is more common in nonagenarians than in younger subjects, and conversely a SNP that results in raised production of IL-6 is rarer in the older than younger subjects. Thus, although it is an essential component of the ability of the body to combat pathogens, inflammation is inimical with longevity.

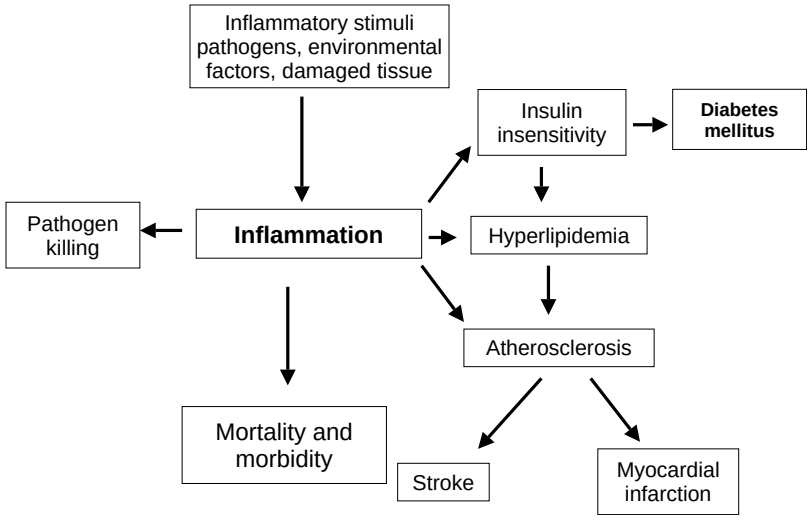


Figure 5 Linkage between inflammation as part of the defense against infection and as a factor in insulin insensitivity and disease processes.

Influence of Nutrients on Cytokine Biology

Proinflammatory cytokines exert widespread effects on metabolism, involving alterations in lipid, carbohydrate, and protein metabolism. In addition, there are substantial changes in micronutrient metabolism. A number of intracellular signaling pathways are activated by the actions of cytokines on target cells, including prostaglandins and leukotrienes, cyclic AMP, and protein kinase C. There are thus many levels at which nutrient intake can modify the intensity and characteristics of the response to inflammatory stimuli. The ability of nutrients to modify inflammation has been used in the treatment of diseases with an inflammatory basis. The interaction between nutritional status and inflammation is also important in public health because it determines the effects of infection on growth and well-being of populations with a poor nutrient intake.

The earliest indications that nutritional status could affect cytokine biology came from studies on malnourished hospital patients. White blood cells from patients had a reduced capacity to produce cytokines. The high mortality rates in these patients highlighted the importance of cytokines in the process of recovery from injury and infection. Protein supplements improved cytokine production and decreased the mortality rate. Since these observations were made, a large number of studies have been conducted in animals and human volunteers that show that fats, amino acids, and micronutrients change the ability of mammals to produce and respond to IL-1, IL-6, and TNF (**Figure 6**). **Figure 6** indicates whether a change in the intake of a nutrient, or nutrient status, alters cytokine production or the response of target tissues to the actions of cytokines.

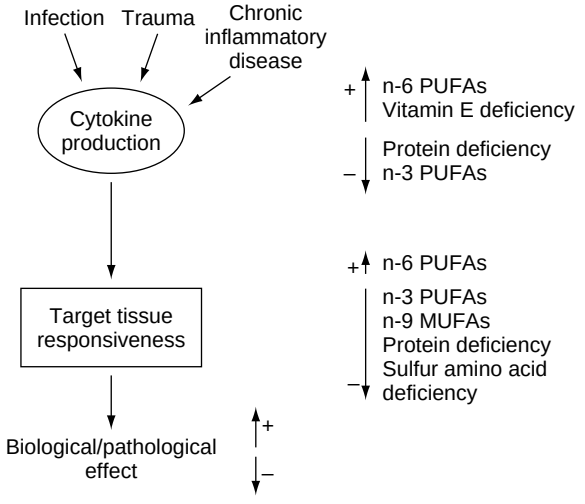


Figure 6 Summary of the effects of nutrients and nutritional conditions on cytokine biology. A stimulatory effect is indicated by a plus sign and an inhibitory influence by a minus sign.

Influence of Fats on Cytokine Production and Effects

Dietary fats can be divided into four main types. Some are rich in n-6 polyunsaturated fatty acids (PUFAs); fats in this group include corn, sunflower, and safflower oils. Some are rich in n-3 PUFAs; these include fats from marine sources. Some are rich in monounsaturated fatty acids; these include olive oil and butter. Some fats are characterized by a high content of saturated fatty acids, usually accompanied by low concentrations of PUFAs; coconut oil, butter, suet, and lard are in this category.

The production and actions of pro-inflammatory cytokines are profoundly influenced by dietary fat intake. There are a number of levels at which fats may modify cytokine biology. Most relate to the ability of fats to change the fatty acid composition of membrane phospholipids. Subsequently, membrane fluidity may be changed, the types and amounts of prostaglandins and leukotrienes produced during inflammation may be altered, and the synthesis of a number of cellular mediators that arise from phospholipids (platelet activating factor, diacylglycerol, and ceramide) may also be changed. As a result of these changes, the binding of cytokines to target tissues and the intensity of the inflammatory response may be altered.

Phospholipids contain two fatty acid chains attached to the remainder of the molecule at positions designated sn1 and sn2. Normally, arachidonic acid (AA C20:4 n-60) is released from this position and provides the parent compound for prostaglandins and leukotrienes. However, the long-chain PUFA eicosapentaenoic acid (EPA C20:5 n-3) may compete with AA for insertion at sn2. Prostaglandins and leukotrienes with a much lower bioactivity may result. This biological effect may account in part for the anti-inflammatory effects of fish oil. Many animal studies indicate that fats rich in n-6 PUFAs exert a pro-inflammatory influence, whereas fats rich in monounsaturated fatty acids or n-3 PUFA have the opposite influence. In human studies, however, evidence for the influence of n-6 PUFA or monounsaturated fatty acids is not so clear-cut. It has been postulated that the major increase in inflammatory disease that has occurred in the past 40 years in industrialized countries is due to a major

increase in the intake of n-6 PUFAs during this time (from approximately 5 to 7% of dietary energy). It has also been postulated that the lower levels of inflammatory disease associated with the habitual consumption of a 'Mediterranean diet' are due in part to high intakes of monounsaturated fatty acids. The evidence for n-3 PUFAs producing an anti-inflammatory effect in humans is much stronger, however. Also, n-3 PUFAs have been shown to produce beneficial effects in inflammatory disease. In many double-blind, randomised controlled clinical trials, fish oil produced significant clinical benefit in patients with rheumatoid arthritis. A number of trials also report beneficial effects of fish oil in the treatment of Crohn's disease. The precise mechanisms for these effects is unclear. A number of studies have demonstrated the ability of fish oil to reduce pro-inflammatory cytokine production and to alter the production of eicosanoids. However, recent studies have indicated a genomic influence on the ability of fish oil to reduce TNF production, thus indicating that fish oil may not be universally effective as an anti-inflammatory agent. A fish oil and vitamin E intervention trial (GISSI) was carried out on 11,324 survivors of a myocardial infarct in Italy. Patients were given 1 g of n-3 PUFA and/or 300 mg vitamin E/d. In the GISSI trial, fish oil supplements were shown to reduce the chance of stroke or a second myocardial infarct by 15%. Because inflammation plays a role in atherosclerosis, it is interesting to note that a trial of fish oil in patients with severe atherosclerosis showed that a supplement of 6 g/day of fish oil for 7 weeks significantly reduced macrophage activity in plaques.

Modulation of Cytokine Biology by Amino Acid and Protein Intake

Substantial increases occur in protein synthesis as the result of infection. It has been estimated that approximately 45 g of protein is required to produce and maintain the increased quantities of white blood cells and acute phase proteins in an infected individual. This demand will have a considerable impact on the availability of amino acids for other processes in the body that involve protein synthesis. The inhibitory effect of infection on growth, pregnancy, and lactation is well recognized. Output of amino acids from skeletal muscle, skin, and bone provides substrate for the synthesis of cells and proteins associated with the response to infection and trauma, as indicated previously. However, the supply may not always match demand, as is evident from the decrease in plasma concentrations of a number of amino acids. In particular, reductions occur in the concentrations of a metabolically related group of amino acids, including glycine, serine, and taurine. All three are metabolically related with the sulfur amino acids. Glycine and serine, together with the sulfur amino acids, are found in high concentrations in many compounds associated with the immune and inflammatory response, most notably comprising 66% of glutathione, 56% of metallothionein, and up to 25% of many acute-phase proteins. Experimental studies have shown that the production of cytokines, acute-phase proteins, and glutathione is influenced by the adequacy of both protein and sulfur amino acid intake. The partitioning of cysteine into glutathione and proteins in the liver may change if dietary sulfur amino acid intake becomes inadequate. This phenomenon is due to the biochemical properties of rate-limiting enzymes in both pathways. Whereas the K_m for γ-glutamyl cysteine synthetase (rate limiting for GSH synthesis) is 0.35 mM, that for amino acid activating enzymes (rate limiting for protein synthesis) is only 0.003 mM. This biochemical characteristic means that the GSH synthesis will fall below maximal rates at much higher intracellular cysteine concentrations than protein synthesis. Thus, at low sulfur amino acid intakes antioxidant defenses will become compromised. Low concentrations of GSH in tissues may have implications for the extent of inflammatory processes in the individual. In animal studies, decreased lung GSH concentrations are associated with the accumulation of inflammatory cells in tissues. In studies on HIV patients given *N*-acetyl cysteine, to improve GSH status, a decrease in plasma IL-6 concentrations has been noted indicating a reduction in inflammation. In view of the effects of NF-κB activation on HIV replication, it is interesting to note that the drug also brought about a reduction in HIV mRNA levels.

Modulation of Cytokine Biology by Micronutrients

Micronutrients play varied and complex roles in the response to infection and trauma. They are incorporated into substances that are synthesized in increased amounts during the response and into components of antioxidant defence, and they also modulate immune function. Trace elements are present in several acute-phase proteins and enzymes associated with antioxidant defense (**Figure 4**). These proteins include metallothionein (Zn), caeruloplasmin (Cu), superoxide dismutases (Mn, Cu, and Zn), and glutathione peroxidase (Se). Deficiencies

in copper impair the ability of rats to increase superoxide dismutase and caeruloplasmin activities in response to inflammatory agents. Deficiencies in zinc impair the ability to increase metallothionein synthesis; furthermore, zinc deficiency has potent suppressive effects on lymphocyte proliferation. Iron status may influence inflammation and immune function in a number of ways. Normally, iron is tightly bound to transport proteins such as transferrin and ferritin. However, following tissue damage and infections such as malaria, which may destroy red blood cells, free iron may be released and exert a proinflammatory effect by catalyzing free radical production. The latter effect may activate NF-κB and upregulate cytokine production. Indeed, iron dextran infusion has been shown to exacerbate inflammatory symptoms in rheumatoid arthritis. Desferrioximine, an iron chelator, suppresses TNF and IL-1 production by rodent macrophages. Iron deficiency also decreases the ability of such cells to produce cytokines. Impairment of immunological defence is commonly found in iron-deficient animals and human populations. Defects occur in T cell proliferation and in the ability of macrophages to engulf and kill bacteria. The latter may relate to the role of iron as part of the NADPH oxidase complex that is responsible for the respiratory burst and generation of hydroxyl radicals that kill bacteria. Myeloperoxidase activity generates hypochlorous acid for bacterial killing, and myeloperoxidase is also a hemoprotein whose activity is decreased by iron deficiency.

Vitamins also exert a number of effects on cytokine biology. These effects may relate to the roles that some of these nutrients play as antioxidants and growth factors (**Figure 4**). Rats deficient in vitamin E exhibit an enhanced inflammatory response to endotoxin; addition of the vitamin to the diet will suppress this effect. In healthy subjects and smokers, a daily dose of 600 IU of vitamin E for 4 weeks reduces the ability of white blood cells to produce TNF and IL-1. Cigarette smoking enhances cytokine production and raises acute-phase protein concentrations. The extent of the elevation is inversely related to vitamin E. Strenuous exercise results in a small increase in plasma concentrations of IL-1 and IL-6; vitamin E supplementation will prevent this effect.

Vitamin A status also influences cytokine production, although the mechanism underlying the effect is unclear. Macrophages taken from Indian children who received a supplement of 100 000 IU of retinol produced seven times the quantity of IL-1 produced by cells of children who had not received supplementation. The effect may be more pharmacological than nutritional in nature. Mice given vitamin A at a dose that was 16 times their requirement had macrophages that produced twice as much IL-1 upon stimulation than cells from unsupplemented animals.

Hormone-like properties have been attributed to vitamin D in relation to its effects on calcium. It is apparent that endocrine effects of the vitamin extend to immune function. Macrophages treated with 1,25-dihydroxyvitamin D_3 produce increased amounts of TNF and were more effective at killing *Mycobacterium avium* than untreated cells.

Vitamin B_6 supplementation has been found to increase lymphocyte proliferation and production of IL-2 in elderly subjects. The effect of the vitamin on pro-inflammatory cytokine production is unknown. Little is known about the effects of other water-soluble vitamins on cytokine biology. Although no effects of vitamin C status on pro-inflammatory cytokine production have been reported, doses of the vitamin reduce the incidence of respiratory infections in long-distance marathon runners.

Conclusions

The objective of the response of the body to infection and trauma is to disadvantage and destroy invading organisms while simultaneously protecting healthy tissues from the damaging influence of compounds produced during the response. Cytokines play a central role in the protection of the animal from damage during the response. The close interrelationship between pro-inflammatory cytokines, oxidant molecules, and antioxidant defenses gives a biological advantage to the host (**Figure 7**).

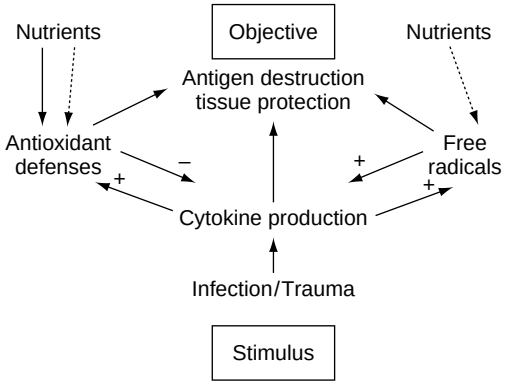


Figure 7 Influence of nutrients on the coordinated inflammatory events for destroying pathogens and protecting the host. Direct and indirect effects of nutrients are shown as solid and broken lines, respectively.

The essence of survival of an individual or species lies in the ability to prioritize physiological processes, particularly those processes that exert a large metabolic demand. Thus, at various times throughout the life cycle mammals will focus metabolic processes on achieving growth, the construction of placenta and fetus, the synthesis of milk components, or the repulsion of invasion by pathogens. For the infected individual, the marshalling of resources to combat the invading pathogen must assume a priority over all other physiological events. These other physiological processes can continue once the invasion has been repulsed and the damage done by the invader has been repaired.

The production of cytokines and other molecules associated with the inflammatory process carries risks of damage to the host as well as a survival advantage. The risk to the host is minimized by a sophisticated range of feedback control systems and synthesis of substances that protect the host. As discussed previously, nutrient intake modulates cytokine biology and the control and protective systems. A wide range of nutrients modulate cytokine biology at the level of production and sensitivity of target tissues (**Figure 6**). As a consequence of the modulation, the extent of depletion of nutrient stores and the risk of damage during the inflammatory response will be changed. The extent of tissue depletion and risk to the host will thus range from mild and transient in nature to severe, chronic, or lethal in effect.

See also: **Amino Acids**: Chemistry and Classification; Metabolism. **Diabetes Mellitus**: Classification and Chemical Pathology. **Fatty Acids**: Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated. **Fish**. **Obesity**: Definition, Etiology and Assessment. **Vitamin A**: Biochemistry and Physiological Role. **Vitamin E**: Metabolism and Requirements. **Zinc**: Physiology.

Further Reading

Beutler B and Cerami A (1986) Tumor necrosis factor as two sides of the same biological coin. *Nature* **320**: 584–588.

Dinarello CA (1988) Biology of IL1. *FASEB Journal* **1**: 108–115.

Douglas RG and Shaw JHF (1989) Metabolic response to sepsis and trauma. *Bristish Journal of Trauma* **76**: 115–122.

GISSI-Prevenzione Investigators (1999) Dietary supplementation with n-3 polyunsaturated fatty acids and vitamin E after myocardial infarction: Results of the GISSI-Prevenzione trial. *Lancet* **354**: 447–455.

Grimble RF (1994) Nutritional antioxidants and the modulation of inflammation: The theory and the practice. *Critical Care Medicine* **2**: 175–185.

Grimble RF (1996) The interaction between nutrients, pro-inflammatory cytokines and inflammation. *Clinical Science* **91**: 121–130.

Grimble RF (2002) Inflammatory status and insulin resistance. *Current Opinion in Clinical Nutrition and Metabolic Care* **5**: 551–559.

Grimble RF (2003) Inflammatory response in the elderly. *Current Opinion in Clinical Nutrition and Metabolic Care* **6**: 21–29.

Grimble RF, Howell WM, O'Reilly G *et al.* (2002) The ability of fish oil to suppress tumor necrosis factor-alpha production by peripheral blood mononuclear cells in healthy men is associated with polymorphisms in genes which influence TNF-alpha production. *American Journal of Clinical Nutrition* **76**: 454–459.

Grunfeld C and Feingold KR (1992) Tumour necrosis factor, interleukin 1 and interferon induce changes in lipid metabolism as part of host defence. *Proceedings of the Society of Experimental Biology and Medicine* **200**: 214–227.

Heinrich PC, Castell JV, and Andus T (1990) Interleukin 6 and the acute phase response. *Biochemical Journal* **265**: 621–636.

Murray MJ and Murray AB (1980) Cachexia: A 'last ditch' mechanism of host defence. *Journal of the Royal College of Physicians (London)* **14**: 197–199.

Newsholme P and Newsholme EA (1989) Rates of utilisation of glucose, glutamine and oleate and end product production by mouse macrophages in culture. *Biochemical Journal* **261**: 211–218.

Paolini-Giacobino A, Grimble R, and Pichard C (2003) Genomic interactions with disease and nutrition. *Clinical Nutrition* **22**: 507–514.

Schreck R, Rieber P, and Baeurerle PA (1991) Reactive oxygen intermediates as apparently widely used messengers of NFκB transcription factor and HIV1. *EMBO Journal* **10**: 2247–2256.

Thies F, Garry JMC, Yaqoob P *et al.* (2003) Association of n-3 polyunsaturated fatty acids with stability of atherosclerotic plaques: A randomized trial. *Lancet* **361**: 477–485.

D

DAIRY PRODUCTS

J Buttriss, British Nutrition Foundation, London, UK

Introduction

Dairy products are traditional dietary items in many parts of the world, in particular regions such as northern Europe where the cooler climate is especially suited to dairying. The history of milk as a food has been documented over the centuries and examples of early dairying are depicted in Egyptian friezes such as that from the sarcophagus of Queen Kawit from Der-al-Bahri, between Luxor and Karnak, dating back 4000 years. There is an even earlier Mesopotamian frieze from the temple of Ninkhasarg, near Ur, which is thought to be 1000 years older.

The popularity of milk as a staple food over the centuries must partly be due to its versatility. Early humans discovered that milk could be churned to make butter and fermented with bacterial cultures to produce cheese and yogurt, all of which were methods of preserving some or all of the nutrients in milk for consumption at a later date.

Variety

The range of dairy products on the market is immense. In most countries, a range of milks with differing fat contents is available. For example, in the UK consumers can choose between Channel Islands milk, with 4.9 g per 100 g fat, whole milk (3.9 g per 100 g), semiskimmed (1.6 g per 100 g fat), and skimmed milk, which has virtually no fat. Similarly, a wide range of cheeses exists with varying fat contents: at one end of the spectrum are soft fresh cheeses, made with skimmed milk, and at the other, hard cheeses such as Cheddar. Also available is cheese made with nonanimal rennet, suitable for vegetarians. In the UK alone, about two hundred different cheeses are produced, and cheese is particularly popular in countries such as France, where an even greater variety is available.

Fermented milk products such as yogurt, smetana, and kefir have always been popular in Middle Eastern countries, but their popularity, particularly that of yogurt, is increasing dramatically in Europe. Again, a wide range of yogurts exists, from very low-fat varieties to the creamier, whole milk or Greek-style product. Today, the range includes set yogurts, stirred yogurts, fruit yogurts, frozen yogurts, drinking yogurts, fromage frais, and the newer 'bio' yogurts with their milder flavor.

Traditional products such as cream and butter are still in demand, in spite of their high fat content, and are being joined by other 'luxury' products such as real dairy ice creams, fresh cream desserts, and luxury mousses. To meet the demand for a spread with a buttery flavor, products have been developed which incorporate butterfat for taste but often have a lower fat and energy content than butter.

A number of products also exist to which nutrients have been added, such as calcium-enriched milks and yogurts fortified with additional vitamins.

Nutrient Composition of Milk and its Products

Milk can be described as one of the most nutritionally complete foods. It provides a wide range of essential nutrients, in particular protein, and a range of vitamins and minerals (**Table 1**). It is, however, a poor source of iron and vitamin D, and contains no starch or dietary fiber. By volume, water is the major constituent of milk, comprising just over 87%. The remainder consists of milk fat and solids-not-fat (SNF)–principally comprising protein, lactose and minerals.

Protein

The principal proteins found in milk and its products are casein, lactalbumin, and lactoglobulin. Milk protein has a high biological value since it contains all of the eight essential amino acids, which cannot be synthesized in the body and so need to be provided by diet. In addition, milk can

Table 1 Nutrient composition per 100 g of pasteurized milk in the UK

	Whole milk	*Skimmed milk*	*Semiskimmed milk*	*Channel Islands milk*
Energy (kcal)	66	33	46	78
Energy (kJ)	275	140	195	327
Protein (g)	3.2	3.3	3.3	3.6
Carbohydrate (g)	4.6	4.8	4.8	4.6
Sugars (g)	4.6	4.8	4.8	4.6
Fat (g)	3.9	0.1	1.6	5.1
Saturates (g)	2.4	0.06	1.0	3.3
Monounsaturates (g)	1.1	Trace	0.5	1.3
Polyunsaturates (g)	0.1	Trace	Trace	0.1
Sodium (mg)	55	55	55	54
Dietary fiber (g)	Nil	Nil	Nil	Nil
Vitamin A (μg)	56	1	23	58
Thiamin (mg)	0.04	0.04	0.04	0.04
Riboflavin (mg)	0.17	0.18	0.18	0.19
Niacin (mg)	0.83	0.87	0.87	0.92
Vitamin B_6 (mg)	0.06	0.06	0.06	0.06
Folic acid (μg)	6	6	6	6
Vitamin B_{12} (μg)	0.4	0.4	0.4	0.4
Pantothenic acid (mg)	0.35	0.32	0.32	0.36
Biotin (μg)	1.9	2.0	2.0	1.9
Vitamin C (mg)	1	1	1	1
Vitamin D (μg)	0.03	Trace	0.01	0.03
Vitamin E (mg)	0.09	Trace	0.03	0.11
Calcium (mg)	115	120	118	130
Chloride (mg)	100	100	100	100
Copper (mg)	Trace	Trace	Trace	Trace
Iodine (μg)	15	(15)	(15)	N
Iron (mg)	0.05	0.05	0.05	0.05
Magnesium (mg)	11	12	11	12
Phosphorus (mg)	92	95	95	100
Potassium (mg)	140	150	150	140
Selenium (μg)	1	(1)	(1)	(1)
Zinc (mg)	0.4	0.4	0.4	0.4

N, no reliable information available; values in parentheses, estimated value.
Adapted from Holland *et al.* (1989) *Milk Products and Eggs*. Fourth supplement of *McCance and Widdowson's The Composition of Foods*, 4th edn. London: Royal Society of Chemistry/MAFF.

improve the overall protein quality of a meal when consumed with foods of lower protein quality such as cereals and pulses.

Carbohydrate

The carbohydrate in milk is in the form of lactose, a disaccharide comprising a molecule of glucose and a molecule of galactose. This sugar is found naturally only in milk and is much less sweet than sucrose (packet sugar).

In the small intestine, lactose is digested by the enzyme lactase to its two component monosaccharides, in readiness for absorption. This enzyme is present in babies, young children, and most European adults. However, in some adults enzymatic activity can decrease, making milk in quantity less well tolerated. Such individuals are described as being lactose intolerant. Most can tolerate small quantities of milk, and fermented milk products appear to be better tolerated. Cheese is also usually well tolerated as it contains only trace amounts of lactose. In the UK, lactose intolerance is relatively rare in people of European descent but is more common in those of Asian, Far Eastern, and African descent, particularly first-generation members of these ethnic groups.

Fat

The fat in milk is in the form of minute droplets, which rise to the top when milk is left to stand. The principal component of milk fat is triacylglycerol, three fatty acids joined to a glycerol backbone. All triacylglycerols contain mixtures of three families of fatty acids: saturated, monounsaturated, and polyunsaturated. The contribution of these three types of fatty acid to milk fat in the UK is 61% saturated, 28% monounsaturated, and 3% polyunsaturated. The percentages do not add up to 100% because milk fat is not composed totally of fatty

acids. When milk is skimmed, it is the amount rather than the type of fat that changes, and so the fatty acid profile remains the same. Milk fat contains small amounts of the two essential fatty acids, linoleic acid (1.4 g per 100 g fatty acids) and linolenic acid (1.5 g per 100 g fatty acids).

Vitamins

All of the known vitamins are to be found in whole milk (**Table 1**), although some are present in small quantities. Milk and milk products make a significant contribution to intakes of a number of these (**Table 2**). The fat-soluble vitamins – A, D, E and K – are removed with the fat when milk is skimmed. Consequently, they are present in only trace amounts in skimmed milk and in reduced amounts in semiskimmed milk. Whole milk is a good source of vitamin A, a pint (568 ml) providing 47% of the adult male and 55% of the adult female UK reference nutrient intake (RNI).

All of the three major types of cows' milk (whole, semiskimmed, and skimmed) are good sources of riboflavin (vitamin B_2) and vitamin B_{12}. Milk can also make an important contribution to intakes of thiamin (vitamin B_1), niacin, and ascorbic acid (vitamin C), particularly where the overall diet is poor.

Some vitamins are sensitive to heat and light. Over 90% of milk for liquid consumption is pasteurized, a mild heat treatment that causes little loss of vitamins other than vitamin C. In general vitamin losses are less than 10%. Milk that has undergone the ultra heat treatment (UHT) process keeps for longer, but the higher temperature used in the processing results in slightly greater losses (10–20%) of some vitamins, particularly vitamins B_6 and B_{12}, vitamin C, and folate. The sterilization process used for milk has a somewhat greater effect: about one-third of the thiamin and half the vitamin C, folate, and B_{12} are destroyed.

Some loss of vitamins is inevitable when milk is stored. The extent of these losses is dependent on the translucency and permeability to light of the container, and the length and conditions of exposure. Milk exposed to bright sunlight on the doorstep will readily lose its vitamin C. Loss of riboflavin is slower but after 4 h, half will be lost and only a third remains after 6 h. Therefore, measures should be taken to limit such exposure. There will also be gradual losses of some vitamins from UHT and sterilized milks, even under ideal storage conditions, because of reactions with small amounts of oxygen remaining in the pack or bottle. Boiling milk also reduces its vitamin content, ranging from a 5% reduction in vitamin B_{12} to a 50% reduction in vitamin C.

Table 2 Contribution of milk and milk products to daily nutrient intake from all food and drink in Great Britain

	Liquid milk		*Cheese*		*All milk and milk products*	
	Average amount provided	*Percentage of total intake*	*Average amount provided*	*Percentage of total intake*	*Average amount provided*	*Percentage of total intake*
Energy (kcal)	152	8.1	56	3.0	247	13.1
Protein (g)	9.3	14.8	3.5	5.6	14.1	22.4
Fat (g)	7.2	8.8	4.7	5.8	13.7	16.8
Saturates (g)	4.3	13.5	2.9	9.1	8.5	26.6
Monounsaturates (g)	2.1	7.0	1.3	4.3	3.9	13.0
Polyunsaturates (g)	0.2	1.4	0.2	1.4	0.5	3.6
Calcium (mg)	330	39.8	93	11.2	471	56.8
Iron (mg)	0.2	2.0	<0.1	0.4	0.3	3.0
Zinc (mg)	1.2	15.4	0.3	3.8	1.7	21.8
Magnesium (mg)	34	14.8	4.0	1.7	40.0	17.5
Potassium (g)	0.43	16.9	0.01	0.4	0.49	19.2
Thiamin (mg)	0.11	8.5	0.01	0.8	0.13	10.1
Riboflavin (mg)	0.5	31.1	0.04	3.9	0.63	39.1
Niacin (mg)	2.5	9.9	0.8	3.2	3.6	14.2
Vitamin B_6 (mg)	0.18	9.6	0.01	0.53	0.21	11.2
Vitamin B_{12} (μg)	1.1	22.9	0.2	4.2	1.4	29.2
Folate (μg)	16	6.6	5	2.1	25	10.4
Vitamin C (mg)	1.8	3.0	–	–	3.0	5.0
Vitamin A (μg)	96	9.3	50	4.8	174	16.9
Vitamin D (μg)	0.05	1.9	0.04	1.5	0.23	8.7

Adapted from Ministry of Agriculture, Fisheries and Food (1995) *National Food Survey 1994*. London: HMSO.

Minerals

Milk makes a contribution to human needs for virtually all the minerals and trace elements known to be essential for health. These are often present in a form that is well absorbed and utilized by the body (high bioavailability), e.g., calcium and zinc. For most people in the Western world, milk and milk products are a major source of calcium. Almost 60% of the calcium in the typical British diet is contributed by milk and milk products (**Table 2**). Milk alone contributes about 40% of the total. Although the contribution to zinc requirements made by milk is relatively low compared with meat (the major contributor), the zinc in milk is in a highly bioavailable form and there is evidence that the combination of milk (or meat) with vegetable foods, in which the zinc is less bioavailable, can enhance the bioavailability of zinc from the whole meal.

Cultural Significance of Milk and Milk Products

In countries where dairying has traditionally been a strong industry, milk and milk product consumption tends to be widespread and makes a significant contribution to nutrient needs. **Table 2** shows the contribution in Great Britain. The average daily intake of milk in the UK is a little under half a pint per person (271 ml). **Table 3** shows the contribution half a pint of whole milk makes to the nutrient and energy needs of a 4-year-old girl and a man. Skimmed or semiskimmed milk would make smaller contributions to intakes of energy and fat-soluble vitamins.

In Scandinavian countries, where dairying is also traditional, liquid milk consumption is typically higher than in the UK. However, in warmer climates, in particular the Indian subcontinent, Africa and South America, climatic conditions lend themselves less readily to cows' needs, so that the development of dairying and, hence, milk-drinking habits has been far more patchy. Exceptions exist, such as the Masai in East Africa, whose culture is dominated by the cow. Similarly, in parts of India, cows have religious significance but have not, until recently, been intensively managed for their milk.

Being rich in calcium needed for skeletal development and maintenance, milk has traditionally been seen as an important food during childhood, pregnancy and lactation, when calcium requirements are particularly high. This view is still supported today. The Departments of Health in the UK advises milk (and water) to be the most suitable drinks for children, both from a nutritional standpoint and in terms of dental health.

Table 3 Contribution of 285 ml (half pint) of whole milk to nutrient needs

	Percentage of UK reference nutrient intake (RNIs)	
	Girls (4–6 years)	*Adult men (19–50 years)*
Energy[a]	12.5	7.6
Fat[a]	19.0	11.5
Saturates[a]	37.3	22.7
Monounsaturates[a]	14.3	8.6
Polyunsaturates[a]	2.7	1.7
Carbohydrate[a]	7.0	4.2
Protein	47.5	16.8
Nonstarch polysaccharide	–	–
Vitamin A	32.8	23.4
Thiamin	16.4	11.5
Riboflavin	62.5	38.5
Niacin	22.1	14.3
Vitamin B_6	19.4	12.5
Folic acid	17.5	8.8
Vitamin B_{12}	143.8	76.7
Vitamin C	10.0	7.5
Calcium	74.8	48.1
Sodium	23.9	10.4
Chloride	26.6	11.7
Copper	Trace	Trace
Iodine	44.0	31.4
Iron	2.4	1.7
Magnesium	26.7	10.7
Phosphorus	77.0	49.0
Potassium	37.3	16.4
Selenium	15.0	4.0
Zinc	17.7	12.1

[a]There is no RNI for these components of food. For energy, the estimated average requirement has been used in the calculation; for fat and carbohydrate, the desirable average intake has been used.

Figures for milk composition from Holland *et al.* (1989) *Milk Products and Eggs*. Fourth supplement of *McCance and Widdowson's The Composition of Foods*, 4th edn. London: Royal Society of Chemistry/MAFF. Figures for RNIs from Department of Health (1991) *Dietary Reference Values for Food Energy and Nutrients*. Report on Health and Social Subjects 41. London: HMSO.

Availability of Subsidized and Free Milk in the UK

Under the school milk subsidy scheme, funded by the European Community (EC), primary school pupils are eligible to receive 250 ml of milk (whole or semiskimmed) each school day at a subsidized price. Until the spring of 1996, secondary school pupils were also eligible for this benefit and a subsidy was available for milk used in school catering (all age groups). Children under the age of 5 years, attending nursery schools or looked after by

registered child minders, are eligible to a third of a pint of milk, free of charge, on each day they attend.

For families on income support, pregnant women, breast-feeding mothers, and children under the age of 5 years are entitled to a pint of milk a day, free of charge. For bottle-fed babies, tokens can be exchanged for infant formula. Others entitled to free milk (a third of a pint per school day) include children attending special schools.

Nutritional Significance of Milk and its Products in the Diets of Children

The benefits of milk fall into three main areas: general provision of a range of nutrients and energy; provision of calcium for bone health; and dental health benefits.

General Benefits

Dairy products such as yogurt, fromage frais, grated cheese, and small amounts of whole milk, e.g., in custard, can be given to babies as they begin to experience a wide range of foods (4–6 months onwards). As the child becomes established on 'solids,' quantities can be increased, e.g., cows' milk may be used to mix infants' cereals. However, cows' milk should not be introduced as a main drink until a child is 1 year old. This is because, unlike infant formulas, cows' milk is relatively low in iron and vitamin D, and there is growing concern about the iron status of older infants and preschool children.

Amounts consumed by babies may be small, but for preschool and school-age children, milk and products such as cheese and yogurt can make a significant contribution to nutrient needs (**Table 3**). In summary, as well as providing energy, milk is a good source of protein, calcium, zinc, and vitamins A (whole milk only), B_2 (riboflavin), and B_{12}. Milk also makes a valuable contribution to intakes of iodine, niacin (a B vitamin), and vitamin B_6.

Given its nutritional credentials, it is not surprising that milk has been made available free of charge or at a subsidized price to children (see above). A study in Scotland of children aged 7–8 years found generally adequate intakes of most vitamins and minerals, although intakes of sugar and fat were high compared with the adult guidelines (no specific guidelines exist for schoolchildren). However, they noted that the nutrient provision of the diet as a whole improved with the amount of milk consumed. There was no significant difference in the proportion of energy provided by fat when the high milk consumers (at least 3 L a week) were compared with the low milk consumers.

Calcium and Bone Health

There is abundant evidence that calcium plays a major role in the development and maintenance of skeletal strength. Achieving adequate intakes of calcium during growth is important to optimizes peak bone density in early adulthood, thus protecting against osteoporosis in later life.

Approximately 45% of the mineral in the adult skeleton is laid down during adolescence. By the age of about 17 years (earlier in girls), over 90% of the maximum amount of bone mineral (e.g., calcium) that will ever be present in the skeleton is already there. Calcium requirements are, therefore, particularly high during adolescence. Yet it is at this age that milk consumption often falls, frequently resulting in dramatic reductions in calcium intake. There is evidence from national UK surveys among schoolchildren and adults that 1 in 4 teenage girls and young women have wholly inadequate calcium intakes, i.e., below the lower reference nutrient intake (LRNI).

Dairy Products and Oral Health

Milk is kind to teeth and dental experts recommend that milk or water are the most suitable drinks for children. A recent national dental health survey in Britain among preschool children revealed that among those aged between $1\frac{1}{2}$ and $3\frac{1}{2}$ years, the main dietary measure differentiating children with and without dental decay was taking a drink containing non-milk sugars (e.g., fruit juice or squash) to bed, as opposed to not drinking in bed or having other kinds of drink (milk or water). A significantly higher proportion of children who consumed drinks containing non-milk sugars in bed had tooth decay (26%) than did children who drank water (11%) or milk (12%).

Cheese is also considered to be important. A number of studies have shown that chewing (not just swallowing) a small cube of cheese (5–10 g) can revert the reduction in pH in the mouth brought about by the acid produced during the fermentation of sugars by bacteria in the mouth. It is this production of acid that is associated with tooth decay. It is thought that this effect is achieved via cheese's ability to stimulate saliva flow. However, other mechanisms may also be important, including cheese's high concentration of calcium and other minerals needed by teeth, and the ability of the protein in cheese (and milk) to buffer acid.

Nutritional Significance of Milk and its Products for Adults

During pregnancy, especially the final months, requirements for a number of nutrients increase. Although the mother's efficiency at absorbing nutrients rises to help meet this need, dietary supply remains important. Milk is seen to have a major part to play here, given the broad range of nutrients it contains (see **Table 1**).

The needs for some nutrients, for example calcium, rise dramatically after the baby is born in women who choose to breast-feed. Dietary requirements for calcium almost double and consumption of several servings of dairy products daily is one of the few ways of easily achieving such intakes (in the UK the RNI is 1200 mg day^{-1}). Breast-feeding teenagers need even more calcium, as their own skeleton is still developing as well as that of the baby.

Milk and milk products remain a nutritional safeguard for other adults, especially those who choose to restrict their choice of foods because they are slimming or following a vegetarian diet. For elderly people, whose overall food intake and dietary variety can be limited, milk, cheese, and yogurt can be of particular importance in meeting nutrient needs, especially with the convenience of doorstep delivery.

Other Health Effects of Dairy Products

Several studies have reported an inverse association between dairy products consumption and the insulin resistance syndrome. One study reported a reduction of insulin resistance of 40% with consumption of 1 serving per day or more. However, this was observed only in men. In the CARDIA study, a reduction of 21% in insulin resistance was demonstrated among 18–30-year-old men who consumed at least one serving of dairy products per day. Similarly, results from at least 10 clinical trials show a beneficial effect of regular consumption of dairy products on cardiovascular disease risk. Taken together, these results indicate a consistent beneficial effect of milk products on risk for insulin resistance and cardiovascular disease, in addition to the already established benefits for osteoporosis. The overall effect threshold for this benefit appears to be 2 or 3 servings of dairy products per day.

The evidence for a beneficial effect of dairy products in weight management is less compelling, and mixed results have been reported. Several studies have shown a beneficial effect of calcium for weight loss, and it is possible that milk products exert their effect on body weight by being an excellent source of this nutrient. More studies are needed to clarify this potential benefit of diary products.

Yogurt and Health

Apart from its contribution to nutrient needs, the perception of yogurt as a 'healthy' food has been augmented by claims of health benefits attributed to specific live bacteria present in some yogurts, in particular *Lactobacillus acidophilus* and bifidobacteria. Both of these types of bacteria are to be found in the human gastrointestinal tract, especially in breast-fed infants, and it has been suggested that these microorganisms may be able to colonize the gut when consumed in yogurt, and protect against pathogens. It has also been speculated that they may be of benefit in a number of intestinal disorders, including those precipitated by antibiotic treatment or by diseases such as cancer and liver or kidney disease. Claims have been made that specific bacteria used in the production of a certain brand of yogurt have the potential to reduce blood levels of low-density lipoproteins (LDL). On the basis of existing research it is not possible to substantiate these various claims, although evidence is increasing and inconsistencies in the findings may in part be explained by differences in strain and species of bacterial cultures, and differences in experimental design.

There is, however, a substantial body of evidence to indicate that fermented dairy products such as yogurt are well tolerated by individuals who are lactose-intolerant. It has been suggested that this is because of the bacterial enzyme β-galactosidase (produced by the culture) in 'live' yogurt. This enzyme, which is able to digest lactose to glucose and galactose, is intracellular and hence is thought to survive gastric digestion. However, as lactose maldigesters tolerate yogurts with varying β-galactosidase activities equally well, it would seem that other factors may also be important, including rate of gastrointestinal transit of yogurt.

Intake of the types of lactic acid-producing organisms found in yogurt has also been postulated to prevent or inhibit intestinal growth of a variety of food-borne, disease-causing organisms. Most of the evidence supporting a role for cultured dairy products, or the bacteria used to make them, in controlling intestinal pathogens comes from experimental animal and *in vitro* studies. Findings are inconsistent, but various mechanisms have been put forward to explain reported protective effects. For example, the ability of lactic acid cultures to lower intestinal pH favors growth of lactic acid bacteria but provides a hostile environment for pathogens. It has also been suggested that lactic acid bacteria

may produce bacteriocins, proteins with a direct antibiotic effect.

In summary, while a fairly clear case has been made for tolerance of yogurt by lactose maldigesters unable to tolerate milk, the potential benefit of yogurt (or specific types of yogurt) in protecting against pathogens, in recolonizing the gut after illness, or in lowering LDL cholesterol concentrations in the blood needs further investigation.

Hygiene and Safety Aspects

Milk, in its raw state, does not stay fresh for very long and is an ideal medium for bacteria to grow. Various types of heat treatment are used to improve the keeping quality of milk and to kill any harmful bacteria present. These techniques are pasteurization, sterilization, and ultra heat treatment, and they are also used for cream.

Pasteurization is named after the French scientist Louis Pasteur and entails the heating of milk to at least 71.7 °C for a minimum of 15 s. After heating, the milk is quickly cooled to less than 10 °C. In the UK, pasteurization accounts for over 90% of all heat-treated milk and has little effect on the taste and nutritional value of milk.

Sterilization is a more severe process, in which milk is heated to a temperature of 115–130 °C for 10–30 min and then poured into sterile plastic or glass bottles. Unopened, sterilized milk will keep for 2–3 months without refrigeration, although once opened it has to be treated in the same way as pasteurized milk and will only keep for 4–5 days in a refrigerator. Sterilized milk has a slight caramel taste because the heat 'cooks' the lactose present in milk. Sterilization also reduces the levels of the heat-labile vitamins (see above).

Ultra heat treatment is a milder form of sterilization in which the milk is held at a temperature of not less than 135 °C for at least 1 s, and is then packed into sterile cartons. Such milk will keep unrefrigerated for many months, but once opened it needs to be refrigerated and used in 4–5 days.

Cheese and yogurt

The fermentation of milk to produce cheese and yogurt are traditional processes for preserving milk's nutrients. Hard cheeses, such as Cheddar, which have a low moisture content and contain salt as a preservative, will last for many months if stored appropriately. Ideally cheese should be eaten fresh, but if properly stored it will retain its flavor for long periods. It should be wrapped in clear film or foil to prevent drying and then stored in a cool larder or refrigerator. Soft, unripened cheeses, such as cottage cheese, are highly perishable, need to be stored in the refrigerator, and have a relatively short shelf life. At low temperatures, microbiological growth will be reduced, as will enzyme action and biochemical changes that might change the flavor, color or texture of the product. Ripened soft cheeses such as Brie should also be kept refrigerated, wrapped in airtight film or aluminum foil.

Nearly all yogurt sold in the UK contains live bacteria (derived from the starter culture used to produce the yogurt). It is necessary to refrigerate the product to restrict the activity of these bacteria, to prevent development of excess acidity and impairment of flavor. At temperatures of about 5 °C, yogurt has a shelf life of about 14 days, after which time acidity levels may rise above acceptable levels. Spoiled yogurts are often referred to as 'blown.' This is because pressure has built up in the pot via the fermentation of the sugar in the yogurt by the growth of yeasts.

A small proportion of yogurts are heat treated to prolong their shelf life. As a result, they no longer contain live bacteria and do not need to be refrigerated.

See also: **Carbohydrates**: Requirements and Dietary Importance. **Food Intolerance**. **Lactation**: Dietary Requirements. **Microbiota of the Intestine**: Prebiotics; Probiotics. **Protein**: Quality and Sources.

Further Reading

Department of Health (1991) *Dietary Reference Values for Food Energy and Nutrients*. Report on Health and Social Subjects **41**. London: HMSO.

Health Education Authority (1995) *Diet and Health in School Age Children*. London: HEA.

Hinds K and Gregory JR (1995) *National Diet and Nutrition Survey: Children aged $1\frac{1}{2}$–$4\frac{1}{2}$ years*, vol. 2. London: HMSO.

Holland B, Unwin ID, and Buss DH (1989) *Milk Products and Eggs*, Fourth supplement of *McCance and Widdowson's The Composition of Foods*. 4th edn. London: Royal Society of Chemistry/MAFF.

Mennen LI, Lafay L, Feskens EJM *et al.* (2000) Possible protective effect of bread and dairy products on the risk of the metabolic syndrome. *Nutrition Research* **20**: 335–347.

Ministry of Agriculture, Fisheries and Food (1994) *The Dietary and Nutritional Survey of British Adults – Further Analysis*. London: HMSO.

Ministry of Agriculture, Fisheries and Food (1995) *Manual of Nutrition*, 10th edn. Reference Book 342. London: HMSO.

Ministry of Agriculture, Fisheries and Food (1995) *National Food Survey 1994*. London: HMSO.

National Dairy Council (1994) *A–Z of Dairy Products*. London: NDC.

Ruxton CHS, Kirk TR, and Belton NR (1996) The contribution of specific dietary patterns to energy and nutrient intakes in 7–8 year old Scottish schoolchildren. I: Milk drinking. *Journal of Human Nutrition and Dietetics* **9**(1): 3–12.

Weaver CM (1996) Calcium and bone health. In: Buttriss J and Hyman K (eds.) *Women in Focus*. A conference held in 1995. London: National Dairy Council.

DEHYDRATION

A W Subudhi, University of Colorado at Colorado Springs, Colorado Springs, CO, USA
E W Askew, University of Utah, Salt Lake City, UT, USA
M J Luetkemeier, Alma College, Alma, MI, USA

Physiological Functions of Water

After oxygen, water is the most essential nutrient needed to sustain human life. In healthy individuals, water comprises between 45 and 70% of total body weight and is responsible for connecting the diverse physiological functions of the body (**Table 1**).

Water is necessary to maintain homeostasis of the internal environment. The most obvious roles of water in the human body are to provide an aqueous medium for transport of material in blood, to dissolve and pass nutrients between blood and cells, to serve as a medium for intracellular reactions, and to transfer metabolic products for redistribution or excretion via urine. Since both the quantity of reactants and the volume of fluid in which they are dissolved influence chemical reaction rates, imbalances in hydration status can alter cellular and tissue function.

Dehydration also adversely affects the body's ability to regulate temperature. Energy transformations during digestion, absorption, and metabolism as well as muscular contraction generate heat. The heat released from the digestion of a mixed meal (thermic effect of food) equals 10–15% of the caloric content of the food ingested. Muscular contraction is dependent on the transformation of chemical energy (ATP) to mechanical energy. Nearly three-fourths of the energy used for muscular contraction is released as heat. Unless localized heat production from metabolism and muscular contraction is dissipated, the heat burden can be structurally damaging to enzymes or other proteins. Water absorbs heat produced at the cellular level and transfers it to the surface of the skin, where it can be dissipated to the external environment (**Figure 1**).

Table 1 Major physiological functions of water

Function	*Example*
Waste product removal	Urea excretion by kidneys
Solvent for chemical reactions	Glycolysis in the cell cytosol
Transport medium	Blood
Lubrication	Synovial fluid of joints
Shock absorber	Disks between vertebrae of spinal column
Temperature regulation	Evaporative sweat loss

The evaporative dissipation of heat through sweating is a two-phase, water-dependent mechanism. Water is removed from capillary blood perfusing sweat glands to produce a thin layer of sweat over the surface of the skin. Simultaneously, the water component of blood carries heat produced from cellular metabolic processes to capillary beds located near the surface of the skin. Heat is transferred by conduction to the skin surface, where it vaporizes sweat coating the skin, thus transferring body heat to the external environment. The heat of vaporization of water is 586 kcal/l (2453 kJ/l) at 20 °C. Approximately 500 ml of sweat is lost per day under average ambient environmental conditions. Such obligatory water loss occurs without visible or tactile sensations and is termed 'insensible' sweat. However, given a sufficient thermal challenge, humans are capable of producing approximately 10 l of 'sensible' sweat per day. Theoretically, if the entire 10 l of sweat was evaporated, more than 5000 kcal (20 930 kJ) of heat per day would be dissipated via the sweating mechanism. Humidity of the air and sweat that drips from the surface of

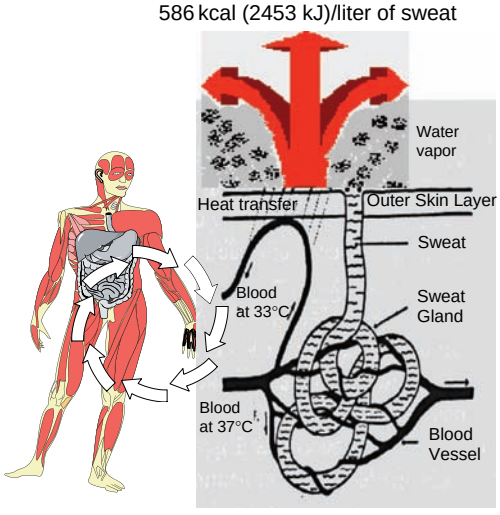


Figure 1 Metabolic heat transfer to the skin and dissipation of heat by evaporation of sweat. The body has more than 2 million sweat glands that secrete sweat to the surface of skin. Blood-perfusing skin capillary beds transfer heat by convection to the surface of the skin. Heat is dissipated by vaporizing the water in sweat. The heat of vaporization of water at 20 °C is 586 kcal/l (2453 kJ).

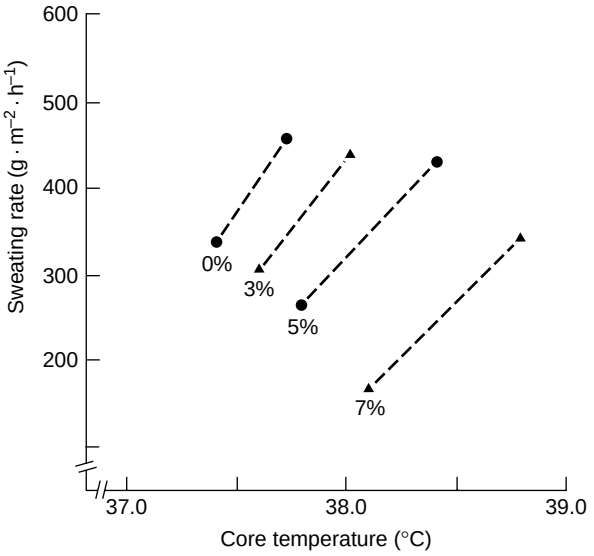


Figure 2 The influence of water loss by dehydration (hypohydration) on the sweating response to exercise following normal hydration (0%) and dehydration equal to 3, 5, and 7% of body weight. The primary stimulus for sweating is the increase in core temperature (thermal drive). Note that dehydration reduces the sweating rate at any given level of thermal drive. Hypohydration compromises exercise by reducing sweat rate and evaporative cooling and increasing body core temperature. (From Sawka MN, Young AJ, Francesconi RP *et al.* (1985) Thermoregulatory and blood responses during exercise at graded hypohydration levels. *Journal of Applied Physiology* **59**: 1394–1401, with permission.)

the skin considerably reduce the potential for evaporative heat dissipation; therefore, actual evaporative cooling is usually less than the theoretical maximum. Since water is the main component of sweat, it is not surprising that dehydration affects the sweat response. The relationship between body water loss by dehydration and the rate of sweating achievable during exercise is shown in **Figure 2**, which illustrates that dehydration reduces sweating rate at any given level of thermal drive (core temperature) during exercise. A diminished sweating response can lead to a dangerous heat buildup unless thermal strain is curtailed by other mechanisms.

Development of Dehydration

In physiological terms, dehydration is the process of progressing from the euhydrated (normally hydrated) to hypohydrated (less water than normal) state. In actual usage, dehydration means losing body water faster than it is replaced. The resultant condition is commonly referred to as the 'dehydrated state' and is associated with hypovolemia (low blood volume).

Contributing Factors

Water is lost through a variety of avenues, including urine, feces, breath, and sweat. In illness or disease, excessive diuresis, diarrhea, and/or vomiting are the main pathways of water loss. During exercise or heat exposure, sweating is the primary mechanism for dehydration. Significant water loss may be stimulated by cold- or altitude-induced diuresis. Additionally, some prescription drugs and over-the-counter herbal products have diuretic effects that exacerbate water loss. Under normal conditions, the body regulates its water contents tightly over a 24-h period (approximately ±200 ml); however, over short periods, water loss can significantly exceed water gain (**Figure 3**).

Body Fluid Balance

Body water losses are rapidly reflected in blood. Volume and electrolyte changes in response to decreased blood water content (increased osmolality) trigger the hypothalamus to stimulate antidiuretic hormone (ADH) release from the posterior lobe of the pituitary gland. ADH acts on the kidney to increase tubular water resorption and maintain plasma volume. Decreased plasma volume also

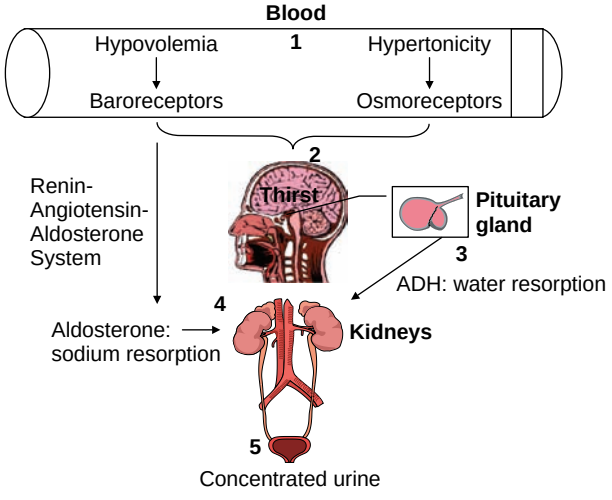


Figure 3 Water and sodium physiology: mechanisms controlling body water gain and loss. As water is lost from the body via sweat, urine, respiration, and feces, (1) plasma osmolality increases and plasma volume decreases with water loss. (2) The increase in osmolality acts on the 'thirst center' in the hypothalamus to secrete ADH and stimulates the conscious desire for water. (3) The release of ADH from the pituitary gland increases tubular resorption of water by the kidney. (4) Aldosterone is formed via a series of reactions involving renin, which is released from the adrenal cortex in response to decreased blood pressure, and a plasma protein, angiotensinogen. Aldosterone promotes sodium resorption by the kidney to maintain plasma volume. (5) These events conserve water and result in the production of concentrated urine.

results in a complex series of events resulting in the release of renin from the kidneys and the subsequent formation of angiotensin II and the minerocorticoid, aldosterone. Angiotensin II is a potent vasoconstrictor and stimulator of thirst. Aldosterone promotes sodium resorption, which allows the blood to retain more water. The net result of these regulatory mechanisms is concentrated urine and maintenance of plasma volume, provided that exogenous fluid intake increases proportionally. If fluid intake is not increased, dehydration will still result.

Thirst

Thirst is not a good short-term regulator of fluid balance. Humans frequently lose up to 2% of their body weight as water before the thirst mechanism is activated. The actual point at which the thirst mechanism is activated varies considerably between individuals. Some athletes are closely attuned to their anticipated fluid needs and develop the habit of drinking before they become dehydrated. However, the majority of individuals do not feel compelled to drink until they have become moderately dehydrated, even though fluids may be available. These individuals are called 'voluntary dehydrators.' Voluntary dehydrators frequently replace only approximately two-thirds of their short-term fluid losses.

Pathophysiology of Dehydration

Dehydration and Human Performance

Natives of desert regions have, over the years, habituated to being chronically dehydrated. A study of the desert inhabitants found that they had a curtailed thirst drive that was associated with excretion of low volumes of concentrated urine and a high incidence of kidney disease (kidney stones). When additional water intake (approximately twice normal) was ingested in a subsample of this population, they were able to exercise 10% longer in the desert environment, presumably due to improved thermoregulation. The results of this and other studies illustrate that humans probably do not adapt to dehydration but can become used to a mild chronic dehydration due to inadequate fluid intake. This is not a true physiological adaptation since there are negative health and performance effects associated with chronic dehydration.

Body Water Deficits

When fluid intakes are insufficient to maintain normal body water content (approximately 60% for males and 50–55% for females), deficits arise in all fluid compartments, with the reduction in plasma volume being of particular concern. Dehydration decreases plasma volume and increases tonicity. Plasma hypertonicity signals the circulatory system to conserve plasma volume for internal organs at the expense of skin blood flow. Reduction in skin blood flow decreases evaporative cooling. Additionally, decreased plasma volume reduces stroke volume and cardiac output, which impairs cooling capacity and exercise performance. The effects of dehydration on heart rate, body temperature, and endurance are shown in **Figure 4**. Consuming water to replace sweat loss while cycling for 6 h at 55% VO_2 max in the heat was associated with lower heart rates and core temperatures compared to a trial in which no water was ingested. The increase in heart rate while cycling without water replacement is a compensatory mechanism to maintain cardiac output in response to reduced plasma volume. Elevated core temperatures in cyclists not consuming water

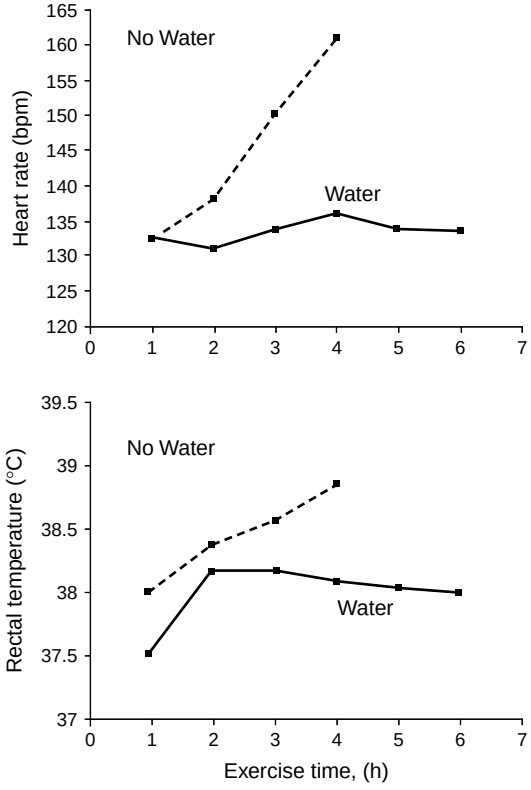


Figure 4 Consequences of consuming or not consuming water during cycle ergometer work (50% VO_2 max, 30 °C, 50% relative humidity). Subjects drank water to replace that lost during exercise (water group) or did not drink during exercise (no water group). Note that the no water group could not exercise as long as the water group. (Adapted from Barr SI, Costill DL and Fink WJ (1991) Fluid replacement during prolonged exercise: Effects of water, saline, or no fluid. *Medicine and Science in Sports and Exercise* **23**: 811–817.)

resulted from reduced skin blood flow and ultimately forced the cessation of exercise.

Dehydration and Heat Illness

If water loss due to sweating is not replaced during exercise, plasma volume and sweat rate will be decreased (**Figure 2**). The combination of reduced peripheral blood flow for heat exchange and reduced sweat volume for evaporative cooling leads to an overall reduction in the ability to dissipate heat.

The consequence of impaired heat dissipation is hyperthermia. Without evaporative cooling, human core temperatures can elevate ~5 °C/h during moderate intensity work. Heat production is proportional to the intensity and duration of work, ranging from ~75 kcal/h (314 kJ) at rest to ~300 kcal/h (1256 kJ) during moderate exercise and ~600 kcal/h (2512 kJ) for maximal sustained work. Brief periods of intense exercise can generate heat at the rate of ~900 kcal/h (3768 kJ).

Hyperthermia can lead to serious or even life-threatening heat injury if left unchecked. Heat injury can result if the rate of heat production is greater than the rate of cooling. When fluid losses are not replaced during activity, heat dissipation mechanisms are compromised. The buildup of heat in blood and tissues adversely affects various physiological systems. Minor heat injury syndromes include prickly heat (skin rash resulting from plugged sweat glands), heat syncope (light headedness due to pooling of blood in the extremities), and heat cramps (muscle cramps related to electrolyte loss). These heat illnesses are of concern but not life-threatening. Major hyperthermia syndromes involving dehydration are heat exhaustion and heat stroke. Conversely, overhydration (replacing large volumes of fluid loss) without sodium provision can, in certain instances, lead to an overdilution of blood sodium (hyponatremia). The symptoms of hyponatremia are similar to those of other forms of heat illness, but the treatment is critically different.

Heat exhaustion Two types of heat exhaustion can be distinguished: water depletion (inadequate consumption of water) and salt depletion (loss of large volumes of sweat that are replaced without adequate sodium intake). Heat exhaustion usually occurs on a continuum somewhere between these two extremes. During heat exhaustion, thermoregulatory mechanisms cannot dissipate heat effectively, primarily because of reduced skin blood flow. People who are unacclimatized to the heat or not in good physical condition are more susceptible to heat exhaustion. Symptoms vary but usually include a temperature of less than 39.5 °C, malaise, weakness, fatigue, headaches, anorexia, nausea, vomiting, diarrhea, and muscle cramps. Although irritability, anxiety, and impaired judgment may be present, the subject is usually alert and capable of responding to questions. If left untreated, heat exhaustion can progress to heat stroke.

Heat stroke Heat stroke is less common than heat exhaustion but is much more serious. Heat stroke is a life-threatening disorder that requires immediate medical treatment. Two forms of heat stroke are generally classified as exertional or classical. Exertional heat stroke generally occurs in young subjects working too hard for too long in the heat. Classical heat stroke is associated with environmental heat waves and primarily afflicts the very young, very old, poor, and debilitated. The pathophysiology of heat stroke involves failure of the body's thermoregulatory mechanisms following a severe heat overload. As core temperature elevates, cell function deteriorates culminating in massive cell damage. Dehydration is often a contributing factor to heat stroke, but the basic pathophysiology is uncontrolled heat overload. Core temperature is higher than that seen in heat exhaustion, generally 41 or 42 °C. Core temperatures higher than 39.5 °C reduce the function of motor centers in the brain and subsequently the ability to recruit motor units required for muscular activity. Extertional heat stroke is characterized by cessation of sweating, hot and dry skin, physical deterioration, confusion, collapse, and seizure. Rhabdomyolysis (muscle fiber destruction) may result from exertional heat stroke. In one reported case, an accelerated rhabdomyolysis resulting from exertional heat stroke occurred during an 8-km fun run when the ambient temperature was higher than 37 °C. This unfortunate runner collapsed with a rectal temperature of 42 °C and suffered acute renal failure as a consequence of an impaired immune system, infection, and decreased clotting ability. He eventually recovered, but it was necessary to amputate one of his legs that became infected following the rhabdomyolysis.

Other Consequences of Dehydration

There are other physiological consequences of dehydration that are not as serious as heat illness but can contribute to decreased performance capacity. Dehydration impairs thermoregulation in both hot and cold environments. Metabolic heat production in the cold may be less efficient in dehydrated individuals. The mechanism is not well understood, but it

may involve a concomitant reduction in energy intake, decrease in resting metabolic rate, impaired shivering response, impaired vasodilation/constriction response, or a combination of these factors. Dehydration also blunts appetite, which in turn may elicit energy and thermoregulatory defects.

Management of Dehydration

Identifying Types of Dehydration

Although prevention is the best management strategy for dehydration, water imbalances may also be treated after the recognition of dehydration by replacement of appropriate fluids, such as water or water containing electrolytes. Dehydration usually occurs along a continuum of fluid and electrolyte loss. The ratio of water to electrolyte loss determines the type of dehydration present. A convenient classification of types of dehydration, their characteristics, and how to treat them is shown in **Table 2**.

It should be noted that hypotonic dehydration is the result of treating isotonic dehydration with nonelectrolyte-containing fluids and can lead to a potentially dangerous condition know as hyponatraemia (low blood sodium). This may be a particular problem in the case of 'overzealous hydrators' such as athletes who overcompensate for sweat losses. Hypotonic dehydration is also seen in infants and children who may be afflicted with gastrointestinal disturbances such as diarrhea, stomach flu, or acute gastroenteritis if water alone is used to hydrate. The American Academy of Pediatrics has published guidelines for oral rehydration therapy of infants and children younger than 5 years old with acute gastroenteritis. **Table 3** gives the Academy's guidelines for the assessment of dehydration. An algorithm for the treatment and hydration of children with acute diarrhea is shown in **Figure 5**.

Clinical standards for assessing dehydration in adults have not been well established. Measurements of plasma osmolarity and urine-specific gravity and osmolarity can be used to assess relative dehydration if baseline (euhydrated) values are known. However, due to significant interindividual variation, the use of absolute specific gravity

Table 2 Types of dehydration

Mild hypovolemia
Fluid intake is insufficient to meet needs, 2–5% body weight loss, yellow urine, dry lips, reduced skin elasticity
Many people are chronically hypovolemic in outdoor environments
Simply need to be reminded to drink, easily corrected by fluids and consuming food

Hypertonic dehydration (hypernatremic dehydration)
Body water losses greater than sodium losses, elevated blood osmolality and hypernatremia
May be accompanied by fever, profuse sweating, and/or evaporative water loss
Acute weight loss; person eats but does not drink
Treatment is provided by additional fluids (water is best)

Isotonic dehydration
Body loses equal amounts of water and sodium from routes other than sweating
Gastrointestinal fluid loss—vomiting, diarrhea
Blood electrolytes normal
Acute weight loss, tachycardia, orthostatic hypotension
Treat by replacing lost fluid *and* electrolytes or hypotonic dehydration may develop

Hypotonic dehydration (dilutional hyponatremia)
Can develop when isotonic dehydration is treated with only water
Hypotonic dehydration can occur anytime the body sodium loss exceeds water loss (e.g., sodium-restricted diets, diuretic use, overzealous hydrators, fluid replacement with only water following repeated vomiting and diarrhea)

Table 3 Guidelines for the assessment of dehydration[a]

Variable	*Dehydration*		
	Mild, 3–5%	*Moderate, 6–9%*	*Severe, >9%*
Blood pressure	Normal	Normal	Normal to reduced
Quality of pulses	Normal	Normal or slightly decreased	Moderately decreased
Heart rate	Normal	Increased	Increased, severe cases bradycardia
Skin turgor	Normal	Decreased	Decreased
Fontanelle	Normal	Sunken	Sunken
Mucous membranes	Slightly dry	Dry	Dry
Eyes	Normal	Sunken orbits	Deeply sunken orbits
Extremities	Warm, normal capillary refill	Delayed capillary refill	Cool, mottled
Mental status	Normal	Normal to listless	Normal to lethargic or comatose
Urine output	Slightly decreased	<1 ml/kg/h	≪1 ml/kg/h
Thirst	Slightly increased	Moderately increased	Very thirsty or too lethargic to indicate

[a]The percentages of body weight loss and their corresponding categorization sometime vary depending on the author.
From the American Academy of Pediatrics (1996) based on Duggan *et al.* (1992) Reproduced by permission of *Pediatrics* and the American Academy of Pediatrics.

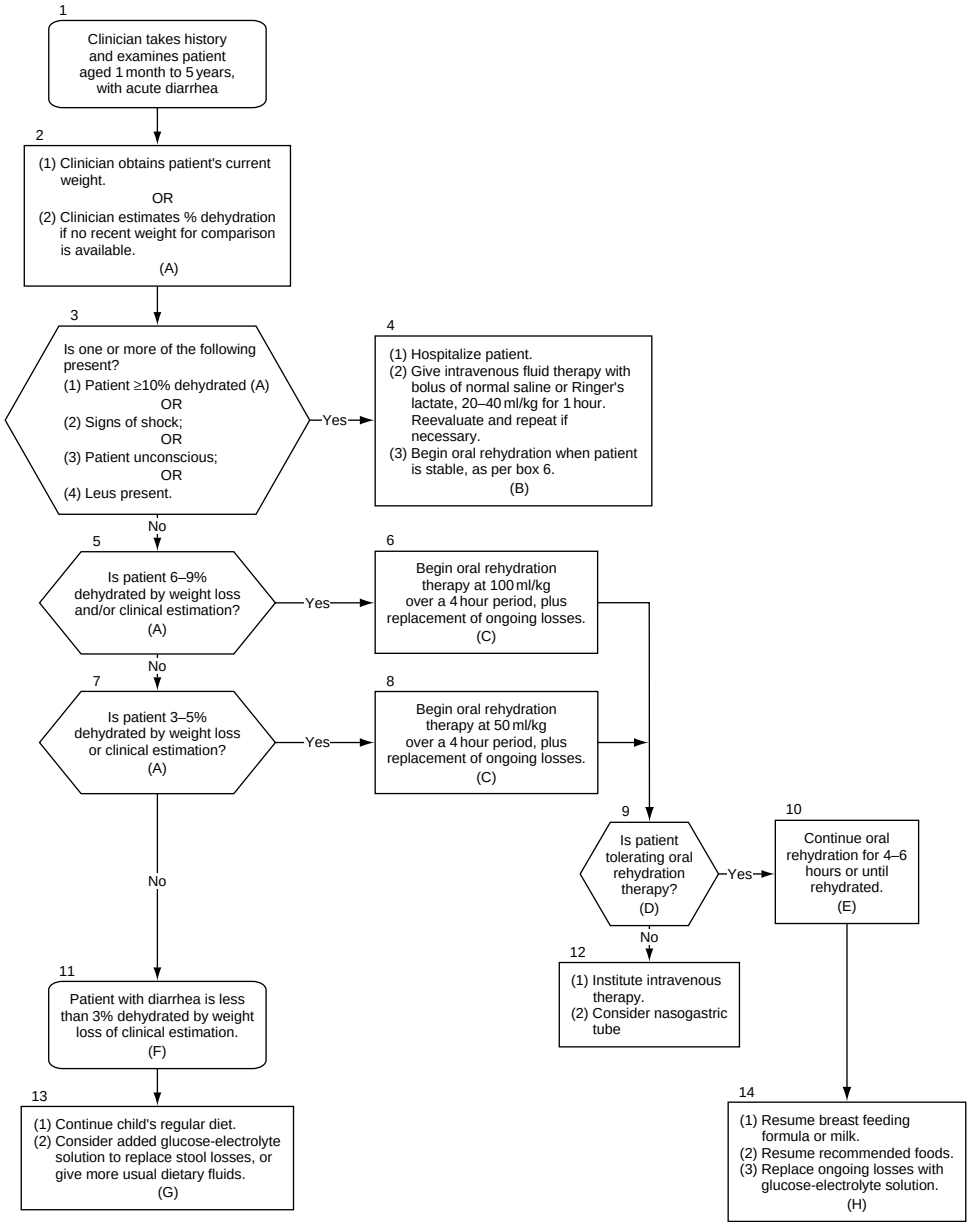


Figure 5 Algorithm for children with dehydration from acute diarrheal disease. The letters at the bottom of the decision boxes refer to the following: (A) See **Table 3** for guidance in the assessment of the degree of dehydration. (B) Restoration of cardiovascular stability is critical and is accomplished by giving bolus i.v. therapy. In the patient who does not respond, consider the possibility of an underlying disorder. When the patient is in stable condition and has achieved satisfactory mental status, oral rehydration therapy (ORT) can be implemented. (C) Solutions containing 45–90 mmol/l sodium should be given in a volume of 100 ml/kg for moderate dehydration and 50 ml/kg for mild dehydration. Giving the child these volumes requires patience and persistence, and progress must be monitored frequently. (D) Intractable, severe vomiting, unconsciousness, and ileus are contraindications to ORT. Persistent refusal to drink may require a trial of i.v. therapy. (E) The rehydration phase usually can be completed in 4 h; reevaluation should occur every 1 or 2 h. See referenced text for guidance to decide when rehydration has been achieved. (F) The type and intensity of therapy will vary with the individual clinical situation. (G) Often, a child has diarrhea but remains adequately hydrated. The parent can be reassured but should be taught to assess hydration and to identify a worsening condition. If the stool output remains modest, ORT may not be required if early, age-appropriate feeding is instituted and increased consumption of usual dietary fluids is encouraged. More significant stool losses can be replaced with an oral rehydrating solution at the rate of 10 ml/kg for each stool. (H) Breast-feeding should be resumed. Nonlactose formula, milk-based formula, or milk may be given, although a small percentage of children will not tolerate lactose-containing fluids. Lactose-containing solutions seem to be tolerated better when combined with complex carbohydrates in weaned children. Children who are eating foods may resume eating, although certain foods are tolerated better than others. Recommended foods include complex carbohydrates (rice, wheat, potatoes, bread, and cereals), lean meats, yogurt, fruit, and vegetables. Avoid fatty foods and foods high in simple sugars (including juices and soft drinks). Supplement feeding with an oral electrolyte solution, 10 ml/kg for each diarrheal stool and the estimated amount vomited for each emesis. (From American Academy of Pediatrics (1996) Practice parameter: The management of acute gastroenteritis in young children. *Pediatrics* **97**: 1–22, with permission.)

and osmolarity values for the diagnosis of dehydration remains questionable. Monitoring urine color has been suggested as a noninvasive way to measure hydration status, where light, pale yellow urine generally indicates a favorable hydration status. Assessing urine color may be a simple method to assess hydration status, but it can also be artificially influenced by dietary intake (i.e., nutritional supplements). Acute change in body weight is the most common and practical method used to assess hydration status. It is assumed that short-term body weight loss is primarily the result of water loss.

Treating Different Types of Dehydration

In the majority of simple, nonsevere dehydration cases, plain water is an adequate rehydration solution. However, there are instances (e.g., children younger than 5 years of age dehydrated by vomiting and diarrhea) when water containing sodium and potassium is the proper hydrating agent. The most effective way of preventing and treating mild to moderate dehydration in infants and children with acute diarrhea is the oral administration of oral rehydration solutions (ORSs). There are a number of commercially available ORSs. These solutions are designed to replace fluid and electrolytes when both water and food intake have been restricted or compromised by diarrheal disease. The World Health Organization recommends the ORS shown in **Table 4** for individuals afflicted with diarrheal disease and vomiting. Oral modes of fluid and electrolyte administration are always preferred in mild (3–5%) to moderate (6–9%) dehydration; however, intravenous fluids may be required in cases of severe dehydration (>9%) and vomiting or if the patient is in a comatose state. When i.v. fluids are administered, 0.45% saline with 5% dextrose is an effective hydrating agent.

In most instances involving heavy sweating, plain water containing 1.25 g of NaCl per liter is a suitable rehydration solution. Increasing the concentration of NaCl to 5 or 6 g per liter may promote the rate of rehydration but may not be palatable for some individuals. Most commercial sports drinks contain 1.2–1.8 g NaCl per liter and are also good rehydration solutions, especially when both fluid and electrolytes have been lost through sweating. Fruit juices can also provide fluid, energy, and electrolytes (e.g., fresh orange juice contains approximately 10 mg of sodium and 2000 mg of potassium per liter) but may be too concentrated and delay gastric emptying. Diluting fruit juices 1:3 with water may yield a more appropriate rehydration solution. The inclusion of carbohydrate in the rehydration solution provides energy for the intestinal sodium pump, which facilitates sodium transport across the intestinal cell wall into the blood, where it in turn exerts a positive osmotic effect on water absorption from the gut. Glucose and electrolyte sports beverages are useful rehydration solutions for sporting activities but are not a good choice for children with diarrhea since these beverages have lower electrolyte and higher carbohydrate concentrations than recommended.

Groups at Risk for Dehydration

Predisposing Factors for Heat Illness

Certain segments of the population are at greater risk for dehydration and subsequent heat illness than others (**Table 5**). The predisposing factors for dehydration and heat illness in these populations are obesity (extra exertion, heat production, and sweating are required to move a larger mass), insufficient heat acclimation (associated with reduced sweating and evaporative cooling and increased cardiovascular and renal stress), socioeconomic barriers to cooling methods (fans, air conditioners, etc.), pyrexial illness (fever), drug and alcohol abuse (interferes with fluid balance and thermoregulation), physical work in environments that contribute to dehydration (heat: sweating; cold: respiratory water loss and diuresis; altitude: respiratory water loss and

Table 4 Composition of recommended WHO/UNICEF oral rehydration solution

Solute	*Content (mmol l^{-1})*
Glucose	75
Sodium	75
Chloride	65
Potassium	20
Citrate	10
Total osmolarity	245

Table 5 Dehydration and heat illness: Populations at risk

Elderly
Poor
Young children
Obese people
Alcoholics
People afflicted with respiratory, cardiovascular, cerebrovascular, renal, or diarrheal disease
Athletes and outdoor workers

diuresis), and athletic competition and training (if athletes do not replace sweat loss). Even athletes who make a conscious attempt to drink during exercise only ingest approximately 300–500 ml fluid per hour; fluid loss through sweating (500–1000 ml/h) can easily surpass this intake of fluid.

Elderly and Children

The very young and the very old are two populations especially susceptible to dehydration. Children have less surface area-to-mass ratio for evaporative cooling and are less inclined to replace fluids; thus, they are less efficient thermoregulators than adults when exposed to high environmental temperatures. In the United States, approximately 9% of all hospitalizations of children younger than 5 years of age are due to diarrhea.

Aging is associated with decreased thirst, sweating, and renal responses that place the elderly at high risk during periods of extreme shifts in environmental temperature. Dehydration is a common cause for hospitalization and death in the aged population. Statistics from a 1991 U.S. survey of Medicare recipients revealed that almost half of the Medicare beneficiaries hospitalized for dehydration died within 1 year of admission. Older men and women may have a higher osmotic operating point (the point at which the thirst sensation is triggered), which may contribute to hypovolemia. Certain behavioral factors may also influence drinking patterns in older adults who may wish to avoid the physical difficulty associated with trips to the bathroom. Besides contributing to an increased risk for hyperthermia, dehydration also alters the effective dosage of medications through plasma volume changes, leading to further medical complications in the elderly. Dehydration in the elderly often accompanies or results from clinical conditions and/or medications.

Prevention of Dehydration

Dehydration resulting from nondisease causes can be easily prevented provided that people are inclined to drink and have access to cool, safe sources of fluids. Drink flavoring, beverage temperature, and sodium chloride content are important promoters of fluid intake in active children. Education of athletic coaches, the general public, and health care providers is necessary to increase awareness of the importance of proper hydration. The American College of Sports Medicine has issued a set of guidelines for fluid replacement (**Table 6**).

Table 6 Fluid replacement: Summary of recommendations of the American College of Sports Medicine

It is recommended that individuals consume a nutritionally balanced diet and drink adequate fluids during the 24-h period before an event, especially during the period that includes the meal prior to exercise, to promote proper hydration before exercise or competition.
It is recommended that individuals drink about 500 ml of fluid about 2 h before exercise to promote adequate hydration and allow time for excretion of excess ingested water.
During exercise, athletes should start drinking early and at regular intervals in an attempt to consume fluids at a rate sufficient to replace all the water lost through sweating, or consume the maximal amount that can be tolerated.
During exercise lasting less than 1 h, there is little evidence of physiological or physical performance differences between consuming a carbohydrate–electrolyte drink and plain water.
Inclusion of sodium (0.5–0.7 g per liter of water) in the rehydration solution ingested during exercise lasting longer than 1 h is recommended since it may be advantageous in enhancing palatability, promoting fluid retention, and possibly preventing hyponatremia in certain individuals who drink excessive quantities of fluid. There is little physiological basis for the presence of sodium in an oral rehydration solution for enhancing intestinal water absorption as long as sodium is sufficiently available from the previous meal.

From the American College of Sports Medicine (1996) Position stand on exercise and fluid replacement. *Medicine and Science in Sports and Exercise* **28**: i–vii.

Simple methods, such as recording body weight before and after exercise to determine fluid loss and observing the color of urine or the turgidity of skin, can be useful for monitoring hydration status. The simplest insurance against dehydration is to consume fluids prior to and during physical activity or heat exposure to match water loss. The amount of fluid needed to maintain a favorable hydration status is variable between individuals but often necessitates drinking in the absence of thirst. Excess fluid consumption is rarely a problem. However, caution should be used to avoid dilutional hyponatraemia from overzealous hydration. Humans can acclimate to work in a hot environment and enhance their ability to thermoregulate and conserve fluid, but they cannot adapt to dehydration. Acute dehydration can decrease physical performance and thermoregulation ability and increase the risk for heat illness. Chronic dehydration can reduce metabolic and thermoregulatory efficiency and increase predisposition to kidney disease. The deleterious effects of dehydration on physiological function are summarized in **Figure 6**.

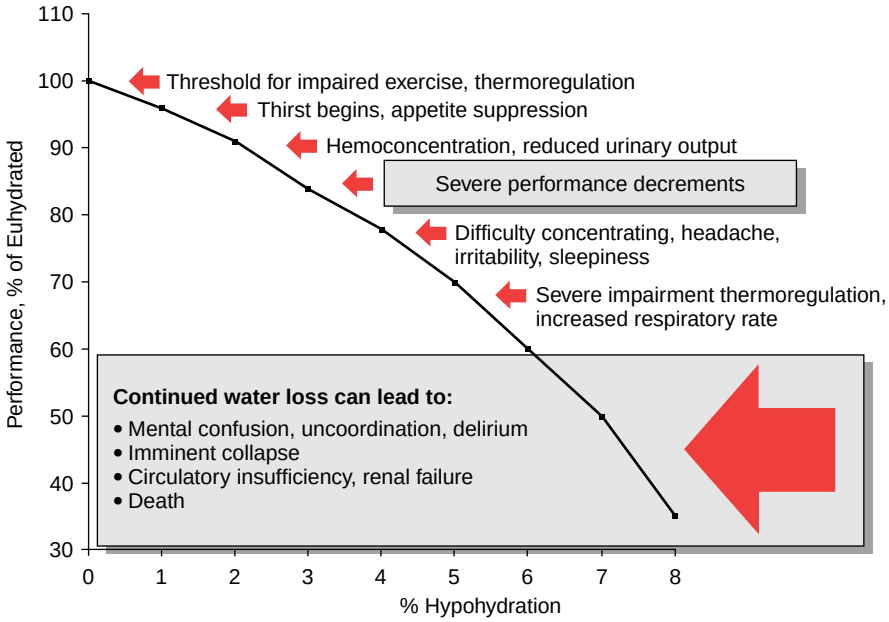


Figure 6 Progressive physiological effects of dehydration on physical performance and pathophysiology of hypohydration. The onset, magnitude, and severity depend on the workload, level of physical fitness, ambient temperature, relative humidity, and degree of heat accumulation of the individual. (From Askew EW (1996) Water. In: Ziegler EE and Filer LJ Jr (eds.) *Present Knowledge in Nutrition*, pp. 98–108. Washington, DC: ILSI Press, with permission.)

See also: **Children**: Nutritional Requirements. **Diarrheal Diseases**. **Electrolytes**: Acid-Base Balance; Water–Electrolyte Balance. **Infants**: Nutritional Requirements. **Older People**: Nutritional Requirements; Nutrition-Related Problems. **Sodium**: Physiology. **Thirst**.

Further Reading

American Academy of Pediatrics (1996) Practice parameter: The management of acute gastroenteritis in young children. *Pediatrics* **97**: 1–22.

American College of Sports Medicine (1996) Position stand on exercise and fluid replacement. *Medicine and Science in Sports and Exercise* **28**: i–vii.

Askew EW (1996) Water. In: Ziegler EE and Filer LJ Jr (eds.) *Present Knowledge in Nutrition*, pp. 98–108. Washington, DC: ILSI Press.

Askew EW (1997) Nutrition and performance in hot, cold and high altitude environments. In: Wolinski I (ed.) *Nutrition in Exercise and Sport*, 3rd edn., pp. 597–619. Boca Raton, FL: CRC Press.

Burke LM and Hawley JA (1997) Fluid balance in team sports. Guidelines for optimal practices. *Sports Medicine* **24**: 38–54.

Buskirk ER and Puhl SM (1996) *Body Fluid Balance*. Boca Raton, FL: CRC Press.

Hawley JA, Dennis SC, and Noaks TD (1995) Carbohydrate, fluid, and electrolyte requirements during prolonged exercise. In: Kies CV and Driskell JA (eds.) *Sports Nutrition: Minerals and Electrolytes*. Boca Raton, FL: CRC Press.

Kenney L and Chiu P (2001) Influence of age on thirst and fluid intake. *Medicine and Science in Sports Exercise* **33**: 1524–1532.

Murray R (1995) Fluid needs in hot and cold environments. *International Journal of Sports Nutrition* **5**: S62–S73.

Noaks TD (1998) Fluid and electrolyte disturbances in heat illness. *International Journal of Sports Nutrition* **19**: S146–149.

Rolls BJ (1998) Homeostatic and non-homeostatic controls of drinking in humans. In: Arnaud MJ (ed.) *Hydration throughout Life*, pp. 19–28. Montrouge, France: John Libbey Eurotext.

Sawka MN, Latzka WA, and Montain SJ (2000) Effects of dehydration and rehydration on performance. In: Maughan RJ (ed.) *The Encyclopedia of Sports Medicine. Vol. VII: Nutrition in Sport*, pp. 216–225. Oxford: Blackwell Science.

Sawka MN, Montain SJ, and Latzka WA (2001) Hydration effects on thermoregulation and performance in the heat. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* **128**: 679–690.

Shirreffs SM and Maughan RJ (2000) Rehydration and recovery of fluid balance after exercise. *Exercise and Sport Science Reviews* **28**: 27–32.

Wilk BS, Kremler H, and Barr-Or O (1998) Consistency in preventing voluntary dehydration in boys who drink a flavored carbohydrate-NaCl beverage during exercise in the heat. *International Journal of Sports Nutrition* **8**: 1–9.

DENTAL DISEASE

R C Cottrell, The Sugar Bureau, London, UK

Introduction

A number of dental diseases are among the most common diseases that are influenced by diet. Dental decay (caries) is experienced to some degree by most people at some time in their lives; dental erosion is thought to be an increasing problem, although data on prevalence are scarce, and dental-enamel defects are still an issue in some communities. Bacterial gum disease, while also common, is not known to be influenced by diet.

Enamel Defects

The structural integrity of the hard tissues in the mouth is influenced by the nutritional status of the mother during fetal development and by that of the child during the early years, when the permanent teeth are being formed (especially important are adequate availabilities of calcium and vitamin D). An excessive intake of fluoride during the period when the enamel of the primary or secondary teeth is being formed will lead to the characteristic defect of fluorosis. Except in rare circumstances, when the intake of fluoride is very high, this condition is usually minor, with no functional significance, and is visible only on professional inspection.

Dental Caries, Erosion, and Gum Disease

Once the teeth have erupted, they may be subject to three main conditions, all of which may threaten their survival: dental caries, tooth wear, and gum disease. Inappropriate dietary habits are a necessary contributing factor to the development of dental caries and may contribute, in part, to tooth wear by causing acid erosion. In contrast, diet has no effect on the common forms of gum disease.

Etiology of Caries

The causes of dental caries and factors influencing their formation have been the subject of research for more than 100 years. The importance of oral bacteria was discovered well before the specific influence of sugars derived from the diet became known in around 1950. While the protective effect of fluoride has also been known for more than 50 years, the mechanism of this effect is still a subject of debate.

Different approaches have been used to try to understand the caries process. Experimental studies have either induced clinically apparent caries or attempted to model the early stages of caries. Ethical limitations on studies that might cause caries in humans and increasing resistance to animal experimentation have stimulated a great deal of imaginative recent work with laboratory modelling.

Direct studies of caries induction are rarely conducted nowadays. But, in the past, important evidence in this field has come from experiments in which caries were induced in laboratory animals and, in one important instance, from a similar experiment in human subjects. The animal experiments are now regarded with some suspicion, since the information gained cannot be readily interpreted in terms of human risk. The animals used differ appreciably from their human counterparts in the structure of their teeth, their way of eating, and other factors such as saliva and oral bacterial populations. These animal experiments have been useful, however, in establishing that all fermentable carbohydrates are capable of inducing caries under appropriate conditions.

A key human experiment was conducted in the 1950s, before it was entirely clear that sugar is capable of causing caries. It was important in that it demonstrated conclusively that the consumption of a large amount of sugar does not necessarily have a discernable influence on caries risk, provided it is eaten at mealtimes, whereas frequent consumption of quite small amounts of sugar had a marked influence. Subjects given $340\,g\,day^{-1}$ of sugar at meal times showed no increase in caries incidence, while subjects given $50\,g\,day^{-1}$ or $100\,g\,day^{-1}$ between meals showed an increase. Typical European intakes of sugar are less than $100\,g\,day^{-1}$.

The subjects in this study had little or no oral hygiene and no access to fluoride. It can therefore be readily concluded that the amount of sugar consumed in the diet, even in countries with high consumption, is unlikely to influence caries risk, especially with the regular use of fluoride toothpaste for oral hygiene. Whether frequent consumption of sugar will influence caries risk in an individual who cleans his or her teeth regularly with a fluoride toothpaste is more controversial. But, given the current state of knowledge, it seems unwise to assume that any dietary behavior would be safe, however outlandish. The current fashion of eating and drinking perpetually and of sipping sugar-containing drinks from a can over long periods seems designed to cause caries and cannot be recommended.

Research into the causes of dental caries has addressed a number of questions. These include why there are large differences in the disease experience of individuals within the same population or even family group, why the prevalence and severity of the disease are so different in different populations, and why these can change so dramatically with time. Entirely satisfactory answers to these questions are still being sought, but much has been learned over the last 100 years about the contributing factors and protective measures that determine the likelihood of this disease developing. This knowledge has been synthesized into the currently held view that clinically significant caries will develop only when a number of circumstances occur simultaneously. Inappropriate dietary habits (frequent consumption of sugars or starches) will allow the selective proliferation of bacteria attached to the tooth surface that are capable of metabolizing sugars to organic acids (especially lactic acid). These acids will facilitate dissolution of the tooth enamel whenever their production is sufficient to lower the local pH below a critical level. The presence of saliva or of other components of the food matrix will influence the pH attained and also the rate at which mineral is lost from the tooth surface.

The formation of dental caries is not, however, a simple unidirectional process of demineralization. Some tooth mineral may be removed almost every time something is eaten or drunk, but this loss will generally be made good by the subsequent accretion of mineral from saliva. Thus, a cavity develops only when the balance of repeated cycles of demineralization and remineralization results in localized overall mineral loss. It is for this reason that caries are most likely to occur at sites where food residues are likely to be trapped and access for saliva is limited (for example, between two closely abutting teeth).

The presence of fluoride not only radically alters both demineralization and remineralization but may also inhibit the activity of the acid-generating bacteria. To date, the most effective methods of reducing the incidence of dental caries have involved the use of fluoride either (at low concentrations) in community water supplies or (at higher concentrations) in toothpaste.

Caries-Causing Bacteria

The surfaces of all teeth are normally covered with a biofilm (plaque) composed of a range of bacterial species embedded in a sticky organic material produced by the metabolic activity of specialized bacteria. Colonization of the surfaces of the teeth starts as soon as they erupt in a baby's mouth (from about 6 months of age) and continues throughout life. There is evidence to suggest that the initial colonization of a baby's teeth with cariogenic bacteria may arise by infection from the mother's mouth. The common practice of sampling the food in a baby's dish, to check that it is not too hot, using the same spoon that is to be used to feed the baby may be a particularly effective way of transferring bacteria from carer to baby. Brushing the teeth with a toothbrush will remove part, but not all, of this film and its accompanying bacterial population. Many of the bacteria present are harmless, but a number of species are capable both of metabolically converting carbohydrates to acids (acidogenic bacteria) and of continuing to be metabolically active when the local pH has become too acid for most bacteria to tolerate. It is these bacteria that cause caries.

Fermentable Carbohydrate

Acidogenic bacteria metabolize (ferment) simple sugars (glucose, fructose, sucrose, lactose, and maltose) to acids. Sugars may be present as a result of their direct consumption or as a result of the enzymatic breakdown of starches within the mouth by salivary amylase. Thus, a substantial proportion of a typical diet will contain a source of fermentable carbohydrate, and many, if not all, eating and drinking occasions will give these bacteria one of these metabolic precursors. The more frequently an individual consumes carbohydrate, the more the acidogenic bacteria thrive and other, less acid tolerant, bacteria are disadvantaged.

A wide variety of foods contain carbohydrate that is capable of giving rise to acids as a result of bacterial metabolism (fermentation) within dental plaque. Of the common dietary sugars, sucrose, fructose, and glucose are found in fruit and fruit juices, soft drinks, jams, honey, chocolate and other confectionary, and an immense variety of composite foods and drinks. Lactose arises naturally in milk and milk products but is also widely used as an ingredient in its own right by the food industry.

Starches are also classed as fermentable carbohydrates because they are partially broken down by amylase in saliva during chewing to maltose and glucose. Residues of starchy foods are frequently caught between the teeth and in the fissures of the molar teeth, where they may be broken down to sugars over long periods. Measurements of the pH of plaque following the ingestion of starches have suggested that the depression of pH may be as great as and last even longer than that produced by some sources of sugars, such as drinks, because of slow clearance. Highly processed starchy products, such

as heat- and pressure-processed extruded snacks, are likely to be more readily converted to sugars than less processed starchy foods, such as bread.

Clearly, the wide range of individual dietary choices and eating habits may influence the risk of developing caries. The physical characteristics of fermentable carbohydrates will affect the rate at which they are cleared from specific sites in the dentition. Foods that are inclined to remain for long periods in stagnation sites (for example, between the teeth), such as toffees or raisins, are likely to give rise to a greater local fall in pH than are those that are rapidly cleared, such as chocolate. Clearance rates are also influenced by the increase in salivary flow that is stimulated by eating or drinking. When salivary flow is greater, for example after consuming a strongly flavored food, clearance will be faster and demineralization is likely to be less than that after consuming a bland food.

Susceptible Sites

Dental caries are more likely to occur at stagnation sites between teeth or in the fissures of molar teeth. Plaque will accumulate in these sites, where it is less likely to be disturbed by tooth brushing. At the same time, the protective buffering of saliva and the remineralization that arises from its mineral content are attenuated by the inaccessibility of such sites, while food debris is retained for longer periods. The reduction of salivary flow during sleep makes food debris remaining in these sites at night particularly damaging to the teeth.

Experimental Models of the Caries Process

Because direct manipulation of the caries process in human subjects is impossible for ethical reasons, a number of techniques have been developed that provide insights without risking clinical damage to the teeth of experimental subjects. Much of the earlier work relied on measurements of the change in plaque pH that followed a single consumption episode of a food or drink containing a source of fermentable carbohydrate. This approach provides an indication of the potential cariogenic challenge of these exposures and addresses the fundamental question of whether pH falls to a level that is expected to give rise to demineralization of the tooth enamel. Plaque pH measurements have thus been used to assess whether a food or drink may be considered safe for teeth. But this technique does not provide any information on the influence of the repair processes that follow exposure to a demineralizing challenge.

Approaches that provide an insight into the balance of demineralization and remineralization episodes over a period of time with naturalistic eating and drinking circumstances have now become more commonly used. These involve placing an enamel sample within a subject's dentition and carefully assessing any changes in the surface of this sample over a period of time. Particular cariogenic challenges can be applied, but, because they are continued for only a limited period of time, the subject's own teeth will not be appreciably affected. In many cases, the enamel sample is not cleaned with fluoride toothpaste, whereas the subject's own teeth are so protected (the additional enamel sample being removed while the teeth are brushed).

These models have provided useful information not only on the relative cariogenic potential of different foods and drinks but also on the protective effects of fluoride toothpaste (even though it may not be applied directly to the enamel sample) and on the influence of stagnation sites on caries risk with different dietary practices. Useful indications of answers to important public-health questions are beginning to emerge from this kind of research, such as the number of exposures to fermentable carbohydrate that can be tolerated without appreciable risk to the teeth, and the influence of fluoride toothpaste use on this number.

Etiology of Tooth Wear

The enamel surfaces of the crowns of the teeth may be damaged by wear arising from abrasion, attrition, or erosion. Abrasion can arise from the action of rubbing a hard substance across the surfaces of the teeth, for example when brushing too vigorously with a hard toothbrush. Attrition involves one tooth surface wearing down because of contact with another. A third form of wear involves the direct erosive action of acids present in foods (such as yoghurt or pickles) or drinks (especially citrus fruit juices). No bacterial metabolism is required for these processes to occur. It is unclear, however, whether the apparent increase in the prevalence of clinically apparent erosion of the teeth is the result of dietary habits or of some other factor. Only recently have dental-health surveys assessed this problem specifically, so it is possible that it has been noticed more, rather than actually occurring more, in these later surveys. It is also not always possible to distinguish acid erosion from other causes of tooth wear, such as over-vigorous tooth brushing. In addition, a common source of acid erosion is not dietary but arises from the regurgitation of the extremely acidic contents of the stomach. This is often seen in young children and, in adults, may be a presenting symptom of bulimia nervosa as a result of repeated vomiting.

Etiology of Gum Disease

Gum disease arises as a result of bacterial infection of the gums, especially at the tooth margins. It is often assumed that excessive accumulation of plaque, arising from inappropriate dietary habits, is a factor in this condition, but there is little evidence for any material influence of diet. The milder forms of gum disease are extremely common in all populations. More severe disease is the most frequent cause of tooth loss in older people. The best form of protection from gum disease is regular tooth brushing.

Protection from and Prevention of Dental Caries

Variations among individuals, and with time, will arise as a result of differences in acid generation from sugars at different localities within the dentition. These variations may be influenced by changing dietary habits and by the extent of the colonization of the relevant tooth surface by acidogenic bacteria. They may also be affected by changes in saliva flow, for example as a result of the use of certain medications or radiotherapy.

These factors may provide a reasonable explanation for many of the differences in caries experience observed between individuals and populations and between different locations within an individual's dentition. They do not explain the dramatic reduction in caries prevalence seen throughout the developed world in the last 30 years. There is no doubt that this improvement has been caused by the introduction of fluoride toothpaste.

Fluoride

Fluoride has provided the great success story in dental public health in the last 30 years. There are two main routes of delivery: water and toothpaste. Both are, in effect, dietary modifications.

The observation that tooth decay was less common in communities whose water supplies naturally contained low concentrations of fluoride led to the introduction of appropriate concentrations of fluoride into many public water supplies that did not naturally contain it. The prevalence of dental caries appeared to fall by between 20% and 50% as a result of this simple public-health measure.

There have been many thorough studies of the general health of populations receiving fluoridated water, which have found no credible evidence of adverse affects, except in a few areas where the fluoride content of the water is naturally very much higher than the level used for caries prevention. Despite vocal opposition to what is seen by some as compulsory medication of the population, many countries (for example, the USA and Ireland) still use this approach widely. Some, however, such as The Netherlands, have discontinued the practice.

Even greater improvements in dental health have followed the introduction of fluoridated toothpastes. The benefits seen at the population level from this innovation have been far greater than those predicted by the controlled clinical trials that preceded the widespread sale of fluoridated toothpastes to the public. Improvements of greater than 60% have been common. Interestingly, caries rates in The Netherlands continued to fall after the discontinuation of fluoridated water supplies, probably as a result of intense dental-health education of the population about the value of regular brushing with fluoridated toothpaste. In contrast, the abandonment of water fluoridation in the UK region of Anglesey was followed by a sharp rise in caries incidence.

A Practical Approach to the Prevention of Caries

The success of fluoridated toothpaste in preventing dental caries has resulted in a change in professional approaches to prevention. Instead of focusing simply on attempts to reverse the main causative factors, attention is now centered on exploiting protective influences. The interaction of the three main causative factors is illustrated in **Figure 1**. Numerous attempts to change the impact of any of these influences on caries have proved ineffective, except, perhaps, under the most extreme situations, such as during war time.

In contrast, exploiting the protective potentials of fluoride, tooth brushing, and salivary stimulation have proved successful. **Figure 2** illustrates the roles of these factors in comparison with the

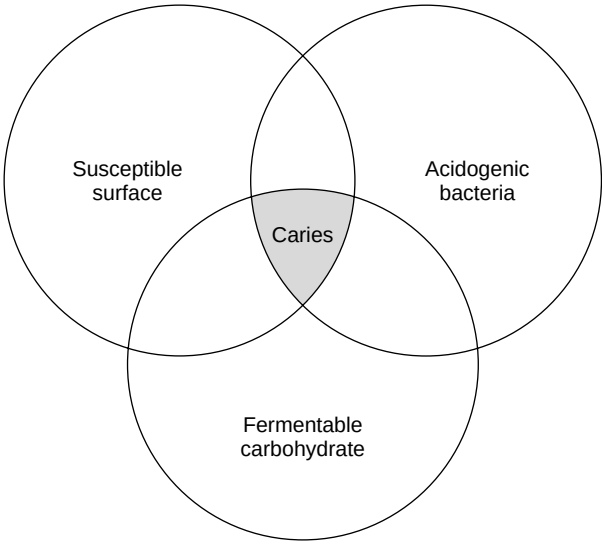


Figure 1 Interacting factors causing tooth decay.

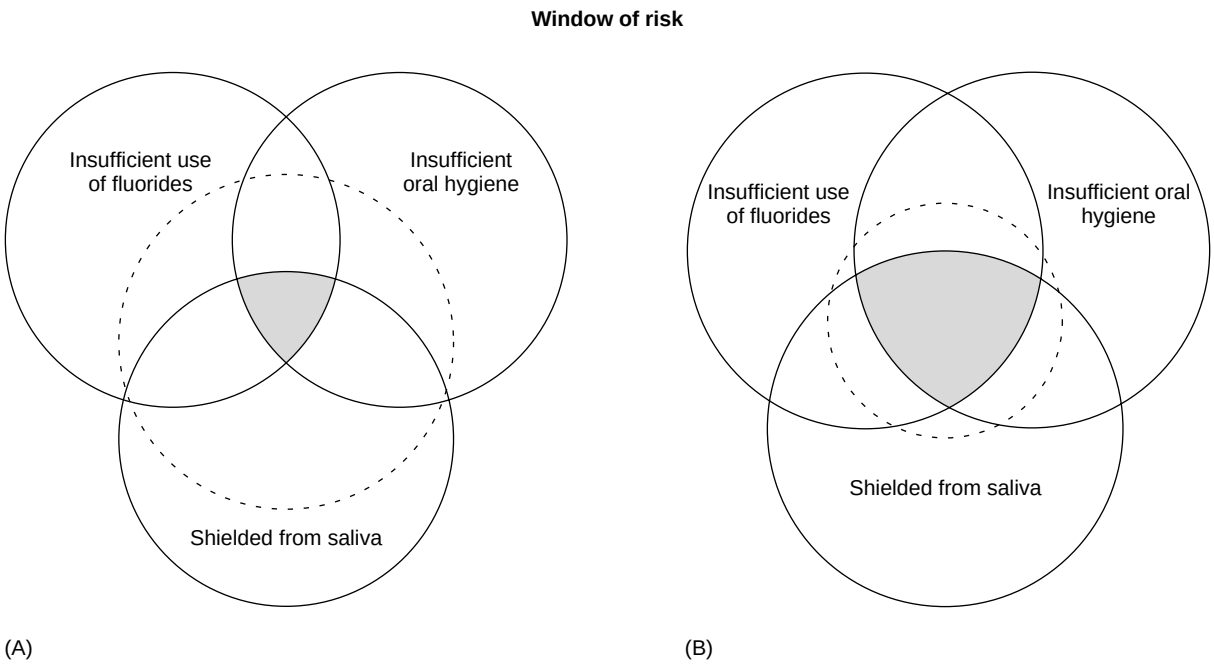


Figure 2 A new model to explain and guide caries prevention. The local factors – insufficient use of fluorides, insufficient oral hygiene, and protection from saliva – form a 'window of risk' through which the circle of cariogenic food (shown dashed) can be seen in the background. (A) In this example it is clear that it would be impossible to reduce the food circle to such an extent that the window is not completely filled (less caries risk). (B) If oral hygiene and, concomitantly, fluoride supply are neglected (large window of risk), a reduction in the burden of cariogenic food could reduce the caries risk. Reproduced with permission from van Loveren CM and Duggal MS (2001) The role of diet in caries prevention. *International Dental Journal* **51**: 399–406.

pervasive challenge of diet. Where a tooth site is shielded from saliva (stagnation site) and oral hygiene and the availability of fluoride are insufficient it is likely that a dietary modification of sufficient magnitude could exert some influence on the final outcome. But where these protective factors are adequate it is highly unlikely that dietary variations will exert any material effect. These predictions are borne out by epidemiological observations. In most developed countries, where fluoride use is adequate, wide variations in dietary exposure to fermentable carbohydrates between individuals are not accompanied by predictable differences in caries experience.

Epidemiology

Studies of Risk Factors

A large number of observational epidemiological studies have been conducted that have attempted to show associations between caries experience and one, or several, of the known risk factors. The large majority of these studies have been of poor design, examining insufficient numbers of subjects and ignoring important confounding influences. Many have been cross-sectional in design and have sought to draw conclusions about the causes of caries by assessing the dietary and other habits of subjects at the same time as measuring their caries experience. Such a study design is somewhat unsatisfactory for this purpose.

The few longitudinal studies in the scientific literature are equally weakened by poor data on dietary habits and, in some cases, idiosyncrasies in caries assessments. Taken together, these studies provide scant evidence about the relative importance of the different etiological factors. It is fortunate that more convincing evidence is available from experimental studies.

National Trends in Caries Prevalence

Data on the prevalence of dental caries within populations are nowadays very reliable as they are collected to internationally recognized standards. Surveys of 12-year-old children are carried out in most countries, and the data are collated by the World Health Organization (see **Table 1**). In contrast, data for adults are scarcer.

The general picture emerging from the repetition of these national surveys is clear. In many countries the prevalence of caries is falling, often dramatically. In poorer countries this is unlikely to be the case, and, even within the richest countries, the dental-health experience of the economically disadvantaged

Table 1 Prevalence of caries by region; the table shows the mean number of teeth with decay experience in 12-year-old children

Lowest DMFT	*Country*	*Year of Survey*	*Highest DMFT*	*Country*	*Year of Survey*
Europe					
0.9	Denmark	2001	7.3	Romania	1998
	The Netherlands	1992			
	Switzerland	2000			
	UK	2000			
Americas					
0.9	Belize	1999	8.1	Guatemala	1987
Africa					
0.3	Tanzania	1994	4.9	Mauritius	1993
	Togo	1986			
	Rwanda	1993			
Southeast Asia					
0.86	India	1993	3.0	North Korea	1991
Eastern Mediterranean					
0.9	Djibouti	1990	3.3	Jordan	1995
	Pakistan	1999			
Western Pacific					
0.6	South Korea	1972	4.9	Brunei Darussalam	1994
0.8	Australia	1999			
	Hong Kong	2001			

DMFT, decayed, missing, or filled permanent teeth.
Data obtained from the WHO Oral Health Country Profile Programme, WHO Collaborating Centre (website http:/www.whocollab.odont.lu.se/index.html).

is significantly poorer than that of those with a higher socioeconomic position. In many countries there is evidence that inequalities in dental health between the rich and the poor have widened.

Attempts to account for these trends are hampered by the unreliability of data on factors that are likely to attenuate caries risk. All assessments of these factors rely on people (often children) accurately remembering and reporting aspects of their everyday behavior, such as whether they clean their teeth and how often and what they have eaten and drunk and when. These data are subjective and notoriously unreliable. More secure conclusions about factors that have influenced caries rates must therefore come from more objective data (see below).

Fluoride Toothpaste

The effect, at a population level, of the introduction and widespread availability of fluoride toothpaste is clear. **Figure 3** shows the falls in caries incidence in 5-year-old and 12-year-old children in the UK seen in successive national representative surveys. A similar picture has been seen in Denmark (**Figure 4**). Fluoride toothpaste was introduced onto the UK market in around 1976 and rapidly became universal. The falls

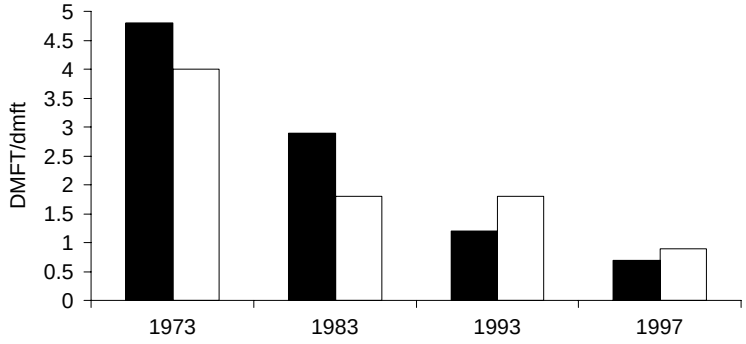


Figure 3 The change in the average decay experience of children in the UK. DMFT (decayed, missing, and filled permanent teeth) in 12-year-olds (filled bars) and dmft (decayed, missing, and filled primary teeth) in 5-year-olds (open bars). Data from OPCS (1973–1993) and NDNS (1997).

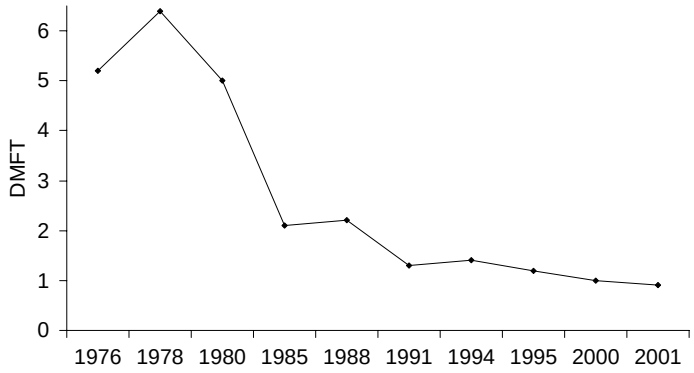


Figure 4 The change in caries experience of 12-year-old children in Denmark. DMFT, decayed, missing, or filled permanent teeth. Data obtained from WHO Collaborating Centre.

in caries prevalence seen at the next survey date (1983) exceeded expectations, based on earlier clinical trials, and led many experts to predict that no further fall would occur. In the event, an even greater decline was seen among 12-year-olds at the next decennial survey (1993). The caries prevalence among 5-year-olds appeared to have reached a plateau by 1993, but later data suggests that further falls in both age groups may have occurred. Regrettably, the self-reported use of fluoride toothpaste (almost certainly an overestimate of actual use) is still not universal, even among children in comfortable socioeconomic conditions.

The variation in caries experience with family income is illustrated for the UK in **Figure 5**. A clear gradient exists, with the poorest dental health seen in the lowest-income families. Trend data indicates that the greatest improvements have occurred among higher-income families and the least among those at the other end of the socioeconomic scale. The reasons for these differences are not entirely clear, but oral hygiene and the use of fluoride toothpaste appear to be important. Evidence of gum disease (an indicator of oral hygiene) is more common among poorer children.

Diet

Attempts to attribute the recent changes in caries prevalence to improvements in dietary habits have been unconvincing. Apart from the difficulty in determining what people are eating and drinking with any accuracy, data on when food and drink have been consumed are needed to assess the overriding dietary influence of frequency of exposure of the teeth to fermentable carbohydrate. These data are rarely collected in surveys and are then of uncertain reliability. All dietary surveys are seriously hampered by the unreliability of the subjective reporting of dietary habits by those surveyed.

The use of nationally aggregated data (such as food-supply data) is hardly more useful, since a large proportion (up to 50%) of the food available for consumption is never actually eaten. Nonetheless, some experts have pointed to changes in caries prevalence following dramatic changes in food supply as evidence of the practical utility of dietary

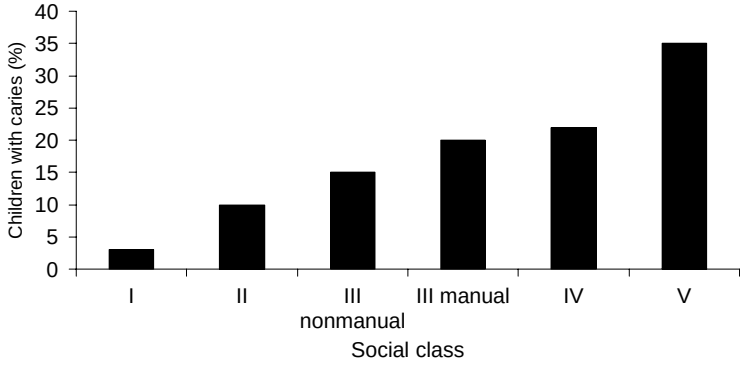


Figure 5 Percentage of children aged 1.5–4.5 years with caries across social-class groups in the UK, calculated from data presented in Hinds and Gregory (1995). Social class I has the highest income, social class V has the lowest.

manipulation as a means of reducing the remaining burden of this disease. The weakness of this argument is that predictable changes in caries prevalence at a population level have rarely been seen except under conditions of extreme dietary change, such as during war time. These changes all occurred before the advent of the widespread use of fluoride. When the food supply of fermentable carbohydrates is severely restricted, changes in the frequency of consumption may occur to a degree that is sufficient to alter caries risk when fluoride is not used. Where fluoride toothpaste and oral hygiene are adequate, even such extreme changes in diet are unlikely to alter caries experience materially. In addition, attempts to use dietary manipulation to reduce the risk of caries in free-living populations have proved unsuccessful.

Other Factors Affecting the Epidemiology of Caries

The influence of other factors that might be expected to have a bearing on caries experience has proved difficult to establish for practical reasons. These include the susceptibility of particular sites within the dentition or in an individual's mouth and local salivary flow rates. Both of these factors are known to be strongly influenced by genetic inheritance. The morphology of the teeth and, especially, the depth and shape of the fissures on the surfaces of the molar teeth are strongly heritable. It is generally difficult to predict, in advance of caries developing, which sites will be particularly susceptible. But one successful preventative approach has been to identify children with deep fissures in their molar teeth at an early age and offer prophylactic treatment in the form of sealants. This addresses the most common site of early childhood caries (the molars) and targets those children most at risk because of unfavorable tooth morphology.

The rate of salivary flow, both at rest and when stimulated by eating or drinking, is also known to be a crucial influence on risk. Patients who have had a salivary duct removed for any reason have a far higher risk of caries than those with normal function. Some people have low salivary flow rates and, again, have a greater risk. Older people are inclined to suffer from a dry mouth, and a number of medications reduce salivary flow.

Epidemiological studies to assess the importance of salivary flow rates in altering the risk of caries have not been carried out because of the practical difficulty of measuring this factor. But the stimulation of salivary flow that accompanies chewing has been successfully exploited to reduce caries risk in experimental studies using chewing gum (usually sugar-free). Reductions in caries incidence were seen when subjects were encouraged to chew the gum, especially between and immediately after meals, while continuing their normal regular oral-hygiene practices. Convincing evidence of an effect at the population level, however, is awaited.

See also: **Calcium**. **Carbohydrates**: Resistant Starch and Oligosaccharides. **Fructose**. **Galactose**. **Glucose**: Chemistry and Dietary Sources. **Sucrose**: Dietary Sucrose and Disease. **Vitamin D**: Rickets and Osteomalacia.

Further Reading

Beighton D (2004) The complex microflora in high risk individuals and groups and its role in the caries process. *Community Dentistry and Oral Epidemiology* **32**: In Press.

Bratthall D, Hansel Petersson G, and Sundberg H (1996) Reasons for the caries decline: what do the experts believe? *European Journal of Oral Science* **104**: 416–422.

Curzon MEJ and Hefferen JJ (2001) Modern methods for assessing the cariogenic and erosive potential of foods. *British Dental Journal* **191**: 41–46.

Duggal MS, Toumba KJ, Amaechi BT *et al.* (2000) Enamel demineralization *in situ* with varying frequency of carbohydrate consumption with and without fluoride toothpaste. *Journal of Dental Research* **80**: 1721–1724.

Food and Agriculture Organization (1998) *Carbohydrates in Human Nutrition.* FAO Food and Nutrition Paper 66. Rome: FAO.

Gibson S and Williams S (1999) Dental caries in pre-school children: associations with social class, toothbrushing habit and consumption of sugars and sugar-containing foods. Further analysis of the national diet and nutrition survey of children aged 1.5–4.5 years. *Caries Research* **33**: 101–113.

Gustaffson BE, Quensel CE, Lanke LS *et al.* (1954) The Vipeholm dental caries study: the effect of different levels of carbohydrate intake on caries activity in 436 individuals over five years. *Acta Odontica Scandanavica* **11**: 232–365.

Hausen H (2003) Fluoride toothpaste prevents caries. *Evidence Based Dentistry* **4**: 28.

Hinds K and Gregory JR (1995) *National Diet and Nutrition Survey: Children Aged 1.5–4.5 Years.* vol. 2. Report of the Dental Survey. London: HMSO.

Marquis RE, Clock SA, and Mota-Meira M (2003) Fluoride and organic acids as modulators of microbial physiology. *FEMS Microbiology Reviews* **26**: 493–510.

Reich E (2001) Trends in caries and periodontal health epidemiology in Europe. *International Dental Journal* **51**: 392–398.

van Loveren CM and Duggal MS (2001) The role of diet in caries prevention. *International Dental Journal* **51**: 399–406.

Woodward M and Walker ARP (1994) Sugar consumption and dental caries: evidence from 90 countries. *British Dental Journal* **176**: 297–302.

World Health Organization (1994) *Fluorides and Oral Health.* WHO Technical Report Series 846. Geneva: WHO.

DIABETES MELLITUS

Contents

Etiology and Epidemiology

J Sudagani and G A Hitman, Queen Mary's, University of London, London, UK

Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by disturbance in glucose metabolism leading to a state of hyperglycemia and is associated with microvascular and macrovascular complications in the long term. Diabetes is the leading cause of noncommunicable diseases worldwide and it is true to say that diabetes has reached epidemic proportions in certain parts of the world and in certain ethnic groups. This has widespread implications for health resources.

Diabetes mellitus is an etiologically and clinically heterogenous group of disorders that share hyperglycemia in common. The two main types of diabetes, type 1 diabetes (T1D) and type 2 diabetes (T2D), are quite distinct from each other in their etiology and epidemiology. T2D is the most common form of diabetes worldwide accounting for 90% of cases globally and affecting approximately 4% of the world's adult population. Type 1 diabetes is an autoimmune disease that results in insulin deficiency. Both T1D and T2D are distinct from 'other causes of diabetes' as defined by the etiological classification of the World Health Organization (WHO); we will not comprehensively review the many types of diabetes but illustrate it with maturity onset diabetes of the young (MODY) and fibrocalculous pancreatic diabetes (FCPD). Gestational diabetes is the fourth category defined by the WHO.

Type 1 Diabetes

Worldwide Prevalence

In 1997 there were 11.5 million people with T1D in the world; this figure is expected to rise to 23.7 million in the year 2010. These increasing figures will have most impact in Asia, where there are currently 4.5 million people with T1D, and this is expected to rise to 12 million by the year 2010. One of the best incidence studies has come from Europe as part of a European collaboration, where the highest incidence of T1D is found in Finland and the lowest rates in Romania (**Table 1**). The incidence of T1D follows a north–south gradient, with the notable exception of Sardinia. The figures from countries such as India are less precise, although one study in Chennai suggested an incidence equivalent to that found in Southern European countries. These different rates of T1D are likely to reflect both the genetic background of individual countries and differences in exposure to environmental agents. In recent years,

Table 1 Extremes of incidence of childhood type 1 diabetes mellitus in different ethnic groups

Higher	*Incidence*[a]	*Lower*	*Incidence*[a]
Sardinia	35–40	Venezuala	0–5
Finland	35–40	Peru	0–5
Sweden	25–30	China	0–5
Canada	20–25	Paraguay	0–5
Norway	20–25	Mauritius	0–5
UK	15–25	Chile	0–5
NewZealand	10–25	Japan	0–5
Portugal	5–20	Barbados	0–5

[a]Age standardized incidence (per 100,000 per year) of type 1 diabetes in children <14 years of age.
Data from Karvonen M, Viik-Kajander M, Moltchanova E *et al* (2000) Incidence of childhood type 1 diabetes worldwide. Diabetes Care **23**: 1516–1526.

the incidence of T1D has been increasing in several different countries. These changes must reflect environmental influences.

Etiology

Type 1 diabetes is due to autoimmune destruction of insulin-secreting pancreatic β cells of islets of Langerhans. T1D typically occurs in young individuals with an age of onset of less than 40 years. The autoimmune reaction is likely to be triggered by an environmental agent *in utero* or in very early life (**Figure 1**). The earliest markers of β cell destruction are the appearance of autoantibodies to glutamic acid decarboxylase(GAD), islet cells, and insulin. Autoantibodies have been detected 10–15 years before the onset of disease and, furthermore, have been known to disappear without T1D occurring in a few individuals. One to two years before onset of the disease, evidence of β cell impairment can be detected, initially evidenced by a reduction in the first phase of insulin response to intravenous glucose and in the later stages by an abnormal oral glucose tolerance. In contrast to the slow β cell destruction, the onset of T1D is acute and is usually measured in weeks. At this stage in the etiological process, it is likely that 70% of β cells have been destroyed and those remaining are inhibited by the action of cytokines.

There is a subgroup of patients who develop diabetes in adult life and do not require insulin during the first few years after diagnosis; they have an autoimmune component to their disease with positive GAD and islet cell antibodies. This condition is named latent autoimmune diabetes (LADA). There are several common features between T1D and LADA, including T cell insulitis, islet antibody positivity, and high rates of HLA DR3 and DR4. The prevalence of LADA in newly diagnosed diabetics has been shown to range from 2.8% to 22.3% in different studies depending on the markers used and characteristics of the patients. Although these patients present with type 2 diabetes, they have been shown to progress to insulin dependency especially if the diabetes is diagnosed at a younger age and the patient is not overweight. Therefore, there may be a role for measuring GAD antibody in newly diagnosed patients with type 2 diabetes to identify the LADA subgroup especially in the younger age groups. This group of patients is now classified as type 1 diabetes by the new WHO classification.

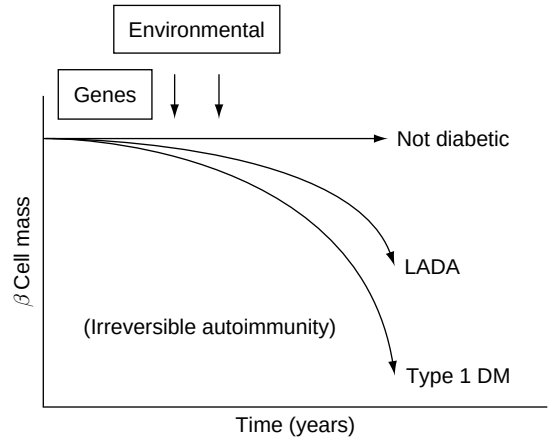


Figure 1 The etiology of type 1 diabetes mellitus. DM, diabetes mellitus; LADA, latent autoimmune diabetes.

Genetics

Type 1 diabetes is a multifactorial disease with both genetic and environmental components. The largest genetic contribution to T1D is determined by genes in the major histocompatibility complex (MHC) located to the short arm of chromosome 6 (IDDM1-HLA, 6p21). Initial associations between T1D and the MHC were described for the HLA class I antigens A1-B8 and B15. With advent of HLA class II serology, closer associations were found with HLA-DR with an increased frequency of DR3 and DR4 and a decreased frequency of DR2 in T1D subjects. At the population level the strongest genetic association with T1D is with HLA-DQ alleles. This is best defined by DNA typing of HLA-DQ1, DQB1, and DRB1. However, due to the strong linkage disequilibrium between these loci it has been very difficult to study the effect of individual HLA-DQ or HLA-DR genes separately. For the individual, susceptibility is best defined by allelic combinations of MHC genes located to all three major regions (classes I, II, and III) called HLA haplotypes. Haplotypes occur because of strong linkage disequilibrium observed in the MHC whereby the combinations of alleles are seen more frequently than would be expected by their individual gene frequencies. An example of a haplotype would be A2, Cw1, B56, TNFa6, DRB1*401, DQA1*0301, DQB1*0302. The haplotypes are likely to relate to functional groups of genes involved in the etiology of T1D. Thus, the critical residues of DR and DQ, accounting for the disease association with T1D, are located in the antigen-binding cleft of the HLA molecule and are likely to influence the binding of antigenic peptides for subsequent presentation to T helper cells. Similarly, polymorphism of the HLA class I molecules are likely to relate to antigen presentation to cytotoxic T cells, and polymorphisms of tumor necrosis factor (TNF) have been associated with differing TNF responses to mitogenic stimulation.

The MHC accounts for approximately 40% of the genetic component to T1D. Evidence from genome scans and candidate gene studies indicates the existence of a large number of putative non-MHC genes contributing to the etiology of T1D, although all of comparatively small effect compared to the MHC. The most reproducible T1D associations have been found with the insulin gene, cytotoxic T lymphocyte antigen 4 gene (*CTLA4*), and the vitamin D receptor. An association between the insulin gene (located on chromosome 11p15.5; *INS*), and T1D was described in the 1980s and subsequently confirmed by linkage studies. The *INS* locus on chromosome 11p15.5 contains a major polymorphism 5′ to the transcription site, which is a variable number of tandem repeats (VNTR) region. One functional hypothesis to explain the association between the insulin gene and T1D is that 'hypersecretors' of insulin determined by the disease-associated polymorphism might induce thymic tolerance to insulin, thus providing protection from the autoimmune reaction. A recent study showed association of the T-cell regulatory gene *CTLA4* with susceptibility to autoimmune disease, including type 1 diabetes. *CTLA4* (gene located on chromosome 2q33) plays an important role in the counter-regulation of CD28 T cell antigen receptor activation of T cells. In the mouse model of T1D, susceptibility was also associated with variation in *CTLA4* gene splicing with reduced production of a splice form encoding a molecule lacking the CD80/CD86 ligand-binding domain. There are associations reported between vitamin D receptor gene polymorphisms and type 1 diabetes. The *VDR* gene is located on chromosome 12q and polymorphisms of the *VDR* gene may be related to T-cell-mediated autoimmune destruction of β cells of pancreas. Vitamin D compounds suppress T cell activation and significantly repress the development of insulitis and diabetes in the nonobese diabetic (NOD) mouse, a mouse model of human type 1 diabetes. Results from other candidate gene studies and a 'total genome' analysis have identified at least another 19 chromosomal regions that may be involved in pathogenesis of the disease. The finding that so many genes are involved in T1D raises the possibility that there are several disease processes that might lead to β cell destruction. Furthermore, evidence is still emerging that the genetic susceptibility to T1D is graded both within and among populations.

Environmental Factors

Environmental factors play a significant part in the etiology of T1D and have been implicated in both initiation and progression of β cell damage. The majority of evidence points to the effects of viruses and/or dietary factors as etiological agents.

Many viruses have been implicated in the pathogenesis of T1D; they may have a direct effect on β cells by infection and cell lysis, or alternatively they may act as triggers to the autoimmune process. Among the viruses that have been implicated in humans are coxsackie A, coxsackie B, rubella, cytomegalovirus, mumps, and Epstein-Barr viruses. The enteroviruses (Coxsackie A, Coxsackie B, and Echovirus) are the most commonly associated viruses with diabetes and serve as a major trigger for T1D in the young possibly by induction of islet cell antibodies. The evidence for viral involvement in type 1 diabetes came from several sources, including anecdotal case reports, epidemiological studies, seasonal incidence studies, and animal models. There is data to support the theory that enterovirus infection either accompanies or precedes the development of T1D in young people in many instances.

Coxsackie B was first implicated in the early 1970s by Gamble, who found an increased titer of coxsackie B antibodies in newly diagnosed T1D patients. More recently, coxsackie virus has been identified in very young onset type 1 diabetes (in patients under 5 years of age) using the polymerase chain reaction. Furthermore, when the coxsackie virus was sequenced, although it had extensive homology to coxsackie virus B4, there was some unique sequence variation indicating an T1D variant. There have also been many anecdotal reports of coxsackie B virus causing T1D, presumably by a direct cytolytic effect on β cells. A previously fit child died in diabetic ketoacidosis 3 days after a flu-like illness. At, necropsy, there was an extensive lymphocytic infiltration into the β cells of pancreas and coxsackie B4 was found in the child's serum. This virus was extracted from the pancreas and, when used to infect mice, led to diabetes.

There is a high incidence of T1D among patients with the congenital rubella syndrome. Clearly, this results from an *in utero* infection, but the diabetes that ensues is indistinguishable from the primary type 1 diabetes: the disease presents in the second decade of life, the onset is preceded by islet cell antibodies, and the genetic predisposition is defined by the same HLA association as T1D. This is likely to be a good example of a virus triggering the immune process.

Dietary factors have also been implicated in the development of T1D. Among the dietary factors indirectly linked to either susceptibility or protection to T1D are cows' milk protein (including bovine serum albumin and β-lactoglobulin), β-cell-toxic drugs (alloxan, streptozotocin, rodenticides), dietary toxins (in particular nitroso-containing compounds), and

others such as coffee and sugar. There is an interesting interplay between vitamin D, vitamin D receptor (VDR), and association with T1D as discussed earlier. The contribution of vitamin D as a potent modulator of the immune system is well recognized. The main sources of vitamin D are ergocalciferol and cholecalciferol found in dietary sources and cholecalciferol produced in the skin by ultraviolet radiation of 7-dehydrocholesterol. Vitamin D deficiency in infancy and VDR polymorphisms may be risk factors for T1D. In nonobese diabetic (NOD) mice, long-term treatment with high doses of vitamin D_3 reduced the incidence of diabetes by changing the cytokine balance at the local pancreatic lesion.

The Future

The working out of the complete genetic basis of type 1 diabetes will lead to a better understanding of disease pathogenesis and, through studies of genetic and environmental interaction, direct evidence of environmental factors. T1D may be the first multifactorial disease to benefit from primary prevention of both insulitis (before immune process has been initiated) and of T1D (once autoantibodies have been detected). In the Diabetes Prevention Trial, low-dose insulin was administered to persons with high risk of T1D as ascertained by family history, islet antibodies, and HLA typing. It was concluded that low-dose insulin does not delay or prevent the onset of T1D. In a European study (European Nicotinamide Diabetes Intervention Trial, ENDIT) high-dose nicotinamide was used to protect the β cells in high-risk individuals for type 1 diabetes; unfortunately this trial also failed to show benefit on active treatment. However, 90% of type 1 diabetes patients do not have a family history of type 1 diabetes. The approach in this latter group might be to identify the genetically susceptible by the use of genetic markers and then test for autoantibodies. If the latter subjects are autoantibody positive then intervention may be considered in the future. The sensitivity and specificity that would be required for such sequential testing would depend on how safe the proposed intervention would be.

Type 2 Diabetes

Worldwide Prevalence

Type 2 diabetes is one of the most common noncommunicable diseases in the world with an estimated 147.2 million people suffering from this disorder; by 2010 this figure is expected to reach 212.9 million. Furthermore, it has been predicted that by the year 2010 over half the people with T2D will be living in Asia. This trend is likely to be due to increasing urbanization and industrialization. According to WHO estimates the figure is likely to double by the year 2025. The prevalence of T2D varies widely from the highest in Pima Indians (almost half of the population affected) to the lowest in Rural Africa (1%). As with T1D, the incidence of diabetes in different countries is likely to reflect the different genetic architecture as well as the differing environment. A good example is afforded by the population of Nauru. In full-blooded Nauruans over the age of 60 years the prevalence of T2D is 83%, whereas in those with genetic admixture as adduced by HLA typing the prevalence is 17%; this clearly reflects the genetic component. However, the rapid increase of T2D in the world in the last few decades, and the rise and a recent decrease in prevalence of T2D in the Nauruan community, can only be ascribed to environmental factors. This illustrates the multifactorial nature of T2D, with strong genetic and environmental contributions.

Etiology

Type 2 diabetes is a multifactorial disease with genetic and environmental factors playing a key role in its pathogenesis. Central to the etiology is a defect in insulin action, hepatic glucose output, and insulin secretion. Although insulin resistance is frequently the first detectable abnormality in the progression of T2D, insulin resistance by itself does not cause the disease, which is only manifested when there is a coexisting insulin secretory defect. T2D typically occurs in middle-aged and elderly people but there is an increasing trend of T2D occurring in young individuals. The main question yet to be answered is whether T2D is one disorder or a group of disorders with hyperglycemia as the end point in disease pathogenesis. Insulin resistance is common to several other disorders, including ischemic heart disease, hypertension, dyslipidemia, central obesity, and coagulation defects; the clustering of these disorders is known as the metabolic syndrome or the insulin resistance syndrome. The interface of T2D with obesity is a complex one, highlighted by the discovery of leptin and adiponectin. The cause of obesity and T2D in the ob mouse is a mutation of the *ob* gene. With administration of the ob gene protein (leptin) the ob mouse decreases its food consumption and increases exercise, leading to a dramatic weight loss; if given early enough it will also prevent diabetes. In contrast, common human obesity is associated with increased leptin levels, and which have been found to correlate with hyperinsulinemia. The newly discovered protein adiponectin signals adipose tissue mass; reduced

levels are found in obese subjects and there is an important interplay between adiponectin, insulin resistance, T2D, and atherosclerosis. Ghrelin is a gut hormone that is a signal of satiety and therefore has a direct effect on obesity. In the obese subject with T2D, there may be interplay between leptin, adiponectin, ghrelin, and insulin, contributing to insulin resistance and the metabolic syndrome.

Genetics

Type 2 diabetes is a complex disease and the heterogeneity both at the phenotypic and pathophysiological level indicates that the genetic component is likely to be heterogeneous with no single locus accounting for the disease. There are many strands of evidence to support a strong genetic component to T2D; these include a near 100% concordance in identical twins, familial clustering, genetic admixture and migration studies, complex segregation analysis, and the detection of gene variants leading to diabetes including the identification of genes responsible for human monogenic diabetes (MODY – see below).

Many groups worldwide have completed the first stages of genome scans for genes that predispose an individual to T2D. Currently, a large international research effort is being directed to those diabetes-associated linkage peaks that are overlapping in several genome scans; specifically on chromosomes 1, 12, and 20. One genome scan has been taken to completion with the identification of the calpain10 gene (located to chromosome 2q37) as a major susceptibility gene in Mexican-Americans. The majority of subsequent studies have confirmed an association between calpain10 and T2D as well as insulin action, insulin secretion, endothelial function, and aspects of adipose metabolism. This putative diabetes susceptibility gene encodes a ubiquitously expressed member of the calpain-like cysteine protease family, calpain-10. Functional studies would suggest a role in insulin secretion, insulin action, and adipocyte metabolism. For instance, reduced levels of calpain10 have been found in skeletal muscle associated with disease-associated polymorphisms and inhibition of calpain affects insulin secretion and translocation of glucose transporter-4 in an adipocyte cell line.

A large number of candidate genes have been studied in T2D. Only a few have produced consistent results, i.e., the genes for the insulin receptor substrate-1, insulin, *KCNJ11,* peroxisome proliferator-activated receptor gamma (*PPARγ*), and are critically dependent on the power of individual studies. Insulin receptor substrate 1 (*IRS-1*) is a protein involved in insulin signaling. After insulin binds to its insulin receptor, it stimulates autophosphorylation of its β chain, which, in turn, leads to phosphorylation of several multisite insulin receptor substrate (*IRS*) docking proteins including *IRS-1*. This then generates one of the signals for insulin action. Several variants of *IRS-1* have been detected, one of which (due to the substitution of a glycine for arginine at position 972 (G972R) in the molecule) is associated with insulin resistance, but only in the presence of obesity. This is a good example of a gene variant that is common in the population (at least 8% in the white population) and will only lead to disease in association with other contributing factors for diabetes. Recently a meta-analysis has confirmed the role of the G972R variant in T2D with an odds ratio of 1.25.

KCNJ11 encodes *Kir6.2*, which is an essential subunit of the pancreatic β cell potassium ATP (K_{ATP}) channel. Rare mutations of this locus lead to the monogenic syndrome of familial hyperinsulinemia, confirming the important role of *KCNJ11* in insulin secretion. Although there are a large number of rare variants of the *KCNJ11* gene only one common variant (E23K) has been associated with T2D, although by no means consistently in all studies. It is likely that the studies mentioned above were underpowered as a meta-analysis, that additionally included a large new study, has demonstrated a significant odds ratio for T2D of 1.23. However, it should be borne in mind that the E23K variant does not change protein function and therefore it is unlikely to be the predisposing mutation.

The *PPARγ* gene is mainly expressed in adipose tissue and is the target of the thiazolidinedione class of drugs used to treat T2D by improving insulin action and secretion. In man, rare mutations of *PPARγ* are associated with a monogenic syndrome of severe insulin resistance, T2D, and hypertension. In contrast, a common amino acid polymorphism (Pro12Ala) in *PPARγ* has been shown to be associated with a increased risk of typical T2D, confirmed by a meta-analysis with an odds ratio for diabetes of 1.25 for the common proline allele. Furthermore, a gene nutrient interaction has been demonstrated with an important interaction of the Pro12Ala variant and the ratio of dietary polyunsaturated fat to saturated fat in the diet. *PPARγ* is a nuclear receptor, which upon activation stimulates the transcription of genes responsible for growth and differentiation of adipocytes. This clearly indicates a role for *PPARγ* in fat cell biology and pathophysiology of obesity, diabetes, and insulin resistance.

One of the first candidate genes to be studied in T2D was the insulin gene with an association described with the class 3 allele. As mentioned in the section on the genetics of T1D the insulin gene hypervariable region is a determinant of insulin secretion. Although consistent associations were found of the insulin gene and T1D, this was not the case for T2D. However, the earlier T2D studies were very much underpowered and the controls and patients not well matched. Using a family-based design the association between T2D and the class 3 allele has been confirmed demonstrating paternal but not maternal transmission. This is in keeping with the fact that the insulin/insulin growth factor II gene locus is maternally imprinted. In addition to the importance of adequately powered studies, the paternal transmission is likely to be another explanation for variable results in a case–control study design, emphasizing the importance of family-based designs as an additional strategy in association studies. The insulin/insulin growth factor II gene locus is a determinant of fetal and postnatal growth, which is also an important factor in susceptibility to T2D.

Environmental Factors

Evidence of a strong environmental element to T2D has come from the studies of Barker and Hales. In a number of separate studies, a strong relationship of the development of glucose intolerance and other associated factors of the insulin resistance syndrome with low birth weight or thinness at birth has been demonstrated. Furthermore, these associations are not confined to those with growth retardation *in utero* but extend to the whole range of birth weights. As a consequence of these epidemiologic studies, the 'thrifty phenotype' hypothesis has been proposed, whereby nutritional deficiencies *in utero* lead to poor fetal and infant growth and the subsequent development of T2D in later life, especially when combined with obesity due to excess food intake and lack of physical activity. These changes are recognized to be due to insulin resistance, which is favorable for survival in the immediate postnatal period but plays a significant role in the progression to T2D and metabolic syndrome, and to a certain extent insulin secretion. While there is much discussion regarding this hypothesis, it illustrates the importance of environmental factors in early life, which might prime the fetus for T2D in later life.

Dietary factors and physical inactivity undoubtedly affect the progression of abnormal glucose tolerance to diabetes in a genetically predisposed individual. The best way to lower the risk of diabetes is to lead a healthy life style by eating a healthy balanced diet, engaging in regular physical activity, and balancing the energy intake with energy expenditure. Indeed, recent evidence would suggest that the adoption of a healthy life style in high-risk subjects can decrease the risk of developing T2D by 60%. There is a close relationship between diabetes and obesity, especially when the latter has central distribution. Apart from obesity, several other nutritional factors affect glucose metabolism and the risk of T2D. Current evidence suggests an association between different types of fats and carbohydrates and insulin resistance and T2D. Diets rich in saturated fats are associated with insulin resistance; a multicentre study in a group of healthy individuals showed that a diet high in saturated fat decreased insulin sensitivity compared with a diet high in monounsaturated fat with the same total fat content. Prospective and cross-sectional studies suggest a role of specific types of fat rather than the total fat content in the development of T2D, where high intake of vegetable oils, oils consisting primarily of polyunsaturated fat, was associated with reduced risk of developing diabetes and a positive association between saturated fat and hyperglycemia or glucose intolerance. In a 20-year follow-up of the Finnish and Dutch cohorts of the Seven Countries Study, it was found that a high intake of fat (in particular saturated fatty acids) contributed to the risk of glucose intolerance and T2D. Dietary carbohydrates are classified into simple or complex carbohydrates depending on their chemical structure. The traditional view is that simple carbohydrates be avoided and substituted with complex (starchy) carbohydrates to reduce postprandial glucose response, but this has been challenged by various studies that recognized that starchy foods such as baked potatoes and white bread produce even higher glycemic responses than simple sugars. Glycemic index (GI) was developed to quantify the different glycemic responses induced by different carbohydrate foods. A low GI diet with a greater amount of fiber and minimally processed whole-grain products seems to improve glycemic and insulin responses and lowers the risk of T2D. This shows that dietary recommendations to prevent and manage diabetes should focus more on the quality of fat and carbohydrate than the quantity alone.

A number of environmental toxins have been shown to cause diabetes in humans, including nitrosated compounds, as well as streptozotocin, the rat poison Vacor, and foods such as smoked mutton; depending on the amount consumed they could lead

to either T1D or T2D, presumably dependent on the amount of direct β cell destruction. It has also been proposed that vitamin D might modulate the diabetic process. Vitamin D deficiency has been shown to reduce insulin secretion. In a UK study of Bangladeshi subjects living in east London who were particularly prone to vitamin D deficiency, vitamin D levels were found to be low in those most at risk of diabetes. Furthermore, there was a correlation between vitamin D levels and 30 min oral glucose tolerance test, blood glucose, insulin, and C-peptide levels.

The Future

Research into the identification of genes involved in T2D is beginning to lead to insights into the pathogenesis of this common condition. With knowledge of the precise biochemical variants involved in disease pathogenesis, we will be in a better position to classify the disease and design more rational therapeutic maneuvers to prevent and ameliorate this condition. Research also needs to be directed at the gene–environment interaction, as this will indicate the appropriate population strategies to combat the increasing incidence of this common noncommunicable disease. Recently, life style interventions (healthy diet and exercise) have demonstrated significant reduction in onset of diabetes in high-risk individuals for T2D. In the future genetic profiling might be used to identify those likely to respond to such strategies including pharmacological treatment.

Other Types of Diabetes

Maturity Onset Diabetes of the Young (MODY)

MODY are a group of monogenic disorders inherited in an autosomal dominant pattern. MODY is characterized by early onset (usually before the age of 25 years) of T2D β cell dysfunction and there being a family history (at least two generations) of early onset diabetes. The defect is in insulin secretion due to mutations in the glucokinase and β cell transcription factor genes (**Table 2**). Hepatocyte nuclear factors *(HNF) 1*α, *1*β, and *4*α, insulin promoter factor (*IPF1*), and neurogenic differentiation-1 (*NEUROD1*) play an important role in the normal development and function of the β cells of the pancreas. In the UK mutations in *HNF1*α is the commonest cause of MODY accounting for 63% of cases, followed by mutations in the glucokinase gene (20% of cases). The clinical presentation and progression of diabetes is different among patients with mutations of glucokinase, *HNF1*α, and *HNF1*β. Subjects with glucokinase mutations are frequently asymptomatic but can be identified when diagnosed with gestational diabetes or with a milder form of diabetes, which is frequently treated with diet alone and is not associated with the complications of diabetes. In contrast subjects with *HNF1*α mutations are more like lean patients with T2D with susceptibility to microvascular complications and progressive loss of β cell function exacerbated by increasing body mass index. In comparison to patients with T2D, subjects with *HNF1*α mutations are very sensitive to sulfonylurea treatment as might be predicted from the genetic defect. Finally, patients with *HNF1*β mutations in addition to T2D have renal cysts that may lead to renal failure and hence such patients are more frequently found in the renal clinic.

Table 2 Maturity onset diabetes of the young

MODY subgroup	*Gene*	*Chromosome*	*MODY frequency*
MODY1	*HNF4*α	20q	Rare
MODY2	*GCK*	7p	10–65%
MODY3	*HNF1*α	12q	20–75%
MODY4	*IPF1*	13q	Rare
MODY5	*HNF1*β	17q	Rare
MODY6	*NEUROD1*	2q	Rare

Gestational Diabetes

Gestational diabetes (GDM) is defined as glucose intolerance first recognized in pregnancy. This therefore, excludes those women with either type 1 diabetes or type 2 diabetes diagnosed before conception. GDM is a relatively common occurrence in pregnancy affecting 1–14% in White European and North American populations and higher in certain ethnic groups such as South Asian and Afro-Caribbean populations.

GDM increases the risk to both mother and fetus although the levels of maternal glycemia that leads to an adverse outcome are not well defined. Furthermore, there is controversy as to who should be screened in pregnancy and the best available diagnostic test that has high sensitivity and specificity and the timing of the test during gestation. This is reflected in the lack of international agreement on diagnostic criteria ranging from the WHO critieria to a more pragmatic approach based on fasting and post prandial glucose levels. Those at particular risk for GDM are ethnic groups with a high prevalence of diabetes, women with a previous history of

delivering large babies, a family history of diabetes, obesity, older women and multiparous women.

Pregnancy is associated with an increase in insulin resistance and an increase in hormones with actions opposing insulin (i.e. cortisol, progesterone, growth hormone and human placental lactogen). GDM develops at a time when the beta cell reserve cannot cope with the prevailing state of insulin resistance. Therefore by definition after pregnancy when the insulin resistance reduces, then the subject becomes normoglycemic. However, GDM can be considered a pre-diabetic condition with an increased future risk of developing type 2 diabetes during their lifetime approaching 20–50%. The treatment of GDM consists of maintaining strict glycemic control. If dietary intervention fails to achieve normoglycemia, the treatment of choice is insulin.

The maternal risk due to GDM include increased risks in pregnancy, accelerated fetal growth leading to macrosomia and increased rates of caesarian section. The fetal risks include stillbirth, congenital malformations, shoulder dystocia, birth trauma and the risk of neonatal hypoglycemia and calcium and bilirubin disturbances in the neonatal period. There is also an increased risk of the child subsequently becoming obese and developing diabetes in adult life as a result of *in utero* hyperglycemia.

Fibrocalculous Pancreatic Diabetes

In tropical countries there is a form of nonalcoholic chronic pancreatitis characterized by pancreatic exocrine and endocrine insufficiency and associated with pancreatic calcification. This disease, tropical calcific pancreatitis, affects young individuals who are malnourished and present with abdominal pain, extreme emaciation characteristic of protein-energy malnutrition, glucose intolerance, and at a later stage diabetes. The diabetic stage of the illness is referred to as fibrocalcific pancreatic diabetes (FCPD). Several reports of FCPD have been reported from the tropical countries and many cases have been reported from the Indian subcontinent. The pathogenesis of the disease is still unclear and is attributed to various possible causes – malnutrition, cassava toxicity, oxidant stress due to micronutrient deficiency, genetic and environmental factors. Recently, a study showed the N34S variant of the *SPINK1* trypsin inhibitor gene as a susceptibility gene for FCPD in the Indian subcontinent. Although by itself it is not sufficient to cause FCPD, it indicates the role of gene–environment interaction in the pathogenesis of diabetes.

Patients with FCPD are at risk of long-term diabetic complications and require insulin to control their hyperglycemia. Given the underlying problem of malnutrition they benefit from high calorie intake, especially the protein content. There is a need for further investigation into the roles of nutritional, environmental, and genetic factors to establish the etiopathogenesis of this illness.

See also: **Diabetes Mellitus**: Classification and Chemical Pathology; Dietary Management. **Glucose**: Metabolism and Maintenance of Blood Glucose Level. **Obesity**: Complications. **World Health Organization**.

Further Reading

Altshuler D, Hirschhorn JN, Klannemark M *et al.* (2000) The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genetics* **26**: 76–80.

Amos AF, McCarty DJ, and Zimmet P (1997) The rising global burden of diabetes and its complications: estimates and projections to the year 2010. *Diabetes in Medicine* **14**(supplement) 5: S1–85.

Feskens EJ, Virtanen SM, Rasanen L *et al.* (1995) Dietary factors determining diabetes and impaired glucose tolerance. A 20-year follow-up of the Finnish and Dutch cohorts of the Seven Countries Study. *Diabetes Care* **18**: 1104–1112.

Forrest JM, Menser MA, and Harley JD (1969) Diabetes mellitus and congenital rubella. *Pediatrics* **44**: 445–447.

Frayling TM, Evans JC, Bulman MP *et al.* (2001) beta-cell genes and diabetes: molecular and clinical characterization of mutations in transcription factors. *Diabetes* **50**(supplement 1): S94–100.

Gloyn AL, Weedon MN, Owen KR *et al.* (2003) Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes* **52**: 568–572.

Hales CN and Barker DJ (1992) Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *Diabetologia* **35**: 595–601.

Hassan Z, Mohan V, Ali L *et al.* (2002) SPINK1 is a susceptibility gene for fibrocalculous pancreatic diabetes in subjects from the Indian subcontinent. *American Journal of Humam Genetics* **71**: 964–968.

Horikawa Y, Oda N, Cox NJ *et al.* (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genetics* **26**: 163–175.

Hyoty H and Taylor KW (2002) The role of viruses in human diabetes. *Diabetologia* **45**: 1353–1361.

Mohan V, Premalatha G, and Pitchumoni CS (2003) Tropical chronic pancreatitis: an update. *Journal of Clinical Gastroenterology* **36**: 337–346.

Pociot F and McDermott MF (2002) Genetics of type 1 diabetes mellitus. *Genes and Immunology* **3**: 235–249.

Ramachandran A, Snehalatha C, Kapur A *et al.* (2001) High prevalence of diabetes and impaired glucose tolerance in India: National Urban Diabetes Survey. *Diabetologia* **44**: 1094–1101.

Tuomilehto J and Lindstrom J (2003) The major diabetes prevention trials. *Current Diabetes Reports* **3**: 115–122.

Zimmet P, Alberti KG, and Shaw J (2001) Global and societal implications of the diabetes epidemic. *Nature* **414**: 782–787.

Classification and Chemical Pathology

K C McCowen, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA
R J Smith, Brown Medical School, Providence, RI, USA

Diabetes mellitus is a common, serious metabolic disorder with diverse causes and multiple complications. In this article, the definition and classification are discussed and diagnostic criteria outlined. Subsequently, an overview of the physiology of normal blood glucose homeostasis and normal insulin action leads to consideration of the pathophysiologic events that occur in uncontrolled diabetes.

Definition

Diabetes mellitus is a chronic disorder that results from a deficiency of the hormone insulin. This occurs either because of an absolute decrease in the amount of insulin produced by the β cells of the islets of Langerhans in the pancreas or because of a relative deficiency of insulin in patients whose tissues are resistant to the hormone. The hallmark of untreated diabetes mellitus is elevated blood glucose concentrations. Frequently, there are associated disturbances of fat and protein metabolism. In addition to reversible acute metabolic abnormalities resulting from inadequate effects of insulin, long-term diabetes is often characterized by the development of irreversible complications that include damage to the kidney, retina, nervous system, and both large and small blood vessels.

The diagnosis of diabetes mellitus is based on the existence of hyperglycemia alone and does not require the presence of any of the associated metabolic or systemic complications. Although patients with diabetes mellitus exhibit a characteristic pattern of metabolic abnormalities and long-term complications, this disease state results from multiple underlying causes that include genetic as well as environmental influences and also certain pancreatic or hormonal conditions.

Diagnostic Criteria

Random Plasma Glucose Determination

In the presence of symptoms of hyperglycemia, a random plasma glucose >11.1 mmol/l is consistent with a diagnosis of diabetes in an ambulatory patient. Classic symptoms of hyperglycemia include thirst, polydipsia, polyuria, and unexplained weight loss.

Fasting Plasma Glucose Determination

Nondiabetic individuals have fasting plasma glucose levels below 5.6 mmol/l. The diagnosis of diabetes mellitus can be made when fasting plasma glucose levels are significantly elevated (>7.0 mmol/l) on at least two occasions. Fasting glucose below 7.0 mmol/l but above 5.6 mmol/l, although not normal, does not meet the criteria for definite diagnosis and is classified as impaired fasting glucose. Although not absolutely predictive of diabetes, individuals with impaired fasting glucose progress to overt diabetes at a rate of approximately 5% per year. Plasma or serum measurements are generally more reliable than whole blood glucose determinations (in which the normal range is lower) because they are independent of the hematocrit and are appropriate for accurate, automated analysis. For accurate measurement, blood samples must be put into tubes containing sodium fluoride (which prevents glycolysis) or be centrifuged within 30 min to remove cells. Fasting plasma glucose is considered the desired method to diagnose diabetes because of its simplicity and reproducibility.

Oral Glucose Tolerance Test

In the absence of an obvious elevation in fasting or random plasma glucose levels, the diagnosis of diabetes mellitus can be made with an oral glucose tolerance test (OGTT). This involves, for the nonpregnant adult, the ingestion of a solution containing 75 g of glucose over 5 min, with a measurement of baseline and 2-h plasma glucose. The criteria used to diagnose diabetes are listed in **Table 1**. The diagnosis can be made if the fasting glucose exceeds 7.0 mmol/l or the 2-h value exceeds 11.1 mmol/l. People with impaired glucose tolerance have normal fasting values but 2-h post-glucose load values above 7.8 mmol/l.

Table 1 Oral glucose tolerance test criteria for diabetes in nonpregnant adults (75 g glucose load)

	Venous plasma glucose (mmol/l)
Diabetes	
Fasting	>7.0
2-h	>11.1
Impaired glucose tolerance	
2-h	7.8–11.1
Impaired fasting glucose	
Fasting	5.6–6.9
2-h	<7.8

The standard OGTT must be performed under certain conditions for the previous thresholds to apply. Subjects need to ingest at least 200 g carbohydrates per day during the 3 days preceding the test, fast overnight (>8 h), not smoke on the day of the test, and have the test performed in the morning. Because glucose tolerance is reduced by bed rest and stressors such as recent surgery or burn injury, subjects must be ambulatory and have been so for at least 1 month prior to the test. Despite this standardization, results are not always precisely reproducible, even in the same person, which may relate in part to variable rates of absorption of glucose from the small intestine. For this reason, elevated fasting glucose is a more reliable diagnostic criterion. In children, if an OGTT is performed, the amount of glucose to be ingested should be determined by body weight (i.e., 1.75 g/kg ideal body weight).

Controversy has existed with regard to proper methods of diagnosis of diabetes in pregnancy. The National Diabetes Data Group (NDDG) of the National Institutes of Health recommends screening between 24 and 28 weeks with a 50 g oral glucose load test. No special preparations are required for this test, and fasting is unnecessary. Blood glucose is measured once only, after 1 h. Women with values above 7.8 mmol/l are evaluated with a full OGTT with a glucose load of 100 g. Clearly, use of a lower threshold (e.g., 7.2 mmol/l) minimizes the occurrence of false-negative tests. Opponents of the use of a lower threshold note that only the milder cases of diabetes in pregnancy are missed using the 7.8 mmol/l criteria to proceed to full OGTT. The NDDG criteria for diagnosis are listed in **Table 2**. The original data used in the determination of normal plasma glucose values during pregnancy have been reevaluated taking into account changes in the methodology for glucose measurement, which led to the American Diabetes Association using even lower threshold values on the 100 g OGTT to diagnose diabetes (**Table 2**, revised criteria). Controversy exists as to which set of values should be used, and cost-effectiveness evaluations of the different criteria are pending.

Table 2 Criteria for the diagnosis of gestational diabetes (100 g glucose load)

	Venous plasma glucose (mmol/l)	
	National Diabetes Data Group	*Modified criteria*
Fasting	>5.8	>5.3
1 h	>10.6	>10.0
2 h	>9.2	>8.6
3 h	>8.1	>7.8

The World Health Organization proposes that the test and criteria for gestational diabetes should be the same as for nonpregnant adults, with the exception that individuals fitting the category of impaired glucose tolerance be treated the same as diabetics because of the potentially harmful effects of hyperglycemia on the fetus.

Glycosuria

Glycosuria may indicate the presence of diabetes, but it is not diagnostic, nor does the absence of glycosuria exclude diabetes. In individuals with a low renal threshold, glucose may be present in the urine in the absence of hyperglycemia. Such "renal glycosuria" is particularly common during the later stages of pregnancy and in some renal tubular disorders. The excretion of other sugars, such as lactose (more common during pregnancy) or fructose, galactose, or xylose (people with inborn errors of metabolism), can yield false-positive results through cross-reactivity in the testing method unless glucose-specific test strips are used. In patients with compromised renal perfusion or function, glycosuria may be absent despite significant hyperglycemia.

Glycosylated Hemoglobin

Glycohemoglobin is formed when a ketoamine reaction occurs between glucose and the N-terminal amino acid of the β chain of hemoglobin. The amount of glycohemoglobin generated is proportional to the mean blood glucose during the 8–10 weeks before the test. Thus, the glycohemoglobin level is a useful indicator of long-term blood glucose control. This is not a useful test for diagnosing diabetes, however, since the normal range is broad, the test is not well standardized between laboratories, and it can be affected by conditions that alter the life span of the red blood cell.

Classification

A new classification system for diabetes mellitus was developed in 1997, which divides patients into four major groups and a number of subgroups, as shown in **Table 3**. It is probable that these categories will be further refined as knowledge of the underlying etiologies of various forms of diabetes progresses.

I. Type 1 Diabetes Mellitus

This form of diabetes is defined by insulin deficiency due to destruction of the β cells of the pancreas. It was formerly designated "insulin-dependent diabetes," but efforts are being made to eliminate this name because many patients with other types of

Table 3 Classification of diabetes mellitus[a]

I. Type 1 diabetes (formerly designated insulin-dependent diabetes)
A. Autoimmune
B. Idiopathic
II. Type 2 diabetes (formerly designated non-insulin-dependent diabetes)
III. Secondary diabetes
A. Genetic defects of β cell function (e.g., maturity onset diabetes of youth)
B. Genetic defects of insulin action pathway
C. Exocrine pancreatic disease
D. Endocrinopathies (e.g., Cushing's syndrome and acromegaly)
E. Drugs or chemicals
F. Infections (e.g., congenital rubella)
G. Other genetic syndromes (e.g., Down's and Klinefelter's syndromes)
IV. Gestational diabetes

[a]Classification proposed by the Expert Committee on the Diagnosis and the Classification of Diabetes Mellitus under the sponsorship of the American Diabetes Association (*Diabetes Care* **27**: S5–S10, 2004).

diabetes also require insulin for adequate control. The predominant cause is believed to be an autoimmune attack against the insulin-producing β cells within the islets of Langerhans (diabetes type 1A). At the time of diagnosis, most patients demonstrate antibodies to certain pancreatic autoantigens, which include antibodies to islet cell cytoplasmic components, glutamic acid decarboxylase, insulin, and tyrosine phosphatases IA-2 and IA-2β. Such autoantibodies, when present, help to confirm the diagnosis. This disease also has strong HLA antigen associations, which may either predispose to or protect from the development of diabetes. In a minor subset of patients classified as idiopathic type 1 diabetes (type 1B), the presentation and clinical course is similar to autoimmune type 1A diabetes, but all tests for autoimmune markers are negative.

Early diagnosis of autoimmune diabetes Type 1 diabetes has a variable presymptomatic phase that may extend for several years, during which time it is possible to make a diagnosis. This form of diagnostic testing is reserved for research purposes because the disease is not sufficiently common to warrant widespread screening strategies and because practical methods for preventing the progression to overt diabetes are not available. Because type 1 diabetes is occasionally familial, screening of individuals with strong family histories can be performed by measuring levels of the specific pancreatic autoantigens described previously; subjects with high titers of antibodies who possess unfavorable HLA subtypes, indicating significant risk of later development of diabetes, may then undergo intravenous glucose tolerance testing with quantitation of the insulin response. Diminution of the early phases of insulin release can be seen even years before the onset of symptoms of disease. Currently, such diagnosis is important only to enable participation in clinical trials of diabetes prevention.

II. Type 2 Diabetes Mellitus

This is a heterogeneous disorder in which there is both resistance to the action of insulin and relative insulin insufficiency. In contrast to type 1 diabetes, endogenous insulin secretion is at least partially preserved and thus most patients are not insulin dependent for acute survival (hence the former name, non-insulin-dependent diabetes). The circulating insulin levels are adequate to protect these patients from ketosis, except during periods of extreme stress. Some patients in this category can be treated with oral agents (sulfonylureas, metformin, and thiazolidinediones), but many are managed with insulin because their pancreases are unable to produce sufficient insulin to overcome their tissue insulin resistance. Obesity is a frequent contributing factor to the insulin resistance in this disorder.

Occasionally, it is difficult to determine whether a patient has type 1 or type 2 diabetes. This is particularly likely in a nonobese person older than 35 years of age who has never had significant ketosis but who has been treated with insulin. Unfortunately, there is no completely reliable diagnostic test. Measurement of autoantibodies in such people may not be helpful because patients with type 1 diabetes lose these markers with time. Several studies have shown that the plasma C-peptide level is a good discriminator between the two forms of diabetes. C-peptide is released during processing of proinsulin to insulin and, thus, is an indicator of endogenous insulin secretion. Values higher than 0.6 nmol/l, either basal or following provocation with a 1 mg glucagon stimulus, indicate sufficient residual insulin secretion for a person to be considered in the type 2 diabetes class.

III. Secondary Diabetes Mellitus/Other Specific Types

This broad category includes multiple disorders that are associated with either extensive pancreatic destruction or significant insulin resistance. Secondary diabetes as a consequence of decreased insulin production can occur following pancreatectomy, chronic pancreatitis, cystic fibrosis, or hemochromatosis. In the absence of pancreatic damage,

secondary diabetes can result from extreme insulin resistance induced by glucocorticoids (Cushing's syndrome); growth hormone (acromegaly); adrenergic hormones (pheochromocytoma); other medical conditions, such as uremia, hepatic cirrhosis, or polycystic ovary syndrome; or medications (diuretics or exogenous glucocorticoids).

Included in this category of secondary diabetes are patients who appear to have type 2 diabetes but in whom monogenic molecular defects in either the glucose-sensing or insulin action pathways have been defined. The best established molecular defects are mutations in the gene coding for the enzyme glucokinase, which has a role in the sensing of glucose by the β cell. Individuals with this autosomal dominant condition usually develop mild diabetes in early adulthood or adolescence. Hence, the condition is known as maturity onset diabetes of the young (MODY). Several other types of MODY have been defined, due to gene defects in β cell transcription factors. Rare causes of diabetes secondary to insulin resistance include various inborn errors of metabolism (e.g., insulin receptor mutations or type 1 glycogen storage disease), chromosomal abnormalities such as Down's and Turner's syndrome, and muscle diseases (e.g., myotonic dystrophy).

IV. Gestational Diabetes Mellitus

This disorder, which is defined as hyperglycemia first detected during pregnancy, occurs in 2–5% of pregnant women. Often, one cannot determine whether glucose intolerance antedated the pregnancy or whether hyperglycemia was provoked by the hormonal milieu associated with pregnancy. Hyperglycemia remits postpartum in 90% of women with gestational diabetes, but these women are at increased risk for subsequent development of diabetes, which is usually type 2. Although most cases of this form of diabetes are detected by blood glucose screening performed as a routine procedure early in the third trimester, the current recommendation is that universal screening is probably unwarranted. A woman younger than age 25 years, of normal body weight, without a family history of diabetes or a personal history of poor pregnancy outcome, and from an ethnic group with low rates of diabetes is at sufficiently low risk of gestational diabetes that glucose testing can be omitted. In contrast, women with clinical features associated with a high risk of gestational diabetes (obesity, positive family history, persistent glycosuria, and prior gestational diabetes) should be screened as early in the pregnancy as is feasible. In women who have documented gestational diabetes, a follow-up glucose tolerance test should be performed 6 weeks postpartum unless overt diabetes is evident.

Other Abnormalities of Glucose Tolerance

Impaired glucose tolerance This is a condition defined by oral glucose tolerance testing and includes nonpregnant individuals with normal fasting blood glucose but modestly elevated postprandial glucose. People with impaired glucose tolerance are at a high risk for subsequent development of diabetes, usually type 2 (approximately 5% per year). Thus, impaired glucose tolerance is a stage in the evolution of diabetes. Until overt diabetes develops, people with impaired glucose tolerance are not believed to have elevated risk of microvascular complications of diabetes. However, impaired glucose tolerance is associated with an increased risk of cardiovascular disease.

Impaired fasting glucose Some patients will have abnormal elevations in fasting plasma glucose, even though 2 h post-glucose challenge values are normal. These people are also at increased risk of developing diabetes, although diabetes incidence rates are highly variable between different populations. Fasting glucose is defined as impaired in the range 5.6–7.0 mM. Until recently, impaired fasting glucose was defined as 6.1–7.0 mM, and the change was recommended by the American Diabetes Association to align better with the category of impaired glucose tolerance discussed previously. However, people in this category who have normal postprandial glycemia, or normal 2-h post-challenge glucose values, have a lower risk of cardiovascular disease than people with impaired glucose tolerance.

Stress hyperglycemia This denotes an individual who is frankly hyperglycemic (>7.8 mmol/l) under conditions of intercurrent illness or during treatment with medications that provoke diabetes. Such people may revert to normal glucose tolerance following removal of the stress. Although not an official category of diabetes, such abnormal glucose values in hospitalized patients cannot be ignored since there is strong evidence that treatment to normoglycemia significantly lowers mortality, at least for patients with acute myocardial infarction or with critical illness in an intensive care unit. Precipitants of stress hyperglycemia are listed in **Table 4**.

Table 4 Risk factors for the development of stress hyperglycemia in critical illness

Factor	*Major mechanism*
Preexisting diabetes mellitus	Insulin deficiency (relative or absolute)
Infusion of catecholamine pressors	Insulin resistance
Glucocorticoid therapy	Insulin resistance
Obesity	Insulin resistance
Increasing APACHE score	Higher counterregulatory hormone levels
Older age	Insulin deficiency
Excessive dextrose administration	Glucose removal rates overwhelmed in the face of ongoing hepatic glucose production
Pancreatitis (acute and chronic)	Insulin deficiency
Sepsis	Insulin resistance
Hypothermia	Insulin deficiency
Hypoxemia	Insulin deficiency
Uremia	Insulin resistance
Cirrhosis	Insulin resistance

APACHE, Acute Physiology and Chronic Health Evaluation.

Pathophysiology of Diabetes

Physiology of Normal Blood Glucose Regulation

The metabolic fate of ingested glucose is determined by the interplay of multiple hormones. Insulin is of major importance in this homeostasis, but glucagon, glucocorticoids, catecholamines, and growth hormone also have significant effects that are interactive with insulin. Glucose ingested with a meal or derived from the digestion of other dietary carbohydrates is rapidly absorbed by the small intestine. It is carried first to the liver by the portal vein, where a substantial portion (30–70%) is removed; the remainder enters the peripheral circulation, where regulated insulin secretion and target tissue responses to insulin contribute to glucose clearance and control of blood glucose levels (**Figure 1**).

Following a meal, insulin is secreted from pancreatic β cells in response to increased circulating glucose concentrations. This direct effect of glucose on β cells is augmented by neural (vagal) and hormonal factors of intestinal origin (e.g., glucose-dependent insulinotropic peptide, cholecystokinin, and glucagon-like peptide 1), such that the insulin secretory response to oral glucose greatly exceeds the response to an equivalent intravenous glucose infusion.

The overall effect of the increase in insulin levels in parallel with increased glucose entry to the circulation is promotion of the net removal of glucose by the liver and stimulation of glucose transport into muscle and adipose tissue, where it is consumed as a metabolic fuel or stored. Insulin also inhibits the catabolism of the alternative energy sources, fat and protein. This is an appropriate response to the abundance of circulating nutrients that occurs after meals. During fasting, insulin levels are low, these processes are reversed, and stored fuel is made available to all tissues.

Liver Glucose enters the liver by facilitated (carrier-mediated) diffusion driven by the concentration gradient that exists in the fed state. A portion of the glucose taken up by the liver is metabolized via glycolytic pathways to produce ATP. A substantial amount is transformed into glycogen and stored. The maximal storage capacity of the liver is approximately 100 g glycogen (400 kcal). The specific molecular effects of insulin in the fed state lead to altered activities of enzymes that trap glucose inside the hepatocyte, promote glycolysis, and enhance glycogen synthesis (**Figure 1**). Insulin also inhibits enzymes important for both glycogenolysis and gluconeogenesis and thus shuts off hepatic glucose production. A portion of the glucose entering the liver is converted into triglyceride and exported to the adipocyte for storage.

Skeletal muscle In skeletal muscle, insulin directly stimulates glucose uptake, which is the rate-limiting step for muscle clearance of glucose. This appears to occur predominantly by causing the rapid translocation of glucose transporters (in particular the Glut-4 transporter) from an as yet undefined intracellular site to the muscle cell surface. Insulin also stimulates glycolysis and the net formation of glycogen in muscle. Even at low insulin concentrations, however, a rise in ambient glucose stimulates substantial glucose clearance by muscle, probably via the Glut-1 transporter. Glycogen stores in muscle (500–600 g glycogen in a 70 kg human) serve as a rapidly mobilized energy source during exercise but do not directly support blood glucose concentrations in the fasted state because muscle lacks the enzyme glucose-6-phosphatase, which is needed for release of free glucose to the circulation. Insulin-stimulated amino acid entry into muscle enhances insulin stimulatory effects on protein synthesis and decreases the availability of circulating amino acids as substrates for hepatic gluconeogenesis. Muscle proteolysis, which yields amino acid precursors that contribute to hepatic gluconeogenesis in the fasted state, is inhibited by insulin.

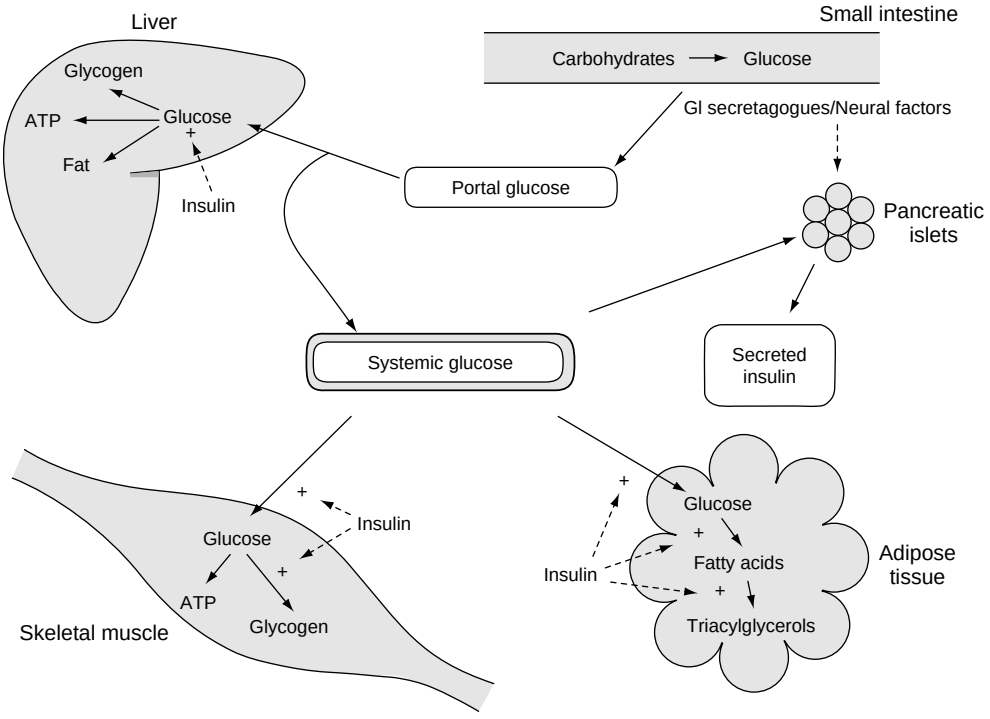


Figure 1 Insulin effects on glucose homeostasis in the fed state.

Adipose tissue In adipose tissue, insulin stimulates glucose uptake via the Glut-4 transporter, providing substrate for energy generation and glycerol synthesis. Even more important effects of insulin in adipose tissue are inhibition of lipolysis and stimulation of FFA uptake and triglyceride synthesis. This limits the availability of fat-derived fuels for other tissues and indirectly contributes to the lowering of blood glucose by favoring glucose utilization in multiple tissues.

From this brief overview, it can be seen that the rise in insulin following a meal has multiple tissue-specific actions that serve to lower blood glucose, prevent hyperglycemia, and inhibit the mobilization of alternative metabolic fuels. Many of the metabolic abnormalities that develop acutely in uncontrolled diabetes can be explained by the loss of these actions of insulin.

Pathophysiology of Uncontrolled Diabetes

Uncontrolled diabetes mellitus occurs when circulating insulin levels are inadequate to lower elevated blood glucose concentrations. This condition includes a spectrum of metabolic abnormalities that range from the effects of mild insulin deficiency (i.e., hyperglycemia) to the effects of marked and prolonged insulinopenia (i.e., ketoacidosis and fluid and electrolyte depletion). Diabetic ketoacidosis, which is the most severe acute manifestation of insulin deficiency, is almost entirely restricted to patients with type 1 diabetes, or those with severe pancreatic disease of other etiologies. In people without absolute insulin deficiency, although the combination of significant insulin resistance and relatively low levels of insulin can result in significant hyperglycemia, ketone body production sufficient to cause ketosis and metabolic acidosis does not occur. Even low levels of insulin, such as are typically present in type 2 diabetes, suffice to restrain lipolysis and limit the availability of free fatty acid precursors for ketone body formation. Otherwise, many of the derangements seen in uncontrolled diabetes are common to all forms of diabetes.

The pathophysiologic events that affect blood glucose levels in states of mild-to-moderate insulin deficiency are classified into two broad categories. First, the normal pathways for glucose clearance after a meal are ineffective; second, body fuel stores are broken down with release of other substrates that lead to inappropriate synthesis of more glucose. These events are brought about by insulinopenia and often are further promoted by the relative abundance of the counterregulatory hormones, glucagon, catecholamines, and, to a lesser extent, cortisol and growth hormone. In addition, hyperglycemia further inhibits pancreatic β cell insulin secretion, compounding the problem ("glucose toxicity").

Following the ingestion of a meal, a substantial portion of the glucose absorbed into the portal circulation is removed by the liver, where it is stored as glycogen, converted to lipid, or consumed via energy-generating pathways. Each of these processes is decreased by insulin deficiency, resulting in increased entry of absorbed glucose to the systemic circulation. Skeletal muscle represents the main tissue site for removal of circulating blood glucose following a meal. In diabetes, insulin deficiency leads to a marked decrease in activity of the Glut-4 glucose transporter largely as a consequence of decreased insulin-stimulated Glut-4 localization to the surface membranes. This decreases the normal postmeal flux of glucose into skeletal muscle. In addition, glucose that does enter muscle is metabolized inefficiently in the absence of insulin. Other insulin-sensitive tissues, such as adipose tissue and myocardium, are affected in a similar manner, with consequent reduction in both glucose uptake and metabolism, although their contribution to glucose clearance is quantitatively less than that of muscle.

In postabsorptive or fasted states, hyperglycemia in uncontrolled diabetes does not resolve and often worsens (**Figure 2**). Abnormally low insulin concentrations lead to an exaggeration of metabolic responses that normally serve to protect against the development of hypoglycemia during fasting. These responses to low insulin and elevated counterregulatory hormones include, initially, the conversion of stored glycogen to glucose. Simultaneously, the hepatic enzymes involved in gluconeogenesis are activated, which results in glucose production from such carbon sources as lactate and pyruvate (by-products of muscle glycolysis), amino acids (from muscle protein breakdown), and glycerol (derived from adipocyte triglyceride stores). With persistent insulin deficiency, glycogen stores are depleted, and hepatic gluconeogenesis becomes the most important contributor to the increasing hyperglycemia. Meanwhile, body stores of protein and fat are being depleted in the futile synthesis of new glucose that cannot be used efficiently and serves to aggravate the existing hyperglycemia.

Excessive glucose accumulation in the circulation and in the extracellular space leads to the movement of water out of cells to maintain osmotic balance, causing intracellular dehydration. The high filtered load of glucose at the renal glomerulus overwhelms the reabsorptive capacity of the renal tubule, and an osmotic diuresis results. Ultimately, this leads not only to water loss along with the glucose but also to excess excretion of potassium, sodium, magnesium, calcium, and phosphate in the urine. The magnitude of the total body electrolyte loss depends on the duration and severity of the hyperglycemia.

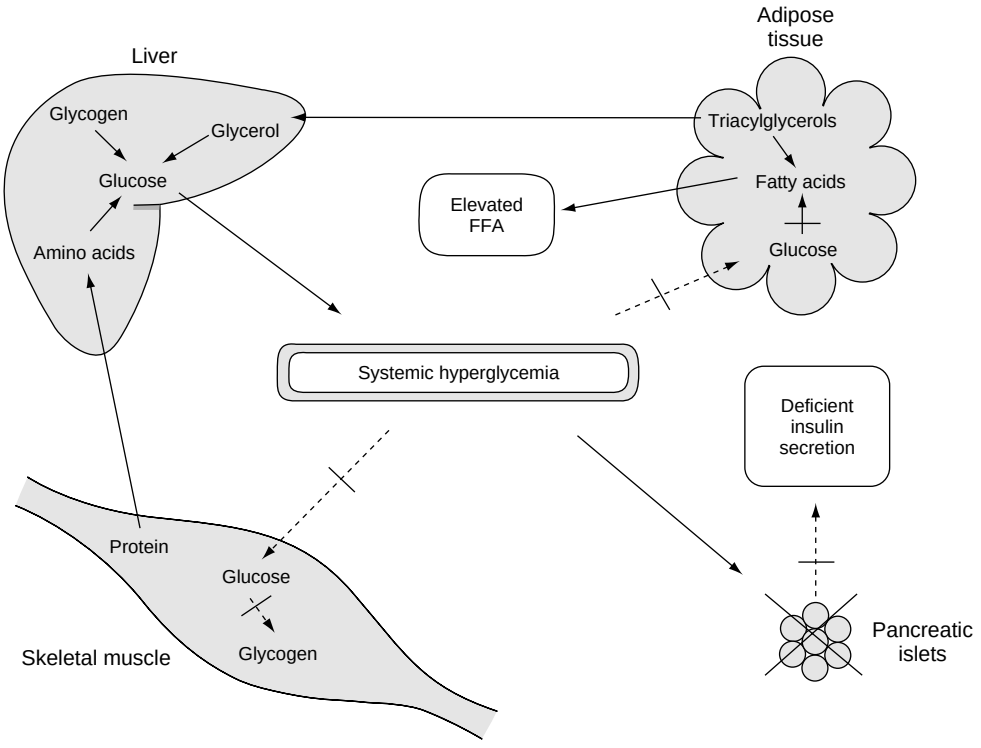


Figure 2 Metabolic events leading to hyperglycemia in the postabsorptive state in uncontrolled diabetes mellitus.

The main symptoms with moderate insulin deficiency are polyuria and consequent thirst and polydipsia. With more severe and prolonged insulin deficiency, loss of large quantities of glucose in the urine can lead to weight loss. If hyperosmolarity is not compensated by an adequate increase in water intake, patients can develop altered mental status and obtundation. In elderly patients with type 2 diabetes, this sequence can lead to the life-threatening state of nonketotic hyperosmolar coma.

In type 1 diabetes, the clinical picture of poor control differs from that described previously in that insulin deficiency is more severe (**Figure 3**). Glucose uptake by muscle is diminished, and glucose production by the liver is augmented. Marked insulinopenia, however, also leads to rapid, uncontrolled lipolysis. Triglyceride breakdown results in accelerated release of free fatty acids and glycerol. The increased delivery of glycerol from adipose tissue to the liver further promotes hepatic gluconeogenesis. In the absence of insulin, the liberated free fatty acids are taken up by the liver and converted at an accelerated rate to ketone bodies (β-hydroxybutyric acid, acetoacetic acid, and acetone).

In the fasting state in nondiabetic individuals, ketone bodies are metabolized under the influence of even low levels of insulin as a source of energy, particularly in skeletal and cardiac muscle. In extreme insulin deficiency states, ketone body utilization is inhibited at the same time that synthesis is increased. With increasing duration of insulinopenia, the ketoacid levels in the bloodstream rise. Ketones, like glucose, spill into the urine, either as free acids or, depending on the pH, as sodium or potassium salts, worsening the osmotic diuresis and electrolyte deficiency. Eventually, the blood buffering capacity for acid is overwhelmed and systemic acidemia occurs. Acidemia has a deleterious effect on all cell membranes and many cellular functions and, when severe, can cause arrhythmias, cardiac depression, and vascular collapse. In combination with the previously described hyperosmolarity and dehydration, diabetic ketoacidosis is a life-threatening situation.

In summary, poor control can lead to dangerous metabolic consequences and, occasionally, death. A primary goal of therapy is insulin replacement, which is needed to reverse the production of glucose and ketoacids by the liver, to promote muscle glucose and ketone body uptake, and to inhibit further breakdown of fat and protein. An equally important goal of therapy should be the replenishment of lost extracellular and intracellular fluids and electrolytes.

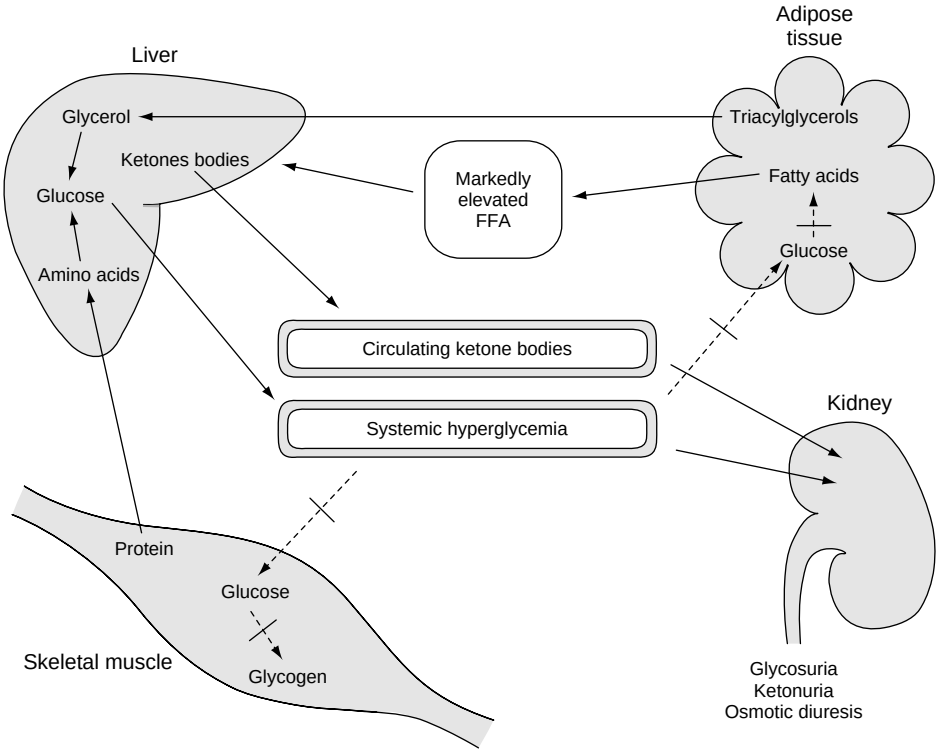


Figure 3 Metabolic events leading to the development of ketoacidosis in uncontrolled diabetes mellitus.

See also: **Diabetes Mellitus**: Etiology and Epidemiology; Dietary Management. **Glucose**: Chemistry and Dietary Sources; Metabolism and Maintenance of Blood Glucose Level; Glucose Tolerance. **Liver Disorders**.

Further Reading

Arner P (1996) Regulation of lipolysis in fat cells. *Diabetes Reviews* **4**: 450–463.

Coustan DR and Carpenter MW (1998) The diagnosis of gestational diabetes. *Diabetes Care* **21**(supplement 2): B5–B8.

Deerochanawong C, Putiyaun C, Wongsuryat M, Serirat S, and Jinayon P (1996) Comparison of National Diabetes Data Group and World Health Organization criteria for detecting gestational diabetes mellitus. *Diabetologia* **39**: 1070–1073.

Delaney MF, Zisman A, and Kettyle WM (2000) Diabetic ketoacidosis and hyperglycemic hyperosmolar nonketotic syndrome. *Endocrinology and Metabolism Clinics of North America* **29**(4): 683–705.

Dinneen S, Gerich J, and Rizza R (1992) Carbohydrate metabolism in non-insulin-dependent diabetes mellitus. *New England Journal of Medicine* **327**: 707–713.

Expert Committee on the Diagnosis and Classification of Diabetes Mellitus (1997) Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care* **20**: 1183–1197.

Expert Committee on the Diagnosis and Classification of Diabetes Mellitus (2003) Follow up report on the diagnosis of diabetes mellitus. *Diabetes Care* **26**: 3160–3167.

Foster DW and McGarry JD (1983) The metabolic derangements and treatment of diabetic ketoacidosis. *New England Journal of Medicine* **309**: 159–169.

Genuth S, Alberti KG, Bennett P *et al.* Expert Committee on the Diagnosis and Classification of Diabetes Mellitus (2003) Follow-up report on the diagnosis of diabetes mellitus. *Diabetes Care* **26**(11): 3160–3167.

Kuzuya T and Matsuda A (1997) Classification of diabetes on the basis of etiologies versus degree of insulin deficiency. *Diabetes Care* **20**: 219–220.

Malmberg K, Ryden L, Efendic S *et al.* (1995) Randomized trial of insulin-glucose infusion followed by subcutaneous insulin treatment in diabetic patients with acute myocardial infarction (DIGAMI study): Effects on mortality at 1 year. *Journal of the American College of Cardiology* **26**(1): 57–65.

McCance D, Hanson R, Pettitt D *et al.* (1997) Diagnosing diabetes mellitus—Do we need new criteria? *Diabetologia* **40**: 247–255.

McCowen KC, Malhotra A, and Bistrian BR (2001) Stress-induced hyperglycemia. *Critical Care Clinics* **17**(1): 107–124.

Mitchell GA, Kassovska-Bratinova S, Boukaftane Y *et al.* (1995) Medical aspects of ketone body metabolism. *Clinical and Investigative Medicine* **18**: 193–216.

National Diabetes Data Group (1979) Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. *Diabetes* **28**: 1039–1057.

Tchobroutsky G (1991) Blood glucose levels in diabetic and non-diabetic subjects. *Diabetologia* **34**: 67–73.

van den Berghe G, Wouters P, Weekers F *et al.* (2001) Intensive insulin therapy in the critically ill patients. *New England Journal of Medicine* **345**(19): 1359–1367.

Dietary Management

C D Saudek, Johns Hopkins University School of Medicine, Baltimore, MD, USA
S H Oh, Johns Hopkins General Clinical Research Center, Baltimore, MD, USA

The successful treatment of diabetes mellitus starts with a sound diet, although the specifics of the diet vary depending on the kind of diabetes being treated and the individual circumstances. Individualization is the hallmark of medical nutrition therapy in diabetes. Since approximately 90% of all people with diabetes have type 2, and approximately 80% of them are obese, weight reduction is often the main therapeutic goal. Many also need treatment of comorbidities such as hypertension or dyslipidemia. People with type 1 diabetes, on the other hand, usually require far more attention to exactly how much carbohydrate they ingest and exactly how their intake matches their insulin dose and activity level. In all cases, education of the patient by a trained nutritionist is essential. Diabetes is rarely well controlled unless patients have at least a basic understanding of what they should eat and why.

Overall Objectives in the Management of Diabetes

Control of Blood Glucose Level

A first and very basic goal of diabetes care is to eliminate the symptoms of hyperglycemia. Treatment is inadequate if the person remains polyuric, thirsty, or continues to lose weight from hyperglycemia. To cause symptoms, however, hyperglycemia usually must average more than 11 mM (200 mg/dl). Since blood glucose in the 7–11 mM (125–200 mg/dl) range is distinctly abnormal and does cause long-term diabetic complications, freedom from symptoms is only the beginning of adequate therapy.

Irrefutable evidence exists that better control of blood glucose concentration reduces the risk of developing long-term complications from diabetes. This is especially true of microvascular complications such as retinopathy (eye disease), nephropathy (kidney disease), and nerve damage in both type 1 and type 2 diabetes. Control of blood glucose also reduces the risk of macrovascular disease (heart disease, stroke, and peripheral vascular disease), although the contribution of blood glucose to these complications is less strong.

Carbohydrate ingestion (rather than fat or protein) is the main determinant of postmeal blood glucose level.

Dietary intake, oral medications, insulin, exercise, and stress all contribute to blood glucose levels in the person with diabetes and must be understood when establishing and implementing medical nutrition therapy.

To determine the efficacy of treating glycemia, blood glucose must be monitored. There are two ways to assess diabetic control: self-monitoring of blood glucose (SMBG) laboratory monitoring of hemoglobin A1c (HbA1c). SMBG, done by obtaining a drop of blood and using a small, handheld meter, measures the blood glucose at the time the measurement is taken. It may be done as often as six to eight times per day or as infrequently as several times per week. The HbA1c is a laboratory test that reflects glycemic control during the previous 60–90 days and should be done every 3–6 months. Target HbA1c is generally considered to be <6.5–7% when the upper limit of normal is <6%.

Prevention or Control of Comorbidities

Morbidity and mortality among people with diabetes are rarely due to acute hyperglycemia or diabetic ketoacidosis. Rather, the long-term complications are either specific to diabetes (e.g., diabetic retinopathy or nephropathy) or accelerated by diabetes (e.g., atherosclerosis). Diabetes significantly increases the risk of coronary artery, cerebrovascular, and peripheral vascular disease, with these cardiovascular complications accounting for approximately 80% of deaths in diabetes. Prudent dietary management of diabetes therefore requires consideration of what can be done to prevent or control the various comorbidities of this disease. For example, all people with diabetes should be on a diet that minimizes the risk of atherosclerosis. At the first clinical sign of hypertension, dietary methods should be implemented to lower blood pressure.

Minimum Intrusion on Quality of Life

To people with diabetes, the 'diabetic diet' can be a fearsome thing, often made worse by the way it is presented. Many modern dieticians refuse even to use the word 'diet' since it conjures up so many bad associations, preferring 'nutrition plan' or 'medical nutrition therapy.' Most patients will not totally abandon their dietary habits of a lifetime, forgoing favorite ethnic flavors and socially accepted foods. Rather, the prescribed diet that intrudes least on a person's quality of life is the most successful nutrition plan. Expert professionals can identify exactly what changes are required and what favorite dishes, spices, or food groups can be built into a good nutrition prescription.

Dietary Approaches to Diabetes

Principles of Dietary Management of Diabetes

Assessment The first step for planning an appropriate nutrition plan is a full assessment of the diabetic patient. Topics covered in the nutritional assessment are included in **Table 1**.

Individualization Individualization is a cardinal principle of medical nutrition therapy for diabetes, facilitating individual lifestyle and behavior changes that will lead to improved metabolic control. Since no one diet fits all, the standard, printed diabetic diet is inadequate. Rather, people with diabetes need to consult a person trained in dietetics, one able to develop and teach an individualized nutritional prescription. **Table 2** indicates the range of goals that may need accommodation among different people with diabetes.

Developing the diabetes nutrition plan With the emphasis on individualization, the meal plan is driven by the diagnosis, pharmacologic treatment, lifestyle, and treatment goals. Important consideration is given to dietary preferences, socioeconomic factors, and the patient's ability to understand and implement instructions. Some patients will need instruction on fine points such as carbohydrate counting; others will benefit from the crudest of prescriptions, such as advice to stop buying concentrated sweets or frequenting fast-food restaurants.

Total energy intake The total energy requirement to maintain constant body weight may be calculated using the Harris–Benedict equation, taking into consideration the patient's activity level. The weight-maintaining requirement is then adjusted according to the therapeutic objective—to accomplish weight loss, maintenance of weight, or weight gain. Examples of how to make these calculations are shown in **Table 3**. Specific conditions such as childhood growth and development, pregnancy, malabsorption, or existing nutritional deficiencies are beyond the scope of this article.

Distribution of energy intake Distribution of carbohydrate, protein, and fat into the total energy target also depends on individual needs and therapeutic objectives. Most guidelines recommend that carbohydrate intake represent up to 50–60% of total energy, protein 15–20%, and fat 20–35%. Another approach groups carbohydrate and monounsaturated fats, recognizing that saturated fat should be restricted. This approach suggests that carbohydrate and monounsaturated fat should together account for 60–70% of energy intake.

Table 1 The Nutritional Assessment

Diet history/nutrition information—can be obtained using dietary assessment tools such as 24-h recalls, food records, food frequency questionnaires, or dietary intake interviews
Meal patterns: Usual distribution of meals and snacks throughout the day, including variations from day to day, weekdays versus weekends, skipped meals, and external influences such as work, school, travel, vacations, and holidays
Food choices: Types and amounts of foods consumed at meals and snacks
Nutritional adequacy: Dietary excess or deficiency; also considers overall dietary balance
Beliefs or misconceptions: Fears or misconceptions of a 'strict' diabetic diet or about certain foods; can also include certain religious beliefs or ethnic beliefs about foods
Personal information
Age, gender, socioeconomic status, ethnicity, occupation, education, and literacy level
Ability and willingness to change (stages of change)
Emotional and mental state if distressed by a new diagnosis of diabetes or other health complications related to diabetes
External stressors that may interfere with compliance
Smoking or drug history
Exercise or activity schedule
Clinical information
Type of diabetes and treatment, such as with insulin, oral hypoglycemic drugs, or diet alone
Physical activity, body weight, and blood pressure
Lab results, A1C, and lipid profile
Other medical conditions
Education
Diabetes education should be an ongoing interactive process between patient and health professional and cannot be given in a single session
Individualism is key to successful nutritional management
Most important aspect is to match the type and level of information to individual needs and abilities
Important to provide written information summarizing key messages that patient can take home and refer to later
Follow-up and monitoring progress
Follow-up and review of progress essential
Frequency will depend on type of treatment, glycemic control, and patient's ability to meet goals
Consider if specific dietary targets have been achieved and/or reasons why targets have not been met and what barriers need to be overcome
Consider acceptability of dietary changes and impact on patient's quality of life
Clinical picture should examine glycemic control, lipid profiles, weight changes, and blood pressure

Adapted from Conner H *et al.* Nutrition Subcommittee of the Diabetes Care Advisory Committee of Diabetes UK (2003) The implementation of nutritional advice for people with diabetes. *Diabetic Medicine* **20**(10): 786–807.

Even these broad goals are the subject of considerable controversy, with some experts recommending a lower carbohydrate intake and higher fat intake, particularly of monounsaturates. A common mistake is for the patients to think they have a diet low in both carbohydrate and fat without being low in total energy intake. It is unlikely that a person's dietary protein intake will exceed 15–30% of all energy consumed; therefore, 70–85% of intake is generally distributed between fat and carbohydrate.

Distribution of energy intake throughout the day may vary, too. Insulin-requiring diabetic patients, for example, may need a more evenly distributed energy intake, even including a bedtime snack to avoid hypoglycemia. This would not necessarily be indicated for someone with type 2 diabetes trying to lose weight, although weight-reducing diets are generally considered more effective if the total energy intake is spread more or less evenly throughout the day so that the patient does not build up a hunger and gorge late in the day. One report of Muslims observing daytime fasting during Ramadan found that more than half did not lose weight, suggesting a major redistribution of caloric intake to nighttime hours. A significant increase in hypoglycemia occurred during the days of Ramadan. Reduced energy intake for prolonged periods is most dangerous for patients taking insulin, but it may also be significant in those taking oral hypoglycemic agents such as sulfonylureas.

The utility of exchange lists There has been a shift on the part of patients and some health professionals away from the use of formal 'exchange lists' for meal planning. The traditional exchanges estimate not only carbohydrate but also certain proportions of fat and protein in similar foods. Food labels make the calculation of specific fat and carbohydrate content easier. The trend, therefore, is to emphasize the carbohydrate and fat awareness by teaching them directly rather than lumping mixed foods together in exchanges.

Gastroparesis An extremely difficult challenge is posed by the patient with diabetic gastroparesis. This

Table 2 Cases illustrating the variable clinical issues affecting people with diabetes and the resulting diversity of their nutritional needs

Type of diabetes	*Type 1*	*Type 1*	*Type 2*	*Type 2*
Age (years)	14	38	56	76
Duration of DM (years)	6	26	6	6
BMI	18	23	27	34
Physical activity	Vigorous	Moderate	Mild	Minimal
Prone to hypoglycemia	Yes	Yes	Yes	No
Prone to hyperglycemia	Yes	Yes	Yes	Yes
Blood lipids	Normal	Normal	High LDL cholesterol	High TG, Low HDL
Blood pressure	Normal	High	Normal	High
Dietary preferences	Likes sweets, snacks	Healthy, little carb awareness	Spicy foods, irregular meals	Fried foods, sweets
Pharmacologic therapy	Multiple-dose insulin	Multiple-dose insulin	Oral agents plus insulin	Oral agents
Life expectancy without diabetes	66 years more	44 years more	26 years more	8 years more
Major nutritional considerations	Adequate caloric intake for growth (see **Table 3**)	Stabilize carb intake, count carbs	Mildly hypocaloric (see **Table 3**)	Low salt, high vegetable for hypertension (DASH diet)
	Recognize carb portions, regularize carb intake	Low salt, high vegetable for hypertension (DASH diet)	Hypolidemic (low saturated fat)	Hypolipemic diet (low saturated fat)
	Avoid excess concentrated sweets		Regularity of meals, consistency of carb and fat intake	Moderately hypocaloric
	Learn factors causing hypoglycemia			Control of dietary carb, especially high-energy concentrated sweets
	Healthy heart diet			

BMI, body mass index; DM, diabetes mellitus; HDL, high-density lipoprotein; LDL, low-density lipoprotein.

condition, a severe autonomic neuropathy reducing gastric motility and gastric emptying time, is often difficult to diagnose by standardized testing, such as gastric emptying studies. Gastroparesis typically causes early satiety, nausea, vomiting, and abdominal pain, with markedly variable food ingestion. Along with pharmacologic management and good glycemic control, the dietary prescription should include small, frequent feeding as tolerated, but the condition can progress to the point that any oral intake is difficult, and tube feeding or a gastrostomy is required. Fortunately, diabetic gastroparesis tends to wax and wane in severity.

Glycemic control and weight gain Research studies have repeatedly found that when a patient with poor glycemic control achieves good glycemic control, there is a strong, almost inevitable, tendency to gain weight. This may simply be due to the retention of energy that was previously lost in the urine as glucosuria, but the patient should be warned of the likelihood of gaining weight when poor diabetic control is adequately treated. As for quitting smoking, the health benefit of glycemic control far outweighs the risk of weight gain.

Nutritional instruction To achieve a stable, healthy diet, the following key educational issues must be considered (adapted from Franz M, Krosnick A, Maschak-Carey BJ *et al.* (1986) *Goals for Diabetes Education*. Chicago: American Diabetes Association):

- Survival skills
  - Relation of food to insulin and activity
  - Importance of good nutrition in the control of blood glucose and lipid levels
  - Necessity of maintaining normal weight
  - Types and amounts of food in meal plan
  - Modification of food intake during brief illnesses
- In-depth counseling
  - Meal planning
  - Types of nutrients, their functions, relation to insulin, and effect on blood glucose and lipid levels
  - Caloric level of meal plan and percentages of carbohydrate, protein, and fat
  - Food sources of fiber
  - Importance of reducing total fat, saturated fat, and cholesterol in the diet
  - Relation of sodium to hypertension
  - Proper serving sizes
  - Changes in food intake based on activity level
  - Eating out and special occasions
  - Label reading and grocery shopping
  - Use of sweeteners, alcohol, and 'dietetic' foods
  - Food modifications for other disorders
  - Incorporation of favorite recipes

Table 3 Sample calculations of energy requirement in differing circumstances using the Harris–Benedict formula to determine caloric requirements for children and adults

Caloric requirements = basal metabolic rate × activity factor × injury factor

Basal metabolic rate (BMR)

For men: BMR = 66 + [13.7 × wt (kg)] + [5 × ht (cm)] − [6.8 × age (years)]

For women: BMR = 655 + [9.6 × wt (kg) + [1.8 × ht (cm)] − [4.7 × age (years)]

Multiply by the following factors:

Activity factors

1. Sedentary (little or no exercise): BMR × 1.2
2. Lightly active (light exercise/sports 1–3 days/week): BMR × 1.375
3. Moderately active (moderate exercise/sports 3–5 days/week): BMR × 1.55
4. Very active (hard exercise/sports 6–7 days/week): BMR × 1.725
5. Extra active (very hard daily exercise/sports and physical job or 2X day training): BMR × 1.9

Injury factors (not used for healthy individuals)

1. Generalized stress: 1.1–1.2
2. Surgery (minor): 1.1–1.5
3. Infection: 1.2–1.5
4. Trauma: 1.14–1.37
5. Cancer: 1.2

For weight loss, use the above calculated formula for caloric requirements and subtract by 500 calories:

Caloric requirements − 500 calories/day = modified calorie requirements

This is for a recommended 0.5–1 pound of weight loss per week

Special aspects: Type 2 diabetes There are two pathophysiologic mechanisms underlying type 2 diabetes: The body's cells are resistant to the action of insulin, and the pancreas is unable to secrete enough insulin to overcome that resistance. Although it is not entirely clear which of these processes occurs first, and although the balance of the two may vary from case to case, the most common cause of insulin resistance is overweight or obesity. Unfortunately, much evidence has shown that people of Asian ethnicity are especially prone to obesity-related type 2 diabetes even when their body weight, by Western standards, is normal. Japanese Americans, for example, show an increase risk of diabetes if their body mass index (BMI) increases to only 24. This excessive risk with even mild degrees of excess body weight may explain the marked rise in diabetes when previously undernourished populations begin to have adequate nutrition. In this sense, diabetes is a disease of prosperity.

Major objectives Approximately 95% of all people with diabetes have type 2, and the major increase in the prevalence of diabetes in recent years is almost entirely accounted for by increase in body weight. It cannot be overemphasized that medical nutrition therapy of type 2 diabetes should address normalization of body weight. In most cases, the focus is on reducing dietary intake of saturated fat and increasing energy expenditure through exercise. By reducing body weight, insulin resistance is reduced, making the patient's endogenous insulin more effective. Given that approximately 85% of people with type 2 diabetes die of cardiovascular cause, the second emphasis of medical nutrition therapy for type 2 diabetics must address dyslipidemia and blood pressure.

Hypoenergetic diets are remarkably effective in controlling hyperglycemia. Indeed, blood glucose levels improve, often dramatically, as soon as a low-energy diet is started, apparently by reducing hepatic glucose production. The correction of insulin resistance is more closely correlated with actual weight loss, which takes much longer. The best strategy for accomplishing and maintaining weight loss is unclear and may vary from person to person depending on the different factors involved, such as willingness to change and other lifestyle behaviors. Dosages of antidiabetic drugs may have to be altered as the person loses weight.

Persistent insulin resistance in type 2 diabetes, together with deteriorating pancreatic insulin secretion over time, means that many people with type 2 diabetes eventually require exogenous insulin therapy. This does not change the diagnosis to type 1 diabetes, which is a disease of entirely different pathogenesis. Because of the insulin resistance, people with type 2 diabetes taking insulin often need high doses, often 50–100 units per day or higher. Insulin requirements will predictably be less when energy intake is reduced.

Recently, in Western societies, many overweight teenagers have presented with type 2 diabetes. It can no longer be assumed that children with diabetes have type 1. Indeed, some reports find that half of all teenagers with diabetes have type 2, a marked shift from prior years. Furthermore, nutrition therapy for children with diabetes must be designed with a clear understanding of what type of diabetes they have. In cases of obesity-related type 2, calorie restriction may be indicated.

Coexisting risk factors Obesity, dyslipidemia, and hypertension are especially prevalent in type 2 diabetes. The constellation of comorbidities has been called metabolic syndrome, 'syndrome X,' or the insulin resistance syndrome (**Table 4**), and some investigators believe that insulin resistance is the primary lesion. Whatever the pathophysiologic mechanisms, it is clear that dyslipidemia and hypertension must be sought and aggressively treated if

Table 4 The Metabolic Syndrome

Three or more of the following components:
Central obesity as measured by waist circumference
Men: >102 cm (40 in.)
Women: >88 cm (35 in.)
Fasting blood triglycerides ≥1.69 mmol/l (150 mg/dl)
Blood HDL cholesterol
Men: <1.04 mmol/l (40 mg/dl)
Women: <1.29 mmol/l (50 mg/dl)
Blood pressure ≥130/85 mmHg
Fasting glucose ≥6.1 mmol/l (110 mg/dl)

present. In fact, most evidence suggests that the management of coexisting risk factors, particularly hypertension, dyslipidemia, and smoking, is more important than the treatment of hyperglycemia in preventing morbidity and mortality.

Special aspects: Type 1 diabetes With type 1 diabetes, there is essentially no endogenous insulin secretion, due to autoimmune destruction of the insulin-producing beta cells of the pancreas. This lack of an essential hormone for life means that insulin must be injected, often multiple times daily. Furthermore, the replacement of a very finely tuned normal insulin secretory mechanism, which provides insulin precisely 'on demand,' cannot be well reproduced by injections, explaining the glycemic lability of type 1 diabetes.

Major objectives Generally, the treatment objective in type 1 diabetes is stabilization of glycemic control in an acceptable range, control of other risk factors, and thus avoidance of long-term complications. This requires close attention not only to diet but also to its interrelationships with insulin dose and timing, activity, stress, and other life factors. In fact, despite the best efforts, almost all people with type 1 diabetes are prone to wide swings of blood glucose, sometimes from 2.8 to 17 mmol/l (50–300 mg/dl) or more during a day.

To control the intrinsic 'brittleness' of type 1 diabetes, the individual needs to learn to stabilize dietary intake, making it as reproducible as possible. If carbohydrate, in particular, varies significantly from day to day and meal to meal, the person must learn to adjust insulin doses to match the changed intake. Carbohydrate counting helps stabilization of the diet or adjustment in insulin doses. It is useful for the nutritionist to understand the various insulin regimens that people with type 2 diabetes are given. Several different typical regimens, with comments on the dietary implications, are shown in **Figure 1**.

In addition to carbohydrate awareness, dietary fat intake should be taken into consideration. Dietary fat is often the main determinant of serum lipids and contributes significantly to total energy intake and thus body weight. It also delays gastric emptying, prolonging the glycemic response to dietary carbohydrate.

Very few people continue to measure and weigh foods, but weighing is a useful tool during the instruction phase. Ultimately, people with type 1 diabetes should become proficient in estimating the carbohydrate content of food so that their food selection becomes second nature.

Energy intake distribution will depend on the type of insulin, the number of injections, and the glycemic targets (very tight blood glucose control or not as tight). Often, small changes in food ingestion can make a significant difference. If, for example, a patient tends to develop hypoglycemia at approximately noon, the skillful dietitian can either emphasize the necessity of eating lunch regularly before noon or suggest the patient consume some of the lunch carbohydrates as an 11 AM snack. These changes may eliminate the need to change insulin dose.

Especially with intensive insulin therapy (three or four daily injections or an external insulin pump), there is some flexibility in the timing of the meals but also a need for more accurate assessment of meal content. Some patients will learn their own ratio of grams of carbohydrate to insulin dose necessary to maintain blood glucose in a good range.

Eating disorders pose a serious problem to the management of type 1 diabetes. Presumably because people with diabetes are often diet conscious, the prevalence of eating disorders is surprisingly high among teenagers with diabetes. The problem is especially dangerous because young people may skip insulin injections in order to induce glucosuria, a sort of 'metabolic purging.' These conditions clearly require prompt professional help.

Growth and development The total daily energy intake of a person with type 1 diabetes should be calculated to maintain normal growth and development in a child and normal weight in an adult. Examples of these calculations are provided in **Table 3**. Since most people with type 1 diabetes are not overweight, most do not need low-energy diets. Indeed, underfeeding is a poor way to maintain blood glucose control. The energy needed to establish and maintain normal weight should be matched with the insulin needed to control glycemia. There is no need for a thin or normal-weight person with type 1 diabetes to be perpetually hungry.

Special aspects of dietary management of other types of diabetes Other types of diabetes include those with relatively well-recognized etiologies, such as

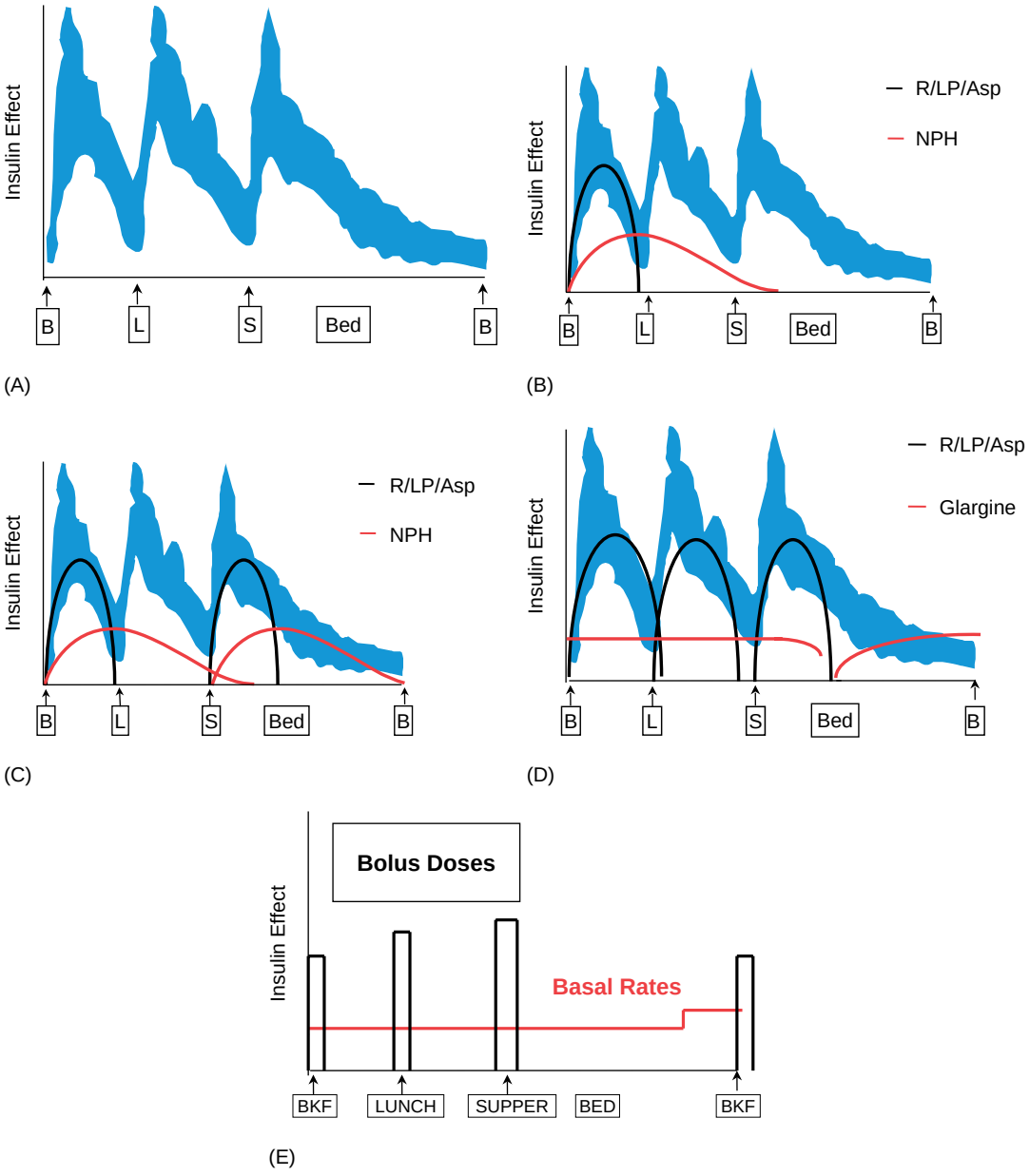


Figure 1 Insulin regimens and notes on the nutritional intake required. (A) The normal insulin response to three meals (breakfast (B), lunch (L), and supper (S)). Note that insulin increases sharply after ingestion of a carbohydrate-containing meal, declining to baseline within several hours. (B) When a combination of short-acting and intermediate-acting insulin is given only at breakfast, the normal response to breakfast is reproduced and the intermediate-acting insulin 'covers' lunch. It is important that the patient ingest a regular breakfast and lunch in order to avoid hypoglycemia from the insulin present at these times. (C) When a combination of short-acting and intermediate-acting insulin is given at breakfast and supper, there is better 'coverage' of the supper meal, but the intermediate-acting insulin peaks near bedtime and the middle of the night, so a bedtime snack may be necessary. (D) A more intensive regimen provides insulin as a 'basal' dose at bedtime, lasting the full 24 h, and short-acting insulin with every meal, for a total of four doses per day. The regimen usually requires patients to monitor their own blood glucose before each meal to adjust their short-acting dose to both the amount of carbohydrate to be ingested and the blood glucose level at the time. The regimen does provide more flexibility of meal timing. (E) Use of an external insulin pump infuses insulin at a precise basal rate, and the patient signals the pump to deliver bolus doses of insulin with each meal. As with D, regular monitoring is required as well as accurate understanding of the content of the meal to be ingested. Basal rate can be adjusted, for example, to avoid nighttime hypoglycemia, and there is flexibility of when meals are eaten.

pancreatectomy-induced diabetes, diabetes due to pancreatitis, cystic fibrosis, iron infiltration of the pancreas (hemochromatosis), or rare syndromes of insulin resistance.

Pancreatitis may be secondary to severe hypertriglyceridemia (triglyceride content >1100 mmol/l (1000 mg/dl)). In this case, a very low-fat diet is often indicated. When there is widespread

destruction of pancreatic cell mass, as with cystic fibrosis, pancreatectomy, or extensive cancer, the exocrine as well as endocrine functions are affected, leading to malabsorption and impaired glucagon secretion. Malabsorption causes steattorhea and may require pancreatic enzyme replacement to avoid marked variability in carbohydrate as well as fat absorption. Lack of the hormone glucagon increases the risk of severe hypoglycemia after insulin administration since there is less counterregulatory ability to raise blood glucose levels after mild hypoglycemia.

Effects of Ingested Nutrients on Blood Glucose

Carbohydrate

Carbohydrate ingestion causes blood glucose to increase. In people without diabetes, the normal increase in blood glucose is approximately 0.5–2.8 mmol/l (10–50 mg/dl) above baseline, returning to baseline within 1–3 h. The pancreatic hormonal response to dietary carbohydrate mediates the return to normal. Insulin is the central mediator of energy metabolism. The basics of insulin-dependent energy metabolism in the fed and the fasting states are depicted in **Figure 2**.

Although carbohydrate intake plays the major role in postprandial blood glucose, there are other factors to consider. The diet is not the only source of glucose in blood; hepatic gluconeogenesis maintains blood glucose in the absence of dietary intake. For example, when a person is ill and dietary intake is curtailed, it would be a mistake to stop insulin administration since hepatic glucose production may in fact be increased. Sick-day instruction is essential for people with diabetes so that they do not simply stop their treatment if they are not eating well. Pharmacologic therapies (insulin or oral agents), of course, also affect blood glucose.

A long-standing debate has surrounded the optimal proportion of intake from carbohydrate, fat, and protein. People with diabetes, especially when insulin is administered, will discover that if they hold back carbohydrate their blood glucose does not increase as much. Holding back carbohydrate, however, unless the diet is hypocaloric, inevitably leads to a high-fat diet, and carbohydrate restriction leaves insulin with no substrate to act on. In our experience, this can cause blood glucose levels to be more unstable, susceptible to swings of hypoglycemia and hyperglycemia. We support the recommendation of most professional guidelines that carbohydrate should make up a substantial percentage (50–60%) of total nutrient intake.

Two areas of controversy and of nutrition research deserve special attention: the glycemic response to oral sucrose (concentrated sweets) versus complex carbohydrates and the so-called 'glycemic index.'

Sucrose versus complex carbohydrate Careful metabolic studies suggest that, gram for gram, sucrose does not increase blood glucose more than complex carbohydrates, either acutely or over a matter of weeks. In these studies, sucrose was isoenergetically substituted for other carbohydrates, mostly under carefully defined research ward conditions in which precise substitutions can be made. Since complex carbohydrates and sucrose are both digested to monosaccharides before they are absorbed, it is not unexpected that each should cause the same glycemic excursion if administered in the same number of grams. It does run counter, however, to the traditional advice that people with diabetes should avoid concentrated sweets.

A number of organizations have cited these research studies in support of a recommendation that allows ingestion of concentrated sweets. The caveat, in the words of the American Diabetes Association, is that "sucrose should be substituted for other carbohydrate sources in the food/meal plan." In our view, there is a practical fallacy in this recommendation: People are unlikely to substitute sucrose for complex carbohydrates in equal amounts. Due simply to taste, concentrated sweets are likely to be taken in far greater quantity than the more filling and less sweet starches. Thus, in reality, people who routinely eat concentrated sweets are likely to have greater and less predictable glycemic excursions than those who stick to complex carbohydrates. There is also the significant risk that excess concentrated sweet intake will cause weight gain (as well as dental caries). However, if a person with diabetes can include a fixed amount of concentrated sweet in his or her diet and can demonstrate that his or her diabetes is well controlled and the postmeal glycemia is not excessive, there is no reason to deny the person the sweet.

Glycemic index The glycemic index (GI) is defined as the area under the 2-h curve of blood glucose after the ingestion of a set amount of carbohydrate compared to ingestion of the same amount of carbohydrate from a reference food (white bread or glucose). The GI is expressed as a percentage of the standard food value:

$$\text{Glycemic index} = \frac{\text{Area under the curve of test food}}{\text{Area under the curve of standard food}} \times 100$$

The glycemic load (GL) is an additional measure in which the amount of carbohydrate in a typical

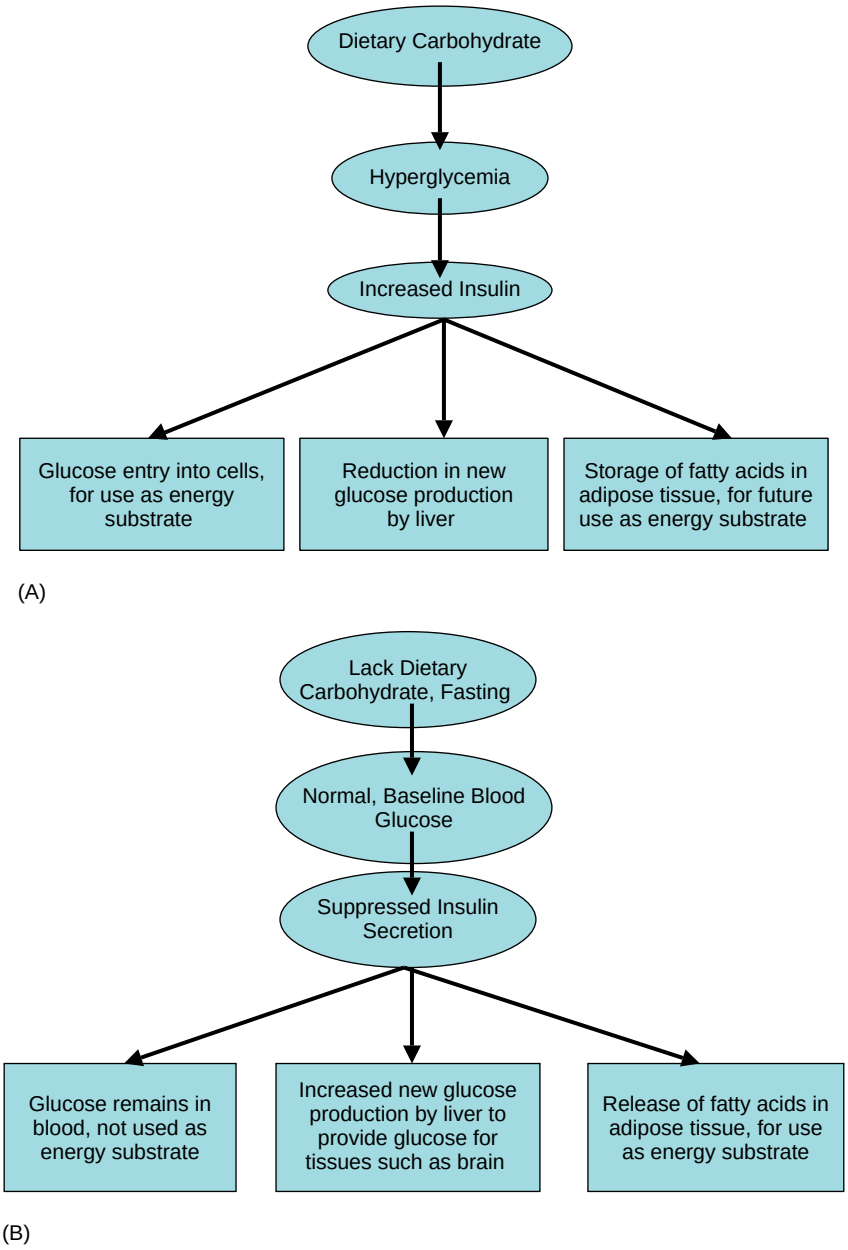


Figure 2 Influence of insulin on basic energy metabolism. (A) With dietary carbohydrate intake, hyperglycemia induces insulin secretion that acts to enhance glucose entry into cells for utilization as metabolic fuel. Simultaneously, insulin decreases new glucose production in the liver, since dietary glucose is already available, and stores excess caloric intake in adipose tissue as fat. (B) With lack of dietary carbohydrate, as in fasting, the reverse occurs: With lower blood glucose, insulin secretion is suppressed. This minimizes entry of glucose into cells but stimulates enough new glucose production from the liver to provide for obligate glucose using tissues such as the brain. Meanwhile, low insulin concentration promotes fatty acid release from adipose tissue to serve as an alternate fuel for metabolism.

portion is taken into account. **Table 5** provides examples of foods high and low in GI and GL. These indices have been calculated for more than 500 different carbohydrates, and values are readily available on the Internet. A number of factors in addition to the reported GI and GL actually affect the blood glucose response to meals, however, because mixed meals are ingested in everyday living. Among these are the fat and fiber content of the meal, type of cooking, the patient's absorptive rate, and micronutrient content.

In our opinion, the concept of the GI is valid in a research sense: Certain carbohydrates, gram for gram, do raise blood glucose levels more, or with different glycemic patterns, than others. However, we believe that basing nutrition plans on the GI and the GL of foods is usually too much of a burden for people with diabetes, who have to closely monitor

Table 5 Examples of foods high and low in glycemic index (GI) and glycemic load (GL)[a]

	GI	*Serving size (g)*	*GL*
Low GI/low GL			
Apple, NS (USA)	40	120	6
Oranges (Sunkist, USA)	48	120	5
Healthy Choice hearty 7-grain bread (USA)	55 ± 6	30	8
Ice cream, premium, French Vanilla—16% fat (Australia)	38 ± 3	50	3
Kidney beans (USA)	23	150	6
Pizza, Super Supreme, thin and crispy—13.2% fat, Pizza Hut (Australia)	30 ± 4	100	7
Low GI/high GL			
Barley (*Hordeum vulgare*) (India)	43 ± 6	150	26
High GI/low GL			
Watermelon, raw (Australia)	72 ± 13	120	4
White wheat flour bread	70	30	10
High GI/high GL			
Cornflakes (Kellogg's, USA)	92	30	24
Bagel, white, frozen (Lenders, Canada)	72	70	25
White rice, type NS, boiled 13 minutes (Italy)	102	150	31

[a]High GI is considered >70 and low <55. High GL is considered >20 and low <10.

Adapted from Foster-Powell K, Holt SH, and Brand-Miller JC (2002) International table of glycemic index and glycemic load values. *American Journal of Clinical Nutrition* **76**(1): 5–56.

the total amount of carbohydrates. It is more practical to encourage people to learn their own glycemic response to different foods from experience. They may learn, for example, that far more insulin is needed before eating pizza or a bagel; they may learn to avoid certain deserts. A general awareness of what preferred foods, in what amounts, raise blood glucose may be more practical than memorizing GI or GL.

Protein

Since the classic experiments by Benedict in the 1910s, it has been known that protein ingestion causes hyperglycemia and glucosuria. The effect of protein ingestion on blood glucose, however, is far less pronounced than the effect of carbohydrate ingestion. A rule of thumb is that a gram of protein raises blood glucose approximately one-third as much as a gram of carbohydrate. In most diets, approximately 50–100 g protein is ingested per day, compared to approximately 200–300 g carbohydrate. Therefore, protein is a calorically less significant part of the diet and far less important in regulating blood glucose. In people with type 2 diabetes, protein does not slow the postprandial absorption of carbohydrate. The same cannot be said about dietary fat.

Fat

Dietary fat has little, if any, immediate effect on blood glucose concentration because the constituent fatty acids do not produce new glucose and the glycerol moieties are insignificant in their contribution to blood glucose. However, there is considerable evidence that circulating free fatty acids promote gluconeogenesis and hyperglycemia. In a normal overnight fast this is good: Fatty acids help maintain normoglycemia. However, in uncontrolled diabetes, when fatty acids can be very high, they significantly worsen hyperglycemia. This has been referred to as 'fat toxicity.'

Fat ingestion slows gastric emptying. The delayed delivery of carbohydrate to the circulation can cause a late, slow postprandial rise in blood glucose, although people who do not self-monitor frequently are unlikely to be aware of this effect of dietary fat.

Nonnutritive Sweeteners

Sweeteners are important to the quality of life of people with diabetes. An essential distinction is to differentiate those with from those without significant energy content. **Tables 6** and **7** provide many of the available nonnutritive and nutritive sweeteners. The nonnutritive sweeteners have no or virtually no energy content, and they can be consumed without concern about their effect on blood glucose.

Table 6 Nonnutritive sweeteners

Type	*US brand names*	*kcal/g*	*Description*
Saccharin	Sweet and Low, Sweet Twin, Sweet 'N Low Brown, Necta Sweet	0	200–700 times sweeter than sucrose; noncarcinogenic and produces no glycemic response
Aspartame	Nutrasweet, Equal, Sugar Twin (blue box)	4	160–220 times sweeter than sucrose; noncarcinogenic and produces limited glycemic response
Acesulfame-K	Sunett, Sweet & Safe, Sweet one	0	200 times sweeter than sucrose; noncarcinogenic and produces no glycemic response
Sucralose	Splenda	0	600 times sweeter than sucrose; noncarcinogenic and produces no glycemic response

Table 7 Polyols and novel sugar sweeteners

Type	*kcal/g*	*Description*
Monosaccharide polyols or novel sugars		
Sorbitol	2.6	50–70% as sweet as sucrose; some people experience a laxative effect from a load ≥50 g
Mannitol	1.6	50–70% as sweet as sucrose; some people experience a laxative effect from a load ≥20 g
Xylitol	2.4	As sweet as sucrose
Erythritol	0.2	60–80% as sweet as sucrose; also acts as a flavor enhancer, formulation aid, humectant, stabilizer and thickener, sequestrant, and texturizer
Disaccharide polyols or novel sugars		
Isomalt	2	45–65% as sweet as sucrose; used as a bulking agent
Lactitol	2	30–40% as sweet as sucrose; used as a bulking agent
Maltitol	2.1	90% as sweet as sucrose; used as a bulking agent
Polysaccharide polyols		
HSH	3	25–50% as sweet as sucrose; other names include hydrogenated starch hydrolysates and maltitol syrup

Adapted with permission from the Journal of the American Diabetic Association Position Paper: Use of Nutritive and Nonnutritive Sweetners.

Many 'diet' sweeteners, such as sorbitol or fructose-based snacks, do cause at least some degree of hyperglycemia. Sugar alcohols (polyols) such as sorbitol, mannitol, and xylitol are classified as hydrogenated monosaccharides, hydrogenated disaccharides, and oligosaccharides. They do contain calories, but because they are only partially absorbed in the small intestine, they have a reduced energy value per gram. Excessive use of sugar alcohols has laxative effects and can cause diarrhea.

It is important for people with diabetes to understand clearly these distinctions because many calories can be ingested with foods labeled as 'diet' under the false assumption that they are without effect on blood glucose.

Trace Elements

There has been recurrent interest in whether such trace elements as chromium, potassium, magnesium, vanadium, and zinc affect blood glucose control in diabetes. It would obviously be attractive if simple oral supplements could facilitate normoglycemia. The evidence, though, is slim and unconvincing that supplementation of any of these trace elements has a beneficial effect except when there is a true deficiency. Such deficiency may occur in an undernourished setting, in the elderly with poor dietary intake, or in certain strict vegetarian diets.

The same conclusion can be reached regarding vitamin supplementation: It is indicated when vitamin deficiency is suspected or likely. For example, populations such as the elderly, those pregnant or lactating, strict vegetarians, or those on a calorie-restricted diet supplementation may require vitamin supplements. Folate supplementation is well documented to improve the outcome of pregnancy, with or without diabetes. However, there is no clear evidence that supplementation is helpful for those eating an adequate diet.

Antioxidants such as vitamins C, E, or A and α-lipoic acid are the object of intensive research. It is unclear whether or which of these actually prevent long-term complications of diabetes, but the literature should be monitored. Vitamins B_1, B_6, and B_{12} are sometimes used to treat diabetic peripheral neuropathy, but without much supporting evidence of benefit. On the other hand, calcium supplementation is indicated, particularly in the elderly, if daily intake is less than 1.0–1.5 g.

In summary, evidence is weak that vitamin or trace element deficiencies occur due to diabetes. Supplementation in more normal circumstances has little or no role in the control of diabetes, and general nutritional guidelines for vitamins and trace elements should be followed.

Major Nonnutrient Factors That Regulate Blood Glucose

No element of diabetes management exists in a vacuum, so it is essential to consider how dietary therapy interacts with other elements of management. Most prominent are insulin, oral hypoglycemic agents, activity, and stress.

Insulin

As two major types of diabetes (1 and 2) differ, so the use of insulin differs for each. As described earlier, in type 1 diabetes the insulin doses must be closely matched to the meals ingested. Too much insulin or too little ingested carbohydrate can cause serious hypoglycemia. Frequently, patients are on

intensive insulin regimens, sometimes four doses per day, and sometimes using an insulin pump. People with well-controlled type 1 diabetes have usually learned to pay close attention to their carbohydrate intake, recognize portion sizes, or even count grams of carbohydrate. They often adjust insulin dose or carbohydrate intake, but this can be done effectively only if they have a good, quantitative understanding of both. Meals skipped or eaten late can be a problem.

Intensive insulin regimens may involve the use of an insulin pump or insulin doses based on a 'sliding scale.' Ordinarily, sliding scales are developed for the patient based on the self-monitored blood glucose at the time of the meal. For higher blood glucose levels, more short-acting insulin is administered. In more intensive regimens, often with insulin pump use, the amount of insulin delivered is also adjusted depending on both blood glucose level at the time and the carbohydrate to be ingested. To be safe and effective, this self-adjusted fine-tuning of insulin dose requires considerable knowledge of diet as well as insulin.

Examples of sliding scales are provided in **Table 8**. Some people, particularly those with intensive insulin regimens, learn to adjust their insulin dose according to both the blood glucose at the start of a meal and the estimated amount that a unit of insulin will reduce their blood glucose. Examples are provided in **Table 9**. Also, although nutritionists do not usually prescribe changes in insulin dosage, it is useful to know the various types of insulin available (**Table 10**) and patterns of insulin action (**Figure 1**).

With type 2 diabetes, since there is usually some insulin secreted on demand from the pancreas and considerable resistance to insulin, dietary intake may not have to be so precisely regulated according to insulin dose. There is a risk of hypoglycemia but it is usually less than that seen in type 1 diabetes. As discussed previously, the nutritional emphasis is usually on reducing the caloric intake, with careful control of fat as well as carbohydrate.

Oral Antidiabetic Agents

Oral antidiabetic agents are not insulin; insulin is delivered only by injection or infusion. The variety of agents in use has escalated dramatically in recent years, so it is worth knowing how the various classes act and how they may interact with diet.

Sulfonylureas (e.g., glyburide, glimepiride, and glipizide) commonly act by stimulation of pancreatic insulin secretion. They therefore can cause hypoglycemia if taken in excess or without normal food intake.

The other most popular oral agent is metformin, which does not stimulate insulin secretion and therefore should not cause hypoglycemia by itself. Metformin can cause bloating and diarrhea, but it can also be mildly weight reducing in conjunction with diet.

The drugs called thiazolidinediones (TZDs), pioglitazone and rosiglitazone, improve insulin sensitivity but do not by themselves cause hypoglycemia. TZDs can, however, cause fluid retention and weight gain, so they are sometimes counterproductive in someone trying to lose weight.

Finally, a class of drugs called α-glucosidase inhibitors (acarbose and miglitol) inhibit digestion and absorption of carbohydrate. They do not cause hypoglycemia, but they may interfere with the treatment of hypoglycemia by oral carbohydrate.

Physical Activity

The effects of exercise on blood glucose levels are complex and sometimes unpredictable. Although moderate, extended aerobic exercise generally causes progressive lowering of blood glucose, intense exercise may transiently increase the blood glucose. We generally recommend modification of diet to accommodate exercise, rather than changing the dose of

Table 8 Examples of sliding scale insulin doses based on blood glucose level at the time of the meal[a]

Meal	*Blood glucose (mmol/l)*	*<3.33*	*3.33–6.7*	*6.8–8.3*	*8.4–11.1*	*11.2–13.9*	*14–16.7*	*>16.7*
	Blood glucose (mg/dl)	*<60*	*60–120*	*121–150*	*151–200*	*201–250*	*250–300*	*>300*
Breakfast	NPH/regular	8/0	8/4	8/5	8/6	10/7	10/8	10/9
Lunch	Regular	0	2	3	4	5	6	7
Supper	Regular	2	4	6	7	8	9	10
Bedtime	NPH/regular	6/0	6/0	7/0	7/0	8/0	8/2	8/3

Example: Prebreakfast blood glucose 10 mmol/l (180 mg/dl): Take 8 units NPH + 6 units regular.

[a]Each patient will have individual insulin requirements and needs. Longer acting insulin (such as glargine) may be used at bedtime, and faster acting insulins (such as lispro or aspart) may be used premeal.

Table 9 Examples of mealtime insulin dose based on blood glucose level *and* amount of carbohydrate to be ingested[a]

Short-acting or fast-acting insulin dose
'Correction factor': 1 unit per 2.8 mmol/l (50 mg/dl) >5.5 mmol/l (100 mg/dl)
plus
'Meal factor': 1 unit per 15 g carbohydrate
Example: Before ingesting a meal estimated to have 45 g of carbohydrate, the blood glucose is measured to be 150. Insulin dose would be 1 unit (for the blood glucose) + 3 units (for the carbohydrate) = 4 units.

[a]Both the correction factor and the meal factor will vary from patient to patient. Long acting insulins are used in addition to these prandial doses.

insulin or oral agents, since the duration and intensity of exercise may be unpredictable. Trained athletes or 'weekend warriors' often learn to take extra carbohydrate before a strenuous workout rather than anticipating exercise by reducing the morning insulin dose. They also learn that the time of greatest hypoglycemic risk may be 6–12 h after exercise.

Stress

Stress in normal life is difficult to quantify or study, but the usual experience is that diabetic control deteriorates under stress. This makes sense, considering the hyperglycemic effects of 'fight or flight' hormones such as adrenaline (epinephrine) and cortisol. However, it is likely that the main problem for people under unusual stress is less hormonal than behavioral—simply neglecting their diet or eating as a stress reliever. Therapists may be most effective by reminding patients to maintain their normal diet even when experiencing emotional stress.

Dietary Prevention and Management of Comorbidities

Accelerated Atherosclerosis

Essentially the same nutritional approaches to the prevention of atherosclerosis apply whether or not a person has diabetes. However, they are even more important for the patient with diabetes since hyperglycemia is a risk factor, and most people with diabetes die of atherosclerotic cardiovascular disease. Therefore, anyone with diabetes should follow a 'heart healthy' diet that focuses on lowering low-density lipoprotein (LDL) cholesterol level, which is a major contributor to the progression of atherosclerosis. Total fat intake can be held to 25–35% of total calories, less than 7% saturated fat and the remainder divided between monounsaturated fat and polyunsaturated fats. The recommendation allows for increased intake of unsaturated fats in place of carbohydrates in people with diabetes. In addition to the antiatherosclerotic diet, there should be routine screening for other specific risk factors, notably hypertension and dyslipidemia. If found, these risk factors, which are even more dangerous in diabetes, should be vigorously treated.

Dyslipidemia

If the dyslipidemia is predominantly elevation of LDL cholesterol, then dietary manipulations are no different in diabetes from those used to treat hypercholesterolemia generally: Low intakes of saturated fat and cholesterol are indicated, with cholesterol-lowering medications (usually statins) given as needed. Target goals for LDL cholesterol are <100 mg/dl, and there have been recommendations to lower LDL

Table 10 Types of available insulin by onset, peak, and duration of action

Category	*Insulin type*	*Approximate onset*	*Approximate peak*	*Approximate duration*
Fast acting	Aspart	5–10 minutes	1–3 h	3–5 h
	Lispro	<15 minutes	0.5–1.5 h	2–4 h
Rapid acting	Regular	0.5–1 h	2–3 h	3–6 h
Intermediate acting	NPH	2–4 h	4–10 h	10–16 h
	Lente	3–4 h	4–12 h	12–18 h
Long acting	Glargine	1.1 h	?	24 h
	UltraLente	6–10 h	18–20 h	20–24 h

Premixes
50/50: 50% NPH, 50% regular
70/30: 70% NPH, 30% regular
Mix 70/30: 70% aspart–protamine suspension (e.g., NPH), 30% aspart
75/25: 75% lispro–protamine suspension (e.g., NPH), 25% lispro

[a]Onset, peak, and duration will vary for each individual. Based on American Diabetes Association (2004) *Diabetes Forecast, Resource Guide*. Alexandria, VA: American Diabetes Association.

cholesterol to as low as <70 mg/dl if there are other risk factors for coronary artery disease.

Hypertriglyceridemia, on the other hand, is the more common dyslipidemia in diabetes, and especially it is dangerous when it is associated with low levels of high-density lipoprotein cholesterol. Insulin–glucose homeostasis is intrinsically and complexly related to triglyceride metabolism. Insulin stimulates both very low-density lipoprotein–triglyceride synthesis in the liver and its clearance via lipoprotein lipase in the periphery.

In extreme cases, reduced chylomicron clearance causes 'diabetic lipemia,' characterized by chylomicronemia and extreme hypertriglyceridemia. More moderate levels of hypertriglyceridemia, on the other hand, are considered to be due to overproduction of hepatic (endogenous) lipid under the influence of hyperinsulinism and peripheral insulin resistance.

Compounding the confusion is the phenomenon of 'carbohydrate induction of hypertriglyceridemia.' Most normal people switched isoenergetically from a low- to a high-carbohydrate diet (i.e., with less dietary fat) will actually increase their fasting triglyceride level. This carbohydrate induction of hypertriglyceridemia may be transient, lasting only a few weeks.

For the person with diabetes, treatment of hypertriglyceridemia begins with optimization of blood glucose control and a diet designed to achieve normal body weight. Weight reduction is often very effective.

If hypertriglyceridemia persists, evidence favors the use of monounsaturated fats when dietary fat increases. Since hypertriacylglyceridemia is sometimes associated with excess alcohol, reducing the amount consumed may be effective. For the unusual condition of fasting chylomicronemia, with triglyceride levels remaining over approximately 11 mmol/l (1000 mg/dl), a low-fat diet is necessary to avoid exacerbating the situation by adding dietary fat and precipitating pancreatitis from the chylomicronemia.

If pharmacologic treatment is necessary to treat hypertriglyceridemia, a fibric acid derivative, such as gemfibrazol or fenofibrate, or nicotinic acid should be used. Statins are usually the first drug of choice when LDL cholesterol is also elevated.

Hypertension

The current nutritional management of hypertension focuses on reducing dietary sodium intake and weight reduction, as well as the recently proven 'DASH' diet. There has been long-standing evidence that in both normal and hypertensive people, a reduction in sodium intake lowers blood pressure. The DASH diet was shown to be effective in a large trial of dietary intervention. It is a fruit and vegetable diet, with balanced consumption of foods emphasizing high fiber, grains, and low-fat dairy products. Potassium, magnesium, and fiber replace some snacks and sweets.

Modest amounts of weight loss and increased activity are also beneficial for the person with hypertension. Thus, overweight and obese individuals should be encouraged to lose weight as part of their medical therapy. In diabetes, ACE inhibitors or ARBs (Angrotensin Receptor Blockers) are usually the first line of medication used when diet and exercise are not effective in controlling blood pressure. Frequently, additional antihypertensives must be added.

Renal Disease

Nutritional therapy of established diabetic nephropathy continues to be studied. In both type 1 and type 2 diabetes, persistent microalbuminuria is a strong predictor of gross proteinuria and developing nephropathy. Evidence suggests that microalbuminuria actually reverses in many cases, but gross proteinuria (over approximately 300 mg/24 h) in most cases eventually progresses to end-stage renal disease. Therefore, treatment focuses on reversing or at least retarding the progression of nephropathy. In recent years, the nutrition recommendation has been to lower dietary protein intake to 0.8–1.0 g/kg of body weight per day for patients with microalbuminuria. For people with overt nephropathy, reducing dietary protein intake to 0.8 g/kg of body weight per day may slow the progression of nephropathy. Protein restriction should not be attempted if serious protein loss from nephrotic-range proteinuria has reduced total serum albumin concentration, in which case it is agreed that a low-protein diet is not indicated. In fact, a large study of the dietary treatment of kidney disease in nondiabetic subjects did not support the value of protein restriction in slowing progression of kidney damage.

The best documented therapies for proteinuria and reducing nephropathy are improved glycemic control and reduction of blood pressure using ACE inhibitors or ARBs. Cessation of smoking is also of established benefit.

Conclusions

Medical nutrition therapy is essential to all people with diabetes, of whatever type or severity. A healthy diet should contain important components, including foods containing carbohydrates from whole grains, fruits, vegetables, vitamins, and low-fat dairy

products. Although blood glucose control and management of coexisting risk factors are overall goals, the implementation of dietary management is a highly complex and individualized process. Principles of medical nutrition therapy that generally apply to all diabetes, and specific features that apply to the various types of diabetes, have been discussed; but it must be emphasized that good nutritional management requires the close interaction of each individual patient with a knowledgeable expert in dietetics.

See also: **Carbohydrates**: Chemistry and Classification; Regulation of Metabolism. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology. **Glucose**: Metabolism and Maintenance of Blood Glucose Level. **Obesity**: Complications. **Sucrose**: Nutritional Role, Absorption and Metabolism; Dietary Sucrose and Disease. **Weight Management**: Approaches.

Further Reading

American Diabetes Association (2004) Position statement: Nutrition principles and recommendations in diabetes. *Diabetes Care* **27**(supplement 1): S36–S46.

Connor H, Annan F, Bunn E *et al.* Nutrition Subcommittee of the Diabetes Care Advisory Committee of Diabetes UK (2003) The implementation of nutritional advice for people with diabetes. *Diabetic Medicine* **20**(10): 786–807.

Eyre H, Kahn R, and Robertson RM (2004) Preventing cancer, cardiovascular disease, and diabetes: A common agenda for the American Cancer Society, the American Diabetes Association, and the American Heart Association. American Cancer Society, the American Diabetes Association, and the American Heart Association Collaborative Writing Committee. *Diabetes Care* **27**(7): 1812–1824.

Franz MJ, Bantle JP, Beebe CA *et al.* (2002) Evidence based nutrition principles and recommendations for the treatment and prevention of diabetes and related complications [Technical review]. *Diabetes Care* **25**: 148–198.

DIARRHEAL DISEASES

A Baqui, R Heinzen, M Santosham and R Black, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Global Burden of Diarrhea and Epidemiological Trends

Diarrheal illnesses in young children are among the leading causes of morbidity and mortality in developing countries. Diarrhea is an important cause of morbidity in developed countries as well. In developing countries, children younger than 5 years old suffer 3–10 episodes of diarrhea per year, whereas in developed countries young children have on average 1 or 2 diarrheal episodes per year. The advent of oral rehydration therapy (ORT) and its use in the past three decades have dramatically reduced the case fatality rate for diarrhea. However, globally the estimated 3 billion annual episodes of diarrhea account for approximately 2 million deaths in children younger than 5 years old. The majority of diarrhea-related mortality occurs in developing countries and the highest rates of diarrhea occur among infants with malnutrition. The case fatality rate is highest among children 6 months to 1 year old. The primary reason is that for most children this is the period when the immune system is not yet fully matured and the maternal antibodies are reduced. In addition, they may receive contaminated foods to complement breast-feeding, and they begin to crawl, potentially to areas where they may have direct contact with human or animal feces.

Although dehydration is the most direct effect of diarrhea, there are many adverse and potentially fatal nutritional consequences when proper nutritional management is not followed. This article provides an overview of the epidemiology of diarrheal diseases, including the interaction of diarrhea and malnutrition, and discusses the treatment of diarrhea, including fluid therapy and dietary management to minimize the nutritional cost of diarrhea.

Diarrhea–Malnutrition Interaction

Diarrheal illnesses are more common, last longer, are clinically more severe, and are more likely to have a fatal outcome for impoverished children in less developed countries because of a complex interaction between infection, protein-energy malnutrition, and micronutrient deficiencies. Diarrhea and malnutrition have a bidirectional relationship in which malnutrition increases the incidence and duration of diarrhea and, conversely, diarrhea exerts a negative effect on nutritional status.

Malnourished children have defects in cell-mediated immune functions and a decrease in IgA-containing cells in the jejunal mucosa. Malnutrition produces morphological and functional changes in virtually all organs. Changes in the intestine include thinning of the gut epithelium, marked flattening and broadening of villi, extensive infiltration of the lamina propria, and diminished secretion of gastric acid. These changes lead to an increased risk of diarrhea. The risk of developing diarrhea may be predicted by a child's ability to respond to skin test antigens. This effect of impaired cellular immunity is independent of age and nutritional status. Studies have shown that patients with a variety of viral and bacterial infections, such as measles, influenza, tuberculosis, and streptococcal infections, develop impaired cell-mediated immunity (anergy). Anergy or reduced responsiveness may also be associated with single nutrient deficiency, such as deficiency of vitamin A, zinc, pyridoxine, folic acid, and iron. Therefore, it may be possible that in an otherwise healthy child, infection or micronutrient deficiencies induce immunodeficiency may increase susceptibility to diarrhea and other infections and lead to a vicious cycle of repeated infections, anergy, and deteriorating nutritional status.

Diarrheal diseases have generally been noted to have an adverse effect on growth of underprivileged children in developing countries. Community-based prospective studies in developing countries have consistently demonstrated a significant negative effect of diarrhea on short-term weight gain, but studies on the effect on short-term height gain are less consistent. On the contrary, some investigators have concluded that the effect of diarrhea on growth is transient and that efforts to control diarrhea are unlikely to improve children's nutritional status. One possible explanation for the discrepant findings is the heterogeneity of diarrheal illnesses. Most studies that have examined the effect of diarrhea on growth have considered diarrhea as a single entity. Studies that have examined the effect of diarrhea on growth by type of diarrheal illness have suggested that certain etiological (e.g., *Shigella* spp.) and clinical-type (dysentery) illnesses are associated with significant growth retardation. It has been estimated that diarrheal illnesses account for 10–80% of growth retardation in the first few years of life, with the magnitude of effect possibly modified by other factors, such as etiology and clinical type, the source and adequacy of dietary intake, treatment and feeding practices during and following illness, and the opportunity for catch-up growth after illness. A cohort study conducted in Bangladesh observed that dysentery had the most deleterious consequences on both ponderal and linear growth, although other types of diarrhea showed a similar but relatively less pronounced negative effect on growth. Losses of approximately 0.5 kg in annual weight gain and 1.25 cm in annual height gain were associated with dysentery in children <5 years old in this population.

Understanding the mechanisms of diarrhea-induced undernutrition and appropriate treatment of diarrhea is important for the immediate illness and also for the long-term well-being of children. Approximately 10% of diarrheal episodes result in persistent diarrhea lasting more than 2 weeks. These infants need specialized treatment in addition to rehydration therapy, such as antibiotics for concurrent infections, micronutrients, and careful dietary management. In many settings, up to 50% of diarrhea-associated deaths occur in children with malnutrition-associated persistent diarrhea.

Clinical Types and Etiology of Diarrhea

Since nutritional costs of diarrhea vary by etiology and clinical type, a discussion of different types of diarrhea is pertinent. Diarrheal episodes can be classified based on clinical presentation as inflammatory (dysentery) and noninflammatory (nondysentery) diarrhea. Therefore, the clinical presentation of diarrheal illnesses may suggest a causative diagnosis. Diarrheal episodes can also be classified based on duration as acute (<14 days) and persistent (≥ 14 days) diarrhea. The diarrhea is generally due to either infectious or noninfectious causes. This article focuses on infectious diarrhea. Pathogens that cause infectious diarrhea include bacteria, viruses, and parasites (**Table 1**).

Most diarrheal episodes are acute. Occasionally, they become prolonged, leading to a vicious cycle of malabsorption, malnutrition, and failure to thrive. Noninfectious diarrhea tends to be persistent because it is often due to a chronic health problem. However, most persistent diarrhea is due to infection or is a sequeale of infection. Infections can lead to persistent diarrhea in the following situations: (i) Some pathogens cause chronic symptoms, usually parasites but sometimes bacteria; (ii) immunosuppressed individuals, such as those with human immunodeficiency virus (HIV) infection, cannot effectively clear pathogens and can develop persistent diarrhea; and (iii) at times the infection clears, but people develop chronic symptoms, such as irritable bowel syndrome, with diarrhea. If the diarrhea is persistent, it can cause dehydration, malnutrition, and systemic infections. Diarrhea in combination with a severe case of malnutrition (e.g., marasmus and kwashiorkor) is very dangerous and can lead to

Table 1 Infectious causes of diarrhea

Bacterial
Shigella species
Salmonella
Campylobacter jejuni
Enteroadherent *Escherichia coli*
Enteroinvasive *Escherichia coli*
Enterohemorrhagic *Escherichia coli*
Yersinia enterocolitica
Staphylococcus aureas
Bacillius cereus
Vibrio cholerae non-O group 1
Vibrio parahaemolyticus
Listeria monocytogenes
Aeromonas hydrophilia
Plesliomonas shigelloides
Viruses
Rotavirus
Norovirus
Enteric adenovirus
Calcivirus
Astrovirus
Coronavirus
Cytomegalovirus
Parasites
Giardia lamblia
Entamoeba histolytica
Cryptosporidium
Cyclospora
Isospora belli
Microsporidia
Strongyloides
Food poisoning
Staphylococcus aureus
Clostridium perfringens
Bacillus cereus
Drugs, especially antibiotic induced
Clostridium difficile

systematic infections and death. Although many different mechanisms may contribute to persistent diarrhea, the result is a similar pathophysiologic syndrome of mucosal atrophy, inflammation, and malabsoption. Therapeutic efforts should concentrate on nutritional rehabilitation.

Factors Influencing Nutritional Decline in Diarrhea

Factors that influence a nutritional decline during a period of diarrhea include reduced food intake, diminished nutrient absorption due to malabsorption of macro- and micronutrients and shorter intestinal transit time, direct loss of protein and other nutrients, and an increase in the body's demand for nutrients. In addition, diarrhea of infectious origins causes cytokine-induced malnutrition, which results from the action of proinflammatory cytokines such as tumor necrosis factor and interleukin-1, -6, and -8.

Food Intake during Diarrhea

Food intake during diarrhea is often reduced due to poor appetite (anorexia), vomiting, deliberate withholding of food, and inappropriate dietary supplementation with diluted food items. Diarrhea can also be associated with fever. Both fever and anorexia have clear effects on the nutritional status of the host. An increase in body temperature of 1 °C causes an increase in the basal metabolic rate of 12–23%. Although the reason for anorexia is not clear, its effect can be important. In a controlled study in Bangladesh, 41 hospitalized children with acute diarrhea consumed only about half of the total calories consumed by healthy children despite an educational intervention.

Malabsorption of Nutrients

Malabsorption in diarrheal illness may result from the epithelial destruction by the pathogen. Diminished nutrient absorption often begins during acute diarrhea. At this time, the body is less able to absorb needed macronutrients, including fats and proteins, as well as some carbohydrates. This is most severe in undernourished children who suffer from persistent diarrhea due to damage to the gut epithelium. When the gut is damaged, food is not properly digested or absorbed. The causes of insufficient nutrient absorption include diminished concentration of bile acids, which are used for fat absorption; damaged epithelial cells, which provide the absorptive surface on the bowel; and a deficiency of disaccharides due to damaged microvilli, which normally produce the needed enzymes. In symptomatic rotavirus infection, the most common cause of acute severe diarheal illness worldwide, there is a 42% decrease in the absorption of nitrogen and fat, a 48% decrease in absorption of carbohydrates, and a 55% decrease in the total carbohydrate absorption. Malabsoption is more severe in both ETEC and shigella diarrhea. In shigellosis, loss of protein and vitamin A is sizeable. Giardiasis leads to malabsorption of vitamin A. A large amount of zinc is lost in diarrhea, usually resulting in a negative net zinc balance during the illness.

Increased Demand for Nutrients

During episodes of diarrhea, the body requires more nutrients than normal. This is due to the need to repair the damage to the gut epithelium, the increased metabolic demand on the body made by a fever, and, during dysentery, the need to replace

serum protein lost by exudation through the damaged intestinal mucosa.

Management of Diarrhea

An invariable accompaniment of diarrhea, particularly persistent diarrhea, is protein-energy malnutrition. However, dehydration is the most immediate complication of diarrhea. Clinical management of acute diarrhea includes four major components: (i) replacement of fluid and electrolyte losses, (ii) zinc therapy, (iii) antimicrobial therapy when indicated, and (iv) continued feeding to supply a sufficient quantity of nutrients to meet both the patient's usual maintenance requirement and the increased needs imposed by infection and malabsoption.

Fluid Therapy

The majority of the diarrhea-associated deaths result from dehydration. Parents should be encouraged to increase fluid intake as soon as diarrhea begins and to give oral rehydration solution if available. Children presenting with diarrhea should be assessed for dehydration. Thirst is an early sign of dehydration in a child. Other signs are mucosal dryness (e.g., dry mouth), sunken eyes, and loss of skin turgor. The coexistence of fever or vomiting exacerbates dehydration. The World Health Organization (WHO) guidelines classify dehydration into two categories—some dehydration and severe dehydration. Weight loss is the main clinical index of degree of dehydration. A vast majority of children with diarrhea present with some dehydration or no clinical signs of dehydration. The cornerstone of treatment for these children is oral rehydration solution (ORS) containing glucose or sucrose and electrolytes. ORS is effective but it must start as soon as diarrhea starts. In children with some dehydration, approximately 100 ml/kg of body weight of ORS should be given within 4 h. Ongoing stool losses should be replaced with ORS. Rehydration and maintenance of hydration in a vomiting child is feasible using ORS by giving small amounts frequently. Severe dehydration is a medical emergency that requires immediate intravenous fluid replacement, and children should preferably be hospitalized. Patients presenting with severe dehydration should receive 40 ml/kg body weight of Ringers lactate or similar intravenous solutions over a 4-h period. ORS should be given as soon as the child is able to drink.

Regardless of etiology, watery diarrhea requires fluid and electrolyte replacement. For more than three decades, an ORS containing 90 mmol/l of sodium and 111 mmol/l of glucose was used throughout the world. This solution has been credited with saving millions of lives. WHO and UNICEF have recommended the use of a new reduced osmolarity ORS formulation consisting of 75 mmol/l of sodium and 75 mmol/l of glucose and total osmolarity of 245 mOSm/l. This recommendation was made on the basis of studies that have demonstrated that the reduced osmolarity ORS was at least as efficacious as the standard ORS containing 90 mmol/l of sodium and an osmolarity of 311/l. In addition, in a meta-analysis, the reduced osmolarity ORS was shown to decrease the need for unscheduled intravenous therapy by 33%, the stool output was reduced by 20%, and the incidence of vomiting was reduced by 30%. However, there is concern that this low osmolar ORS may lead to asymptomatic and symptomatic hyponatremia in adults with severe diarrhea. This issue needs to be evaluated in large-scale effectiveness trials. Despite the proven efficacy of ORS, only approximately 20% of children receive appropriate ORS therapy during diarrheal episodes. The barriers to use of ORS include lack of knowledge of the importance of rehydration therapy, lack of access to ORS, and the perception that ORS is not a medicine since it does not stop the diarrhea.

The management of a child with persistent diarrhea is often difficult due to other related heath issues. These children are more likely to be severely undernourished due to micronutrient and protein-energy malnutrition as well as more prone to systematic infections. Due to the systematic infections, appropriate antibiotic therapy is needed.

Dietary Management of Diarrhea, Including Persistent Diarrhea

Data suggest that continued feeding during diarrhea is generally well tolerated and it minimizes the nutritional cost associated with diarrhea. A child should receive the same type of food during an episode of diarrhea as when the child is healthy. Feeding is usually tolerated, with the occasional exception of lactose intolerance. A small subgroup of children exclusively receiving nonhuman milk may have a higher rate of complications. These children should be closely supervised and provided with alternatives if needed. Full feedings will help to minimize growth faltering and a decline in nutritional status. Growth faltering may still occur, especially in severely undernourished children, due to poor nutrient absorption.

Importance of Continued Breast Feeding

Breast feeding should be continued for as long as the child can tolerate it during episodes of acute

diarrhea as well as during persistent diarrhea. It is normally well tolerated during such episodes and has been shown to reduce stool output and decrease the duration of the illness compared to those of non-breast-fed children.

Milk Intolerance: Lactose Intolerance

The majority of children can tolerate lactose during a diarrheal episode. A small proportion of children with diarrhea may not be able to digest lactose and are therefore not tolerant of milk- and lactose-containing formulas. This is more likely to occur among young children who only receive animal milk or formula in their diet and who have persistent diarrhea, and it rarely occurs in children on a diet of breast milk. In a lactose-intolerant child, milk- and lactose-containing formulas result in a significant increase in stool output. Stool output reduces dramatically when the milk- or lactose-containing formula is stopped. The warning signs of lactose intolerance include deterioration of the child's clinical condition, signs of dehydration, and an increase in the stool volume when milk feedings are given. However, only when the child is not gaining weight, eating less, and not fully alert is this a real cause for concern. This condition can be managed by continuing breast feeding. If the child is not yet in the weaning period but takes animal milk, yogurt or diluted milk (equal water and milk) or soy milk can be used as a substitute given in small feedings. The child should be taken to a health care provider if the condition does not improve in 2 days.

If the child eats soft or solid foods, the lactose in the diet should be substituted in the same manner as for the infant (with diluted milk or soy milk) but mixed with cooked cereals and vegetables. If this does not improve the condition of the child, all animal milk should be excluded from the diet, and protein- and energy-rich foods such as finely ground chicken should be given. The treatment should be continued for a few days after the cessation of the diarrhea, when the milk is slowly replaced in the diet.

Soft or Solid Food, Energy Density of Diet, and Protein and Energy Requirements

A child's diet during the period of diarrhea should not be drastically different from his or her normal healthy diet. Therefore, for children who are currently breast-feeding, they should continue to do so, and for children who are in the weaning period and have a mixed diet, they should continue to have a mixed diet of soft or solid food. If the child on a mixed diet is dehydrated, his or her soft and solid foods should be temporarily stopped for a period of approximately 4 h when he or she rehydrates. However, it should then be resumed. For children who are in the weaning stages, WHO recommends small, frequent feedings (six or more times a day) to increase nutrient absorption. The type of food should be energy rich, low in bulk, locally available, and nutritious. The diet should contain complementary protein sources and easily digestible fats, and complex carbohydrates should be avoided. All the foods should be well cooked. Easily digestible staple foods that can be easily mashed include rice, corn, potatoes, and noodles. These staple foods should be mixed with vegetables as well as sources of protein if possible. It is also important to ensure adequate rehydration. In addition, the consumption of fresh fruit juices and mashed bananas is highly encouraged because they provide a good source of potassium.

Feeding during the Convalescent Period

The convalescent period is the recovery period for the body during which the child's diarrhea has stopped but the body has not yet fully recovered to its initial condition. During the first few weeks, the child's appetite will be returning and the child may consume up to twice as much as usual. This is a necessary part of the process because even if the child was fully fed during diarrhea, he or she most likely did not absorb sufficient nutrients. During this time, the child's nutritional state should return to at least the level before the child became ill. The desired energy intake ranges from 100 to 160 J/kg per day, which is achieved with a high-energy, low-bulk, and low-viscosity diet. This is needed for a catch-up growth period and rapid nutritional recovery.

Supplementation with Micronutrients

ORT reduces mortality from dehydrating diarrhea, but it does not decrease the duration of episodes or their consequences, such as malnutrition. In addition, adherence to recommendations regarding ORT in children is poor because caregivers want to reduce the duration of illness. This often leads to use of antibiotics and other treatment of no proven value. In addition, there are indications that knowledge and use of appropriate home therapies, including ORT, to manage diarrhea successfully may be declining in some countries. The limitations of ORT and continued high diarrhoeal morbidity, mortality, and associated malnutrition led to a search for adjunct therapies. Zinc and vitamin A are essential to repair the intestinal mucosa and boost

immunological responses. These supplements should be given during periods of diarrhea.

Zinc

Based on the results of a large number of randomized controlled trials that have demonstrated the therapeutic benefits of zinc supplementation during diarrhea, WHO and UNICEF have recommended the use of zinc supplementation at a dose of approximately 2 RDAs per day (10–20 mg) for 10–14 days. This strategy has the advantage of having a good delivery mechanism (i.e., during the delivery of ORS packets). The effectiveness of different delivery strategies is being evaluated in large-scale trials.

A study from India evaluated the efficacy of zinc-fortified ORS and reported that zinc–ORS was moderately efficacious in reducing the severity of acute diarrhea. If these results are confirmed by other studies, this strategy has the advantage of reducing the need to deliver the zinc supplementation separate from the ORS packets. However, it will probably result in increased cost of ORS production and not all countries will be able to rapidly scale up the production of zinc–ORS in the near future. Additional potential disadvantages of this strategy are that coverage with ORS is low, ORS is usually needed only for 2 or 3 days, and the volume of ORS taken is generally low. Therefore, it will be difficult to ensure adequate zinc intake.

Vitamin A

The body's ability to absorb vitamin A during diarrhea is reduced, which may lead to acute vitamin A deficiency. Repeated episodes may lead to blindness and signs of xerophthalmia. If these are recognized, 200 000 units of vitamin A should be orally administered to children (100 000 units for infants). There should be two subsequent doses, one the following day and another in 4 weeks. In areas where vitamin A deficiency is a problem, any foods that are rich in carotene should be administered (including dark leafy vegetables and yellow or orange fruits and vegetables). Vitamin A deficiency is most common among severely undernourished children and among children who have recently recovered from the measles.

Medications

Antimicrobials and antiparasitics should not be regularly used. Most episodes do not benefit from these treatments, with the following exceptions: in suspect cases of cholera, in cases of persistent diarrhea when cysts or trophozoites of Giardia are identified in the feces, and antibiotics that are effective for Shigella are only used for cases of dysentery (blood in the stool).

Management of Diarrhea in Children with Severe Malnutrition

Children who are severely undernourished and have diarrhea often have other infections. Infections can cause hypothermia as opposed to fever. An appropriate antibiotic should be give if an infection is identified.

Assessment of Dehydration in a Severely Malnourished Child

Children who are severely undernourished have different, unreliable signs and symptoms to assess the status of their hydration. For example, children with marasmus have poor skin elasticity even though they are not dehydrated, and sunken eyes are also not a reliable sign. Irritability in these children may be a sign of systemic infection rather than dehydration. The skin of children with kwashiorkor, on the other hand, may appear to be normal even if they are dehydrated. These children generally have apathetic attitudes. Undernourished children do not readily cry, so determining the absence of tears is also a challenge. Signs that prove to be more indicative to the status of a child's level of hydration include cool and clammy extremities, eagerness to drink, dry mouth and tongue, and a weak radial pulse.

Rehydration

Severely undernourished children should be hospitalized. ORS should be started as soon as possible. A standard ORS treatment with additional potassium should be given orally (preferred method) or using a nasogastric tube. An intravenous solution should be avoided since fluid overload may potentially cause heart failure and increase the risk of septicemia. By dissolving 7.5 g of potassium chloride in 100 ml of water, it is possible to prepare 1 mmol of potassium per milliliter of solution. Every 24 h, 4 ml/kg body weight should be given, mixed with food, for 14 days.

Feeding

Children with marasmus must limit their food intake to approximately 110 kcal/kg/day for the first week. On the other hand, children with kwashiorkor must begin a slow feeding treatment starting as low as 50 kcal/kg/day and work their way up to approximately 110 kcal/kg/day after approximately 1 week. This can be a difficult task since these children are often apathetic and have

severe anorexia. The initial diets of all severely undernourished children must be given in small, frequent (every 2 h day and night), semiliquid doses. An example of an initial diet packet is 8 g skim milk powder, 6 g vegetable oil, 5 g sugar, and 100 ml water to make a high-energy (100 kcal/100 ml) meal with additional minerals, including 60 mg iron, 100 mg folic acid, and 200 000 units of vitamin A and vitamins B (complex), C, and D.

Conclusion

Diarrhea, a disease of fluid and electrolyte imbalance, is an important worldwide cause of morbidity and mortality among infants and children, especially in developing countries. However, it is also very much a nutritional disease. This is primarily because during periods of diarrhea, nutrient intake and absorption are dramatically decreased, which results in undernutrition even when sufficient food is available. The losses of nutrients affect growth rates, and where diarrhea occurs frequently the child may not grow properly. This is a cyclical pattern in that undernourishment in children makes them more prone to diarrhea. Their immune systems are less robust and the episodes affect them more than well-nourished children. Undernourishment and diarrhea can be a fatal combination that can result in a vicious cycle. This cycle requires intervention, sometimes at a treatment center if the case is severe enough. Therefore, is a matter of not only replacing the fluids and electrolytes but also of managing good feeding practices at all times before, during, and after the illness.

Sixty percent of the 10 million deaths among children younger than 5 years old are associated with malnutrition. Approximately 2 million of the deaths are due to diarrhea. Repeated episodes of diarrhea result in malnutrition, which in turn puts the child at an increased risk of recurrent infections, including diarrhea. To break this cycle, diarrheal episodes should be managed with appropriate fluid and nutritional therapy.

See also: **Colon**: Disorders; Nutritional Management of Disorders. **Lactose Intolerance**. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **United Nations Children's Fund**. **Vitamin A**: Biochemistry and Physiological Role. **World Health Organization**. **Zinc**: Deficiency in Developing Countries, Intervention Studies.

Further Reading

Ahmed T, Ali M, Ullah MM *et al.* (1999) Mortality in severely malnourished children with diarrhoea and use of a standardised management protocol. *Lancet* **353**(9168): 1919–1922.

Alam DS, Marks GC, Baqui AH *et al.* (2000) Association between clinical type of diarrhea and growth of children younger than 5 years old in rural Bangladesh. *International Journal of Epidemiology* **29**(5): 916–921.

Baqui AH, Black RE, Arifeen SE *et al.* (2002) Community randomized trial of zinc supplementation started during diarrhoea reduces morbidity and mortality in Bangladeshi children. *British Medical Journal* **325**(7372): 1059.

Baqui AH, Black RE, Sack RB *et al.* (1993) Malnutrition, cell-mediated immune deficiency and diarrhoea: A community-based longitudinal study in rural Bangladeshi children. *American Journal of Epidemiology* **137**(3): 355–365.

Baqui AH, Sack RB, Black RE *et al.* (1992) Enteropathogens associated with acute and persistent diarrhoea in rural Bangladeshi children. *Journal of Infectious Disease* **166**: 792–796.

Baqui AH, Sack RB, Black RE *et al.* (1993) Cell-mediated immune deficiency and malnutrition are independent risk factors for persistent diarrhea in Bangladeshi children. *American Journal of Clinical Nutrition* **58**: 453–458.

Baqui AH, Zaman K, Persson LA *et al.* (2003) Simultaneous weekly supplementation of iron and zinc is associated with lower morbidity due to diarrhea and acute lower respiratory infection in Bangladeshi infants. *Journal of Nutrition* **133**(12): 4150–4157.

Black RE, Morris SS, and Bryce J (2003) Where and why are 10 million children dying every year? *Lancet* **361**(9376): 2226–2234.

Brown KH (2003) Diarrhea and malnutrition. *Journal of Nutrition* **133**(1): 328S–332S.

Kosek M, Bern C, and Guerrant RL (2003) The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. *Bulletin of the World Health Organization* **81**(3): 197–204.

Lutter CK, Habicht J-P, Rivera JA *et al.* (2004) *Clinical Management of Acute Diarrhoea*, WHO/UNICEF Joint Statement (WHO/FCH/CAH/04.7). New York: United Nations Children's Fund/World Health Organization.

Rahman MM, Vermund SH, Wahed MA *et al.* (2001) Effect of simultaneous zinc and vitamin A supplementation on diarrhoea and acute lower respiratory infection in Bangladeshi children: A randomized double-blind placebo controlled trial. *British Medical Journal* **323**: 314–318.

Schrimshaw NS (2003) *Nutrition and Diarrhoea: Fifty Years Experience. Keynote Lecture 1, Nutrition and Diarrhoea.* Presented at the 10th Asian Conference on Diarrhoeal Diseases and Nutrition (ASCODD), December 7–9, Abstract Book, ICDDR, B, Dhaka.

Thapar N and Sanderson IR (2004) Diarrhoea in children: An interface between developing and developed countries. *Lancet* **363**(9409): 641–653.

World Health Organization (2000) *IMCI Model Handbook*, WHO/FCH/CAH/00.12. Geneva, WHO.

DIETARY FIBER

Contents

Physiological Effects and Effects on Absorption

I T Johnson, Institute of Food Research, Norwich, UK

Introduction

It has long been recognized that both animal feedstuffs and human foods contain poorly digestible components, which do not contribute to nutrition in the classical sense of providing essential substances or metabolic energy. With the development of scientific approaches to animal husbandry in the nineteenth century, the term 'crude fiber' was coined to describe the material that remained after rigorous nonenzymic hydrolysis of feeds. During the twentieth century, various strands of thought concerning the virtues of 'whole' foods, derived from plant components that had undergone only minimal processing, began to converge, leading eventually to the dietary fiber hypothesis. Put simply, this states that the nondigestible components of plant cell walls are essential for the maintenance of human health.

In the early 1970s the physician and epidemiologist Hugh Trowell recognized that the crude fiber figures available at the time for foods had little physiological significance and were of no practical value in the context of human diets. He was amongst the first to use the term dietary fiber to describe the 'remnants of plant cell walls resistant to hydrolysis (digestion) by the alimentary enzymes of man.' This definition was later refined and given the more quantitative form: "The sum of lignin and the plant polysaccharides that are not digested by the endogenous secretions of the mammalian digestive tract." This definition paved the way for the development of analytical methods that could be used to define the fiber content of human foods. Broadly, these techniques are based on enzymic removal of the digestible elements in food, followed by either gravimetric analysis of the residue ('Southgate' and Association of Analytical Chemists (AOAC) methods), which results in the retention of some undigested starch, or chemical analysis ('Englyst' method), which enables a more precise separation of starch from the structural polysaccharides of the cell wall. In the latter case, the cell wall components are defined as 'nonstarch polysaccharides' (NSP). Whatever analytical approach is used, both 'dietary fiber' and nonstarch polysaccharides are shorthand terms for large and complex mixtures of polysaccharides. The components of such mixtures vary widely among foods and they often share few properties other than resistance to digestion in the small intestine. A summary of the main types of plant cell polysaccharides contained in the general definition of dietary fiber is given in **Table 1**.

In recent years this problem has been made more complex in some ways because of the explosion of interest in functional foods for gastrointestinal health. These often contain high levels of novel oligo- and polysaccharides, which might perhaps be regarded as analogs of dietary fiber. Fructose oligosaccharides, which are nondigestible but highly fermentable, are now often added to foods as prebiotic substrates for the colonic microflora. Such materials may not fit the original definition of dietary fiber, but it is certainly not helpful to exclude them from the contemporary concept, which needs to expand to accommodate modern developments.

The presence of large undigested cell wall fragments, finely dispersed particulates, or soluble polysaccharides can alter physiological processes

Table 1 Major components of dietary fiber

Food source	*Polysaccharides and related substances*
Fruits and vegetables	Cellulose, xyloglucans, arabinogalactans, pectic substances, glycoproteins
Cereals	Cellulose, arabinoxylans, glucoarabinoxylans, β-D-glucans, lignin, and phenolic esters
Legume seeds	Cellulose, xyloglucans, galactomannans, pectic substances
Manufactured products	Gums (guar gum, gum arabic), alginates, carrageenan, modified cellulose gums (methyl cellulose, carboxymethyl cellulose)

throughout the gut. The effects of different fiber components depend upon their varied physical and chemical properties during digestion, and also upon their susceptibility to degradation by bacterial enzymes in the colon. The complex nature of the various substances covered by the general definition of dietary fiber means that a single analytical value for the fiber content of a food is a poor guide to its physiological effects. This article will review the main mechanisms of action of resistant polysaccharides in the alimentary tract and their implications for human health.

Sources and Types of Dietary Fiber

The main sources of dietary fiber in most Western diets are well characterized, and high-quality data are available for both food composition and dietary intakes. This is not always true for diets in developing countries, however, and this problem bedevils attempts to investigate the importance of fiber by making international comparisons of diet and disease. Another problem is that different analytical approaches give slightly different values for the dietary fiber content of foods, and do not reflect the physical and chemical properties of the different polysaccharide components. The use of enzymic hydrolysis to determine the 'unavailable carbohydrate' content of foods was refined by Southgate, and his technique was used for the 4th edition of the UK standard food tables, *The Composition of Foods* published in 1978. The 6th edition, published in 2002, contains values for nonstarch polysaccharides, derived using the Englyst technique, but recommends use of AOAC methods for food labeling purposes. A comparison of values for nonstarch polysaccharides and dietary fiber values obtained by the AOAC method is given in **Table 2**.

Table 2 A Comparison of values for nonstarch polysaccharides and dietary fiber

Food source	*Nonstarch Polysaccharides (Englyst method)*	*Total dietary fiber (AOAC method)*
White bread	2.1	2.9
Brown bread	3.5	5.0
Wholemeal bread	5.0	7.0
Green vegetables	2.7	3.3
Potatoes	1.9	2.4
Fresh fruit	1.4	1.9
Nuts	6.6	8.8

Data modified from McCance and Widdowson's (2002) *The Composition of Foods*, 6th Edition, Cambridge: Royal Society of Chemistry.

In the UK about 47% of dietary fiber is obtained from cereal products, including bread and breakfast cereals. The level of cell wall polysaccharides in a product made from flour depends on the extraction rate, which is the proportion of the original grain present in the flour after milling. Thus a 'white' flour with an extraction rate of 70% usually contains about 3% NSP, whereas a 'wholemeal' flour with an extraction rate of 100% contains about 10% NSP. The terms 'soluble' and 'insoluble' fiber have been coined in order to partially overcome the problem of the lack of correspondence between the total analytical value for fiber and the physical properties of the measured polysaccharides. By adopting the Englyst technique for the separation and chemical analysis of nonstarch polysaccharides it is possible to specify both the soluble and insoluble fiber content of foods. Some representative values for soluble and insoluble fiber in cereal foods are given in **Table 3**, and those for fruits and vegetables, which provide a further 45% of the fiber in UK diets, are given in **Table 4**.

Fiber in the Digestive Tract

The primary function of the alimentary tract is to break down the complex organic macromolecules of which other organisms are composed into smaller molecules, which can then be selectively absorbed into the circulation by specialized mucosal epithelial

Table 3 Soluble and insoluble nonstarch polysaccharides in some cereal products and nuts

Food source	*Nonstarch polysaccharides (g per 100 g fresh weight)*		
	Total NSP	*Soluble NSP*	*Insoluble NSP*
Sliced white bread	1.5	0.9	0.6
Sliced brown bread	3.6	1.1	2.5
Wholemeal bread	4.8	1.6	3.2
Spaghetti	1.2	0.6	0.6
Rye biscuits	11.7	3.9	7.8
Cornflakes	0.9	0.4	0.5
Crunchy oat cereal	6.0	3.3	2.7
Walnuts	3.5	1.5	2.0
Hazelnuts	6.5	2.5	4.0
Peanuts	6.2	1.9	4.3
Brazil nuts	4.3	1.3	3.0

Data modified from Englyst HN, Bingham SA, Runswick SS, Collinson E, and Cummings JH (1989) Dietary fibre (non-starch polysaccharides) in cereal products. *Journal of Human Nutrition and Dietetics* **2**: 253–271 and Englyst HN, Bingham SA, Runswick SS, Collinson E, Cummings JH (1989) Dietary fibre (non-starch polysaccharides) in fruit vegetables and nuts. *Journal of Human Nutrition and Dietetics* **1**: 247–286.

Table 4 Soluble and insoluble nonstarch polysaccharides in some vegetables and fruits

Food source	*Nonstarch polysaccharides (g per 100 g fresh weight)*		
	Total NSP	*Soluble NSP*	*Insoluble NSP*
Apples (Cox)	1.7	0.7	1.0
Oranges	2.1	1.4	0.7
Plums	1.8	1.2	0.6
Bananas	1.1	0.7	0.4
Potatoes	1.1	0.6	0.5
Sprouts	4.8	2.5	2.3
Peas (frozen)	5.2	1.6	3.6
Carrots	2.5	1.4	1.1
Courgettes	1.2	0.6	0.6
Runner beans	2.3	0.9	1.4
Baked beans	3.5	2.1	1.4
Tomato	1.1	0.4	0.7
Lettuce	1.2	0.6	0.6
Onion	1.7	0.9	0.8
Celery	1.3	0.6	0.7

Data modified from Englyst HN, Bingham SA, Runswick SS, Collinson E, and Cummings JH (1989) Dietary fibre (non-starch polysaccharides) in fruit vegetables and nuts. *Journal of Human Nutrition and Dietetics* **1**: 247–286.

cells. Food is conveyed progressively through the alimentary tract, stored at intervals, and broken down mechanically as required, by a tightly controlled system of rhythmic muscular contractions. The digestive enzymes are released into the lumen at the appropriate stages to facilitate the decomposition of carbohydrates, proteins, and complex lipids. By definition, the polysaccharides that comprise dietary fiber are not digested by endogenous enzymes, though they are often fermented to a greater or lesser degree by bacterial enzymes in the large intestine.

The Mouth and Pharynx

The earliest stages of digestion begin in the mouth, where food particles are reduced in size, lubricated with saliva, and prepared for swallowing. The saliva also contains the digestive enzyme salivary amylase, which begins the hydrolysis of starch molecules. Cell wall polysaccharides are an important determinant of food texture, and they exert an indirect effect on the degree of mechanical breakdown of plant foods prior to swallowing. Hard foods tend to be chewed more thoroughly than soft ones, and hence the presence of dietary fiber in unrefined foods may begin to regulate digestion at a very early stage.

The Stomach

The first delay in the transit of food through the digestive tract occurs in the stomach, where large food fragments are further degraded by rigorous muscular activity in the presence of hydrochloric acid and proteolytic enzymes. The need to disrupt and disperse intractable food particles and cell walls appears to delay the digestive process significantly. For example, the absorption of sugar from whole apples is significantly slower than from apple juice. Similarly, the rate at which the starch is digested and absorbed from cubes of cooked potato has been shown to be much slower when they are swallowed whole than when they are chewed normally. Thus, simple mechanical factors can limit the rate at which glucose from carbohydrate foods enters the circulation.

The Small Intestine

The small intestine is the main site of nutrient absorption, and it is in fact the largest of the digestive organs in terms of surface area. The semi-liquid products of gastric digestion are released periodically into the duodenum, and then propelled downstream by peristaltic movements, at about 1 cm per minute. The hydrolysis of proteins, triglycerides, and starch continues within the duodenum and upper jejunum, under the influence of pancreatic enzymes. The final stages of hydrolysis of dietary macromolecules occur under the influence of extracellular enzymes at the mucosal surface. The released products are absorbed into the circulation, along with water and electrolytes, via the specialized epithelial cells of the intestinal villi. Muscular activity in the small intestinal wall, together with rhythmic contractions of the villi, ensures that the partially digested chyme is well stirred. In adults, the first fermentable residues from a meal containing complex carbohydrates enter the colon approximately 4.5 h after ingestion. When a solution containing indigestible sugar is swallowed without food it reaches the colon about 1.5 h earlier than when the same material is added to a solid meal containing dietary fiber. The presence of solid food residues slows transit, probably by delaying gastric emptying and perhaps also by increasing the viscosity of the chyme so that it tends to resist the peristaltic flow. Soluble polysaccharides such as guar gum, pectin, and β-glucan from oats increase mouth to cecum transit time still further.

In creating the dietary fiber hypothesis, Trowell's principal interest was its role in the prevention of metabolic disorders. In particular, he believed that dietary fiber was a major factor in the prevention of diabetes mellitus, which, he argued, was probably unknown in Western Europe prior to the introduction of mechanized flour milling. In earlier times the near-universal consumption of unrefined

carbohydrate foods would have ensured that intact indigestible cell wall polysaccharides were present throughout the upper alimentary tract during digestion. This, according to Trowell and others, favored slow absorption of glucose, which in turn placed less strain upon the ability of the pancreas to maintain glucose homeostasis. There is no doubt that type 2 diabetes has become more common in Western countries as prosperity, and an excess of energy consumption over expenditure, has grown. It is not established that rapid absorption of glucose due to consumption of refined starches is a primary cause of diabetes, but the control of glucose assimilation is certainly a key factor in its management. Cell wall polysaccharides influence the digestion and absorption of carbohydrates in a variety of ways, and are a major determinant of the 'glycemic index,' which is essentially a quantitative expression of the quantity of glucose appearing in the bloodstream after ingestion of a carbohydrate-rich food. To calculate the index, fasted subjects are given a test meal of the experimental food containing a standardized quantity of carbohydrate. The change in concentration of glucose in the blood is then measured over a period of time. The ratio of the area under the blood-glucose curve in response to the test meal to that produced by an equal quantity of a standard reference food is then calculated and expressed as a percentage. When glucose is used as the standard, most complex starchy foods have glycemic indices lower than 100%.

The physical resistance of plant cell walls during their passage through the gut varies considerably from one food to another. Cell walls that remain intact in the small intestine will impede the access of pancreatic amylase to starch. This is particularly true of the cells of legume seeds, which have been shown to retain much of their integrity during digestion. Legume-based foods such as lentils and chilli beans have glycemic indices that are amongst the lowest of all complex carbohydrate foods. Even when enzymes and their substrates do come into contact, the presence of cell wall polysaccharides may slow the diffusion of hydrolytic products through the partially digested matrix in the gut lumen. These effects of dietary fiber on carbohydrate metabolism emphasize once more that physiological effects cannot be predicted from simple analytical values for total fiber, because they are consequences of cellular structure, rather than the absolute quantity of cell wall polysaccharides within the food.

Many studies on postprandial glycemia have been conducted using isolated fiber supplements such, as pectin or guar gum added to glucose testmeals or to low-fiber sources of starch. They demonstrate that, contrary to Trowell's original hypothesis, wheat bran and other insoluble cell wall materials have little effect on glucose metabolism. However, certain soluble polysaccharides, such as guar gum, pectin, and oat β-glucan, which form viscous solutions in the stomach and small intestine, do slow the absorption of glucose. Highly viscous food components may delay gastric emptying and inhibit the dispersion of the digesta along the small intestine, but the primary mechanism of action appears to be suppression of convective stirring in the fluid layer adjacent to the mucosal surface. The rapid uptake of monosaccharides by the epithelial cells tends to reduce the concentration of glucose in this boundary layer, so that absorption from the gut lumen becomes rate-limited by the relatively slow process of diffusion. The overall effect is to delay the assimilation of glucose and hence suppress the glycemic response to glucose or starchy foods in both healthy volunteers and in people with diabetes. A similar mechanism probably inhibits the reabsorption of cholesterol and bile salts in the distal ileum and this may account for the ability of some viscous types of soluble dietary fiber such as guar gum and β-glucan to reduce plasma cholesterol levels in humans.

One of the main reasons for developing analytical methods to distinguish between soluble and insoluble components of dietary fiber was to provide a means of assessing the capacity of fiber-rich foods to influence carbohydrate and lipid metabolism. There is evidence that diets that provide 30–50% of their fiber in the form of soluble polysaccharides are associated with lower cholesterol levels and better glycemic control than diets that contain mostly insoluble fiber. Several officially recognized sets of guidelines for patients with impaired glucose metabolism and its complications (syndrome X) now specifically recommend a high intake of carbohydrate foods that are rich in soluble fiber.

Some of the interactions between cell wall polysaccharides and other food components in the small intestine are much more specific. There has been considerable interest over a number of years in the possibility that the polysaccharides and complex phenolic components of cell walls contain polar groups that could interact with and bind ionized species in the gastrointestinal contents, thereby reducing their availability for absorption. Intraluminal binding of heavy metals, toxins, and carcinogens might be a valuable protective mechanism, but binding of micronutrients could seriously compromise nutritional status.

Interactions of this type can be shown to occur *in vitro*, and studies with animals and human

ileostomists suggest that charged polysaccharides such as pectin can displace cations into the colon under experimental conditions. However, there is little objective evidence that dietary fiber *per se* has much of an adverse effect on mineral metabolism in humans. Indeed, highly fermentable polysaccharides and fructose oligosaccharides have recently been shown to promote the absorption of calcium and magnesium in both animal and human studies. The mechanism for the effect is not entirely clear, but it is probably a consequence of fermentation acidifying the luminal contents of the colon and enhancing carrier-mediated transport of minerals across the colonic mucosa.

In unprocessed legume seeds, oats, and other cereals phytate (myo-inositol hexaphosphate) is often present in close association with cell wall polysaccharides. Unlike the polysaccharides themselves, phytate does exert a potent binding effect on minerals, and has been shown to significantly reduce the availability of magnesium, zinc, and calcium for absorption in humans. Phytate levels in foods can be reduced by the activity of endogenous phytase, by hydrolysis with exogenous enzymes, or by fermentation. Dephytinized products may therefore be of benefit to individuals at risk of suboptimal mineral status. However, there are indications from animal and *in vitro* studies that phytate is an anticarcinogen that may contribute to the protective effects of complex fiber-rich foods. The overall significance of phytate in the diet therefore requires further assessment in human trials.

The Large Intestine

Microorganisms occur throughout the alimentary tract but in healthy individuals their numbers and diversity are maintained within strict limits by the combined effects of intraluminal conditions, rapid transit, and host immunity. The colon and rectum, however, are adapted to facilitate bacterial colonization, and the typical adult human colonic microflora has been estimated to contain about 400 different bacterial species. The largest single groups present are Gram-negative anaerobes of the genus *Bacteroides*, and Gram-positive organisms including bifidobacteria, eubacteria, lactobacilli, and clostridia. However, a large proportion of the species present cannot be cultured *in vitro* and are very poorly characterized.

The proximal colon, which receives undigested food residues, intestinal secretions, and the remnants of exfoliated enterocytes from the distal ileum, contains around 200 g of bacteria and substrates in a semiliquid state. These conditions are ideal for bacterial fermentation. Most of the bacteria of the human colon utilize carbohydrate as a source of energy, although not all can degrade polysaccharides directly. Many that are ultimately dependent upon dietary carbohydrate residues for energy are adapted to utilize the initial degradation products of the polysaccharide utilizers, rather than the polymers themselves. It has been estimated that somewhere between 20 and 80 g of carbohydrate enter the human colon every day, about half of which is undigested starch. Around 30 g of bacteria are produced for every 100 g of carbohydrate fermented.

Apart from dietary fiber, there are three major sources of unabsorbed carbohydrate for the colonic microflora. Perhaps the most important is resistant starch, which consists of retrograded starch polymers and starch granules enclosed within intact plant cell walls. Nondigestible sugars, sugar alcohols, and oligosaccharides such as fructooligosaccharides and galactooligosaccharides occur only sparingly in most plant foods, but they are now of great commercial interest because they can be used as prebiotics to selectively manipulate the numbers of bifidobacteria in the human colon. Endogenous substrates including mucus are also important for the colonic microflora. Mucus is an aqueous dispersion of a complex group of glycoproteins containing oligosaccharide side-chains, which are a major source of fermentable substrates. Even when the colon is surgically isolated and has no access to exogenous substrates it still supports a complex microflora.

The beneficial effects of dietary fiber on the alimentary tract were emphasized by another of the founders of the dietary fiber hypothesis, Denis Burkitt, who based his arguments largely on the concept of fecal bulk, developed as a result of field observations in rural Africa, where cancer and other chronic bowel diseases were rare. His hypothesis was that populations consuming the traditional rural diets, rich in vegetables and cereal foods, produced bulkier, more frequent stools than persons eating the refined diets typical of industrialized societies. Chronic constipation was thought to cause straining of abdominal muscles during passage of stool, leading to prolonged high pressures within the colonic lumen and the lower abdomen. This in turn was thought to increase the risk of various diseases of muscular degeneration including varicose veins, hemorrhoids, hiatus hernia, and colonic diverticulas. Colorectal neoplasia was also thought to result from infrequent defecation, because it caused prolonged exposure of the colonic epithelial cells to mutagenic chemicals, which could initiate cancer. Burkitt's overall hypothesis for the beneficial effects of fecal

bulk has never really been refuted, and epidemiological evidence continues to support a protective role of fiber against colorectal cancer, particularly within Europe. However, the origins of intestinal neoplasia are now known to be far more complex than Burkitt was able to envisage, and there is little evidence to suggest a direct causal link between chronic constipation and colorectal cancer. Indeed, in one recent prevention trial, the risk of recurrence of colorectal polyps was slightly increased by prolonged supplementation with a bulk laxative based on one specialized source of cell wall polysaccharides.

Whatever the relationship to disease, it is certainly true that the consumption of dietary fiber is one major determinant of both fecal bulk and the frequency of defecation (bowel habit). However, the magnitude of the effect depends upon the type of fiber consumed. Soluble cell wall polysaccharides such as pectin are readily fermented by the microflora, whereas lignified tissues such as wheat bran tend to remain at least partially intact in the feces. Both classes of dietary fiber can contribute to fecal bulk but by different mechanisms. The increment in stool mass caused by wheat bran depends to some extent on particle size, but in healthy Western populations it has been shown that for every 1 g of wheat bran consumed per day, the output of stool is increased by between 3 and 5 g. Other sources of dietary fiber also favor water retention. For example, isphagula, a mucilaginous material derived from *Psyllium*, is used pharmaceutically as a bulk laxative. Soluble polysaccharides such as guar and oat β-glucan are readily fermented by anaerobic bacteria, but solubility is no guarantee of fermentability, as is illustrated by modified cellulose gums such as methylcellulose, which is highly resistant to degradation in the human gut. Fermentation reduces the mass and water-holding capacity of soluble polysaccharides considerably, but the bacterial cells derived from them do make some contribution to total fecal output. Thus, although all forms of dietary fiber are mild laxatives, the single analytical measurement of total fiber content again provides no simple predictive measure of physiological effect.

Although fermentation of fiber tends to reduce its effectiveness as a source of fecal bulk, it has other very important benefits. The absorption and metabolism of short-chain fatty acids derived from carbohydrate fermentation provides the route for the recovery of energy from undigested polysaccharides. Butyrate functions as the preferred source of energy for the colonic mucosal cells, whilst propionate and acetate are absorbed and metabolized systemically. There continues to be much debate about the importance of butyrate for the colon. *In vitro*, butyrate causes differentiation of tumor cells, suppresses cell division, and induces programed cell death (apoptosis). These effects are thought likely to suppress the development of cancer, but it is not yet entirely clear whether they also occur in the intact intestine. Research continues on the importance of butyrate and other short-chain fatty acids for human health.

The other major breakdown products of carbohydrate fermentation are hydrogen, methane, and carbon dioxide, which together comprise flatus gas. Excess gas production can cause distension and pain in some individuals, especially if they attempt to increase their fiber consumption too abruptly. In most cases, however, extreme flatus is probably caused more by fermentation of oligosaccharides such as stachyose and verbascose, which are found principally in legume seeds, rather than the cell wall polysaccharides themselves.

Conclusion

Several decades of research have confirmed that cell wall polysaccharides modify physiological mechanisms throughout the alimentary tract. Delayed absorption of glucose and lipids in the small intestine makes an important contribution to metabolic control in type 2 diabetes, and certain types of hypercholesterolemia, respectively. Any loss of carbohydrates to the colon will lead to increased fermentative activity, and through this pathway, most of the unabsorbed energy will be recovered as short-chain fatty acids. Unfermented cell wall polysaccharides and increased bacterial mass contribute to fecal bulk. All these established physiological effects, coupled with the possibility of using oligosaccharides as prebiotics to modify the colonic microflora, have greatly stimulated interest in nondigestible carbohydrates amongst food manufacturers and consumers in the past few years. There is little to suggest that conventional sources of fiber compromise micronutrient metabolism in otherwise healthy individuals, but the possibility of this and other adverse effects needs to be considered, as the use of novel polysaccharides as sources or analogs of dietary fiber, both for conventional products and for functional foods, continues to expand.

See also: **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Carbohydrates**: Resistant Starch and Oligosaccharides. **Cereal Grains**. **Colon**: Disorders. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Dietary Fiber**: Potential Role in Etiology of Disease; Role in Nutritional Management of

Disease. **Functional Foods**: Health Effects and Clinical Applications; Regulatory Aspects.

Further Reading

Bell S, Goldman VM, Bistrian BR, Arnold AH, Ostroff G, and Forse RA (1999) Effect of beta-glucan from oats and yeast on serum lipids. *Critical Reviews in Food Science and Nutrition* **39**: 189–202.

Burkitt DP and Trowell HC (eds.) (1975) *Refined Carbohydrate Foods: Some Implications of Dietary Fibre*. London: Academic Press.

Englyst HN, Bingham SA, Runswick SS, Collinson E, and Cummings JH (1989) Dietary fibre (non-starch polysaccharides) in cereal products. *Journal of Human Nutrition and Dietetics* **2**: 253–271.

Englyst HN, Bingham SA, Runswick SS, Collinson E, and Cummings JH (1989) Dietary fibre (non-starch polysaccharides) in fruit vegetables and nuts. *Journal of Human Nutrition and Dietetics* **1**: 247–286.

Englyst HN, Bingham SA, Runswick SS, Collinson E, and Cummings JH (1989) Dietary fibre (non-starch polysaccharides) in fruit vegetables and nuts. *Journal of Human Nutrition and Dietetics* **1**: 247–286.

Food Standards Agency (2002) *McCance and Widdowson's The Composition of Foods*, 6th summary edn Cambridge: The Royal Society of Chemistry.

Johnson IT and Southgate DAT (1994) *Dietary Fibre and Related Substances*. London: Chapman Hall.

Kushi LH, Meyer KA, and Jacobs Jr DR (1999) Cereals legumes and chronic disease risk reduction: Evidence from epidemiologic studies. *American Journal of Clinical Nutrition* **70**: 451S–458S.

McCance and Widdowson's (2002) *The Composition of Foods*, 6th Edition, Cambridge: Royal Society of Chemistry.

Nutrition Society (2003) Symposium on "Dietary Fibre in Health and Disease". *Proceedings of the Nutrition Society* **62**: 1–249.

Scholz-Ahrens KE and Schrezenmeir J (2002) Inulin, oligofructose and mineral metabolism – experimental data and mechanism. *British Journal of Nutrition* **87**(suppl 2): S179–186.

Southgate DAT (1992) *Determination of Food Carbohydrates*, 2nd edn. London: Elsevier Applied Science Publishers.

Potential Role in Etiology of Disease

D L Topping and L Cobiac, CSIRO Health Sciences and Nutrition, Adelaide, SA, Australia

Noninfectious diseases cause much morbidity and mortality in developed countries in Europe, the Americas, Asia, and Australasia. They are expected to increase due to an alarming increase in obesity, with its attendant risk of diabetes, coronary heart disease (CHD), and some cancers. Equally important, they are becoming an issue in developing countries through greater affluence. In every case, they have serious negative socioeconomic impacts, and their prevention through appropriate dietary and lifestyle change is the optimal strategy to minimize personal and community costs. This strategy is believed to have contributed substantially to economic growth in countries where it has been applied.

Obesity is a fairly visible problem that tends to overshadow other considerations. Energy intake in excess of expenditure is the root cause of obesity, and dietary carbohydrates are implicated specifically in the development of overweight. Exclusion of carbohydrates is attracting attention as a weight control strategy but this ignores the fact that digestible carbohydrates (including starch) provide the same amount of energy per gram as protein and less than 45% of the energy of fat and 60% of the energy of alcohol. Furthermore, it overlooks the many early comparative population studies that showed that several low-risk groups ate high-starch diets compared to high-risk populations that consumed processed foods high in refined carbohydrates and fat and low in fiber. With time, dietary fiber (rather than the whole diet) received specific attention, and many studies were conducted on the health benefits. This may have contributed to some of the current lack of clarity of the role of fiber and complex carbohydrates in health promotion. It may be compounded by the inadequacy of population data linking fiber (and also starches) to disease processes for important conditions such as colorectal cancer. Only recently have good population data emerged for a protective role for fiber in this condition. This is in marked contrast to the well-established therapeutic and preventive action of fiber in constipation and diverticular disease. Furthermore, it is critical to determine what is meant by the term 'dietary fiber' because other food components may contribute to the major effects ascribed to fiber. This is important when considering the apparent protection conferred by whole grains against disease risk, especially because a health claim is permitted in the United States for their consumption.

Associations between Dietary Fiber and Disease Processes

The early population studies of Walker, Burkitt, and Trowell in Africa in the 1950s showed that serum cholesterol concentrations were low in South African Bantus (at low risk of CHD) who consumed a diet apparently high in fiber and low in fat. Additionally, colon cancer was virtually

unknown among the latter, in contrast to white South Africans. Dietary fiber was known to resist digestion by human intestinal enzymes, which helped to explain the greater fecal bulk seen with higher fiber intakes. This was thought to lower colonic exposure to carcinogens through a simple dilution effect with fiber consumption. Subsequently, it was suggested that diabetes may be related to a deficiency of fiber in the diet whereas other epidemiological studies have shown associations between more dietary fiber consumption and lower risk of some of the hormone-dependent cancers (prostate and breast). Many of these observational population studies are limited by their reliance on reported food intakes which may be compromised in turn by food compositional data because the latter can be limited by the analytical methodology used. Multinational comparisons may be affected by the fact that food sources and processing vary between countries. There are other potential confounders. For example, diets high in fiber-rich foods may contain other protective agents (e.g., phyotoestrogens folate and antioxidants) that could be the actual mediators of protection. Finally, some of the experimental studies and human interventions, especially in colorectal cancer, have given ambiguous or negative outcomes due to limitations of study design.

Dietary Fiber, Complex Carbohydrates, and Health Outcomes: A Need for Fiber Equivalents?

Technology has proved to be a significant issue in human fiber research. Early studies were limited by the relatively simple analytical methods then current. These were designed to measure the fiber components of forage consumed by important ruminant farm animals. Forage foods are high in insoluble polysaccharides and contain lignin (which is not a carbohydrate but a complex polyphenolic ether) and look 'fibrous,' so dietary fiber was equated with roughage and was defined as "those structural and exudative components of plants that were resistant to digestion by human gut enzymes." The methods used initially were quite severe and, with increasing sophistication of analytical methodology (notably chromatography), it became apparent that lignin was only a minor component of fiber compared with nonstarch polysaccharides (NSPs). Technological advances have revealed the importance of fractions such as soluble NSPs. As their name suggests, these dissolve in water but not necessarily under gut conditions. Fiber was then redefined as NSPs plus lignin, which has moved the concept of away from the roughage model. Recently, there has been a further substantial revision of the view of what exactly constitutes dietary fiber with the emerging recognition of the contribution of resistant starch (RS) and oligosaccharides (OSs) to 'fiber' action. It was thought that all of dietary starch consumed in cooked foods was digested in the human small intestine. It is becoming clear that this is not so, and that a substantial fraction of ingested starch, RS, escapes from the small intestine and enters the colon of healthy humans. In terms of the dietary polysaccharides entering that viscus, RS may exceed NSPs in quantity and in the range of its actions. Indeed, some populations that were thought to consume high-fiber diets actually eat less than higher risk groups. An example is the native Africans studied by Burkitt and colleagues who eat unrefined diets based on maize but in fact eat less fiber than the high-risk whites.

Unlike NSPs (which are intrinsically resistant to human digestive enzymes), RS is influenced greatly by physiological and physical factors. Thus, raw starches are highly resistant and gelatinized starches (cooked in the presence of water) are much more digestible due to hydration and the loss of granular structure. However, cooling of cooked starchy foods leads to the generation of RS through retrogradation (which is a realignment of the polysaccharide chains). Other factors, including the relative proportions of amylose and amylopectin and the presence of other food components such as NSPs and lipids, influence RS. High-fiber foods tend to be higher in RS, whereas fats can form complexes with starch that resist digestion. Mastication, transit time, and gender can influence the amount of RS, which means that purely chemical determinations probably give an underestimate of the quantity of starch entering the large bowel. Physical breakdown of foods (especially whole grains) may be particularly important because access to upper intestinal digestive enzymes is limited in large particles, especially in the presence of NSPs. RS has been classified into four types based on the main factors that influence its presence, including physical inaccessibility, cooking, and retrogradation (RS types RS_{1-3}). Of particular interest is the emergence of chemically modified starches (RS_4) as RS because these are used widely in food processing.

In the large bowel NSPs and RS are fermented by the microflora, yielding metabolic end products, principally short-chain fatty acids (SCFAs), which may mediate some of the health benefits ascribed to the carbohydrates. Undigested protein (resistant protein) and other nondigested carbohydrates

(e.g., OSs) also contribute to large bowel fermentation. These nondigested fractions contribute to dietary fiber via fermentation and could be considered in net dietary fiber intake. This problem could be overcome relatively easily by classifying them (and other non-NSP carbohydrates) as fiber equivalents in which their actions are compared against an agreed standard. This is similar to the situation with other nutrients such as vitamin A, where retinol equivalents include carotenoids, which are retinol precursors. It follows that classifications based on chemical composition alone appear to be quite inadequate if one considers as improved health and diminished disease risk as the most important issues.

Dietary Fiber and the Etiology of Coronary Heart Disease

Population Studies

Consumption of unrefined plant foods has been related to lower risk of CHD for some time, but the hypothesis that dietary fiber (i.e., NSP) intake could protect directly against the disease is relatively recent. The suggestion has been supported by a number of epidemiological studies linking higher intakes with lower risk. Vegetarians who consume more plant foods tend to have lower plasma lipids and blood pressure than age- and gender-matched omnivores. However, the strongest evidence derives not from these studies but from a number of very large cohort studies in several countries showing a consistent protective effect of whole grain consumption and CHD risk. Whole grain cereal consumption has been related to substantially lower risk of CHD in both men and women. The evidence for the latter is considered to be sufficiently strong for the US Food and Drug Administration to permit a health claim for consumption of whole grain cereal foods and lowering of the risk of CHD. Similar claims are being considered in Europe. However, these relationships are a long way from proving a specific protective effect of dietary fiber. A study of the relationship of long-term intake of dietary fiber by 68 782 women showed a substantial lowering of relative risk of 0.53 for women in the highest quintile of fiber consumption (22.9 g/day) compared with the lowest (11.5 g/day). These intakes are low compared to those recommended by health authorities. Nevertheless, they do support the view that fiber is protective against CHD. Only the effect of cereal fiber was significant; that of fruits and vegetables was not. However, they leave unanswered the question of the relationship of other contributors to the effects of dietary fiber (e.g., RS) and CHD risk.

Potential Mechanisms Indicating a Role in the Etiology of Coronary Heart Disease

The mechanism for risk reduction and the fiber components responsible need resolution. Elevated plasma total and low-density lipoprotein (LDL) cholesterol concentrations are established risk factors for coronary morbidity and mortality. There are abundant human and animal data showing that diets high in soluble fiber lower plasma cholesterol. One population study has shown a significant negative relationship between viscous (soluble) fiber intake and carotid artery atherogensis as measured by intima–media thickness. This association was significant statistically even though average fiber intakes were not particularly high. When dietary fiber intakes have been related to measures of actual disease outcomes, the evidence is less convincing. A protective effect is often observed on univariate analyses, but once confounding variables are added, dietary fiber intake tends not to be a significant independent predictor of risk for developing CHD. However, in one 12-year follow-up study of men and women, a 6-g increase in daily dietary fiber intake was associated with a 25% reduction in the risk of developing CHD. The most likely direct protective role for dietary fiber in CHD etiology is through plasma lipid lowering. The effect appears to be specific for plasma total and LDL cholesterol, and, possibly, triacylglycerols (TAG). Of the main fiber components, soluble NSPs seem to be effective, but insoluble NSPs and RS (and probably OSs) are not. Indeed, it appears that some insoluble NSP preparations, such as wheat bran, may raise plasma cholesterol slightly. There is good evidence from animal and human studies to support a hypocholesterolemic effect of soluble NSPs either in enriched plant fractions (e.g., oat bran) or as natural (e.g., pectins and guar gum) or synthetic isolates (e.g., hydroxypropylmethylcellulose). The magnitude of the effect varies with dose, but reductions of approximately 5–10% at intakes of 6–12 g of NSPs/day appear to be reasonable. This lowering response approaches that seen with certain drugs, such as cholestyramine, used to manage hypercholesterolemia. Some studies have also shown a reduction in TAGs with soluble fibers such as oat bran. However, it is important to recognize that many of the demonstrations of plasma cholesterol lowering by soluble fiber products are against insoluble NSPs such as wheat bran.

There are several hypotheses to explain the NSP action on plasma cholesterol, including enhanced bile acid and neutral sterol excretion, the slowing of fat and cholesterol absorption and direct inhibition of hepatic cholesterol synthesis by propionate formed by large bowel fermentation of NSPs. Whole body cholesterol homoeostasis represents a balance between influx and loss. Cholesterol influx can come from dietary intake and *de novo* synthesis. Losses occur through the sloughing of epithelial cells and through the fecal excretion of nonabsorbed dietary cholesterol and biliary steroids (bile acids and neutral sterols). Bile acids are generally recovered in the ileum, and those that are not absorbed are excreted in the feces. Any increase in bile acid excretion leads to enhanced hepatic uptake of cholesterol and its conversion to bile acids with a consequent depletion of the plasma cholesterol pool.

It was initially thought that fiber could bind some bile acids selectively, in a similar manner to cholestyramine, an ion exchange resin that binds bile acids. Bile acid binding *in vitro* by insoluble fiber preparations appears to be an artefact. Cholestyramine is strongly charged, whereas most NSPs with cholesterol-lowering potential are neutral or even acidic (e.g., pectins). Neutrality is not consistent with ionic binding and uronic acid residues would repel bile acids at the pH of the small intestine. The property that appears to mediate the increased steroid excretion is the viscosity in solution. Most (but not necessarily all) NSPs that lower cholesterol form viscous solutions in water. Presumably, bile acids are lost from the ileum through a form of entrapment in a viscous gel. This would also contribute to the loss of cholesterol and the slower digestion of fat seen with ingestion of NSPs. Abundant animal and human data show that feeding soluble NSPs increases fecal steroid excretion. However, the major problem with these relationships is that although soluble fibers may lower plasma cholesterol, the strongest evidence of a protective effect is for insoluble fibers which do not lower plasma cholesterol. It may be that other components in the grain are actually mediating the effect and fiber is the surrogate marker for their intake.

Dietary Fiber and the Etiology of Cancers—Colon and Rectum

Population Studies

This is one long-standing association that has been surprisingly problematic. Early studies on native Africans who consumed an unrefined diet showed them to have a very low incidence of this cancer. Although subsequent studies have shown a negative association between greater fiber intake and lowered risk, it has proved to be relatively weak. Indeed, in one US study there was no real association between fiber intake and cancer susceptibility. Some of the loss of significance seen in this evaluation may reflect the lack of allowance for confounding variables. For example, in a 6-year follow-up of women, the association between low fiber intake and the incidence of colon cancer disappeared after adjustment was made for meat intake. In another study of men, low fiber intake was an independent risk factor for the incidence of adenomatous polyps during a 2-year follow-up period.

Fruit and vegetable fiber has been consistently associated with a lower risk of colon cancer, but the relationship with cereal fiber is less clear. However, whole grain cereals appear to be protective—a further anomaly in the relationships between plant foods and disease risk. These discrepancies may be in the process of resolution. First, it seems that the early observational data were confounded by the analytical technologies available, and the perception that native populations consuming unrefined diets had high fiber intakes is incorrect. It seems likely that they ate relatively little fiber but had high intakes of RS. Population studies have shown a protective effect of apparent RS intake and colorectal cancer risk. The word 'apparent' is pivotal because there is currently no accepted method for RS determination and thus, there are no reliable data on dietary intakes. There are also issues regarding the intakes of dietary fiber and cancer risk. Part of the problem inherent in the study of colonic cancer is that, in contrast to CHD (in which there are easily measurable risk markers such as plasma cholesterol that can be modified by diet), the only indices for colon cancer are not easily measurable: the appearance of aberrant crypts, adenomatous polyps, or the disease itself. Hitherto, animal studies have largely been confined to rodents treated with chemical carcinogens (usually dimethylhydrazine), and they suggest that dietary fiber from wheat bran and cellulose may afford greater protection against the development of colon cancer when associated with a low-fat diet compared with soluble NSPs. These data stand in contrast to observational studies but are supported by interventions in humans with familial adenomatous polyposis. These people are at genetically greater risk of colonic cancer and represent one means of assessing risk modification through dietary intervention and monitoring polyp size and frequency through colonoscopy. In the Australian

Polyp Prevention Trial, subjects consumed 25 g of wheat bran per day and there was a decrease in dysplasia and total adenoma surface area when the diet was also low in fat. This supports epidemiological studies that show that increased fat and protein intakes increase risk. Other prevention trials have examined the effects of increasing fiber intake on the recurrence of polyps following a polypectomy. In a Canadian study of 201 men and women, a high-fiber, low-fat diet protected against polyp recurrence in women but in men there was actually an increase. A third trial examined the effects of diet on the prevalence of rectal polyps in 64 people with familial polyposis coli who had a total colectomy. Those who received and actually took the high-fiber (22.5 g fiber as a breakfast cereal) showed a reduction in polyps. These data are not conclusive but are reasonably consistent with overall knowledge.

Complex Carbohydrates and Colorectal Cancer

An obvious factor for the inconsistent results of the effect of different intakes of dietary fiber on colorectal cancer is the variation in the analytical methodology used in different studies. There is also increasing evidence that total dietary complex carbohydrates may be as important as fiber. Analysis of stool weight from 20 populations in 12 countries showed that larger stools were correlated with a lower incidence of colon cancer. Intakes of starch and dietary fiber (rather than fiber alone) were the best dietary correlates with stool weight. A subsequent meta-analysis showed that greater consumption of starch (but not of NSPs) was associated with low risk of colorectal cancer in 12 populations. The examination also showed that fat and protein intakes correlated positively with risk. This meta-analysis is probably the first of its kind to suggest a protective role for starch in large bowel cancer and underscores the need to consider complex carbohydrates as fiber equivalents and not just as NSPs and starch. The need for better information on dietary intake data and risk is underscored by the data from the European Prospective Investigation of Cancer and Nutrition, which showed a substantial reduction in risk with increasing fiber intake. This multinational study is important because it has sufficient power (expressed as a range of fiber intakes and individuals observed) to give confidence in the observations. Follow-up of 1 939 011 person-years throughout 10 countries showed that a doubling of fiber intake from foods could reduce risk by 40%.

Potential Mechanisms Indicating a Role in the Etiology of Colorectal Cancer

Colorectal tumorigenesis is a multistep process. These steps involve a number of genetic alterations that convert a normal epithelium to a hyperproliferative state and then to early adenomas, later adenomas, and, finally, frank carcinoma and metastasis. Fiber may, and probably does, play a role in all of these stages, and several mechanisms have been proposed by which it could play a role in the etiology of the disease (**Table 1**).

A number of agents may induce genetic damage in the colonocyte, including mono- and diacylglycerols, nonesterified fatty acids, secondary bile acids, aryl hydrocarbons and other pyrrolytic products of high-temperature cooking, and ammonia and amines and other products of large bowel bacterial protein degradation. One of the simplest protective mechanisms for dietary fiber is purely physical. By increasing fecal bulk, fiber could produce a more rapid transit time as well as act as a diluent and thus reduce exposure to potential mutagenic agents. It is also possible that fiber components could bind mutagens. However, because this appears to be unlikely for bile acids, the same may apply to other carcinogens.

Table 1 Effects of dietary fiber and resistant starch that could impact on the etiology of colorectal cancer

Increased stool bulk (mainly insoluble NSPs)
Decreases transit time, minimizing contact between colonocytes and luminal carcinogens
Reduces exposure through dilution of carcinogens
Binding of bile acids and other potential carcinogens (mainly insoluble NSPs)
Lowers free concentrations of mutagens
Modifying fecal flora and increasing bacterial numbers (soluble and insoluble NSPs and RS)
Decreases secondary bile acids, which are potential carcinogens
Lowers colonic NH_3 (a cytotoxic agent) by fixing nitrogen in the bacterial mass
Lowering fecal pH through SCFA production (NSPs but mainly RS)
Inhibits growth of pH-sensitive, potentially pathogenic species, which may degrade food constituents, and endogenous secretions to potential carcinogens
Lowers absorption of toxic alkaline compounds (e.g., amines)
Lowers solubility of secondary bile acids
Fermentation to SCFAs (NSPs but mainly RS)
Depending on source, raises butyrate which is a preferred substrate for normal colonocytes, and (*in vitro*) promotes a normal cell phenotype, retards the growth of cancer cells, and facilitates DNA repair

NSPs, nonstarch polysaccharides; RS, resistant starch; SCFAs, short-chain Fatty acids.

Production of SCFAs by the resident microflora induces a number of general changes in the colonic environment, including a lowering of pH. Case–control studies show that pH is higher in patients with cancer compared to controls but this may reflect altered dietary habits rather than long-term risk. However, at lower pH, basic toxins are ionized while secondary bile acids are less soluble so that the absorption of both would be reduced. The activities of both of the enzymes 7α-dehydroxylase and glucuronidase are decreased at lower pH. These changes would diminish the conversion of primary to secondary bile acids and the hydrolysis of glucuronide conjugates, respectively and thus limit their carcinogenic potential. However, there is consensus that the effects of SCFAs may be rather more specific and mediated through one acid—butyrate. Butyrate is a preferred substrate for normal colonocytes and numerous studies *in vitro* have shown that it has several actions that promote a normal cell phenotype. Cell studies show that butyrate induces hyperacylation of histones, leading to downregulation of gene expression and arrest of proliferation. Other actions include DNA hypermethylation which would have similar effects on tumor cell growth. Butyrate also has favorable effects on apoptosis so that a normal program of cell death is maintained. One marker of a differentiated colonocyte is its ability to produce alkaline phosphatase and butyrate is a powerful promoter of alkaline phosphatase *in vitro*. There is reciprocal downregulation of various oncogenes in colorectal cancer cell lines. These data are very promising for a direct role of butyrate in protecting against colonic cancer but there is an emerging paradox. In the presence of butyrate, there is either increased proliferation or no effect in normal cells but the proliferation of neoplastic cells is reduced. The differentiation of the normal cells is unchanged or suppressed with butyrate but is induced in cancer cells. These differing effects may be explained by neoplastic alterations (perhaps as a result of mutations in oncogenes) in cell signal systems.

It must be emphasized that none of the effects of butyrate *in vitro* have been duplicated *in vivo*, but they are of great promise and supportive evidence continues to accumulate. This is especially true for RS which appears to produce relatively more butyrate than other nondigestible carbohydrates. However, consideration may also need to be given to the existence of interindividual differences in the fermentative capacity of the microflora, the fact that RS from different sources may be fermented to different extents, and the actual colonic site at which fermentation takes place (i.e., whether in the proximal or distal colon).

Inter alia, the data suggest that protection against colorectal cancer is due to several mechanisms and that these can interact. One factor of considerable importance is the issue of overweight which is an independent risk factor for colorectal cancer. Obesity may have to be taken into account much more than has been the case in earlier studies. It appears that some of the effect may be mediated through raised plasma insulin and insulin-like growth factors (which may well be influenced by dietary carbohydrates).

Dietary Fiber and the Etiology of Hormone-Dependent Cancers

Population Studies

Cancers of the breast, endometrium, ovary, and prostate fall into the hormone-dependent classification. An association between hormonal status and cancer risk arose from observations of oestrogen deprivation and breast cancer and testosterone deprivation and prostate cancer. Nutritional influences on breast cancer have been studied extensively and several (but not all) studies show diminished risk with greater intakes of dietary fiber. The situation for other cancers, especially prostate cancer, appears to be rather unclear, but given the commonality of the proposed protective mechanisms, it is reasonable to expect that some linkage may be found. Male vegetarians have been reported to have lower testosterone and oestradiol plasma concentrations compared to omnivores, and inverse correlations of testosterone and oestradiol with fiber intake have been reported.

Potential Mechanisms Indicating a Role in the Etiology of Hormone-Dependent Cancers

There are many published studies that have produced mixed and inconsistent results on the potential mechanisms involved. Dietary fiber could act by reducing circulating concentrations of oestrogen and testosterone. Such an effect would not be unexpected in view of the fact that soluble NSPs can increase bile acid and neutral steroid excretion and fecal steroid outputs are higher in vegetarians than in omnivores. However, one anomaly is the finding that wheat bran (which does not enhance biliary steroid excretion) lowers circulating and urinary oestrogens. It is possible that fiber acts rather differently on hormones than on bile acids and neutral sterols. For example, the colonic flora may be modified so as to increase deconjugation of the sex hormone precursors or their conversion to other

metabolites. Direct binding of sex hormones is possible but is subject to the same concerns as were raised for cholesterol reduction. In addition, it is possible that other components in, or associated with, fiber (phytooestrogens or antioxidants) may be responsible for any observed protective effect. Soy phytooestrogens are believed to play a role in lowering the risk of breast cancer in Asian populations. Lycopenes are antioxidant carotenoids from tomatoes, and their intake has been correlated with a lower risk of prostate cancer.

Dietary Fiber, Obesity, and the Etiology of Diabetes

In 1975, Trowell suggested that the etiology of diabetes might be related to a dietary fiber deficiency. This is supported by several key pieces of evidence. Vegetarians who consume a high-fiber lacto-ovo vegetarian diet appear to have a lower risk of mortality from diabetes-related causes compared to nonvegetarians. Consumption of whole grain cereals is associated with a lower risk of diabetes. Importantly, the same dietary pattern appears to lower the risk of obesity, itself an independent risk factor in the etiology of type 2 diabetes. Obesity is emerging as a problem of epidemic proportions in affluent and developing countries. Consumption of whole grain cereal products lowers the risk of diabetes. A report showed that in 91 249 women questioned about dietary habits in 1991, greater cereal fiber intake was significantly related to lowered risk of type 2 diabetes. In this study, glycemic index (but not glycemic load) was also a significant risk factor, and this interacted with a low-fiber diet to increase risk. These results provide epidemiological evidence of a role of fiber in the etiology of diabetes.

Potential Mechanisms Indicating a Role in the Etiology of Diabetes

It can be hypothesized that a reduction in the general and postprandial glycemic and insulinemic response may delay the development of insulin resistance and thus the development of diabetes (NIDDM) although there is very little direct evidence to support this hypothesis. However, diets high in both carbohydrate and dietary fiber have been reported to improve insulin sensitivity. Much of the research in this area has studied the effect of dietary fiber on the management rather than the prevention or etiology of diabetes.

There is good evidence that diminished glucose absorption lowers the insulin response to a meal. The action of fiber in this regard may be through slowing the digestion of starch and other nutrients. It seems that soluble fiber may play a role because large amounts of soluble dietary fiber have been shown to reduce postprandial glucose concentration and insulinemic responses after a single meal in both normal and diabetic subjects. However, the effect appears to be dependent on viscosity rather than on solubility *per se*. The very viscous gum, guar gum, gum tragacanth, and oat gum are all very effective whereas psyllium and some pectins are less viscous and less effective. One suggested mechanism for reducing the glycemic response is an impairment in the convective movement of glucose and water in the intestinal lumen due to the formation of a viscous gel: Glucose is trapped in the gel matrix, such that there is less movement toward the absorptive brush border of the surface of the intestinal wall and the glucose needs to be squeezed out by the intestinal motor activity of the intestine. However, other factors may also be important. There may be some impairment in digestive activity in the lumen, an alteration in hormonal secretion by cells in the gut mucosa, and a reduced gut motility that delays transit time. In the case of whole grains, there is scope for the fiber to interfere with the physical accessibility of starch to small intestinal α-amylase. Clearly, there is also potential for foods of low glycemic index to be high in RS and this does seem to be the case. A specific instance is a novel barley cultivar that exhibits both characteristics. It should be noted that there are reports of a second meal effect (i.e., the dietary fiber ingested at one meal can affect the glucose rise after the subsequent meal). The mechanism for this is unknown.

A Role for Fiber in the Etiology of Other Diseases?

Although much of the earlier observational studies in native African populations were wide ranging, most attention has subsequently focused on CHD and cancer. Probably this is a reflection of the socioeconomic importance of these conditions in economically developed societies. However, fiber has a role in the prevention and management of other conditions, but much of the relevant information has come from interventions, not from case–control or cross-sectional studies.

Constipation, diverticular disease, and laxation Unquestionably, fiber is of direct benefit in relieving the symptoms of constipation and diverticular disease but there is little information about its role in the etiology of these conditions. Numerous

interventions have shown that foods high in insoluble NSPs (e.g., certain cereal brans) and some soluble NSP preparations (e.g., psyllium) are very effective at controlling constipation and diverticular disease and enhancing laxation. The actual effect can vary with source. Wheat bran increases undigested residue, and fiber from fruits and vegetables and soluble polysaccharides tend to be fermented extensively and are more likely to increase microbial cell mass. Some NSP (and OS) preparations retain water in the colon. The physical form of the fiber is also important: Coarsely ground wheat bran is a very effective source of fiber to increase fecal bulk, whereas finely ground wheat bran has little or no effect and may even be constipating. RS appears to be a mild laxative and seems to complement the laxative effects of NSPs. The effective dose appears to be approximately 20–30 g of total fiber/day consumed either in food or as a supplement. In addition, animal studies show that NSPs and RS appear to prevent colonic atrophy seen in low-fiber diets. The mechanism of action appears to be greater fecal bulking and fermentation and the generation of SCFAs, which is necessary to prevent atrophy.

Diarrhea Colonic SCFA absorption stimulates fluid and electrolyte uptake in the colon and thus can assist in reducing diarrhea. Complex carbohydrates may also play a role in modifying the colonic microflora thus reducing the number of pathogens. An etiological role for fiber is unknown, but there is good evidence that RS can act to minimize the fluid losses that occur in serious conditions such as cholera.

Inflammatory bowel diseases (colitis and Crohn's disease) Clearly, inflammatory conditions have an immune component. In the case of Crohn's disease, there appears to be no established therapeutic or etiological role for fiber. The situation is slightly different for distal ulcerative colitis, in which fiber intake seems unrelated to incidence. However, rectal infusion of SCFAs (especially butyrate) has been reported to lead to remission, so it appears that either the generation of these acids or their delivery to the distal colon may be the issue.

See also: **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Cereal Grains**. **Colon**: Disorders; Nutritional Management of Disorders. **Coronary Heart Disease**: Prevention. **Diabetes Mellitus**: Etiology and Epidemiology. **Diarrheal Diseases**. **Dietary Fiber**: Physiological Effects and Effects on Absorption. **Food Safety**: Bacterial Contamination. **Obesity**: Prevention. **Vegetarian Diets**.

Further Reading

Baghurst PA, Baghurst KI, and Record SJ (1996) Dietary fibre, non-starch polysaccharides and resistant starch—A review. *Food Australia* **48**(supplement): S3–S35.

Bingham SA, Day NE, Luben R *et al.* European Prospective Investigation into Cancer and Nutrition (2003) Dietary fibre in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC): An observational study. *Lancet* **361**: 1496–1501.

Ellis PR, Rayment P, and Wang Q (1996) A physico-chemical perspective of plant polysaccharides in relation to glucose absorption, insulin secretion and the entero-insular axis. *Proceedings of the Nutrition Society* **55**: 881–898.

Giovannucci E (2001) Insulin, insulin-like growth factors and colon cancer: A review of the evidence. *Journal of Nutrition* **131**(supplement 3): 109S–120S.

Olson BH, Anderson SM, Becker MP *et al.* (1997) Psyllium-enriched cereals lower blood total cholesterol and LDL cholesterol, but not HDL cholesterol in hypercholesterolemic adults: Results of a meta-analysis. *Journal of Nutrition* **127**: 1973–1980.

Richardson DP (2003) Whole grain health claims in Europe. *Proceedings of the Nutrition Society* **62**: 161–169.

Schulze MB, Liu S, Rimm EB *et al.* (2004) Glycemic index, glycemic load, and dietary fiber intake and incidence of type 2 diabetes in younger and middle-aged women. *American Journal of Clinical Nutrition* **80**: 243–244.

Slavin JL (2000) Mechanisms for the impact of whole grain foods on cancer risk. *Journal of the American College of Nutrition* **19**: 300S–307S.

Stamler J, Caggiula AW, Cutler JA *et al.* (1997) Dietary and nutritional methods and findings: The Multiple Risk Factor Intervention Trial (MRFIT). *American Journal of Clinical Nutrition* **65**(1 supplement): 183S–402S.

Topping DL and Clifton PM (2001) Short-chain fatty acids and human colonic function: Roles of resistant starch and nonstarch polysaccharides. *Physiological Reviews* **81**: 1031–1064.

Topping DL, Morell MK, King RA *et al.* (2003) Resistant starch and health – *Himalaya 292*, a novel barley cultivar to deliver benefits to consumers. *Starch/stärke* **53**: 539–545.

Truswell AS (2002) Cereal grains and coronary heart disease. *European Journal of Clinical Nutrition* **56**: 1–14.

Wu H, Dwyer KM, Fan Z *et al.* (2003) Dietary fiber and progression of atherosclerosis: The Los Angeles Atherosclerosis Study. *American Journal of Clinical Nutrition* **78**: 1085–1091.

Role in Nutritional Management of Disease

A R Leeds, King's College London, London, UK

Introduction

Dietary Fiber was an unknown phrase to all but a handful of individuals in the early years of the 1970s when a wide range of potential therapeutic applications were suggested by Hugh Trowell, Denis Burkitt, and Alexander Walker. Twenty-five years later there can hardly be an ordinary mortal who has not heard the term, though he may not be able to define it. In some cases the claims remain largely unsubstantiated but in three areas, hyperlipidemia, diabetes, and bowel function, there is sufficient evidence to allow dietary advice to be given.

Hyperlipidemia

Some forms of dietary fiber lower blood lipids, notably total cholesterol and low-density lipoprotein (LDL) cholesterol. The earliest observations on fiber preparations and blood lipids date from the mid 1930s when there was a fairly extensive investigation of the effects of pectin (polygalacturonic acid). The next period of investigation dates from 1974 when extracted and purified dietary fiber preparations such as guar gum – a glucomannan – were tested in normal subjects, diabetics, and hyperlipidemic subjects and were found to lower blood cholesterol when given in sufficient quantities. In very large doses these materials increase fecal excretion of fat and sterol compounds and would be expected to reduce the body bile salt pool. Subsequent work has shown that at lower doses preparations of soluble dietary fiber have a mild cholestyramine-like effect: they bind bile salts rendering them unavailable for reabsorption in the terminal ileum, thus interfering with the normal entero-hepatic cycle of bile salts and depleting the bile salt pool. Total and LDL cholesterol fall as cholesterol is diverted for the resynthesis of lost bile salts. There have been few direct clinical applications of the early experimental work on pectin and guar gum. No pectin compounds have been developed commercially, but there are a few pharmaceutical preparations of guar gum presented primarily as adjuncts to dietary therapy in diabetes rather than for lipid lowering. Dietetic food products containing guar gum have been developed, again for use in controlling diabetes.

Preparations of soluble dietary fiber have been shown to lower blood cholesterol whereas most preparations of predominantly insoluble fiber, such as wheat bran, have little or no effect. The major food sources of soluble fiber are oats, beans, lentils, rye, and barley, and these foods have naturally become the subject of investigations. The addition of oats to the diet in normolipidemic and hyperlipidemic subjects following either their normal diets or where pretreated with low-fat diets has been the subject of extensive research. In sufficient quantity oats, oat products, and oat β-glucan (providing at least 3 g oat β-beta glucan per day) lower blood total cholesterol and LDL cholesterol (usually by 5–10%) while leaving triglycerides and HDL cholesterol largely unchanged. A sufficiently large number of good-quality studies have now been done on oats that the Food and Drug Administration (FDA) has allowed the first ever food-specific health claim: "Soluble fiber from oatmeal, as part of a low saturated fat, low cholesterol diet, may reduce the risk of heart disease." Products that are labeled with this claim must provide at least 0.75 g of soluble fiber (as β-glucan) per serving. When considering the above claim the FDA reviewed 37 studies and found that a sufficient number provided convincing evidence of efficacy. An earlier meta-analysis of some of those trials had shown that the efficacy of oats and oat products was influenced by the initial values of blood cholesterol in the subjects: patients with high starting values (over 6.7 mmol per liter total cholesterol) showed the greatest reductions when treated with oats, while healthy young subjects with low–normal starting values showed little response. There was a dose effect: food products providing more than 3 g soluble fiber per day had a greater blood cholesterol lowering effect than diets that provided less than 3 g per day.

Other soluble fiber-containing products have been shown to lower blood cholesterol. Recent extensive studies on psyllium (*Plantago ovata*) presented both as a pharmaceutical preparation and as a food product (a ready to eat breakfast cereal) have shown blood cholesterol-lowering properties where the dose–effect relationship is such that a useful additional therapeutically meaningful lipid-lowering effect can be achieved by prescribing a daily portion of psyllium-fortified breakfast cereal. Products of this type are now marketed in the US and Australia, and the US FDA has now allowed a food specific health claim for psyllium.

There is also a small literature on the effects of beans on blood lipids and the findings of a blood cholesterol-lowering effect are as expected.

Virtually all of the reports of the effects of soluble fiber products on blood lipids report lowering effects on total cholesterol and LDL cholesterol without any effect on HDL cholesterol or triglycerides – this contrasts with the effects of some drugs that may cause slight rises of triglycerides and falls of HDL cholesterol. The relationship between lowering of blood cholesterol and lowering of risk of heart disease is now generally accepted and a proven lipid-lowering effect is taken to mean a beneficial effect on risk of coronary heart disease. This means that in clinical practice it is perfectly reasonable to include advice on use of foods high in soluble dietary fiber in a lipid-lowering diet, and perfectly proper to emphasize the benefits of oats and oat products. Generally, a high-soluble-fiber diet is more acceptable when the soluble fiber is drawn from smaller quantities of a larger range of foods; thus the diet includes beans, lentils, rye breads, and barley as well as generous use of oats. A range of foods containing mycoprotein and fungal mycelial cell walls (chitin) may also help to lower blood cholesterol.

Diabetes

Diabetes mellitus is characterized by either an absolute or relative lack of insulin, which has short-term and long-term consequences. Diabetic people may develop both microvascular complications (mainly affecting the eyes, kidneys, and nerves) and macrovascular complications (essentially accelerated development of atherosclerosis presenting mainly as heart attack and peripheral vascular disease). Medical management aims to replace the insulin, or modulate its production or efficacy using oral (hypoglycemic) drugs, in a metabolic environment enhanced by good control of diet and body composition. Medical management also aims to achieve early detection of complications and other risk factors for cardiovascular disease by regular testing of blood and urine biochemical variables and blood pressure and by regular physical examination of the eyes, neurological, and cardiovascular systems.

Control of dietary energy intake (in relation to the varying demands for growth, maintenance, physical activity, etc.) remains the key feature of dietary control affecting metabolic fluxes, blood glucose levels, and body weight. Views on the appropriate proportional sources of energy from fat, carbohydrate, and protein have changed enormously over the last century from seriously energy-restricted high-fat diets (with percentage energy from fat as high as 70% raising some doubts about the level of compliance) through to very high-carbohydrate diets (sometimes 60–65% energy from carbohydrate) used in specialist centers in the US. Today, for most diabetic patients in most countries the target is to achieve 50–55% energy from carbohydrate sources. Prior to the 1970s, when the move towards high-carbohydrate diets began, the high fat content of the diet along with less tight blood glucose (and urine glucose) control than is customary today was partly responsible for the high relative mortality from cardiovascular disease seen among diabetic patients. At that time young male diabetics were up to nine times more likely to die from heart attack than matched nondiabetic individuals. Reduction of fat in the diet and achievement of an optimal distribution from saturated, monounsaturated, and polyunsaturated sources (<10%, 10–20%, and no more than 10%, respectively, for patients with diabetes in the UK) remain a major aspect of dietary management of diabetic people in order to reduce the risk of developing coronary heart disease.

Control of blood glucose is critical in order to achieve avoidance of prolonged periods of hyperglycemia, which is associated with glycation of proteins and the risk of development of microvascular complications, and avoidance of hypoglycemia with its attendant risks of coma. In day-to-day practice, the avoidance of hypoglycemia is very important to patients and any new method of achieving normalization of blood glucose profiles is an advance. Dietary fiber offered such an advance from the mid 1970s when some forms (notably isolated polysaccharides such as guar gum, a glucomannan, and pectin, polygalacturonic acid) were shown to reduce the area under the blood glucose and insulin curves after acute test meals. Subsequent long-term (6-week) clinical trials showed that diets high in foods containing soluble dietary fiber, such as beans, oats, and barley, were more effective in reducing the area under the 24-h blood glucose profiles than diets containing more high-fiber foods based on wheat products.

Research in this area led David Jenkins to describe (in 1981) the concept of the 'glycemic index' (GI) which is a numerical expression of the ability of a food to raise blood glucose levels. In practice it is measured by comparing the blood glucose response to a 50-g carbohydrate portion of food with the response to 50 g glucose (in some papers the comparison is with a 50-g carbohydrate portion of bread). The dietary fiber (especially soluble fiber) content of a food slows down the rate of digestion and absorption of starch in foods giving flatter blood glucose responses and a lower GI; however, the structure of the starch (whether amylose or

amylopectin) influences its rate of degradation and the extent to which the starch granules are hydrated by processing (including cooking) is also important. The physical structure of the food (particularly the extent to which plant cells are intact), the presence of fat, which may slow gastric emptying, and the presence of some 'antinutrient' substances may all influence the GI. Low-GI diets have been shown in many clinical trials to improve important variables that are secondary indicators of blood glucose control, and to reduce blood lipids. Low-GI diets may be particularly helpful to patients who are frequently troubled by episodes of hypoglycemia though adequate proof of this is still awaited. Low-GI diets are not just relevant to treatment of diabetes but have been shown in two large-scale epidemiological surveys published in 1997 to result in a significant reduction in the risk of development of maturity onset (type 2) diabetes in middle-aged American men and women. Thus, there is good reason to believe that there should be greater emphasis on the GI of diabetic diets and the fiber content, as well as emphasis on GI for those at risk of developing diabetes, especially the older obese person. Expert committees in many developed countries of the world have set target values for dietary fiber intake for diabetic patients (e.., the American Diabetes Association (ADA) recommends 20–35 g day^{-1} total dietary fiber by the AOAC method) and many, especially the Australian Diabetes Association and with the notable exception of the ADA, have recommended an increase in low-GI foods. In 2003 even Diabetes UK (the UK Diabetes Association) noted that there might be merit in taking account of GI in dietary management for those with diabetes. Some physicians believe that the GI of foods is too complex an issue for patients to grasp, but in essence simply requires a partial substitution of bread and potatoes with pasta products, an increased use of high-fiber breakfast cereals including oats, increased use of beans and lentils, and emphasis on the use of temperate fruits (e.g., apples and pears).

Obesity (body mass index (weight in kilograms divided by height in meters squared) in excess of 30 kg m^{-2}) is becoming more prevalent in developing countries and attracts an increased risk of the development of diabetes mellitus; a high proportion of established type 2 diabetics are obese and overweight. In the popular diet book 'The F-Plan Diet,' published in 1982, Audrey Eyton claimed that dietary fiber would help people lose weight by a number of mechanisms including reducing the efficiency of dietary energy absorption and by making people feel full for longer after meals thus having an overall effect on reducing food intake. At the time of publication these ideas were hypothetical - subsequent investigation has shown that increasing fiber intake two- or threefold by a variety of dietary changes can increase fecal energy losses by 75–100 kcal day^{-1}. Studies on the effects of dietary fiber on postprandial satiety where experimental meals are carefully designed to differ little except for fiber content have given variable results. However, there is a clear effect of fiber on chewing (the number of chews necessary to eat the same energy equivalent of food) where high- and low-fiber types of commonly consumed foods are eaten and this may have an important satiating effect. Clinical trials of high-fiber weight loss regimens have given variable results. Double-blind placebo-controlled trials using pressed barley fiber and pectin tablets compared to a starch control have been undertaken in Scandinavia and have demonstrated statistically significantly greater weight losses in the fiber-treated groups up to 26 weeks of treatment. It seems reasonable to conclude that under some conditions the right kind of high-fiber diet can facilitate weight loss, but may not always do so.

Diabetic people are more likely to have dyslipidemia than nondiabetic people. When control of diabetes is lost, patients may demonstrate gross hypertriglyceridemia due to increased production of very-low-density lipoprotein (VLDL) particles in the liver as a consequence of the increased flux of free fatty acids from the peripheral tissues. At the same time total and LDL cholesterol may be raised. Improvement in diabetic control often achieves normalization of blood lipids, but where hyperlipidemia persists there may be a place for use of dietary fiber, especially soluble fiber, and especially oat β-glucan-containing foods as an adjunct to dietary and pharmacological therapy (see above).

Bowel Disorders

Denis Burkitt first suggested a role for dietary fiber in bowel disorders in 1971. In the intervening period understanding of the normal physiology and pathophysiology of the colon have improved enormously. During the same period methods of analysis have been refined and a distinction is drawn between dietary fiber (as determined by the AOAC gravimetric method) and nonstarch polysaccharide (NSP; determined by GLC analysis of component sugars), and starch not digested in the small gut is now defined as being resistant. Three types of resistant starch have been described. These advances in analysis have helped physiologists appreciate the

contributions of various substrates to colonic fermentation and stool bulking.

The intake of dietary fiber (nonstarch polysaccharides) is directly related to the amount of wet stool passed each day in large population groups. An average wet stool weight for the UK is about 105 g day^{-1}, which corresponds roughly to a nonstarch polysaccharide (Englyst method) intact of 12.5 g day^{-1}. Nearly half of the members of groups studied in the UK have stool weights of less than 100 g day^{-1} below which complaints of constipation are common. Stool weight has been shown to be clearly inversely related to colon cancer incidence in population groups: a mean daily stool weight of 105 g corresponding to a relatively high population colon cancer incidence of about 22 per 100 000 per annum. An incidence rate of 11 per 100 000 per annum corresponds to a mean daily stool weight of about 175 g day^{-1}. This information was used as the numerical basis for calculating the UK's dietary reference value (DRV) for nonstarch polysaccharide (NSP) in the late 1980s. In the UK the population is urged to increase NSP intake by 50% to a population averageof 18 g day^{-1} in order to shift the distribution of wet stool weight upwards.

Constipation is generally considered to be infrequent opening of the bowels with straining to pass stools (less than three defecations per week and straining and or the passing of hard stools in more than one in four defecations). Constipation is sometimes caused by other specific disease of either an endocrine nature (e.g., myxoedema – reduced thyroid function) or physical obstructive nature (e.g., colon cancer). Where constipation has developed recently in a previously nonconstipated individual over the age of 40 years colon cancer must be excluded as the cause of the change of bowel habit. In the absence of evidence that the constipation is secondary it is probably due to dietary and life-style factors. The mucosa of the lower colon has a great capacity to desiccate its contents. If the call to stool does not occur or is ignored residual material drys out and individual fecal pellets become smaller. There is experimental evidence to suggest that greater abdominal pressures are needed to expel pellets that are 1 cm in diameter than those that are 2 cm in diameter. Thus, factors that result in the call to stool being ignored, like not allowing sufficient time for defecation after a stimulus such as breakfast or the walk to the station or being unprepared to defecate anywhere except at home (a common characteristic consistent with mammalian behavior), are likely to cause constipation. Simple solutions include going to bed earlier and getting up earlier in the morning, and finding another acceptable location for defecation at the workplace. Increasing fiber in the diet, most easily achieved by making breakfast a high-fiber meal with either high-fiber breakfast cereals or high-fiber breads, will increase stool bulk, shorten transit time (the time for a marker to pass from the mouth and be passed in the stool), and alleviate symptoms in many cases. The importance of exercise in maintaining normal colon function is gradually being recognized – the importance of brisk walking should not be underestimated. However, some specific types of simple constipation have been identified which do not necessarily respond to high-fiber diets. Grossly prolonged transit times reflecting seriously slow colonic motility has been seen particularly in young women and do not respond well to high-fiber diets, and some 'outflow abnormalities,' which sometimes have a basis in abnormal rectal conformation, may also not respond.

Diverticular disease of the colon, characterized by the development of protrusions of mucosa through the bowel wall, is common and usually asymptomatic. It has been shown to be less likely to develop in those following a high-fiber diet, and once acquired can be managed, in many cases, by ensuring an adequate amount of fiber in the diet. Experimentally, various fiber supplements and 'bulking agents' have been shown to reduce the abnormally high peak intracolonic pressures that are characteristic of diverticular disease. Sometimes 10–20 g of coarse wheat bran as a supplement is all that is required, but some patients develop flatulence and distension at least initially. Other fiber supplements such as ispaghula husk (psyllium) may be as effective, without the initial adverse side effects. Sometimes, simple dietary changes to achieve an adequate total daily intake of dietary fiber particularly from wheat-based foods are effective. Diverticulitis (inflammation of the diverticula) is a complication requiring medical management, which will usually include a short period of abstention from food. Many patients remain largely without symptoms once the right 'fiber' regimen has been determined.

The irritable bowel syndrome (IBS) is a 'functional' disorder of the bowel, which is said to affect up to 15% of the population and is characterized by some, but not necessarily all, of a range of symptoms including abdominal pain relieved by constipation, alternating diarrhea and constipation, recurrent abdominal pain, and urgent or frequent defecation. An important part of management is the exclusion of other serious organic disease such as inflammatory bowel diseases. In IBS the gut is abnormally sensitive to distension, and symptoms may be related to or

exacerbated by external emotional events. The role of high-fiber diets in IBS has been investigated and not surprisingly is only of benefit in some cases: in those patients in whom the predominant feature is constipation. In some patients high-fiber diets may make their symptoms worse.

In inflammatory bowel disease (IBD) high fiber diets have no special part to play in the management of Crohn's disease where enteral feeding (with formula low-residue, low-fiber preparations) is especially beneficial where there is acute extensive small bowel disease. In ulcerative colitis specific dietary advice is usually unnecessary though fiber supplements may be of benefit in patients whose disease is limited to proctitis (inflammation of the rectum).

The treatment of newly diagnosed colon cancer does not include diet therapy, but treatment of those at increased risk of developing colon cancer by dietary and other means will become increasingly common as more information about the effects of high fiber diets and supplements on colon function becomes available. The critical step in the adenoma-carcinoma sequence in the human large bowel is the enlargement of the small adenoma (which has a low risk of malignant transformation) to a large adenoma (which has a high risk of malignant transformation); dietary factors, including low amounts of fiber in the diet, enhance adenoma growth. Bile acids are strongly linked to adenoma growth and bile acid concentrations in the colon are influenced by dietary fat and dietary fiber. Other effects of fiber may also be protective: bulking the stool and accelerating material through the colon, and provision of substrate for fermentation particularly with production of butyrate, which may have antineoplastic properties. However, despite a great deal of epidemiological and experimental work the potential role of dietary fiber in modulating the risk of colon cancer remains controversial.

See also: **Cereal Grains**. **Cholesterol**: Factors Determining Blood Levels. **Colon**: Nutritional Management of Disorders. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Dietary Fiber**: Physiological Effects and Effects on Absorption; Potential Role in Etiology of Disease. **Glucose**: Metabolism and Maintenance of Blood Glucose Level. **Glycemic Index**. **Hyperlipidemia**: Overview; Nutritional Management. **Lipids**: Chemistry and Classification. **Lipoproteins**.

Further Reading

Committee on Medical Aspects of Food Policy (1991) *Dietary Reference Values for Food Energy and Nutrients for the United Kingdom: Non Starch Polysaccharides*, pp. 61–71. London: HMSO.

Cummings JH (1997) *The Large Intestine in Nutrition and Disease. Danone Chair Monograph*. Brussels: Institute Danone.

Diabetes UK Dietary Guidelines (2003) The implementation of nutritional advice for people with diabetes. *Diabetic Medicine* **20**: 786–807.

Food and Agriculture Organization (1998) *Carbohydrates in Human Nutrition*. FAO Food and Nutrition Paper 66. Rome: FAO.

Jenkins DJA *et al.* (2003) The garden of Eden – plant based diets, the genetic drive to conserve cholesterol and its implications for heart disease in the 21st century. *Comparative Biochemistry and Physiology Part A* **136**: 141–151.

Relevant Websites

http://www.fda.gov – FDA health claim for psyllium on reducing risk of heart disease.

http://www.cfsan.fda.gov – FDA health claim for soluble fiber from whole oats and risk of coronary heart disease.

http://www.jhci.org.uk – JHCI final health claim for whole-grain foods and heart health.

http://www.jhci.org.uk – JHCI generic health claim for whole oats and reduction of blood cholesterol.

DIETARY GUIDELINES, INTERNATIONAL PERSPECTIVES

B Schneeman, University of California—Davis, Davis, CA, USA

Introduction

The use of food-based dietary guidelines (FBDG) has emerged as an important food and nutrition policy and education program since the late 1970s and is valuable for addressing issues of both nutrient adequacy and excess. Importantly, FBDG communicate nutritional principles in a manner that is relevant to the population. Most FBDG encourage energy balance, physical activity, a healthful variety of foods including fruits and vegetables, whole-grain products, food sources of protein, calcium, and unsaturated fatty acids, and safe food handling. Cautionary messages focus on excess energy intake, saturated and *trans* fatty acids, added sugars, salt, and alcohol. To develop relevant FBDG, each country must identify local public health issues and appropriate diet-related strategies for the population.

Historical Background

Throughout the ages, religious and philosophical writings have included dietary recommendations, and this is reflected in an oft-quoted line from Hippocrates: 'Let thy food be thy medicine'. In the late 1800s, modern science began to influence the nature of recommendations regarding foods and beverages. The original focus of guidelines developed from Pasteur's discoveries of the disease-causing organisms that could be present in foods such as milk; thus, recommendations emphasized sanitation in food handling. In the early 1900s the discovery of vitamins and minerals led to the realization that foods contain factors that are essential for health. This 'vitamin theory of disease' led to research throughout the first half of the twentieth century to discover these factors and determine their essential functions in the treatment of deficiency diseases. The knowledge that food was important in the prevention of diseases that were major public health problems, such as scurvy, beri-beri, night blindness, and pellagra, led to early efforts to develop and promote dietary recommendations or guidelines, even though the specific curative factors in foods had not been identified. Among the earliest examples, Egyptians were known to promote the use of liver to correct night blindness, the British Navy used lemons or limes to prevent scurvy, and alkali treatment of corn was associated with a lower incidence of pellagra in Mexico. The understanding of the linkage between certain foods and the prevention of disease resulted in the development of food guides or groups illustrating a pattern of food choices that was most likely to prevent deficiency diseases. As the chemical nature of the factors in food that prevented or cured nutritional deficiencies became known, it was possible to determine the specific amount required in the diet to maintain health. These studies led to the development of recommended dietary allowances (RDA), which are numeric recommendations of the nutrient intakes that will meet the needs of the majority of the population. RDA have also been used to evaluate the adequacy of diets in many populations.

By the second half of the twentieth century, it had become clear that in many developed countries the primary causes of disease were shifting from dietary deficiencies to those associated with dietary excess. As noted by the Surgeon General of the USA, by 1988 micronutrient deficiencies were no longer major public health problems in the USA, and diseases associated with excess intakes of energy, saturated fat, total fat, cholesterol, alcohol, and sodium, in conjunction with inadequate fiber intake, were the major causes of death in the USA. The economic transition experienced by many developing countries has led to a similar pattern; this is sometimes referred to as the double burden of disease. While nutritional deficiencies continue to be prevalent in large segments of the population, an increasing proportion of the population is at risk of developing diet-related chronic diseases, such as obesity, cardiovascular disease, cancer, and diabetes. This emerging pattern of disease has resulted in the development of FBDG, which recommend dietary patterns that are adequate in nutrient content and encourage food choices to lower the risk of noncommunicable diet-related diseases.

Types of Guidelines

This section outlines the evolution of three interrelated general types of nutrition recommendations that are developed and used in most areas of the world by national or regional government agencies: technical recommendations, which provide specific numeric criteria for nutrient intake; FBDG, which outline strategies to lower the risk of chronic disease; and food guides, which illustrate dietary patterns or food choices to encourage individuals to meet the recommended nutrient requirements and to follow the advice in dietary guidelines. The more technical quantitative guidelines are typically used by health professionals to develop educational materials and evaluate the adequacy of diets. Food guides and FBDG are important components of educational materials for healthy individuals. Although this section will focus on FBDG, it is important to understand how they are related to the other types of nutritional recommendations, and also that recommendations categorized as FBDG ideally are related to and supportive of other types of nutritional recommendations that are part of national or regional health policy. Frequently, non-governmental groups develop food guides or FBDG to suit a specific purpose (such as weight loss, treatment of cardiovascular disease, or promotion of a food culture) or a specific population (such as older individuals); however, before accepting these recommendations, it is important to determine how they have been validated in terms of other criteria such as the dietary reference intakes. These non-governmental recommendations do not have the same policy status as recommendations developed by government agencies or through government-sponsored scientific organizations and may be suitable only for a specific targeted function.

In 1992 a recommendation of the International Conference on Nutrition organized by the World Health Organization (WHO) and the Food and Agriculture Organization (FAO) was that each country should to develop nutritional recommendations that included FBDG. To encourage this activity, FAO and WHO convened a group of experts to recommend a process for developing FBDG; their findings are published in a WHO technical report.

The Development of Food-Based Dietary Guidelines

FBDG express the principles of nutrition education in terms of the food and food choices available to the population rather than in terms of specific nutrients or food components. Scientifically, these guidelines are based on the association between dietary patterns and the risk of diet-related diseases and incorporate recommendations that address major diet-related public health issues. In addition to communicating scientific knowledge about the association between food, dietary patterns, and health, development of FBDG provides an opportunity to strengthen consensus among various government and non-government organizations on important nutrition recommendations to be incorporated into educational programs. In addition, by expressing scientific principles in terms of food, FBDG recognize the consumer awareness of food rather than nutrients and emphasize to consumers the importance of meeting nutrient needs with foods. Thus, both the content of the FBDG and the process of development are important.

Researchers often focus their studies on a specific nutrient or food component that may alter the risk of developing a disease. These studies are reviewed in the development of FBDG, but the information must be reorientated from a nutrient-based focus to a food recommendation by addressing the questions in **Table 1**. As indicated by these questions, the process is driven by the identification of diet-related public health issues and the development of food-based strategies that are relevant to the target population.

The process for developing FBDG is based on building consensus among various sectors and groups involved in public health. **Table 2** provides a general outline of the steps in the process, which can be adapted to the specific needs of a country or region. The goal is to have a set of guiding principles for food-based recommendations that lay out the overall policy agreed by various agencies and groups.

The product of the working group is likely to be a document that outlines recommendations and includes background information on the rationale for the guidelines as well as guidance on implementing the recommendations. The guidelines from three countries are shown in **Table 3** as an example of the types of message developed during this process. In all cases, the messages are accompanied by a document containing background information. **Table 4** presents common themes emerging from the FBDG that have been developed in a variety of countries. Based on foods available and cultural practices, the types of fruits, vegetables, and whole grains and the specific types of food that are emphasized as sources of protein, calcium, or unsaturated fatty acids may vary. In

Table 1 Reorientating from nutrients and food components to foods

What are the important public health issues for the population? Do they have diet-related factors?
Health statistics will indicate the major causes of morbidity and mortality in a population. Diet-related diseases include nutritional-deficiency diseases and noncommunicable diseases such as obesity, type 2 diabetes, certain types of cancer, and cardiovascular disease. It is important to determine whether nutrition is the primary cause of the disease or secondary to some other more prevalent problem (e.g., smoking, infectious agents)

What are the target nutrients linked to the major public health issues? Are there related nutrients or other factors?
In many nutrition-related problems several nutrients or food factors may interact to cause the nutritional problem. For example, the fat content of the diet affects absorption of fat-soluble vitamins, obesity can be related to either excess energy intake or inadequate expenditure, multiple factors contribute to adequate bone formation, folic acid can mask vitamin B_{12} deficiencies, etc. Simply increasing the intake of a target nutrient and ignoring these other factors may not address the problem adequately

What foods are high in the nutrient(s) or consumed in sufficient quantity to be a significant source of the nutrient(s)?
Using both food-composition databases and food-consumption data, foods that are good sources of the nutrient and foods that are consumed in sufficient quantity to meet the target intake can be identified. Likewise, dietary patterns that lower the risk for the public health problem and are associated with adequate intake of the nutrient can be identified

What is likely to be acceptable by the target audience?
For nutrition interventions to achieve success, recommendations must target food choices that can be integrated into the diet based on cost and acceptability of the foods

How do diet strategies integrate with other food policies?
Economic, agricultural, and trade analysis is useful to determine which diet strategies are sustainable

Table 2 Steps in the development of food-based dietary guidelines

1. Develop support from key government agencies
The successful implementation of FBDG will depend on support from key ministries such as health, agriculture, education, sports, and recreation. Building consensus among these agencies will result in consistent messages regarding diet, health, and lifestyle for the public. Examples of support include technical support for data analysis or a Secretariat to maintain and coordinate activities

2. Form a working group of experts
The working group should include diverse expertise in areas such as public health, nutrition, food science, agriculture, and behavioral sciences

3. Solicit public comment and input
The expert panel needs to gather and evaluate scientific information to determine the guidelines that are most relevant to the target population. This information can be obtained from the scientific literature. In addition, professional groups may have important information to submit to the panel for consideration. Solicitation of information is consistent with an open process; however, the panel is responsible for evaluating the relevance of the information submitted

4. Review and identify key public health issues and evaluate the diet–health relationships of concern for the population, determine the critical health, food, and nutrition issues to be targeted in the FBDG, and define the purpose, target groups, and content of the FBDG
Even if data are limited, it is important for the working group to identify the key public health issues. This step may be especially important in countries in which both under-nutrition and over-nutrition are of concern. Identification of the public health issues allows the working group to address the questions in **Table 1**

5. Develop and draft the main messages for the FBDG
The working group will need to decide whether the draft document will be targeted primarily at health professionals, and hence may be more technical, or will be targeted toward the general public. In developing the main messages, they may identify consumer-orientated materials, such as a food guide that will be useful in communicating the FBDG to the public

6. Assess the cultural and economic appropriateness and credibility of the messages as perceived by the target groups
Through focus groups or other types of consumer testing the effectiveness of the FBDG can be assessed. This information can be used to revise the guidelines before developing the final draft

7. Release and implement the FBDG
It is valuable to have government leaders from key ministries involved in the release and implementation of the FBDG so that there is a commitment to integrate the guidelines into departmental policies. In addition, the implementation can require development of educational materials for different target groups as well as public–private partnerships to aid in dissemination of the messages to the public

8. Monitoring and revision
Monitoring can be used to assess the impact and implementation of the FBDG. In addition, monitoring data are useful for making appropriate revisions and updates to the guidelines on a periodic basis

FBDG, food-based dietary guidelines.

Table 3 Dietary-guideline messages from three countries

USA	*China*	*Thailand*
Aim for fitness		
Aim for a healthy weight Be physically active each day	Eat a variety of foods, with cereals as the staple Consume plenty of vegetables, fruits, and tubers	Eat a variety of foods from each of the five food groups, and maintain proper body weight Eat adequate amount of rice or alternative carbohydrate sources
Build a healthy base		
Let the pyramid guide your food choices Choose a variety of grains daily, especially whole grains Choose a variety of fruits and vegetables daily Keep food safe to eat	Consume milk, beans, or diary or bean products every day Consume appropriate amounts of fish, poultry, eggs, and lean meat; reduce fatty meat and animal fat in the diet	Eat plenty of vegetables and fruits regularly Eat fish, lean meats, eggs, legumes, and pulses regularly
Choose sensibly		
Choose a diet low in saturated fat and cholesterol and moderate in total fat Choose beverages and foods to moderate your intake of sugars If you drink alcoholic beverages, do so in moderation	Balance food intake with physical activity to maintain a healthy body weight Choose a light diet that is also low in salt If you drink alcoholic beverages, do so in limited amounts Avoid unsanitary and spoiled foods	Drink milk in appropriate quality and quantity for one's age Eat a diet containing the appropriate amounts of fat Avoid sweet and salty foods Eat clean and safe food Avoid or reduce the consumption of alcoholic beverages

Table 4 Common themes for food-based dietary guidelines

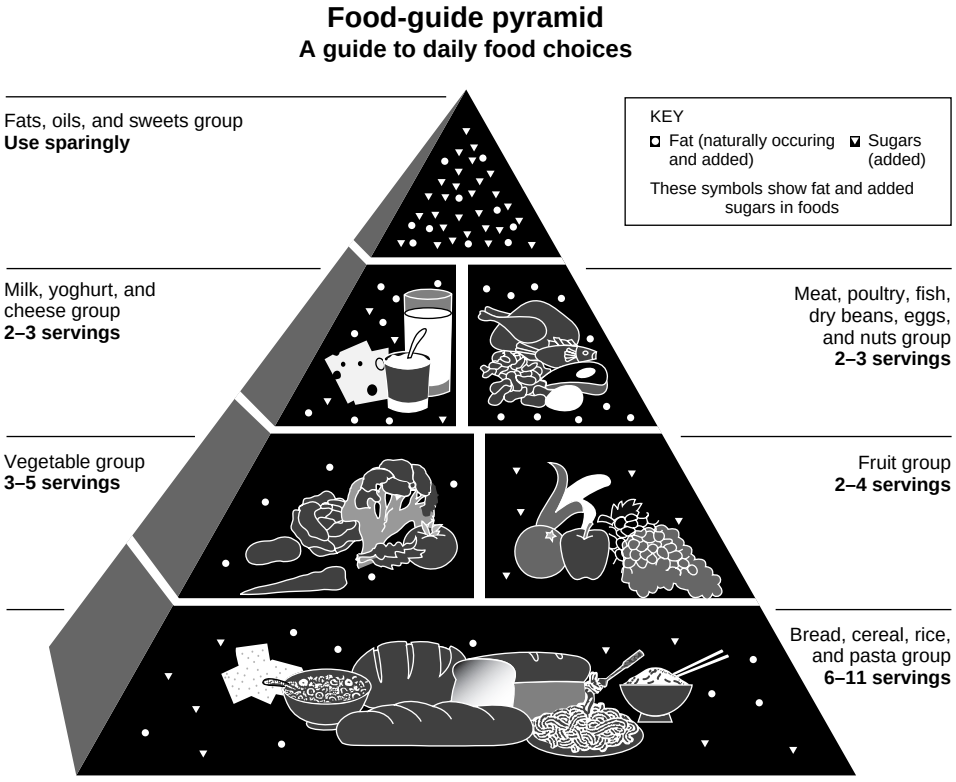
Foods or behaviors that are encouraged	*Cautionary messages*
Energy balance Includes physical activity Encouraging a healthful variety of foods Fruits and vegetables Use of whole grains Protein-based foods Foods that are calcium sources Sources of unsaturated fatty acids Safe food handling	Saturated fatty acids and *trans* fatty acids Energy balance Total energy from fat Consumption of foods high in added sugar Use of salt and salty foods Alcohol

all countries concerns about the increasing incidence of obesity have placed greater focus on energy balance, in terms of both food selection and physical activity. As a part of their effort to support the development of FBDG, the FAO launched a public information initiative for consumers entitled 'get the best from your food'. This initiative promoted four simple principles (**Table 5**) that can be adapted for educational programs in a variety of settings.

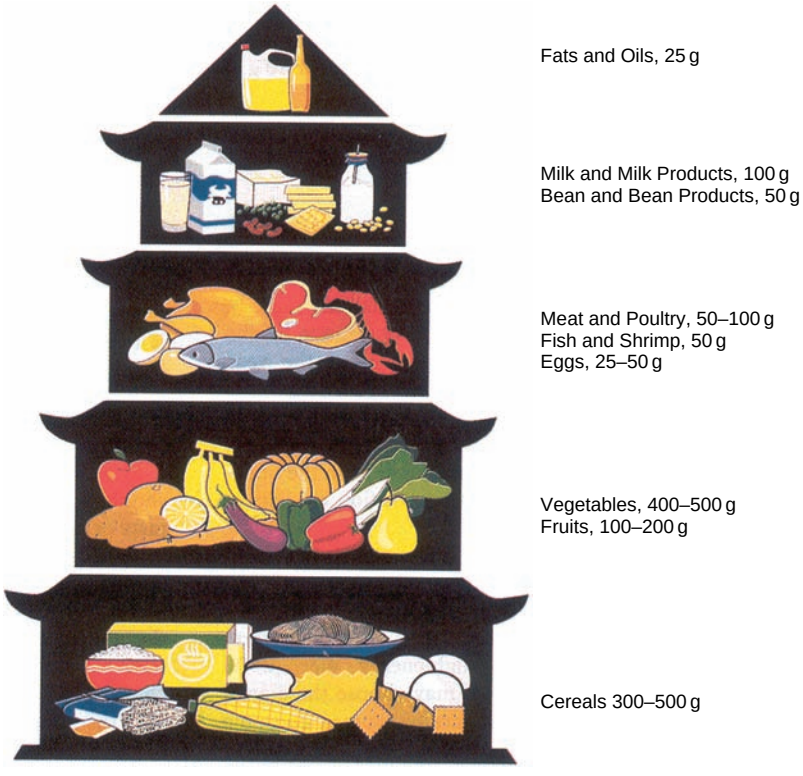
Most countries that have developed FBDG have also developed a food guide to accompany the messages in the guidelines. The food guide is typically a simple graphic illustration of food choices and dietary patterns. The food guides that accompany the FBDG shown in **Table 4** are illustrated in **Figure 1**. Criteria for a food guide should include representation of foods common to the population, consistency with the FBDG, use of simple graphics that are meaningful to the target population, and developing a food pattern that meets the nutrient requirements of the population. Although a simple graphic is useful for visual communication, it should be clear that proper use of the food guide depends on understanding the more complete information in the FBDG.

Table 5 Food and Agriculture Organization initiative: get the best from your food

	Key concept
Enjoy a variety of food	Recognizing the importance of food in understanding nutrient requirements, nutrient and non-nutrient interactions, and diet–health relationships
Eat to meet your needs	Importance of energy balance and different needs across the life cycle
Protect the quality and safety of your food	Recognizing the importance of food and water sanitation, especially in developing countries
Keep active and stay fit	Importance of physical activity in maintaining well-being



(A)



(B)

Figure 1 The food guides that accompany the FBDG in (A) the USA, (B) China, and (C) Thailand.

(C)

Figure 1 *Continued.*

Conclusion

The use of food-based dietary guidelines (FBDG) has emerged as an important food and nutrition policy and education program since the late 1970s and is valuable for addressing issues of both nutrient adequacy and excess. Importantly, FBDG communicate nutritional principles in a manner that is relevant to the population. As a policy document, they should be revised periodically so that the information reflects current science on food and nutrition factors that promote health and prevent disease. Additionally, it is important for each country to develop their own set of FBDG so that the recommendations and presentation are relevant to the local population.

See also: **Nutritional Surveillance**: Developed Countries; Developing Countries. **World Health Organization**.

Further Reading

Centers for Disease Control (1999) Achievements in public health, 1900–1999: safer and healthier foods. *MMWR Weekly* **48**: 905–913.

Chinese Nutrition Society (1999) Dietary guidelines and the food guide pagoda for Chinese residents: balanced diet, rational nutrition, and health promotion. *Nutrition Today* **34**: 106–115.

Dietary Guidelines Advisory Committee (2000) *2000 Report*. Washington, DC: US Government Printing Office.

Food and Agriculture Organization *Get the Best From Your Food*. Rome: FAO.

Food and Agriculture Organization and World Health Organization (1998) *Preparation and Use of Food-Based Dietary Guidelines*. WHO Technical Report Series 880. Geneva: WHO.

National Research Council (1989) *Diet and Health: Implications for Reducing Chronic Disease*. Washington, DC: National Academy of Sciences Press.

Schneeman BO (2001) Preparation and use of food-based dietary guidelines: lessons from Thailand and the Philippine Islands. *Food, Nutrition and Agriculture* **28**: 55–62.

Schneeman BO and Mendelson R (2002) Dietary guidelines: past experiences and new approaches. *Journal of the American Dietetic Association* **102**: 1498–1500.

Surgeon General (1988) *Surgeon General's Report on Diet and Health* Washington, DC: United States Government Printing Office.

US Department of Agriculture and Department of Health and Human Services (2000) *Nutrition and Your Health: Dietary Guidelines for Americans*, 5th edn. Home and Garden Bulletin 232. Washington, DC: US Government Printing Office.

Welsh S, Davis C, and Shaw A (1992) A brief history of food guides in the United States. *Nutrition Today* Nov/Dec 1992: 6–11.

Working Group on Food-Based Dietary Guideline for Thai *Food Based Dietary Guideline for Thai*. Institute of Nutrition, Mahidol University, and Food and Agriculture Organization.

World Health Organization and Food and Agriculture Organization Expert Consultation (2003) *Diet, Nutrition and the Prevention of Chronic Disease*. WHO Technical Report Series 916. Geneva: WHO.

DIETARY INTAKE MEASUREMENT

Contents
Methodology
Validation

Methodology

A A Welch, University of Cambridge, Cambridge, UK

Introduction

Dietary intake measurements are used to assess food or nutrient intake of individuals, groups, or populations. The purpose of collection of measurements varies from individual assessments in clinical situations (nutrition screening) or the adequacy of intake of population groups (nutrition surveillance) to use in research relating diet to health status, particularly in epidemiology. Measurements are also used to establish exposure to food-borne contaminants, in the evaluation of nutritional intervention programs, and to develop nutritional guidelines for governmental health policy.

This article describes the dietary intake measurements available, issues associated with data collection, conversion to nutrients and food types, measurement error when using dietary intake methods, validation and calibration of dietary methods, and future developments.

Dietary Intake Measurements

Table 1 describes the advantages and limitations of the main types of dietary methods, which are suitable for different purposes.

Of the individual methods weighed records, estimated food records, 24-h recalls (24-h), and dietary histories are more intensive. The quantity of food consumed may be weighed directly or estimated using household measures such as cups and spoons, photographs, standard units, or average portions (see **Table 2**). For all methods the amount consumed can be measured or described either including or excluding wastage material usually discarded during food preparation, e.g., outer leaves and peel from vegetables or bones from cuts of meat.

Some considerations when choosing a dietary method are shown in **Table 3**.

Methods for Measuring Food Consumption at the National Level

Food Balance Sheets

The FAO (Food and Agriculture Organization) publishes food balance sheets (FBSs) for around 200

Table 1 Names and characteristics of dietary methods used for estimating food and nutrient intake

Name of method	*Advantages*	*Limitations*
National level		
Food balance sheets	Available for 200 countries; suitable for monitoring change	Per caput not individual intake; intake overestimated as nutrient losses during storage and preparation not accounted for; should not be used to provide estimates of nutritional adequacy of particular regions
Household level		
Food account method	Low respondent burden; relatively inexpensive	No estimates of change in larder stocks; measurements confined to food brought into the home (unless method modified to measure food consumed outside the home, which can be quite large); consumption of confectionery, alcoholic, and soft drinks excluded
Inventory method	Low respondent burden; relatively inexpensive	Consumption of confectionery, alcoholic, and soft drinks excluded
Household record	Suitable for populations with high proportion of home-made foods; useful if literacy levels are low; provides direct measure of food available for consumption	High input from field workers or interviewers
List recall methods	Relatively rapid and inexpensive; only 1 interview required; suitable for populations with higher proportion of purchased than home-produced food	Advance warning of interview may distort food consumption patterns; subject may fail to record items from memory; no record of foods eaten outside the home
Individual level		
Retrospective methods		
24-Hour recall (24-HR) (single or multiple days)	If interviewed, respondent literacy not important; not reliant on long-term memory; providing not forewarned, individuals do not alter food consumption; interview length 20–45 min	Single 24-h should not be used for estimating intake of individuals but can be used for group assessments
Diet history	Respondent literacy not required	Report of past intake is influenced by current diet; trained interviewers required; average interview length 1–1.5 h; high processing costs
Food frequency questionnaire (FFQ) (if includes portion estimates termed semiquantitative FFQ)	Useful for large numbers; relatively straightforward to complete; administration simpler and less costly than other individual methods; more rapid data processing	Needs to be developed for specific population group to ensure important food items are covered and requires updating to accommodate changes to supply of foods; less flexible for later analysis as food lists are fixed; responses governed by cognitive, numeric, and literacy abilities of respondents also by length and complexity of the food list
Current methods		
Weighed food record (weighed inventory technique)	No requirement for memory retrieval as it records current intake; food intake weighed so estimates of quantity consumed not required	Literate, cooperative respondents required as burden is high; possible that respondents change usual eating patterns to simplify the record; high processing costs
Food record with estimated weights	No requirement for memory retrieval as it records current intake	Literate, cooperative respondents required as burden is high; possible that respondents change usual eating patterns to simplify the record; necessary to find values for estimates of quantity of food consumed; high processing costs
Duplicate analysis	Greater accuracy	Highly labor intensive; requires laboratory to do food composition analysis; limited applicability in population studies

In all methods 'foods' refers to consumption of foods, beverages, and snacks both inside and outside of the home.

countries. Food balance sheets present a comprehensive picture of the pattern of a country's food supply during a specified reference period. Food balance sheets may also be termed national food accounts, food moving into consumption, food consumption statistics, food disappearance data, and consumption level estimates, reflecting differences in the method of calculation but providing similar information.

Table 2 Types of portion used for methods using estimated portions

Portion types
Average or small, medium, large portions, weights – available from studies of weighed intake
Photographs (ideally should be 5 or more representing the population range of intake)
Household measures (spoons, cups, mugs, liquid measures)
Standard units (1 apple, 1 banana)
Food models/replicas (three-dimensional models representing foods)

Data should be derived from weighed intakes, government surveys, and research groups in populations similar to the one to be studied.

The supply available during a period is calculated from the total quantity of foodstuffs produced in a country, added to the total quantity imported and modified for any change in stocks that may have occurred. Calculation of quantities used for purposes other than human consumption (exports, livestock, used for seed, nonfood uses) and losses during storage and transportation are made. The per caput supply of each food item available for human consumption is calculated by dividing the total of available food by the number of the population actually consuming it and expressed in terms of quantity and nutrients. Estimates from FBSs include household wastage material, plate waste, and food fed to pets. Nutrient losses during storage, preparation, and cooking are not calculated and so figures for available food are greater than those reported by individual dietary surveys.

Table 3 Factors determining choice or suitability of method

Size and scale of the data collection
Screening, clinical, research, surveillance purposes?
Literacy or numeracy of the population
Age of the individual or population (the very young or very old may need assistance with completion)
Intended or potential use of the data (immediate short-term assessment versus prospective research)
Requirement for group or individual estimates for nutrient intake
Requirements from the data for nutrients, food groups, or phytonutrients
Detail and comprehensiveness of the information to be extracted for analysis (if information only required for particular nutrient or food type, shortened questionnaires may be administered)
Has repeatability of the method been assessed?
Have previous validation studies performed on the method by other researchers in similar population group to be studied?
Availability of resources for interviewers and including training
Availability of suitable coding program (record and recall methods require greater resources than frequency methods but frequency programs are more complex to develop)

Food balance sheets can be used to formulate agricultural policies concerned with production, distribution, and consumption of foods and as a basis for monitoring changes and forecasting food consumption patterns, as well as to provide intercountry comparisons of available supplies.

Methods for Estimating Dietary Intake at the Household Level: Household Budget Surveys

Techniques for estimating intake at the household level include the food account method, the inventory method, the household record, and the list recall method. These methods measure all foods and beverages available for consumption by a household or family group during a specified time period of between 1 and 4 weeks, although some last for 2–3 months. Wastage factors are sometimes applied. Household surveys provide data for per capita consumption of foods or nutrients, not intake for specific individuals. Data are calculated irrespective of the age and gender distribution in the household. These methods provide population data for annual mean food consumption and selection patterns, and are used for analyzing trends in intake. Household budget surveys are used more widely in Europe than elsewhere. As countries may not produce compatible data the Data Food Networking Project (DAFNE) has developed the methodology to allow the data from 11 European countries to be combined and compared.

Food Account Method

A record is made by a respondent of details of all quantities of food entering the household (purchased, home grown, or received over a period), usually over a period of 7 days. Changes in larder stocks are not estimated as on average some households will gain and some will use up stocks. Estimates of losses and wastage during preparation are made. This method is used for the UK Expenditure and Food Survey (until 2001 the National Food Survey), and has included consumption of food, confectionery, soft drinks, and alcohol outside the home since 1992. As consumption outside the home now accounts for a substantial proportion of dietary intake in the UK the method was modified in 2001 to include the use of till receipts and individual 2-week diaries for each household member aged 7 years or older. This method can be used to measure seasonal variation in intake over 1 year.

Inventory Method

The inventory method is similar to the food account method and respondents record all foods coming into the household. A wastage factor is often applied and a larder inventory is included at the beginning and end of the survey period.

Household Record

Foods available for consumption (either raw or processed) are weighed or estimated. Foods for each meal are recorded separately to give a total for the household. Waste is measured directly or estimated. Interviewers visit the household early in the day to determine the quantity of food used to prepare the first meal and the number of individuals who consumed it. The midday meal may be weighed or recorded using estimated measures. A further interview is required later in the day. This method is appropriate for use in pre-industrial societies where literacy is low and units for buying foods not standardized.

List Recall Methods

The respondent is asked by a trained interviewer to recall the amount and cost of food obtained for household use over a period, usually of 1 week. The method takes into account food use, purchases, and acquired food, but not waste. Quantities consumed are weighed or estimated using household measures. The interview can take up to 2.5 h. Response rates are usually high. Information on the age and sex of people in the household and the number of meals eaten both in and outside the home, income, and other socioeconomic characteristics may be collected. It is helpful to notify the respondent in advance so that records of purchases can be kept prior to the interview. This method was used by the United States Department of Agriculture (USDA) National Food Consumption Survey between 1931 and 1988.

Individual Dietary Intake Methods

Many methods are available for estimating individual dietary intake measures and can be divided into two types: retrospective measures of intake such 24-h recalls (24-HR), dietary history or food frequency questionnaires (FFQs), or current measures of intake such as weighed or estimated food records. Qualitative information is available from all methods but quantitative estimates for nutrient consumption are possible only if data for weighed or estimated portion weights are available. Most methods may be either self-completed or completed by a surrogate. Surrogates may be required if study individuals are too young, old, or infirm but data will be less reliable than when reported directly.

24-HR and FFQs may be self completed or interview administered either face-to-face or by telephone and can be mailed. Data collection costs can be reduced if questionnaires can be self-completed or mailed.

The number of days of report required for adequate measures of nutrients using 24-h recalls, weighed, or estimated records varies depending on the day to day variability of nutrient consumption. The number of days is partly dependent on the variation in nutrient concentration in foodstuffs. The concentration of macronutrients such as protein and carbohydrate in foods varies less than micronutrients such as vitamin C or iron. The number of days required to classify individuals into the correct third of the percentage distribution for usual intake, for 80% of individuals, has been calculated in British and Swedish populations. Up to 7 days of recall would be required for energy, protein, sugars, and calcium. Nutrients with greater variability and requiring between 4 and 14 days of records were alcohol, vitamin C, riboflavin, and iron.

24-Hour Recalls

24-HR determine intake during the preceding 24 h. Interviews can be recorded on paper or using interactive computerized software. Day-to-day variability in nutrient intake is large and a single day will not categorize individuals correctly within a distribution of intake. Therefore, single 24-HR are better used for group assessments than estimates for individuals. However, multiple 24-HR can be used to overcome this problem. The sampling protocol for studies should include an equal proportion of all days of the week and coverage of all four seasons.

Diet History

The diet history consists either of an interview administered 24-HR or establishing usual eating pattern over a 1-week period, followed by a frequency questionnaire to provide additional information. The dietary history provides a representative pattern of usual intake and is interview administered only.

Food Frequency Questionnaires

FFQs consist of a list of specific foods or food types associated with frequency of consumption. They are termed semi-quantitative if portions are included. Most questionnaires specify a frequency response in relation to an average or medium portion but some request records of specific portions. The period

of record is usually the previous month or year. FFQs provide an indication of usual intake and can be used to obtain population estimates of frequency of consumption of food types.

FFQs need to be developed for specific population groups otherwise important foods may be missed. FFQs may become outdated if the supply of foodstuffs changes. FFQs consist of a fixed food list, which may be a disadvantage for prospective studies as hypotheses to be tested are limited by the list. Factors that affect the response to FFQs are the literacy and numeracy of respondents as some mathematical ability is necessary to calculate relative frequencies, the length and complexity of the food list, and the influence of current diet. Not all respondents will relate frequency to portion size accurately.

In the US examples of FFQs are the Block and Willett questionnaires. In Europe FFQs were developed for the European Prospective Investigations into Cancer and Nutrition (EPIC) study in the Netherlands, Germany, Greece, Italy, Denmark, France, and the UK.

Weighed Food Record Inventory and Estimated Food Record

For weighed food records (WRs) all food consumed over a period is weighed and recorded with details of food type and method of preparation, on preprinted forms or booklets, to obtain consumption over a period of days. Portable scales need to be supplied. WRs may include some estimated items eaten out of the home. Leftover food should be weighed and deducted. The recommended time period for records is 4–7 days or more, although the number of days depends on the nutrient of interest, study population, and objectives of the study. As some populations have different eating habits at weekends, weekend days should be included proportionately.

For estimated food records all foods consumed over a period are recorded with details of food type, method of preparation, and estimated portions over a period of days (see **Table 2**). If recorded over 7 days, this may be called a '7-day diary'.

Both these methods have a high respondent burden and need cooperative, literate respondents. Respondents require training in the level of detail needed to describe foods. It is also possible that respondents may change usual eating patterns to simplify the process of the record. It is also beneficial to include a review of weighed records during the period of recording either after the first day or at the end.

Duplicate sample technique

Duplicate samples of all foods consumed are made and the nutrient content analyzed. This method is used for metabolic studies and though providing greater accuracy than other methods, its use is not feasible for most purposes.

Further Information

Although nomenclature for dietary methodology is reasonably consistent, care should be taken when reading the literature as methods with the same name may have been applied differently. The final decision over which method to choose will depend on the aims of the study, the population for study, the potential burden on respondents, and the resources available. Household surveys and food balance sheets provide data for per caput but not individual intake. In general, individual and the more intensive methods are associated with higher costs and respondent burden, whereas household methods are more economical and have a lower respondent burden.

Clinical Practice

Dietary methodology for clinical practice requires rapid assessments of nutritional intake in order to prescribe dietary change or to improve nutritional status. Traditionally, 24HR of 'usual' intake or diet histories have been used for this purpose. Food frequency questionnaires and weighed or estimated food records are not generally used due to the more intensive burden on respondents and on the resources required for coding and processing the data.

There is considerable discussion over the optimum method to use for establishing individual dietary intake and studies designed to measure the validity of methods suggest that those that are more intensive and detailed lead to greater measurement precision, justifying the greater cost. Confirmation of these findings is required. Despite these potential benefits if resources are unavailable less intensive methods tend to be used.

Factors Affecting Individual Ability to Report Intake Accurately

Factors governing individual accuracy and quality of reports are respondent's literacy and numeracy skills; preconceived ideas on the purpose of the enquiry and, for list-based methods, the interpretation and meaning of food names. Individuals may make errors when measuring and recording food weights or estimating weights of foods consumed.

There is also respondent variation in the perception of the size of portions represented by photographs.

Interviewers

The aim of using interviewers with dietary methods is to obtain a complete, accurate, and detailed record of what respondents eat. Therefore, it is important for interviewers to be well trained and have an awareness of food composition and preparation techniques. Ideally, interviewers should be educated in nutrition (dietitians or nutritionists), although nonnutritionists can be trained to standardized techniques, and come from the same cultural or ethnic background as the study population. Interviewer protocols should be developed.

Computerized Interview Procedures

Computerized interview systems can aid interviewers by prompting for specific questions to elicit sufficient and specific detail and reduce the burden on interviewers. Examples are the Minnesota Nutrition Data System and the EPIC-SOFT systems, used in the US and a number of European countries. Although computerized interviews have advantages in improving accuracy and standardization, and in saving time and effort when recording and coding data, interviewers do have to be competent with computers and the resources required to develop systems are high.

Using Dietary Methods in Different Populations

Ethnic subpopulations may consume different food types than a main population and baseline surveys will be required to establish what types of foods and method of preparation are common. This information would be required before list-based methods such as the FFQ could be developed.

Recall of Remote Diet

Investigators may wish to recall diet in the remote past, perhaps of many years. However, interpretation of remotely recalled dietary data is complex as recalled diet is heavily influenced by current dietary habit. Some studies have found that the correlations between recalled past diet and current diet were higher than the correlations between actual past diet and recall of past diet. The onset of diseases such as cancer may affect the appetite and dietary intake of study participants and as recall of remote diet is strongly related to current diet, may affect recall of remote diet. As diet prior to the onset of disease is the measure of interest, it is preferable to collect dietary information prospectively, that is before disease onset. Case–control studies in which the diet of cases with disease is compared with controls may be affected by altered perception of recalled diet, particularly by cases.

Reproducibility of Dietary Methods

The reproducibility of a method may also be referred to as reliability, repeatability, or precision and is a measure of the extent to which the same results can be obtained when repeated under the same conditions. Repeated measures provide an estimate of the within-person variability of intake. However, interpretation of the repeatability of measures is difficult as a lack of consistency may be due to genuine change over a time period or a lack of sensitivity or specificity of the method used to measure intake.

Use of Data and Conversion of Reported Intake to Nutrients and Food Types

Qualitative Analysis

Dietary method data can be used qualitatively, for instance during the process of reviewing nutritional intake for the purpose of dietary treatment as in clinical practice. Data on frequency of consumption may also be collected and analyzed by the FFQ method without conversion to nutrient intakes. However, even for qualitative analyses it is likely that paper-based dietary methods will require conversion to an electronic format. The majority of uses of dietary methods are targeted towards quantitative analyses.

Quantitative Analysis

The data collected by dietary methods are converted into food and nutrient consumption by calculating the amount of food eaten and linking this to a database with values for the nutrient composition of foods.

The databases of nutrient composition of foods are provided by the governments of many countries. They consist of nutrient composition data for the average composition of commonly consumed foodstuffs and are usually available as printed publications, computerized databases, or as part of software packages. Values in nutrient composition databases are expressed as either per 100 g of food or per common household measure. Nutrient databases vary in the coverage and comprehensiveness of the foods and nutrients. They are revised periodically to cover newer foods of different nutrient compositions or to modify or extend the nutrient coverage. Some issues concerning the choice of nutrient databases are shown in **Table 4**. It is important to read the

Table 4 Factors to consider when choosing a nutrient database to calculate nutrient intakes

Comprehensiveness of food item and beverage coverage?
Does the database contain entries for important foods consumed by the population to be studied?
How comprehensive is the coverage of nutrients?
Does the database contain data for mixed or multiple ingredient recipes or dishes?
What analytical techniques were used to derive nutrients in the database? (There can be differences in nutrients measured by different techniques)
Are the data officially evaluated?
What compilation methods were used to construct the database?
Which conversion factors are used to calculate metabolizable energy content of foods for protein, fat, carbohydrate, and alcohol?
What proportion of missing values exists within the database? (Missing values are counted as zero in calculations, resulting in systematic underestimates of intake.)
For international studies or comparisons how do the analytical methods for determining nutrient composition and compilation techniques affect the resulting data?

information distributed with the printed or electronic versions of databases to determine the uses and limitations of the data.

Several steps are involved in calculating nutrient intake (also known as coding or processing). The first is to choose an item in the database, which corresponds most closely with the food consumed. If the food consumed is not in the database a suitable alternative can be chosen by considering food type, general characteristics, and likely nutrient profile. Once the food has been chosen the nutrient composition of the food quoted in the database is multiplied by the amount of food eaten, e.g., for 60 g food the nutrients would be multiplied by 0.6 (where nutrients are expressed per 100 g of food).

To calculate daily intake for an individual the contribution of each food is calculated and all the foods for a day summated. If more than one day's data have been collected it is usual to calculate the average of the number of days recorded. Data from FFQs are usually computed to consumption per day but can also be computed per week.

Although it is possible to compute intake by hand, using a calculator and a printed copy of a nutrient database, this is very labor intensive and in practice for most purposes has been superseded by computerization.

Data Processing and Computing Dietary Intake

The same care as that invested in data collection should be applied to data processing as errors of great magnitude may be introduced.

Estimated Food Quantities

To obtain quantitative information for nutrients or food groups, actual or estimated food weights are used. For methods using estimated food weights, values also need to be found for foods described such as standard units, average portions, or household measures. Sources of data are national publications, surveys of weighed dietary intakes, and food manufacturers. Data may also be included in nutrient calculation programs. Portion weights need to be population specific and, if unavailable, studies to establish values will be needed. Intensive methods used for large-scale surveys will require databases of more than 20 000 values for portion weights.

Data Entry and Nutrient Calculation Systems

A number of computerized data entry systems and nutrient calculation programs exist; factors that need to be considered when choosing a system are given in **Table 5**. The features required depend on the intended use of the data but as a minimum should include a list of foods, weights of portions, and a nutrient composition database. Ideally, systems should enable entry of data in sufficient detail to fulfil hypotheses for investigation and include measures to ensure consistent entry by staff such as defaults for inadequately reported foods, portions, or mixed component foods. They should also include a method for entering newer foods with different nutrient composition from the existing nutrient database. This is particularly important, as the range of new foodstuffs and products with different nutritional characteristics is ever increasing.

Computerized systems and nutrient databases become out-dated and for large-scale prospective studies it is desirable to develop systems with a flexible approach to updating by using database technology.

Data Processing Errors

Errors arising during the coding (data entry) and processing of individual dietary methods (24-HR,

Table 5 Factors to consider when choosing a computerized entry or interviewing program

Speed of the assessment
Requirements of the study for detailed or general data
Food composition database used
Food portion database used
Cost of the system
Facilities for organization of data
Ability to extract nutrients or food groups from the system
How up to date are the nutrient composition databases included in the system?
Commercial availability

diet history, weighed and estimated records) need to be avoided. Misclassification can arise due to human error if incorrect foods are chosen during coding, for instance, if milk was consumed in the full-fat form but was coded for skimmed milk. This may also arise where a food has local or alternative food names, which may be unknown to the coder. It is important to have a qualified nutritionist available to develop a protocol for training staff, answering queries, and dealing with ambiguities. Coders should have knowledge of food composition and food preparation techniques. It is difficult to control entry of incorrectly matched foods but careful checking and staff training are crucial in preventing this. Other potential errors are entry of incorrect quantities or multiplication factors for portion weights and missed items, problems that can occur even with structured computer programs. So, systematic post-entry checks to identify extremes of portion weights or nutrient values and the verification and correction of data are necessary.

Issues Associated with Measurement of Dietary Intake

Measurement Error

There is potential for the occurrence of measurement error with the measurement of any exposure such as when using dietary methods to measure nutritional intake. Errors may arise as a result of flaws in the design of the measurement instrument or during data collection or processing. Measurement error may also occur as a result of individual characteristics of participants in studies. Measurement error can be defined as the difference between the measured exposure (or measure of dietary intake) and the true exposure. All measurement of dietary exposures is subject to some degree of measurement error making it difficult to achieve measurements of true intake.

Efforts to reduce measurement error during data collection and processing should be introduced into the protocol of all studies, however, even if preventative measures are taken it is impossible to eliminate it altogether. It is difficult to identify the type and structure of measurement error associated with dietary intake. Measurement error may occur because of inaccurate reporting by respondents. It may also vary according to dietary method, for instance, food items within record methods may be intentionally or unintentionally omitted and with FFQs frequency of consumption may be inaccurately reported. Systematic bias, interviewer bias, recall bias, and social desirability bias have been identified but there are likely to be other sources of error. (Bias can be defined as the modification of a method of measurement by a factor, which influences the measurement in one or more directions.) Measurement error associated with dietary methods may consist of one or more types of error.

Measurement Error in Data Collection and Processing

Systematic bias Systematic bias is a systematic mismeasurement of data and can occur, for instance, if equipment such as weighing scales under- or overestimates values or if an interviewer consistently fails to use questions to probe for consumption of snacks and additional foods. If systematic bias can be identified solutions can be found, for instance, by calibrating equipment or training and monitoring interviewers.

Interviewer bias The behavior of an interviewer can influence the response of interviewees leading to interviewer bias. The degree of rapport between interviewer and respondent also influences results. Bias may occur if interviewers omit responses or record them incorrectly. Trained interviewers should ask open-ended questions in a neutral or nonleading manner, and not imply that a food or beverage should or should not have been consumed and avoid value judgments.

Social desirability bias Social desirability bias can influence dietary measures as respondents strive to report what they think is required not what was actually consumed, for example, reporting less alcohol consumption than is the case or greater consumption of foods with perceived health benefits such as fish, fruit, or vegetables. This is likely to be the cause of mis-reporting, under-reporting, or low energy reporting, which occurs in certain respondents. It is possible to predict how much energy a respondent should report, as this is the amount required to maintain a stable weight. (Weight will be either gained or lost if more or less energy is consumed than required.) As energy intake should equate to energy expenditure, expenditure effectively measures intake. Techniques for measurement of energy expenditure such as whole body calorimetry and doubly labeled water can be used. Using these techniques those individuals classified as low energy reporters are likely to be older, more overweight, and of lower educational and socioeconomic status than the rest of the population. Low energy reporters tend to have lower consumption of foods in the groups cookies, cakes, puddings, confectionery (candy) and

sugary foods and, in some populations, lower consumption of spreads, cooking fats, and potato chips. Interviewers should be aware of low energy reporting, aim to be entirely nonjudgemental, and also request participants make complete records of food intake.

Impact of Measurement Error

As the proportion of error within a measurement increases, the accuracy of the measurement decreases and the results using the measurement will become less interpretable. Hence, greater measurement error reduces the likelihood that the truth has been measured with accuracy and increases the likelihood that analyses relating diet to disease status will tend towards null results. The effect of measurement error is to mis-classify an individual within a range of intake.

Validation of Dietary Methods

Validation is used to quantify the measurement error that occurs when measuring dietary intake exposures. It requires two measures: a main measurement and a second measurement subject to less measurement error than the first. The errors of the two measurements should be independent. Validation is used to estimate the proportion of measurement error within the main method by modeling the differences between the main and the secondary measurement. It had been considered that dietary methods had errors independent of each other and that record methods such as 24-HR could be used, but it is now known that the errors are not independent as individuals report in the same way with different methods. Therefore, it is better to use biological variables measurable in blood or urine (also known as biomarkers) as the second measure for dietary validation. Examples are vitamins, minerals, and individual fatty acids in blood such as vitamin C and carotenoids or urinary excretion over 24-h of nitrogen, potassium, and sodium. Examples of validation studies are those performed within EPIC-Europe and the Observing Protein and Energy Nutrition Study (OPEN) in the US. Work is ongoing to extend the number of biomarkers available and to define further and elicit the structure of measurement error.

Use of Calibration Methods to Adjust for Measurement Error

In contrast to validation, which attempts to identify the type and scale of measurement error, calibration is designed to adjust for systematic over- or underestimation in dietary intakes within populations. It may also be used at the individual level to attempt to correct for attenuation bias (or dilution) in relative risk due to errors in dietary measurements. Calibration of data has been proposed for large multicentre nutritional studies that have used different dietary methods to capture population-specific diets. Calibration studies require a highly standardized second dietary measure to be used in a representative sub-sample from each cohort to form a common reference measurement across populations. An example of this approach has been used by the European EPIC (European Prospective Investigations into Cancer and Nutrition) Study using a computerized, standardized 24-h recall in ten countries.

Future Developments

Future developments in methodology are likely to use computing, digital, and Internet technology, such as videos of food eaten and online programs for self-reported intake. Use of Dictaphones and combinations of weighing and other recording equipment are also possible.

The number of foodstuffs available, particularly of manufactured foods and ready-made meals, will continue to increase, presenting challenges for those attempting to estimate nutrient intake. Nutrient databases will continue to be expanded and updated to incorporate newer food items and nutrient measurements available using improved analytical techniques.

In some populations more than 40% of individuals have been shown to consume supplements and as very few comprehensive databases of vitamin and mineral supplements exist these need to be developed, as supplements can make a major contribution to nutrient intakes.

See also: **Dietary Intake Measurement**: Validation. **Dietary Surveys**. **Food Composition Data**.

Further Reading

Bingham SA (1987) The dietary assessment of individuals; methods, accuracy, new techniques and recommendations. *Nutrition Abstracts and Reviews (Series A)* **57**: 705–742.

Cade J, Thompson R, Burley V, and Warm D (2002) Development, validation and utilisation of food-frequency questionnaires – a review. *Public Health Nutrition* **5**(4): 567–587.

Cameron ME and Van Staveren WA (1988) *Manual on Methodology for Food Consumption Studies*. Oxford: Oxford University Press.

Feskanich D, Sielaff BH, Chong K, and Buzzard IM (1989) Computerized collection and analysis of dietary intake information. *Computer Methods and Programs in Biomedicine* **30**: 47–57.

Freudenheim J (2003) Biomarkers of nutritional exposure and nutritional status. *The Journal of Nutrition* **133** no. 873S–874S.

Gibson RS (1990) *Principles of Nutritional Assessment*. Oxford: Oxford University Press.

Margetts BM and Nelson M (1997) *Design Concepts in Nutritional Epidemiology*, 2nd edn. Oxford: Oxford University Press.

Pao EM and Cypel YS (1996) Estimation of dietary intake. In: Zeigler EH and Filer LJ (eds.) *Present Knowledge in Nutrition*, 7th edn, pp. 498–507. Washington, DC: ILSI Press.

Riboli E, Hunt KJ, Slimani N *et al.* (2002) European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutrition* 5(6B): 1113–1124.

Slimani N, Deharveng G, Charrondiere RU *et al.* (1999) Structure of the standardized computerized 24-h diet recall interview used as reference method in the 22 centers participating in the EPIC project. Prospective Investigation into Cancer and Nutrition. *Computer Methods and Programs in Biomedicine* 58: 251–266.

Smith AF (1993) Cognitive psychological issues of relevance to the validity of reports of dietary intake by college men and women. *European Journal of Clinical Nutrition* **47**(supplement 2): S6–S18.

Thompson FE and Byers T (1994) Dietary assessment resource manual. *The Journal of Nutrition* **124** no. 11S.

Welch AA, McTaggart A, Mulligan AA *et al.* (2001) DINER (Data Into Nutrients for Epidemiological Research) – a new data-entry program for nutritional analysis in the EPIC-Norfolk cohort and the 7-day diary method. *Public Health Nutrition* **4**(6): 1253–1265.

Willett W (1998) *Nutritional Epidemiology*, 2nd edn. Oxford: Oxford University Press.

Relevant Websites

http://www.fao.org – INFOODS information for nutrient database compilers and suppliers.

Validation

M Nelson, King's College London, London, UK

Although the basic approaches to dietary assessment have changed little in the past 20 years, there has been a growing awareness of the ways in which errors in dietary assessment may undermine an understanding of diet–disease relationships. There have been two stages in this process: (i) acceptance of the fact that every measure of food consumption is likely to be influenced by the reporting process and (ii) the improvements in the methods for estimating the size of the difference between what is observed and the likely true values for countries, households, or individual subjects. This article examines the techniques for coping with the sources of error in the assessment of diet, particularly through the use of biochemical and statistical techniques that are available to evaluate the quality or enhance the interpretation of dietary data.

Reproducibility and Validity of Dietary Intake Measurements

Reproducibility, Repeatability, and Reliability

'Reproducibility' is the extent to which a tool is capable of producing the same result when used repeatedly in the same circumstances. The terms 'repeatability' and 'reliability' are often used synonymously with 'reproducibility.' All of these terms are equivalent to the word 'precision,' often used by biochemists to describe the variation of a measurement based on repetition of a particular assay using a single piece of equipment.

A measurement may have good reproducibility and yet have poor validity, but a measurement that has good validity cannot have poor reproducibility.

In reality, 'the same circumstances' can never exist in relation to dietary measurements because diet (whether of individuals, households, or countries) varies over time, be it on a daily, weekly, seasonal, or annual basis. In epidemiological studies, the aim is usually to assess 'usual' intake. Part of the variation in any dietary measurement will thus relate to genuine variability of diet. The remaining variation will relate to biases associated with the method. Due consideration must be given to these time-related factors when evaluating the reproducibility and validity of dietary measures, and a well-designed validation study will separate the variability associated with reproducibility (the error in the method) from that associated with genuine biological variation over time.

Validity

Validity is an expression of the degree to which a measurement is a true and accurate measure of what it purports to measure. Establishing validity requires an external reference measure against which the 'test' measurement (the one being used in the main survey or research activity) can be compared. In nutrition, there is no absolute reference measure of the truth. Every measurement of dietary intake includes some element of bias. The best that can be managed is to assess relative or congruent validity of measurements, comparing results obtained with the test instrument with what are believed to be more accurate measures of food or nutrient intake obtained, for example, using a biological marker.

There are two main problems arising from inaccurate measurements:

1. Incorrect positioning of a country, household, or person in relation to an external reference measure (e.g., dietary reference values)
2. incorrect ranking of countries, households, or people in relation to one another

The first type of error can result in inappropriate investigations or actions being taken to remedy an apparent dietary deficit or excess that does not really exist. Alternatively, no action may be taken when some is needed (e.g., a true deficit is not detected because diet is overestimated). The second type of error tends to undermine the ability to assess relationships between diet and health. For example, if someone who properly belongs in the top quarter of the distribution of saturated fat intake (associated with an increased risk of heart disease) is classified in the bottom quarter because he or she underreports his or her true fat intake, the relationship between fat intake and outcome (risk of heart disease) will not be shown. Again, this can lead either to inappropriate recommendations for improving health in the population or, more often, to a failure to take action because the true relationship between diet and health is obscured by measurement error.

Table 1 lists some of the sources of measurement error, their principal effects, some ways of taking errors into account in analysis, and ideas for dealing with them in practice. At the national level, errors in reporting of food production, imports, exports, food moving in and out of stocks, and nonfood uses (e.g., sugar used in the brewing industry) tend to produce overreporting of food and nutrient availability in economically developed countries and underreporting in developing countries. The consequences for between-country comparisons will therefore vary nutrient by nutrient. Where true values are higher in developed countries and lower in developing countries (e.g., energy), a tendency in wealthier countries to overstate consumption and in poorer countries to understate consumption will exaggerate the true differences between countries (**Figure 1**). Where true intakes tend to be higher in developing countries and lower in developed countries (e.g., nonstarch polysaccharides), such biases in measurement will minimize any apparent differences between countries. These types of errors do not lend themselves easily to correction or adjustment. The best solution is to try to improve reporting mechanisms that reduce duplication or omission, although this may be particularly problematic in developing countries, in which a high percentage of the population is engaged in subsistence agriculture.

At the household level, information can be obtained from records of food acquisitions (expenditure and/or amounts), often referred to as household budget surveys. Errors associated with the recording process (omission and misrecording) will contribute to the error of estimates of consumption. Keeping a record of purchases can lead householders to alter their usual purchasing patterns or encourage purchasing in excess of true requirements (especially easily storable items such as flour, sugar, oil, and cooking or spreading fats). Some form of cross-check is necessary—either an internal reference measure such as a cross-check list of foods purchased or used at mealtimes or an external reference measure such as an independent assessment of food consumption by all household members. There are problems related to the estimation of waste and the amount of food given to pets. The true amount of food wasted cannot be known, and direct measures of waste are likely to introduce bias into the measurements. Adjustments for waste range from 4 to 10%, and 6% is a typical value for average waste.

In household budget surveys, the amounts of food eaten away from home (or, more correctly, derived from outside the household food supply) are sometimes measured directly by individual household members who keep records of their food and drink consumption away from home. Alternatively, the amount of food eaten from outside the household food supply can be measured indirectly by asking the main household respondent to record which meals have been consumed from the household food supply and the ages and genders of all people present at those meal, including any visitors. The menu records do not usually include information about snacks between meals. Whether based on the direct or indirect method, household surveys tend to underestimate the amount of food eaten away from home and thus to overestimate the proportion of the diet consumed from the household food supply.

The estimate of the nutrient content of the diet in household-level surveys is based on nutrient conversion factors. These differ from food composition data. Food composition data reflect the chemically determined nutrient content of foods. Nutrient conversion factors modify these values to allow for preparation and cooking losses so that the final estimate of the nutrient content of the diet relates the amount of food initially purchased to the amount of food and nutrient likely to be consumed. There are many assumptions in the determination of the nutrient conversion factors (e.g., the proportion of water-soluble vitamins lost in the cooking process). The result may be an over- or underestimate of

Table 1 Sources of error at different levels of dietary assessment, their effects, ways to adjust for measurment error in analysis, and ways to minimize error in data collection

Level of assessment	*Source of error*	*Effect of error*	*Statistical adjustment*	*Ways to minimize*
National: Food balance sheets	Inaccurate values of food production, imports, exports, storage, nonfood uses	Over- or underemphasizes apparent differences between countries	Not possible; error likely to vary nonsystematically between countries	Improve reporting mechanisms and estimates of waste and production losses
Household: Household budget surveys, with or without larder inventories	Omission of food items	Underestimation of consumption, especially for certain foods (e.g., sweets)	Inflate estimate of intake to compensate for known underestimation	Cross-check food purchasing list at final interview
	Misrecording of food items	Misrepresentation of true consumption pattern	Not possible; likely to differ between households for different foodstuffs	Cross-check food purchasing list at final interview
	Distortion due to effect of survey (e.g., selection of items believed to be 'healthy')	Misrepresentation of true consumption pattern	Not possible; likely to differ between households for different foodstuffs	Cross-check food purchasing list at final interview
	Overpurchasing stimulated by measurement process	Overestimation of consumption, especially for certain foods (e.g., flour and sugar)	Deflate estimate of intake to compensate for known overestimation	Validate purchasing against reference measure of actual consumption at individual level (for all household members)
	Estimate of amount wasted or given to pets	If amount is underestimated, leads to overestimation of consumption (and vice versa)	Deflate estimate of intake to compensate for known overestimation (or inflate)	Improve estimates of waste through surveys
	Estimate of amounts of food eaten away from home or consumed by visitors, based on menu records	If amount is underestimated, leads to underestimate of total consumption (if overestimated, leads to overestimation)	Inflate estimate of intake to compensate for known underestimation (or deflate)	Household members keep records of amounts eaten away from home (but see problems with records kept by individuals)
	Use of nutrient conversion factors	Unknown biases introduced into calculation of the nutrient content of the diet	Not possible; errors not systematic between foods	Improve estimates of nutrient content and of preparation and cooking losses
Individual: prospective or retrospective methods	Underreporting	Underestimate of usual intake	Calibrate result against valid measure of intake and adjust accordingly	Use more than one measure of intake, including valid reference measure
	Differential misclassification	The error in measurement not consistent between subgroups, making valid comparisons impossible	Identify likely error according to characteristics on which subgroup comparisons are to be based, and adjust accordingly	Undertake appropriate validation studies within subgroups
	Distortion of usual diet (prospective recording or retrospective recall)	Overestimation of intake of some foods and nutrients and underestimation of others	Not possible unless nature of distortion is known for different foods and for different subgroups	Use biomarkers of intake to confirm dietary measures; use more than one measure of intake, including valid reference measure

Adapted from Nelson M (1998) Methods and validity of dietary assessment. In: Garrow J and James WPT (eds.) *Human Nutrition and Dietetics*, 10th edn. London: Churchill Livingstone.

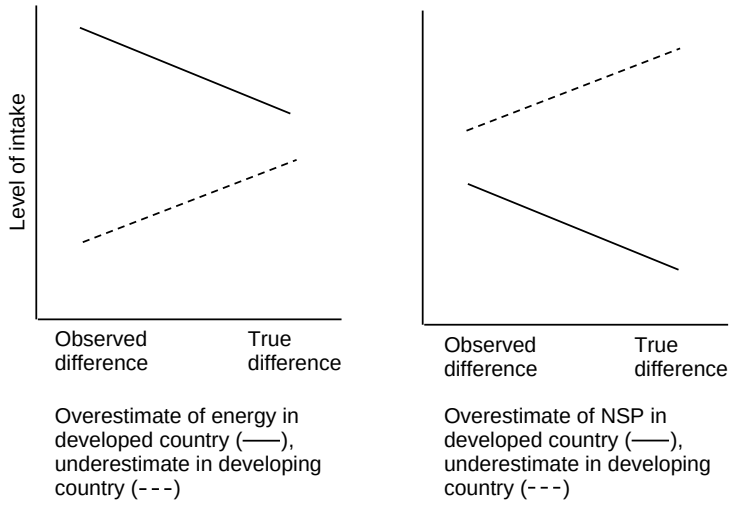


Figure 1 Effects of measurement error on between-country comparisons of nutrient availability.

the actual nutrient content of the diet consumed, depending on the nutrient concerned.

At the individual level, the high reproducibility of results using a given technique led researchers for many years to assume that the observed results were valid. Reproducibility was confused with validity. Researchers argued that the results from their chosen technique were correct, but they were unable to explain adequately why measurements of the same diets using different techniques yielded different results. It is now clear that different techniques introduce different types of errors in assessment. Moreover, one can argue that the main sources of error in the assessment of diet at the individual level (e.g., errors in perception of frequency of consumption and of food portion sizes; the desire to increase the reporting of those foods perceived as being 'healthy' or to reduce the reporting of foods perceived as being 'unhealthy'; and items forgotten, intentionally or unintentionally) are not readily amenable to correction.

The most common error in the assessment of diet at the individual level is underreporting. Thus, the mean estimate of energy intake in a group of subjects that includes people who underreport their true consumption is also likely to be an underestimate. Moreover, any comparison of subgroups that include different proportions of underreporters will yield a false picture of the relationships between subgroups. For example, when comparing energy intake based on 7-day weighed records between normal weight and overweight subjects, it is often observed that the overweight subjects (whose energy requirements are higher) will report energy intakes equal to or less than those of the normal weight subjects. This is because overweight subjects show a strong tendency to underreport their usual consumption levels, particularly of foods with a high proportion of fat. It is worth noting that there is also a small group of subjects who overreport their usual consumption, but their characteristics are less readily defined.

In many validation studies, comparisons are made between methods to establish 'concurrent validity,' often between a new instrument such as a food frequency questionnaire and a more established technique such as a weighed inventory of diet. **Table 2** summarizes the problems associated with the use of specific dietary reference measures. The

Table 2 Limitations of dietary reference methods appropriate for validation of dietary assessment measures

Reference method	*Limitations*
Weighed records and household measures	Underreporting Unrepresentative of 'usual' diet over sufficient number of days Distortion of food habits due to recording process
Diet history	Interviewer bias Inaccuracy of portion size reporting due to conceptualization and memory errors Errors in reporting of frequency, especially overreporting of related foods listed separately (e.g., individual fruits and vegetables) Requires regular eating habits
Repeat 24-h recalls	Under- or overreporting of foods due to reporting process (e.g., alcohol underreported and fruit overreported) Unrepresentative of 'usual' diet over sufficient number of days Inaccuracy of portion size reporting due to conceptualization and memory errors

principal error is likely to be underreporting within certain subgroups, and any validation study needs to address this. Issues regarding distortion of diet are more difficult to identify, however. Work has been carried out to identify some of the errors associated with perception, conceptualization, and memory in relation to assessment of food portion sizes. Agreement between methods is no guarantee of validity. If the two methods are biased in the same way (e.g., using a diet history to validate a short food frequency questionnaire), then the observed level of agreement is likely to overstate the validity of the new method. If the validation is based on repeated measures and the errors in the reference assessment are repeated (e.g., validating a food frequency questionnaire against repeat 24-h recalls), then the level of agreement between the two methods is likely to underestimate the true correlation between the test measure and the truth because the true variance of diet will be underestimated by the reference measure.

Some types of bias may influence reporting in a similar way in all subjects. For example, if a food frequency questionnaire does not ask for sufficient detail about tropical fruit consumption in a survey of West Indian families living in London, then all responses are likely to be biased in the same way. This will result in nondifferential misclassification: Everyone's intake of tropical fruit (and hence vitamin C and carotene intake) is likely to be similarly misclassified. If, on the other hand, the nature of the bias is likely to be related to individual characteristics—for example, West Indian women overreport their tropical fruit consumption compared to West Indian men—then the tropical fruit intake of men and women will be misclassified in different ways in different subgroups. This is known as differential misclassification. If the risks for men and women are being compared in a diet–disease analysis, the errors in the estimates of those risks will not be the same for both groups.

There are many factors in addition to body size or gender that may influence the ability of a dietary measure to estimate accurately the level of food consumption or nutrient intake. **Table 3** lists the potential confounders of good dietary reporting that may affect validation studies. If there is concern about misclassification in one group (e.g., smokers) being different from misclassification in another group (e.g., nonsmokers), then validation must identify the extent of misclassification in the two groups independently so that it can be compensated for in any subsequent analyses. It is therefore important to measure all of the potential confounders that may be a source of misclassification (differential or nondifferential).

Table 3 Factors that may influence the reliability and validity of dietary assessments and that may need to be measured in a validation study

Related to the subject	*Related to the measuring process*
Age	Portion size
Gender	Interviewer
Height, weight, etc.	Learning effects
Region	Recency effects
Social class	Lag time
Education	Number of foods on an FFQ
Language	Nutrient database
Culture and ethnic background	
Smoking	
Social approval	
Social desirability	

The design of a validation study must also reflect the purpose for which the dietary assessment is being carried out. Most dietary assessments measure current or recent consumption. However, if the aim is to estimate past diet related to the induction of cancer 10 years before the time of the assessment (e.g., in a case–control study), then the validation process should in theory address the relevant time period. In practice, this may be very difficult (i.e., if there is no robust reference measure of diet relating to the relevant time period). In such circumstances, the weaknesses of an unvalidated test measure should not be overlooked nor understated.

Use of Biological Markers to Validate Dietary Intake Measurements

Nutritional biomarkers are those elements or compounds in biological samples capable of reflecting relationships between diet, nutritional status, and disease processes. Not all biomarkers are suitable for use in dietary validation studies. One of the key features of a marker should be its ability to reflect intake over a wide range of intakes.

Figure 2 shows the sequence whereby food or drink containing potential biomarkers (nutrients and nonnutrients) may be ingested, absorbed, distributed, and excreted. The stages in the top of **Figure 2** are not usually amenable to sampling (e.g., taking samples of gastric contents), and it is only in the later stages in which compounds are in circulation, present or stored in tissues, or excreted that they are more readily sampled. The complexity of the relationship between intake and the measured levels of biomarkers in these lower stages, however, may limit the usefulness of certain compounds in validation studies. **Figure 3** illustrates for four nutrients the varying relationships between tissue levels

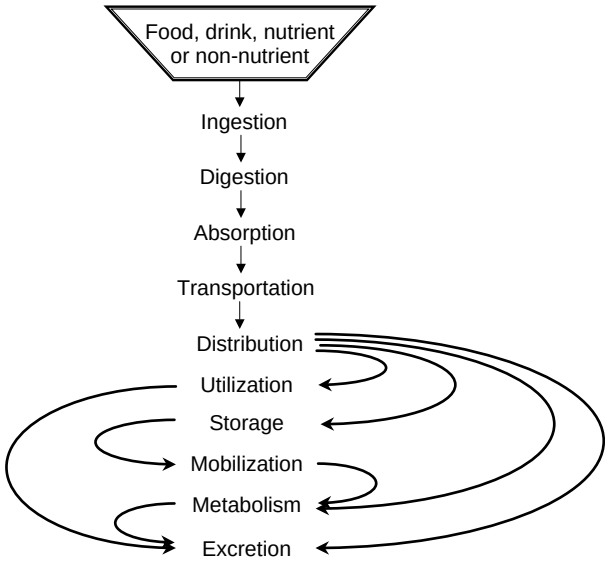


Figure 2 Stages in the pathway between intake and measurement of biomarkers.

and levels in diet. Vitamin E shows a more or less linear relationship between blood and dietary levels across a broad range of intakes. Riboflavin appears in urine only after tissues are saturated and any excess is excreted. An alternative measure to assess riboflavin, erythrocyte glutathione reductase activity coefficient), is a sensitive biomarker of intake only at low levels. Vitamin C in blood is a poor marker at low levels of intake, increases in sensitivity as a marker of intake across the middle range of intakes, and is poor once again at high levels of intake where excess vitamin is excreted in the urine. Retinol is stored in the liver and its level in blood is controlled homeostatically. It is therefore a poor marker above relatively low intakes. A further problem is the point in time or span of intake reflected by the marker. **Figure 4** shows that some markers relate to intakes days or weeks prior to the sampling point (e.g., energy and doubly labeled water), whereas other may reflect intake over months (iron intake reflected by ferritin) or years (calcium intake reflected by bone mass). The influence of other factors relating marker to intake (e.g., hem versus non-hem iron in the diet, and the influence of vitamin C or dietary fiber on absorption) may undermine the ability to conclude that a low measurement of a marker is necessarily a reflection of low dietary intakes. Thus, the value of markers in assessing the validity of intake measurements is often limited to specific ranges of intake in diets of known composition. This may not be sufficient for epidemiological purposes, where the entire range of intake may be of interest in relation to disease risk.

The two most widespread uses of biomarkers for the assessment of the validity of measures of diet are to identify under- or overreporters and to assess the correctness of ranking of individuals according to their nutrient intake.

Techniques for Identifying Under- and Overreporters

Doubly labeled water The scientific basis that underlies the doubly labeled water method for estimating energy expenditure relies on the differential rates of loss of hydrogen and oxygen from the body at different levels of energy expenditure. Hydrogen is lost primarily in water, whereas oxygen is lost in

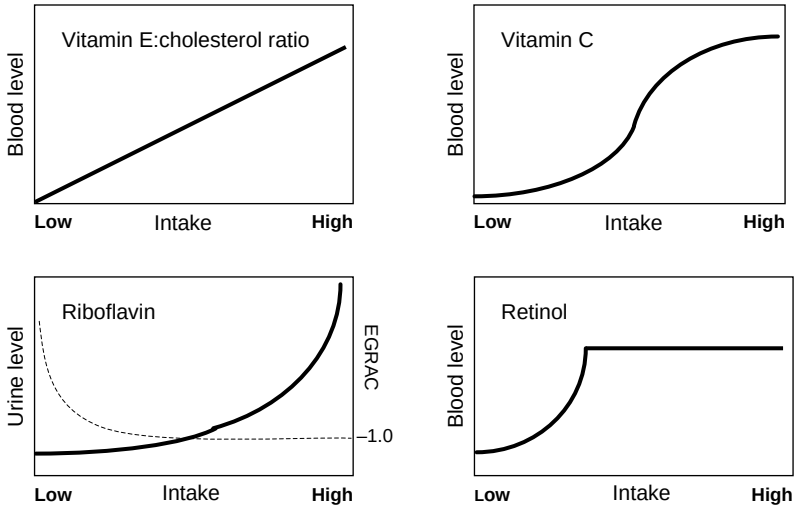


Figure 3 Associations between intake and biomarker over a wide range of intakes: vitamin E:cholesterol ration, vitamin C, riboflavin, and retinol. EGRAC, erythrocyte glutathione reductase activity coefficient. (Adapted from Kohlmeier L (1991) What you should know about your biomarker. In: Kok FJ and van't Veer P (eds.) *Biomarkers of Dietary Exposure*. London: Smith Gordon.)

	Days	Weeks	Months	Years
Energy	Doubly labeled water		Body weight	
Fatty acids	Cholesterol esters	Erythrocyte membranes	Adipose tissue	
Tocopherols	Serum		Adipose tissue	
Retinol			Liver tissue	
Carotenoids	Plasma		Adipose tissue	
Vitamin C	Urine Leucocytes Plasma			
Iron		Hemoglobin	Ferritin	
Calcium	Urine			Bone mass
Selenium	Plasma	Erythrocyte glutathione peroxidase	Toenails	

Figure 4 Time scale over which different biomarkers may reflect the relationship with intake.

both water and carbon dioxide. The relative rates of loss can be used to estimate energy expenditure in free-living subjects over a period of approximately 2 weeks, thereby providing a reference measure for energy intakes over a similar period. Provided the subject is in energy balance (neither gaining nor losing weight due to changes in body composition), the measures of expenditure and intake should agree. The technique allows for identification of both under- and overreporters. The boundaries of acceptability of the test measures (e.g., to within $\pm 10\%$ of habitual energy intake) need to be chosen according to the needs of the study in which the validity of the test measure is being assessed. The level of agreement between test and reference measure will dictate both the precision of the estimate of mean intake for an individual or subgroup and the extent to which subjects will be misclassified when ranked according to level of intake.

The main disadvantage of the doubly labeled water technique is its high cost. In a large-scale study, it is not feasible to use doubly labeled water with every subject in order to assess the completeness of dietary records. The technique is therefore usually used to assess validity of the test measure in a sample of subjects who are taken to be representative of the sample for the main study.

Another disadvantage is that doubly labeled water provides a marker for energy only. The diet recorded could differ substantially from the subject's usual diet but have an energy content in agreement with the estimate of energy expenditure. In the absence of additional information about usual patterns of food consumption, such a record would be regarded as valid. A further problem is that not all food consumption or nutrient intake correlates strongly with energy intake. For example, fruits and vegetables and their associated nutrients (e.g., vitamin C, beta-carotene, and potassium) may be overreported in a dietary assessment in which energy intake agrees well with energy expenditure, but the overreporting would not be identified. These comments are summarized in **Table 4**.

Urinary nitrogen and potassium excretion A second technique for identifying under- or overreporters is to collect 24-h urine samples and compare the amounts of nitrogen and potassium excreted with the amount ingested. Allowing for incomplete absorption and losses of nitrogen from the gastrointestinal tract (digestive juices and shed epithelial cells), hair, skin, and sweat, the amount of nitrogen excreted should be approximately 81% of the nitrogen ingested. Allowing for daily variations in intake and excretion, if daily recorded intake of nitrogen is less than 70% of the corresponding urinary nitrogen excretion over the following 24 h, the respondent is likely to have underreported his or her usual consumption; anyone whose recorded intake of nitrogen is more than 100% of their excretion is likely to have overreported their consumption. The more days of intake and excretion data that are collected, the better the agreement over the recording period should be for subjects who are in nitrogen balance. If urinary nitrogen is to be used as a marker for the completeness of dietary recording, it is helpful to have at least 4 days' worth of data (diet and urine). For potassium, the expected urinary excretion is 95% of the intake, with limits of 80 and 110% for 'good' reporting.

As with doubly labeled water, it is assumed that the subject is in balance, neither losing nor gaining body nitrogen or potassium.

Table 4 Limitations of biological reference methods appropriate for validation of dietary assessment measures

Reference method	*Limitations*
Doubly labeled water	Energy only
	Assumptions of model regarding water partitioning may not apply in cases of gross obesity, high alcohol intake, or use of diuretics
	Very expensive
	Analysis technically demanding
Urinary nitrogen: completeness of samples confirmed using PABA	Protein only
	PABA analysis affected by paracetamol and related products
Urinary nitrogen only	Protein only
	Danger of incomplete samples
Biochemical measurements of nutrients in blood or other tissues	Complex relationship with intake mediated by digestion, absorption, uptake, utilization, metabolism, excretion, and homeostatic mechanisms
	Cost and precision of assays
	Invasive
Energy intake:BMR ratio	Imprecision of estimate of BMR based on body weight and regression equations
	Single cutoff point (e.g., EI:BMR <1.1) will not identify low-energy reporters with higher habitual energy expenditures
	Higher estimates of cutoff (e.g., EI:BMR <1.2) captures more true low energy reporters but also more good reporters

It is important to ensure that the urine collections are complete. This necessitates the use of an inert metabolic marker (para-amino benzoic acid (PABA)), which is rapidly absorbed and excreted. Subjects take a divided dose of 240 mg PABA throughout the day. At least 85% of the PABA should be recovered in the urine in a 24-h collection. If the amount recovered is less than 85%, the urine sample may be regarded as incomplete and therefore not suitable for analysis of nitrogen in order to check the completeness of the dietary record. Because paracetamol and related compounds interfere with the PABA assay, a measure of excretion over 115% of the administered dose would be suspect.

As with doubly labeled water, the principal weakness of urinary nitrogen as a marker for the completeness of dietary records is that many foods contain low levels of nitrogen but may be important sources of other nutrients. Any check for the completeness of dietary records based on nitrogen will not assess the presence or absence of these other foods. Also, the issue of dietary distortion is not addressed. Potassium is more widespread in foods, although the largest contributors are usually fruits and vegetables. Using urinary nitrogen and potassium in combination gives a better assessment of the completeness of the recording than any single marker.

Ratio of energy intake to basal metabolic rate Doubly labeled water and urinary nitrogen excretion are particularly useful for assessing the validity of prospectively recorded diets because the time frame of the test and reference measures can be made to coincide. A third technique for assessing validity can be used with both prospective records and recalls of diet. It is based on energy and thus has the limitations of a validating marker relating to a single dietary factor. It has the advantage, however, of being able to be applied to all subjects in a dietary survey because no external reference measure is needed. It is a biomarker in the sense that it relies on a biological measure (body weight) and is best applied when measures of physical activity at the individual level are also available. The assumption is that there should be reasonable agreement between estimated requirement and estimated intake.

Schofield equations can be used to estimate basal metabolic rate (BMR) based on age, gender, and body weight. An individual whose reported energy intake is below the level of energy expenditure likely to be needed to carry out day-to-day activities has probably underreported his or her diet. A typical cutoff point for an acceptable ratio of the energy intake to BMR ratio in an individual is 1.2, taking into account daily variations in energy intake over a period of 7 days of dietary recording and allowing for the inaccuracies of the estimate of BMR based on the Schofield equations. A cutoff of 1.2 will identify only those subjects who underreport and whose levels of activity are low. For subjects with higher levels of activity (e.g., estimated from questionnaire responses), proportionately higher cutoff points are appropriate. It is also possible to estimate an upper probable level of energy expenditure (e.g., 2.5 times BMR, depending on habitual level of activity) and subjects with reported levels of intake over this value may be regarded as being overreporters.

Statistical Assessment of Validity

There are two broad approaches to establishing validity between test and reference measures: comparison of mean values and correlation. The use of mean values is appropriate where group intakes are to be determined or where an absolute measure of intake is required. This is especially important where a threshold value of intake will be used to make recommendations (e.g., recommending an increase in potassium intake because of its association with lower blood pressure and reduced risk of myocardial infarction—there would be no point in recommending additional consumption of potassium for individuals who were identified as having intakes already above the levels that were seen to be protective). The correlation technique (plotting the observed measure against the reference measure) is appropriate where it is important to classify subjects according to high or low intakes because differences in intake are associated with different levels of disease risk. In relation to disease risk, a steeper or shallower slope of observed intake in relation to true intake will have a profound effect on the relative risk estimates in relation to disease outcome. Moreover, the correlation coefficient describes only one aspect of agreement. In practice, both ranking and absolute levels are important in establishing correct diet–disease relationships.

Mathematically, disease risk can be modeled by the expression

$$R(D|T) = \alpha_0 + \alpha_1 T$$

where $R(D|T)$ is the risk of disease D, T is the unobservable true long-term habitual intake of a given food or nutrient relevant to disease risk, α_0 is the underlying risk of disease in the population independent of dietary exposure, and α_1 is the log relative risk (RR) that may be positive (for predisposing factors) or negative (for protective factors).

Because we cannot measure the true dietary exposure, we approximate it using the dietary test measure that is the focus of the validation process. The expression for describing disease risk then becomes

$$R(D|Q) = \alpha_0 + \alpha^* Q$$

where Q is the observed intake, and α^* is the observed log relative risk. In most circumstances, it can be argued that $\alpha^* = \lambda\alpha_1$, where λ is known as the 'attenuation' factor and is equal to the slope of the regression line of T plotted against Q. In most dietary studies, the value for λ is between 0 and 1, but in cases of differential misclassification it may also be negative. The consequence is that the estimate of disease risk in relation to diet is likely to be different from the true risk (usually tending toward a relative risk of unity).

For a given individual i, the observed measure Q_i will be given by the expression

$$Q_i = (T_i)B + a + \epsilon_i + e_i$$

where T_i is the true measure, and the attenuation factor λ is a function of B (proportional bias), a (constant bias), ϵ_i (random error within a subject, such that the mean of the random error across all subjects is equal to zero), and e_i (bias in the ith subject, such that the mean of the individual biases across all subjects is not equal to zero—this is the consequence of differential misclassification). In practical terms, the aim of a validation study is to quantify these sources of error (see **Table 3** for the main likely sources) and to estimate the value for λ so that the true relative risk α_1 can be estimated using the expression $\alpha_1 = \alpha^*/\lambda$.

It is probable that both the test measure and the reference measure are positively correlated with the truth. This is represented by r_{QT} and r_{RT} in **Figure 5**. There will also be a relationship between the test and reference measures, r_{QR}, given by the expression

$$r_{QR} = r_{QT} \times r_R$$

The relationship between the test measure and the truth can be estimated by solving for r_{QT}:

$$r_{QT} = r_{QR}/r_{RT}$$

Assuming that the reference measures are unbiased, r_{RT} can be estimated by knowing the relationship between within- and between-subject variance in a group of subjects whose records or measures are assumed to be valid (e.g., from whom

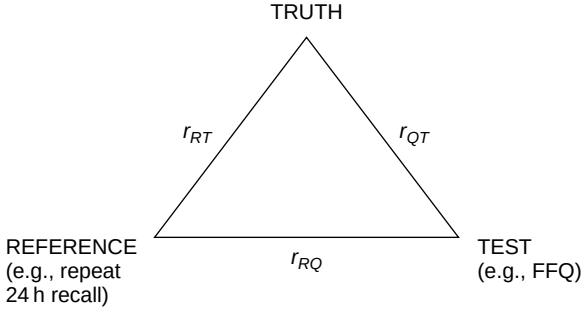


Figure 5 The relationship between test and reference measure and the truth. (Source: Nelson M (1997) The validation of dietary assessments. In: Margetts BM and Nelson M (eds.) *Design Concepts in Nutritional Epidemiology*, 2nd edn. Oxford: Oxford University Press.)

likely under- or overreporters have been excluded) using the expression

$$\sqrt{n/(n + (s_w^2/s_b^2))}$$

where n is the number of repeat observations within one subject, and s_w^2 and s_b^2 are the within- and between-subject variances, respectively. In this way, the likely relationship between the test measure and the truth can be estimated, and the relative risk can be adjusted to account for misclassification of subjects based on the test measure alone.

This approach has two weaknesses. First, if the reference measure is a dietary measure, it does not address the problem of correlation of errors (the tendency for an individual to misreport diet in the same way using the test and reference measures). If errors are correlated between methods, then the observed r_{RQ} is likely to overestimate $r_{RT} \times r_{QT}$; it will appear that the test method is performing better than it actually is performing. If the errors are correlated within methods (e.g., if the same types of within-person bias are occurring from day to day using repeat 24-h recall), then the observed r_{RQ} is likely to underestimate $r_{RT} \times r_{QT}$. The second weakness of this approach is that it does not address the problem of differential misclassification.

A similar technique is the method of triads (**Figure 6**), in which no assumption need be made about the relationships between reference measures and the truth. Instead, the relationships between three measures can be used to estimate values for ρ, which in theory approximate the correlation between each of the measures and the truth. As in the technique described previously, valid estimates of ρ are based on the assumption that the errors in the methods are uncorrelated and that the errors are random and not differentially biased between subjects.

Because the validation process helps to identify subjects who are likely to be misreporting their diet, the temptation may be to exclude from analysis those subjects who have misreported their diet. It may be, however, that the very subjects who are most likely to misreport their food consumption (e.g., people who are overweight) are also those who are at increased risk of disease (e.g., hypertension, heart disease, and colon cancer). In estimating disease risk, therefore, the aim must be to retain all of the subjects in the analysis.

Estimating the components of error and finding appropriate values for λ is the best way to address this issue. A special case is to adjust nutrient intakes to allow for misreporting in some subjects by assuming that true energy intake and true nutrient intake in all subjects are well correlated. Thus, if a subject underreports energy intake, it is assumed that other nutrients will be underreported to a similar extent. By estimating nutrient intake in relation to reported energy intake, subjects can be ranked according to whether, for a given level of energy intake, their nutrient intake was above or below the average. This is known as energy adjustment. To find the energy-adjusted estimate of nutrient intake, the nutrient intakes should be plotted against energy intakes and the regression line and the residual values derived (**Figure 7**). Energy-adjusted nutrient intakes are then computed by adding the residual to the mean nutrient intake. This approach allows all subjects to be included in an analysis, and it provides realistic estimates of intake (unlike computations of nutrient density in which each subject's nutrient intake is divided by his or her energy intake). Like doubly labeled water, however, the

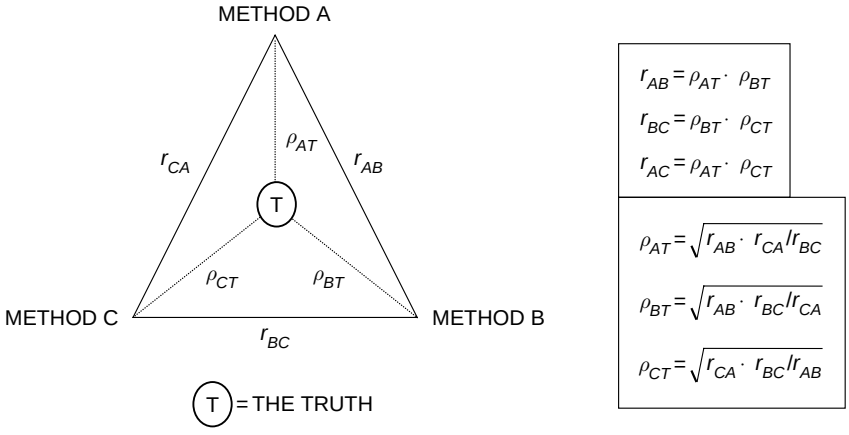


Figure 6 Graphic representation of the method of triads. (Source: Ocké M and Kaaks R (1997) Biochemical markers as additional measurements in dietary validity studies: Application of the method of triads with examples from the European Prospective Investigation into Cancer and Nutrition. *American Journal of Clinical Nutrition* **65**:1240S–1245S.)

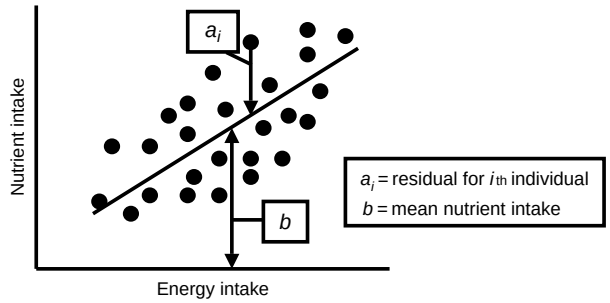


Figure 7 Energy-adjusted nutrient intake for the *i*th individual = $a_i + b$. (Adapted from Willett WD, Howe GR and Kushi LH (1997) Adjustment for total energy intake in epidemiologic studies. *American Journal of Clinical Nutrition* **65**:1220S–1228S.)

weakness of energy adjustment lies in the fact that not all nutrient intakes are well correlated with energy intake. Overreporting of fruits and vegetables consumption leading to an overestimate of vitamin C intake, for example, would not be appropriately compensated for using energy adjustment if the person was at the same time underreporting his or her fat consumption.

Conclusions

The validation of dietary assessment measures is a necessary part of any dietary investigation. The use of unvalidated instruments is likely to lead to misinterpretation of diet–disease relationships. In most circumstances, unvalidated measures of diet will lead to the conclusion that there is no diet–disease relationship when in reality one exists (bias toward the null). In circumstances in which differential misclassification is at play, it may also lead to the conclusion that there is a relationship when in fact none exists. In order not to waste resources, it is vital to ensure either that dietary assessments are valid or that the errors associated with them are clearly identified and taken into account in the interpretation of diet–disease relationships.

In summary,

- Never take dietary measurements at face value.
- Always include an analysis of errors of dietary data in epidemiological studies.
- Ensure that the validation study sample is representative of the population in the main study.
- Quantify all significant sources of error.
- Obtain measures of within-subject reliability.
- Adjustment of estimates of intake is preferable to excluding misreporters from the diet–disease analysis.

Finally, it is worth noting that virtually all of the discussion in this article relates to nutrient intakes. There is an urgent need to improve our understanding of the validity of measurements of the consumption of foods. Advice on healthy eating is given primarily in terms of foods, not nutrients.

See also: **Dietary Intake Measurement**: Methodology. **Dietary Surveys**. **Energy Expenditure**: Doubly Labeled Water.

Further Reading

Bates CJ, Thurnham DI, Bingham SA *et al.* (1997) Biochemical markers of nutrient intake. In: Margetts BM and Nelson M (eds.) *Design Concepts in Nutritional Epidemiology*, 2nd edn. Oxford: Oxford University Press.

Bingham SA, Cassidy A, Cole T *et al.* (1995) Validation of weighed records and other methods of dietary assessment using the 24 h urine technique and other biological markers. *British Journal of Nutrition* **73**: 531–550.

Black AE (2000) Critical evaluation of energy intake using the Goldberg cut-off for energy intake: Basal metabolic rate. A practical guide to its calculation, use and limitations. *International Journal of Obesity* **24**: 1119–1130.

Cameron ME and van Staveren WA (eds.) (1988) *Manual on Methodology for Food Consumption Studies*. Oxford: Oxford University Press.

Clayton D and Gill T (1997) Measurement error. In: Margetts BM and Nelson M (eds.) *Design Concepts in Nutritional Epidemiology*, 2nd edn. Oxford: Oxford University Press.

Department of Health (1991) *Report on Health and Social Subjects, 41. Dietary Reference Values for Food Energy and Nutrients for the United Kingdom. Committee on Medical Aspects of Food Policy*. London: HMSO.

Goldberg GR, Black AE, Jebb SA *et al.* (1991) Critical evaluation of energy intake data using fundamental principles of energy physiology: 1. Derivation of cut-off limits to identify underrecording. *European Journal of Nutrition* **45**: 569–581.

Kohlmeier L (1991) What you should know about your biomarker. In: Kok FJ and van't Veer P (eds.) *Biomarkers of Dietary Exposure*. London: Smith Gordon.

Margetts BM and Nelson M (1997) *Design Concepts in Nutritional Epidemiology*, 2nd edn. Oxford: Oxford University Press.

Millward DJ (1997) Urine nitrogen as an independent validatory measure of protein intake: Potential errors due to variation in magnitude and type of protein intake. *British Journal of Nutrition 77: 141–144.*

Nelson M (1997) The validation of dietary assessments. In: Margetts BM and Nelson M (eds.) *Design Concepts in Nutritional Epidemiology*, 2nd edn. Oxford: Oxford University Press.

Nelson M, Atkinson M, and Meyer J (1997) *A Photographic Atlas of Food Portion Sizes* London: Ministry of Agriculture, Fisheries and Food.

Prentice AM, Black AE, Coward WA *et al.* (1986) High levels of energy expenditure in obese women. *British Medical Journal* **292**: 983–987.

Schatzkin A, Kipnis V, Carroll RJ *et al.* (2003) A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: Results from the biomarker-based Observing Protein and Energy Nutrition (OPEN) study. *International Journal of Epidemiology* **32**: 1054–1062.

DIETARY SURVEYS

K L Tucker, Tufts University, Boston, MA, USA

Purpose and Design

Dietary surveys are used for multiple purposes and range from measurement of food disappearance at the national level, to food use at the household level, to detailed multiple assessments of individual intake for linkage with health outcomes. Each of these methods has strengths and limitations, depending on the survey purpose (see **Table 1**).

Research Questions and Data Needs

At the national level, information on food use and dietary intake is needed for economic and agricultural policy decision making. For policy makers to advise on food production, food imports, pricing of staple foods, and other factors that affect food availability, they require information on the production, inflow, and outflow of food commodities and products at the national level. Most countries use food balance sheets to measure these flows, and total available nutrients are estimated in relation to the size and composition of the population. These surveys measure overall national food production, imports, and available food stocks, and subtract exports, food used for animals rather than humans, and losses that occur during production, storage, and manufacturing. The FAO has compiled food balance sheets for many countries since 1949, thus allowing useful intercountry comparisons of food availability. However, the aggregate information obtained with food balance sheets does not allow consideration of food distribution within a country and does not quantify food intake or needs of subgroups of the population.

Most countries need more information on household level food use to target food nutrition policies toward groups at need. Household food surveys capture the amounts and types of food that enter a household, and *per capita* intake equivalents are calculated by dividing the total nutrients available in the household from the edible portion of entering foods by the numbers of household members, weighted by specific age and sex. With information on the average *per capita* intake of specific households, it is possible to consider groups at risk of inadequate intake of energy or of specific macro- or micronutrients. For example, these surveys may highlight rural–urban differences, inland–coastal differences, differences by socioeconomic strata, and so on. Such surveys provide critical information within countries for the development and targeting of economic, agricultural, and nutritional policies specific to regions or other subgroups of the population.

Table 1 Advantages and disadvantages of different types of dietary survey

Level	*Survey type*	*Advantages*	*Disadvantages*
National	Food balance sheets	Inexpensive	Crude estimate; no consideration of wastage; does not allow disaggregation to sublevels
Household	Food account method	Inexpensive	Does not account for food consumed away from home, inventories or wastage
	Interviewer-administered list-recall	More detail obtained on foods than in the food account method	List may limit responses; Waste usually not accounted for
	Household diet record	Usually covers one week with great detail; most accurate of household methods	High respondent burden; expensive
Individual	24-h Dietary recall	Detailed information on food intake, good estimate of means by subgroup	Misclassifies individuals; not useful for correlative investigation
	Multiple recalls	Average of multiple days can give good quantitative estimate of usual intake	Expensive
	Food frequency questionnaire	Inexpensive, measures usual intake	Semiquantitative, dependence on food list and recipe assumptions may lead to error in estimation of intake in some subgroups

They do not, however, provide information on individual intakes within the household, and are not useful for understanding age- and sex-specific intakes.

For a detailed understanding of food and nutrient intake by age, sex, and physiologic state, data are needed at the individual level. Surveys of individual dietary intake use methods that range from qualitative food checklists to multiple detailed records of food intake, with quantification of preparation methods and portion sizes. Data at the individual level are used for a variety of purposes and the design of the survey and level of detail utilized will depend on the primary data needs. At the national level, a primary objective is to identify subgroups at risk of inadequate intake of energy or specific nutrients. The important advantage of individual level data is that target age and sex groups may be identified in addition to groups identified by region or other household level characteristics. A further objective is to determine the extent of undernutrition in relation to energy or specific nutrients in the total and subpopulations. This requires consideration of the distribution of intakes in specific age, sex, and physiologic status groups. A more ambitious objective of some national or targeted dietary surveys is to associate aspects of individual dietary intake with the existence of health conditions. This objective requires that survey data be valid and reliable at the individual level and requires the estimation of usual intake by individuals.

Issues in Survey Design

National data on food availability is generally collected with food balance sheets. While not a survey in the formal sense, this is a collection of data from the food sector regarding wholesale distribution. After adjusting for expected losses and wastage, these data are compared to nutrient values and then to the size and composition of the population to calculate *per capita* nutrient availability. Because this is a crude assessment, it generally does not account for all losses or waste and therefore tends to overestimate availability.

In order for a household level survey to be nationally representative, one must carefully consider the sampling design. This is generally done by multilevel selection of regions, then sub-regions, then households, in such a way that the resulting data may be generalized to the national level. In cases where results are required at the regional level, coverage of all regions is necessary, although this will usually increase the cost of the survey. When the objective is more specific than national description, target areas may be selected, based on risk status or relevance to the question being addressed.

Similarly, for individual level data to be representative of the greater population of individuals, complex sample design is employed to be sure individuals are selected randomly. Decisions on sampling design will generally be a balance between equal opportunity for subject inclusion against logistic and cost considerations of full randomization. For that reason, multi-level complex sampling design is usually employed. This design may be similar to that of the household level survey, with the added step of randomly selecting individuals within households. Although some surveys do select households and then interview all members of the household, this decreases the generalizability of the individual data due to the lack of independence of the observations. Members in the same family, for example, will consume similar foods and therefore will be more like each other than like others in their community. Although this lack of independence can be adjusted in the analysis design, it will require larger total numbers of interviews to achieve representative stability of data estimates and is therefore not usually the most effective design. While the multistage approach of region, subregion, and community also leads to reduced power, this is corrected by consideration of the 'design effect,' which can be calculated by comparing the variation in intake within versus between sampling units at each level. Although the design effect may demand higher overall numbers of surveyed individuals, this is generally considerably less expensive than expanding coverage to all locations.

In addition to the multistage selection of respondents for representation of the general population, many surveys are concerned with subgroups that will not be well represented unless specifically over-sampled. Examples may include pregnant women, ethnic subgroups, or low-income groups. In these cases, individuals that meet the specified characteristic are identified within the existing sampling design, but are then selected in larger numbers than would be representative of the entire population. This allows sufficient sample size to present valid estimates for these groups. When included in measures of the total population, the extent of over-sampling by subgroups can be adjusted using weights that correct for what would otherwise be an over-representation of these groups.

Another design consideration relevant to accurate representation of dietary intake is the timing of the survey. In many countries, intake may vary

considerably by season, and it is therefore important that all seasons are represented. Although logistic and cost constraints often limit ideal design planning, it is also optimal if data from all seasons is collected in all survey locations, as opposed to certain regions being collected in the summer and others in the winter. If the latter is the case, comparisons across regions may be compromised. Similarly, intakes are known to differ by day of the week, and overall intakes may be misrepresented if certain days of the week are not included in the data collection plan.

Selection of Dietary Assessment Measure

Household Level

There are several alternative methods of dietary assessment that may be selected to assess intake. At the household level, a commonly used approach may be referred to as the food account method. A person in the household who is responsible for the acquisition and/or use of food is selected to keep a daily record of all the food that enters the household for a specified period – often 1 week. This includes household food purchases, food production, and food received as gifts during that period. This provides a general picture of the food that passes through the household in a given week. There are several limitations to this approach, including the assumption of constant food stores, which may not be the case.

In some locations, it is not feasible for many individuals to record this information accurately. In this case, an interviewer-administered list-recall method is often used. The interviewer asks the responsible household member to recall food purchases, production, or other receipt in the household during a specified period of preceding days, following a list of major foods that are relevant for that country or location. Additional information on age and sex of household members and number of meals each consume at home is collected to calculate adult equivalent *per capita* food availability for the household. Although edible portions of foods are generally considered in quantifying availability, most such surveys do not account for wastage, loss, or use by animals and therefore may overestimate household food use and availability. On the other hand, they do not generally account for food consumed away from the home, thereby underestimating food intake. While useful for economic and food commodity flow information, this type of survey is therefore limited with respect to nutritional intake assessment.

For the purpose of better understanding the dietary intake within households, more elaborate methods are needed. One approach is to use a household diet record. In this case the household respondent is asked not only to report inflows of food, but also to record actual use and preparation of foods in the household over a specified period of time. Food consumed outside the home may also be recorded for each household member as well as the number of individuals, including guests, who are present at each meal. While much more demanding for the respondent, this approach provides a better estimation of the total food consumed by the household than do the inventory methods described above. Estimation of waste is included in some but not all such surveys and is a limitation of most. Because of the heavy respondent burden, nonresponse or incomplete response is also a major problem. With high proportions of nonresponders, the generalizability of the survey is threatened.

Individual Level

A wide variety of methods are available for use at the individual level, and their selection depends on the questions to be addressed balanced with cost considerations. As noted above, major uses of individual intake data from dietary surveys include the description of mean intakes by subgroup, description of the proportions of the population with inadequate intake of specific foods or nutrients, and comparison of dietary intake with individual characteristics, including health status.

For the purpose of describing mean intake of groups and subgroups, the most efficient method is the use of a single 24-h dietary recall per selected individual. For estimating group means, it is most effective to include sufficient numbers of individuals per subgroup to be represented as opposed to completing multiple recalls on a smaller number of individuals. This is the methodology that was, until recently, used in the US National Health and Nutrition Examination Survey (NHANES), providing a good description of average intakes of nutrients by age, sex and ethnic groups. As an aggregate measure, this design has worked very well. However, there are limitations to these estimates and validation against quantified energy expenditure measurements has shown that most people tend to under-report intakes with the 24-h dietary recall method. Further research is needed to better understand whether this tendency to under-report is random or, as some investigations have shown, associated with individual characteristics

such as obesity, restrained eating behavior, or social desirability bias in reporting.

The major limitation in the use of single 24-h recalls in a dietary survey is the misclassification of individuals that results from day-to-day variation in individual intake. An individual who on average is a heavy consumer of energy and fat, for example, may on any single day, eat uncharacteristically lightly or vice versa, leading to severe misclassification of individuals in the intake distribution. Although the mean intake is reliable and reasonably valid, the tails of the intake distribution with a single day of intake per person are extended, leading to overestimation of proportions either above or below a specified cut-off point, relative to what is seen when usual intake is assessed as the average of multiple days. The misclassification of individuals relative to their actual usual intake also severely limits the ability to correlate intake data with any individual characteristics, including health status and biomarkers.

Because of the importance of using national or regional survey data to identify the extent of inadequate nutrient intake, there has been considerable discussion on how to assess diet efficiently, yet estimate prevalence of inadequacy. As noted above, the distribution of intake obtained with a single day per individual is extended due to day-to-day variability in intakes. However, with repeated recalls on a representative subset of the population, this day-to-day variability may be quantified and used to adjust the distribution to better represent usual intake. Although they require specialized training to use, statistical methods have been developed to adjust distributions for this purpose. By adjusting the distribution, we are able to pull in the tails and get a more realistic distribution of usual intake and, thereby, a more accurate estimate of the proportion that falls below or above a specified cut-off point.

Because recommended dietary allowances (RDAs) are set as a guideline for nutrient intake that will meet the requirements of most healthy individuals in the population, a cut-off point of two-thirds of the RDA has frequently been used for assessing the proportion with low intake. A more precise way to estimate the relationship between intake and actual requirements was proposed that uses information on the probability that a specific nutrient is inadequate. This method requires information on the requirement distribution for that nutrient, which is then compared with an intake distribution. In actuality, we still have only limited data on the distribution of requirements. In the most current US nutrient intake recommendations, estimated average requirements (EARs) are proposed as well as RDAs. In general, the proportion of individuals who fall below this EAR, using an intake distribution adjusted for day-to-day variability, are a good estimate of the proportion of the population with inadequate intake.

A third important objective of dietary survey data is to gain a better understanding of the correlates of nutrient intake, but with respect to individual characteristics that may be associated with lower versus higher intake and the extent to which intake is associated with indicators of health. For many nutrients, the day-to-day variation in intake is considerable, and multiple days would be required to achieve stable estimates of intake at the individual level. Without this, the misclassification of individuals in the distribution leads to a weakening in the ability to see associations that may truly be there. An extreme example is vitamin A, which tends to be concentrated in a few foods. If one frequently has liver and carrots, but happened not to on the day of the recall, that individual would be classified as having low vitamin A intake when their usual intake is quite large. Conversely, one who almost never eats these foods, but happened to have liver on the day of the recall would be placed at the upper end of the vitamin A distribution despite the fact that this may have been the only time that year they consumed such a high vitamin A source food. The effect of this is to weaken correlations or regression coefficients so that no association may be seen between intake and outcomes such as plasma retinal or eye health.

There are two ways that this major limitation may be handled in dietary surveys. First, with information from multiple recalls on a subset, it is possible to calculate the ratio of the intra-individual variance, or day-to-day intake variation within individuals, to the inter-individual variance, or difference in intake across individuals. To the extent that intra/inter-individual variance ratios are large, as in the case of vitamin A, the ability to see associations will be severely limited. When inter-individual differences exceed day-to-day variation within subjects, the ability to correlate intakes with other factors is stronger. Unfortunately, in most cases, this variance ratio is sufficiently large that a single day of intake will not allow correlational analyses. The collection and averaging of multiple days of intake will greatly improve this situation. For most nutrients 3 or 4 days are acceptable. However, for some nutrients of interest, including vitamin A and vitamin B_{12}, the variability ratios are so high that an unrealistic number of days is needed for stable estimates. For this reason, many researchers choose to use food frequency questionnaires when correlational analysis is a major objective.

The food frequency questionnaire asks respondents to report the frequency of consumption of a prespecified list of specific foods. Additional questions on portion size and on preparation methods are added in differing ways to different questionnaires. Food frequency questionnaires provide a lower cost alternative to multiple recall days, but also have limitations. Because they rely on a food list, their validity is dependent on the representativeness of that list, and of portion size and recipe assumptions. Most food frequencies in wide use have been developed using data that represent the major sources of nutrient intake in a population. However, individuals with divergent eating patterns will not be well represented using this tool.

When single recalls are available along with information on the variance ratios for each nutrient, another approach to improving correlational estimates is to use what is called 'deattenuation' methods to correct for the effect of day-to-day variability. Because the effect of this variability is assumed to be random in the population, the ratios provide a mathematical basis from which to estimate what the 'true' association between the nutrient intake and correlated indicator may have been, after accounting for this variability.

Data Analysis and Limitations

Whatever dietary assessment measure is used, the utility of the data is dependent on the translation of reported food intake to nutrient intake. This requires detailed and accurate nutrient databases. The US Department of Agriculture has the most extensive nutrient database in the world, allowing for good estimation of dietary intakes in the US. Most other countries have not conducted this level of food composition analysis for their own locations. Therefore, most databases used throughout the world have obtained at least some of their values from the US nutrient database, adding information as possible from locally analyzed products. However, because the nutrient composition of many foods, including fruits, vegetables, and even animal products can vary widely by growing conditions and specific subvariety, most available nutrient databases remain inadequate. Many use extrapolated values from similar foods when chemical analysis has not been completed. Furthermore, it is common for many country-specific databases to include information only on macronutrients and a few selected vitamins and minerals. The continual arrival of new manufactured products also complicates the upkeep and management of food composition databases. Consequently, considerable database work remains to expand the utility of worldwide dietary surveys.

Once the nutrient data are calculated, data are generally tabulated to present age, sex and, sometimes, ethnic specific mean intakes and standard deviations. Further disaggregation by region, socioeconomic group, or other group characteristic can by very helpful in understanding the macro-distribution of nutrient intake and for targeting specific groups with nutrition intervention programs. If a complex survey design was used, or if systematic nonparticipation was observed, sampling weights must be applied to adjust the means and standard errors.

Using the methods described above, estimates of the population with low intakes of specific nutrients are also calculated. Beyond these descriptive measures, comparison of nutrient intakes with individual characteristics and health measures generally requires multiple regression analysis with appropriate adjustment for potentially confounding variables. Again, when complex survey designs have been used, the inclusion of sample weights and appropriate adjustment of variances is needed. Specialized statistical software for use with survey data is available. In cases where a single recall has been used, substantial weakening, or attenuation, of associations is likely, but use of 'deattenuation' methods can, at a minimum, provide information on the likely extent of this attenuation. When food frequency data are used, it is important to include some validation methods, preferably with comparison to key biomarkers of nutritional status, but at least to multiple recalls on a subset. This is particularly true when a new questionnaire is being used, but is also important over time as the food supply and food habits change in the population.

See also: **Dietary Intake Measurement**: Methodology; Validation. **Nutritional Assessment**: Anthropometry; Biochemical Indices. **Nutritional Surveillance**: Developed Countries; Developing Countries. **Vitamin A**: Biochemistry and Physiological Role.

Further Reading

Briefel RB and Sempos CT (eds.) (1992) *Dietary Methodology Workshop for the Third National Health and Nutrition Examination Survey*. Vital Health Stat 4 (27). Hyattsville, MD: National Center for Health Statistics.

Food and Nutrition Board, Institute of Medicine (2001) *Dietary Reference Intakes: Applications in Dietary Assessment*. Washington, DC: National Academy Press.

Gibson RS (1990) *Principles of Nutritional Assessment*. New York: Oxford University Press.

Interagency Board for Nutrition Monitoring and Related Research (1995) *Third Report on Nutrition Monitoring in*

the United States, vol. 2. Washington, DC: US Department of Agriculture.
Murphy SP (2003) Collection and analysis of intake data from the Integrated Survey. *Journal of Nutrition* **133**: 585S–589S.
Subcommittee on Criteria for Dietary Evaluation, National Research Council (1986) *Nutrient Adequacy: Assessment Using Food Consumption Surveys*. Washington, DC: National Academy Press.
Thompson FE and Byers T (1994) Dietary assessment resource manual. *Journal of Nutrition* **124**: 2245S–2317S.
Willett W (1998) *Nutritional Epidemiology*, 2nd edn. New York: Oxford University Press.

DIETETICS

P A Judd, University of Central Lancashire, Preston, UK

This article briefly relates the history of the development of dietetics, discusses the changing roles of the dietitian, and outlines the current involvement of dietitians in some general and specialist areas of practice.

Definition and History of Dietetics

Dietetics is defined as 'the application of the science of nutrition to the human being in health and disease.' However, the term 'dietitian,' used to describe a practitioner of dietetics, was in use long before the science of nutrition had become an accepted discipline. The first use of the title of dietitian was recorded in 1899 in the United States when the dietitian was described as 'a person working in a hospital who provided nutritious meals to patients.' The earliest dietitians were therefore mainly concerned with provision of food and usually trained as home economists. The role of the dietitian has changed markedly in the past 50 years, and the dietitian is now accepted as the expert in the planning and evaluation of nutritional care for patients requiring therapeutic dietary regimens as well as for the population in general.

The profession of dietetics is a relatively young one, first formalized in the United States in 1917 with the foundation of the American Dietetic Association (ADA). In the United Kingdom, the first dietitians were nurses and the first dietetic department opened in the Edinburgh Royal Infirmary in 1924. The British Dietetic Association (BDA) was established in 1936. The profession developed rapidly in other countries, and in 2004 there were 23 dietetic associations registered with the European Federation of the Association of Dietitians and 36 national dietetic associations registered with the International Committee of Dietetic Associations.

The Role of the Dietitian

The first dietitians (with the exception of those concerned mainly with food service provision) worked mainly in hospitals. Clinical dietetics and the acute hospital service still claim a large proportion of the graduates from dietetics but other areas of work are increasingly becoming more important. In the United Kingdom, changes in the emphasis of health care, particularly the change from acute (hospital) care to care in the primary health care setting, has resulted in a marked increase in the number of dietitians based in primary care.

Dietitians also have many other roles outside the health services. Increasingly practitioners work with government agencies, for example, in dietary surveys of the population, in execution and evaluation of nutrition intervention programs, and advising on the practical application of policy. In industry, they may work as advisors to food companies, wholesale and retail suppliers of food, and with companies producing specialized dietary products. In addition, dietitians are increasingly working independently as consultants, for example, in private practice, journalism, and sports nutrition. The scope of the dietitians' work is illustrated by **Table 1**, which lists the special interest groups for dietitians in the United Kingdom.

Whatever aspect of work a dietitian chooses, one of his or her primary roles will be that of an educator, whether this be in assisting individuals to understand and apply a therapeutic regimen; teaching doctors, nurses, or other health professionals about nutrition and dietetics so that they may carry out their own role more effectively; teaching groups of people about aspects of preventative nutrition; or writing an article for the scientific or lay press. The ability to communicate is therefore central to every dietitian's role.

Table 1 Specialist groups and special interest groups in the British Diatetic Association (2004)

Specialist groups	*Special interest groups*
Community Nutrition Group	Burns Interest Group
Diabetes Management and Education Group	D – Liver (Dietitians Working in Liver Disease)
Dietitians in HIV and AIDS	Dietitians in Neurological Therapy
Freelance Dietitians Group	Dietitians in Sport and Exercise Nutrition Group
Mental Health Group	Dietitians Working in Obesity Management (UK)
Multicultural Nutrition Group	Gastroenterology Interest Group
Food Counts	Northern Eating Disorders Dietitians Interest Group
Nutritional Advisory Group for Elderly People	Nutritionists in Industry
National Dietetic Managers Group	Pediatric Renal Interest Group
Oncology Group	UK Dietitians Cystic Fibrosis (CF) Interest Group
Pediatric Group	
Parenteral and Enteral Nutrition Group	
Renal Nutrition Group	
UK Heart and Thoracic Dietitians	

The Dietitian's Role in Food Service

In the United States and countries that follow the US model hospital dietitians work in either administrative or clinical (therapeutic) areas. Administrative dietitians manage the provision of food services for all patients and staff. They are responsible for food production and quality control in the delivery of the hospital meal service as well as ensuring their nutritional adequacy. They are also often responsible for budgeting and staffing of the dietary departments and usually relate to other administrators and managers, having little or no direct contact with patients or medical staff. The clinical dietitian is the person who has direct contact with patients and the medical and paramedical staff involved in their care.

In the United Kingdom, very few dietitians have overall responsibility for food service. However, there is usually close liaison between the dietitians and the catering manager in hospital practice to ensure the provision of nutritionally sound selective menus. The dietetic manager will also be consulted on matters of policy, such as the implications of changes in food preparation systems or the introduction of healthy eating policies. It is accepted now that many patients (especially elderly people) are malnourished when they enter the hospital and that this malnourishment may get worse during the hospital stay. Dietitians are therefore very much involved in attempts to ensure that every patient has a nutritionally adequate diet.

Clinical or Therapeutic Dietetics

The term 'therapeutic dietetics' is used in the United Kingdom to describe the work of the dietitian in his or her direct dealings with patients who require 'special diets' for various reasons. In the United States, the corresponding term is 'medical nutrition therapy' and the importance of this is recognized by its acceptance as allowable treatment by medical insurance companies. Medical nutrition therapy includes nutritional diagnostic, therapeutic, and counselling services provided by a registered dietitian and can

effectively treat and manage disease conditions;
reduce or eliminate the need for prescription drug use;
help reduce complications in patients with disease; and
improve patients' overall health and quality of life.

The role of the clinical dietitian has broadened in the past decade, with respect to both the range of conditions that are encountered and the setting in which the work is done. In the past, the role of the therapeutic dietitian was to calculate, teach, and facilitate compliance to a range of dietary regimens prescribed by medical or surgical practitioners for specific disorders. These functions are still important and are extending as improvements in medicine enable patients who would previously have died in childhood to survive to adulthood with continuing needs for nutritional care. For example, patients with cystic fibrosis require nutritional support as their lungs deteriorate and women with hyperphenylalaninemia need dietary advice in order to help them achieve successful pregnancies. The dietitian is an important member of the multidisciplinary team dealing with many clinical conditions.

In addition to these roles, dietitians are increasingly involved in the assessment and support of patients not traditionally seen as requiring a therapeutic diet. For example, planning and implementing feeding for patients who are nutritionally compromised as a result of aging, trauma, surgery, or chronic illnesses such as cancer or AIDS is now an important part of the workload of the dietitian working in both acute and community settings.

Clinical dietitians are increasingly involved in decisions about the appropriateness of particular dietary

regimens or the type of nutritional support required and in many institutions are responsible for prescribing the patients' diet in consultation with the physician or surgeon, who recognises the dietitian as the expert. In acute care a dietitian is an important member of the nutrition support team, working alongside nurses, pharmacists, and medical staff to advise on the feeding of all patients who need nutritional support.

The dietitian must be aware of advances in clinical practice, and many dietitians are now involved both in research on the development of new treatments and in evaluating current practice. In the past decade, the requirement to demonstrate evidence-based practice and satisfy the demands of clinical governance within acute care has increased the need for understanding and contribution to the development of the evidence base. Dietitians are increasingly obtaining research qualifications and initiating their own projects, working alongside experienced investigators and in multidisciplinary teams.

In the United Kingdom, consultant posts for Allied Health Professionals, who are considered to be 'experts in a specialist clinical field, bringing innovation, personal mastery, and influence to clinical leadership and strategic direction,' have recently been introduced. The consultants are expected to have exceptional skills and advanced levels of clinical judgment, knowledge, and experience and be able to enhance quality in areas of assessment, diagnosis, management, and evaluation, improving patient outcomes and extending the dietitian's role. Although in early 2004 there were only two such dietetic posts (in diabetes and oncology), the number will doubtless increase.

In the United Kingdom a shortage of dietitians and the recognition that many of the tasks carried out by dietitians did not require such highly qualified practitioners have resulted in the recent official recognition of the role of assistant practitioner by the BDA. Dietetic assistants (or community nutrition assistants) are now working under the supervision of registered dietitians, carrying out many of the tasks that do not require specialist input. In the United States and some other countries, diet technicians have been undertaking these roles for some years and the American Dietetic Association has recognized training programs for them; such training programs are being developed in the United Kingdom.

Examples of Specialist Roles in Dietetics

Renal Dietetics

Renal dietitians are usually attached to specialized renal units and are an integral part of the team involved in the treatment of people suffering from varying degrees of renal impairment, whether acute or chronic. In the United States, there is a legal requirement related to funding of patient care that states that a qualified dietitian must be part of the professional team that develops long- and short-term care plans for renal patients. The dietitian, together with the nephrologist, has responsibility for nutritional assessment, the diet prescription, and for monitoring responses to treatment. In addition, he or she must be able to devise appropriate individualised dietary plans, taking into account any other ongoing disease processes or conditions (e.g., diabetes mellitus), and will teach the patient and family how to manipulate the diet. The dietitian will also monitor dietary compliance. A thorough knowledge of physiology and the pathological processes involved in the various kidney diseases and an ability to interpret the patient's biochemical data are therefore essential for the renal dietitian.

Some disorders may resolve with treatment, whereas others may become chronic and result in permanent kidney failure. In progressive renal disease, patients may initially be managed using diet and drugs alone but as the kidneys fail will require replacement therapy, such as peritoneal dialysis, hemodialysis, or transplantation. Each of these stages requires different dietary treatment. The dietitian will deal with a variety of patients with different types of disease and at different stages of progression, with different needs with respect to diet, and will also have to teach the patient how to cope with changes in diet that follow as he or she changes from one treatment to another.

Nutritional Support

Nutritional support of patients who are unable to feed themselves adequately by the normal oral route is an important area of practice. The dietitian's involvement will include the assessment of the nutritional status of the patient, decisions about the most suitable method of nutritional support, and advising on provision of appropriate nutrition. This may range from prescribing oral supplements for the patient who cannot eat enough to designing and advising on complete parenteral nutrition regimens for the unconscious patient in intensive care. Between these two extremes will be the patients who need enteral feeds to provide complete or supplementary nutrition for a variety of reasons and for periods varying from a few days to a lifetime.

Patients requiring nutritional support may be acutely ill or may require long-term feeding, sometimes at home. The dietitian must be able to assess the nutritional requirements of each individual and

design appropriate feeding regimens in all circumstances. Many patients will be sent home while still being tube fed, either enterally or parenterally, and the community dietitian as well as the dietitian in the acute hospital will both be closely involved in the patient's care and in monitoring progress.

Diabetes Care and Education

Traditionally, the dietitian always had an important role in the treatment of patients with diabetes mellitus (DM), and the radical changes in dietary approaches during the past 15–20 years have emphasized this role. In the latter part of the twentieth century there was a move away from the use of diets low in carbohydrate in the treatment of diabetes and the basis for the treatment of all people with DM, whether young and insulin dependent or older and treated by diet alone or diet and hypoglycemic drugs, tended to rely on the supply of an appropriate amount of energy as a low-fat diet with at least 50% of energy from foods rich in complex carbohydrates and nonstarch polysaccharides. In addition, the recognition that similar amounts of carbohydrate from different foods have different effects on blood glucose levels, and that this response may be further effected by other foods eaten with them, has led to less stress on absolute intakes of carbohydrate and more toward a qualitative approach to the diet.

Many established diabetics found the change in dietary treatment confusing and have needed help in switching to the new regimen, and newly diagnosed patients also need help in learning to manipulate and control this lifelong disorder. Many dietitians have also been involved in the research that has underpinned the progression of dietary treatment and are now evaluating its effects.

HIV and AIDS

The dietetic care of patients with HIV and AIDS has become increasingly important in the past decade. Dietitians may work with people who are HIV positive to help them optimize their nutritional status and resist the opportunistic infections that eventually cause death in patients with frank AIDS. Once the person develops AIDS, the dietitian's role becomes both therapeutic and palliative, devising and implementing with the patient regimens that enable him or her to satisfy nutritional requirements when, for example, disease of the gastrointestinal tract results in multiple malabsorption or cancer results in weight loss and anorexia. The advent of new, multidrug regimens that require careful planning of meals to match their absorption characteristics and that have side effects affecting nutritional status has made the dietitian's role even more important.

Pediatric Dietetics

The pediatric dietitian has a unique role in that they have to combine the metabolic requirements of the disease process or condition with the normal requirements for growth and development. With the advances in early diagnosis of many complex metabolic conditions, children may require complicated diets that are very different from those of the rest of their family and peers, need constant modification as the child grows, and may be lifelong. The dietitian is responsible for modifying the diet as necessary to take account of the patient's metabolic requirements, any feeding difficulties, mechanical or physiological, and the patient's food preferences and dislikes as he or she grows. The dietitian is an essential part of the support system for children with inborn errors of metabolism such as phenylketonuria and cystic fibrosis, conditions such as renal or heart disease, food allergies, diabetes, and many others, being able to tailor the diet to the patient's specific needs and having access to information about special foods and products of which the care provider may be unaware or have difficulty locating.

The increase in overweight and obese children and adolescents and the consequent increase in type 2 diabetes in young people is a new challenge for the pediatric dietitian.

The pediatric dietitian has an important role as an educator, often teaching the child's parents initially and later the child how to cope with the constraints of a special diet both at home and at school. As is the case with adult patients, the dietitian will often be able to put the child and family in contact with support groups in which newly diagnosed patients or parents will be helped by others with firsthand experience of the disease and its treatment.

Other Areas of Specialization

There are many other areas of clinical dietetics in which individuals may specialize, including obesity, oncology, liver disease, gastroenterology, eating disorders, gerontology, and care of the mentally ill or mentally handicapped. Many of the activities in these areas and those described in the previous sections are not confined to the hospital but require input from the dietitian in the community.

Dietetics in the Community

Community dietitians fulfill a variety of roles that may range from working mainly in clinical dietetics in the community setting (e.g., advising a variety of

patients in a general practitioner's clinic) to being a public health nutritionist advising the local health or social services on aspects of food policy. In recent years, due to the changing emphasis in health care, the number of community dietitians in the United Kingdom has increased markedly. The main focus of the increase has been in supplying clinical (therapeutic) care to the primary health care setting and approximately 21% of UK community dietitians surveyed in 1997 stated that this was their only role. Approximately half of the respondents to the survey were involved in health promotion activity as well as clinical work, but only 9% worked solely in health promotion.

Many dietitians working in health promotion achieve their objectives by educating other professional groups, such as doctors, nurses, health visitors, and midwives, who will then pass the specific knowledge on to the individual or groups of patients. Prevention of diseases that may be diet related has recently become a much more important issue, and dietitians are working with schools, health education departments, and industry to try to educate the public to consume a healthier diet. Dietitians also work as advisors in government departments and are therefore involved in planning nutrition policies for the country as a whole.

Dietitians in Research and Education

The advances in all areas of nutritional knowledge, both in terms of achieving optimal health and prevention of disease and in therapeutic nutrition, have led to an increasing number of dietitians working in research. The combination of nutritional and medical knowledge and the ability to translate these into terms of foods eaten means that the dietitian has a unique role to play. Dietitians often approach research from a deductive perspective (i.e., in order to understand or solve difficulties observed in practice) and the results of this research will often have practical significance and can be incorporated into treatments. Evaluation of practice can also be considered as part of this deductive process and is essential in the current health care climate, in which increasing reliance is put on measuring effectiveness and the use of evidence-based medicine. In addition, dietitians are increasingly involved in the basic experimental and analytical scientific research that is essential for nutrition and dietetics to advance in both clinical and nonclinical areas. Involvement in research has led to registration for higher degrees and the number of dietitians with masters' degrees or doctorates is now considerable in countries such as the United States, United Kingdom, Canada, and Australia.

Research is also seen as an important part of the role of those dietitians employed in universities and colleges to teach dietetic and other students. In the United Kingdom and other countries, there is a requirement that each dietetic training course has registered dietitians on the staff, and in many cases these people also work in the NHS in order to keep up-to-date with current practice. Dietitians have been involved for many years in the education of other professional groups, including nurses, midwives, and pharmacists. Recently, advances have been made in convincing those in charge of medical education at undergraduate and postgraduate levels of the importance of nutrition in medical education, and this is also seen as an important area in which dietitians should be involved.

Dietetics Education and Training

The following quotation from the introduction of *The Manual of Dietetic Practice* (Thomas, 2001) summarizes some of the skills needed by a dietitian—all of which must be acquired during preregistration and continuing professional development. The emphasis here is on the role of the clinical dietitian; additional skills will be acquired by the dietitian working in public health and policy:

> While principles of care can be standardised, the way in which they are applied has to vary to take account of individual needs, problems, lifestyle, associated health risks, and readiness to change. In order to provide effective care the modern dietitian has to exercise considerable clinical judgement in deciding how a specific set of circumstances may be most appropriately managed. This requires more than just nutritional knowledge. The modern-day dietitian has to make a global risk assessment when setting nutritional goals, have an understanding of human behaviour in order to achieve dietary change, acquire the interviewing and counselling skills necessary for meaningful dialogue between patient and professional, and have the ability to evaluate whether objectives have been achieved.
>
> (page x)

It is now therefore accepted that the practice of dietetics requires a wide range of knowledge and skills that are achieved by both academic study and practice learning. As the scope of dietetic practice has expanded over the years, the preregistration education and training programs have continually adapted to ensure that the practitioner has the current knowledge and skills required. The education and training of a dietitian usually comprises a degree program (either BSc or MSc), based in a university, including or followed by a period of practical

training (or internship) based in recognized hospital dietetic departments.

Preregistration programs include coverage of basic and applied sciences (chemistry, biochemistry, physiology, nutrition, and microbiology) as well as social sciences (psychology and sociology). In addition, because dietetics is concerned with feeding people, a knowledge of the food habits of populations together with detailed knowledge of food composition and food preparation is essential. To this basic foundation is added knowledge of medicine, pathology, and the therapeutic uses of dietary treatment and, increasingly emphasized, the development of skills required to communicate with all types of people whether counselling individuals or teaching groups.

During the practical training or internship, the student dietitian learns to apply the theory learned at university with individuals or groups of people. The training covers all aspects of dietetic practice and the students spend time in different settings, including community care and, in some countries, large-scale catering establishments. In order to become a registered practitioner, the students must demonstrate that they have both good theoretical knowledge and are competent practically.

In the United States, United Kingdom, Canada, Australia, New Zealand, South Africa, The Netherlands, and many other countries training programs are regulated by bodies external to the educational establishments and successful completion of such a regulated training allows registration as a dietitian. In the United States, regulation of courses and training programs is carried out by the ADA and in the United Kingdom by the Health Professions's Council (HPC), in conjunction with the Quality Assurance Agency of the Higher Education Funding Council. In the United Kingdom, only registered dietitians may be employed in the National Health Service. The registration body, in each case, produces a statement of conduct that describes the role and responsibilities of the registered dietitian, and failure to work within this statement of conduct may result in disciplinary action and removal from the register. In the United Kingdom, since 2003 this code of conduct has been presented as 'standards of proficiency' and on registration the registered dietitian must sign a document that involves taking responsibility to work only in areas in which he or she is competent to practice.

Registration in one country does not automatically mean that a dietitian can work elsewhere in the world because levels of education and training are not always comparable from country to country. Within Europe, for example, the education level and skills of dietitians vary widely and there is currently a move within the European Federation of Dietitians to develop benchmarks for dietetic qualifications. The registering body will therefore consider applications from dietitians from other countries and suggest further training if appropriate.

Continuing education and demonstration of continuing competence to practice are increasingly being seen as vital in this rapidly changing profession; in the United States there has long been a requirement to demonstrate continuing education, and continuing registration is dependent on this. In the United Kingdom, it will soon become a requirement to demonstrate continued competence to practice for continued registration with the HPC. The BDA, the professional association for dietitians, has well-developed systems for assisting dietitians in both accessing and recording continuous professional development. The provision of validated specialist courses, the development of a Diploma in Advanced Dietetic Practice, which recognizes CPD over a 5-year period, and most recently the support for a Masters Course in Advanced Dietetic Practice are examples of this. In Australia, continuing professional development is recognized by the status of Accredited Practising Dietitian.

See also: **Arthritis**. **Burns Patients**. **Children**: Nutritional Requirements; Nutritional Problems. **Celiac Disease**. **Colon**: Nutritional Management of Disorders. **Cystic Fibrosis**. **Diabetes Mellitus**: Dietary Management. **Food Allergies**: Diagnosis and Management. **Gall Bladder Disorders**. **Gout**. **Handicap**: Down's Syndrome. **Hyperlipidemia**: Nutritional Management. **Hypertension**: Nutritional Management. **Inborn Errors of Metabolism**: Classification and Biochemical Aspects; Nutritional Management of Phenylketonuria. **Infection**: Nutritional Management in Adults. **Low Birthweight and Preterm Infants**: Nutritional Management. **Obesity**: Prevention; Treatment. **Older People**: Nutritional Management of Geriatric Patients. **Stroke, Nutritional Management**. **Surgery**: Long-term Nutritional Management.

Further Reading

American Dietetic Association (1995) Position of the American Dietetic Association: Cost-effectiveness of medical nutrition therapy. *Journal of the American Dietetic Association* **95**(1): 88–91.

American Dietetic Association (2003) Position of the American Dietetic Association: Integration of medical nutrition therapy and pharmacotherapy. *Journal of the American Dietetic Association* **100**(10): 1363–1370.

Bateman EC (1986) *A History of the British Dietetic Association. The Second Twenty-Five Years 1936–1986*. Sunderland, UK: Edward Thompson.

Council for Professions (2000) *Supplementary to Medicine: Dietitian's Board Pre-registration Education and Training*. London: Council for Professions.

Department of Health (1994) *Targeting Practice: The Contribution of State Registered Dietitians. Health of the Nation*. London: HMSO.

Department of Health (1994) *Nutrition Core Curriculum for Nutrition Education of Health Professionals*. London: HMSO.

Fox C (1999) *Community Dietetics: Supporting the Future*. Birmingham, UK: Community Nutrition Group of the British Dietetic Association.

Judd PA, Butson S, Hunt P *et al.* (1997) Pre-registration training for dietitians—Report of the Dietitians Board/BDA working group on pre-registration training. *Journal of Human Nutrition and Dietetics* **10**: 157–162.

Thomas B (ed.) (2001) *The Manual of Dietetic Practice*, 3rd edn. Oxford: Blackwell Science.

Digestibility *see* **Bioavailability**

DRUG–NUTRIENT INTERACTIONS

K G Conner, Johns Hopkins Hospital, Baltimore, MD, USA

Introduction

Understanding the interactions between dietary constituents and pharmacological compounds is essential to monitor drug therapy correctly and to assess the potential nutritional impact of medications. Most therapeutic agents exhibit some form of interaction that ultimately affects the nutritional status of the host, by altering absorption or utilization of nutrients. Frequently, these changes are not readily identified or may be obscured by the underlying disease.

The interactions between therapeutic agents and nutrients are part of the large number of interactions occurring between nutritional and non-nutritional constituents of the human diet. These constituents include all substances added to the food chain – incidentally or deliberately – during harvesting, processing, packaging, distribution, and preparation of foods. Some examples are pesticides, food additives, antibiotics, hormones, and environmental toxins.

Drug–nutrient interactions operate in two directions: drugs can have a significant impact on nutrient absorption and utilization, and the nutritional status of the host affects the drug's ability to be absorbed and transported and to exert an effect on the target tissues.

Drug–nutrient interactions can be broadly classified into two categories: direct physicochemical interaction and physiological or functional interaction. Drug–nutrient interactions can also be classified according to their site of occurrence: within the food matrix, in the gastrointestinal (GI) tract, or during transport, metabolism, and excretion. The mechanisms and sites of drug–nutrient interactions are listed in **Table 1**.

Table 1 Mechanisms and sites of drug–nutrient interactions

Site	*Mechanism*	*Effect*
Food matrix	Binding and chelation	Decreases bioavailability
Gastrointestinal tract	Changes in gastrointestinal motility, binding and chelation, bile-acid concentration, and gastric pH	Increase in transit time reduces absorption, decreases bioavailability, and reduces absorption of fat-soluble nutrients
Circulation	Albumin concentration	Affects absorption of iron, vitamin B_{12}, and other substances
	Competitors for albumin binding	Decreases transport of bound substances; displaces albumin-bound nutrients (fatty acids, tryptophan, etc.)
Target tissues	Antagonistic effects	May increase requirements for antagonized nutrients
	Enzyme activities	Reduced concentration of enzyme product
Excretion	Renal function	Increased excretion may lower nutrient levels, increasing requirements
	Sequestration	As above

Physicochemical Interactions

Physicochemical interactions usually involve some form of molecular interaction between the drug and a nutrient, and occur primarily during digestion and absorption. The usual consequence of this interaction is a reduction in the bioavailability of the drug and/or the nutrient. A well-known example of this is the binding of metals by the antibiotic tetracycline.

Functional Interactions

Functional Interactions in the Gastrointestinal Tract

Functional interactions in the GI tract are particularly significant because alterations in GI function are likely to affect the digestion and absorption both of the drug and of a number of nutrients. The most common GI functional effects are as follows.

Changes in GI motility A reduction in transit time may lead to decreased absorption. There are a large number of drugs that affect gut motility, whether this is their primary therapeutic effect or not. Conversely, food composition also affects motility. Dietary fiber not only increases motility but also may trap other nutrients and drugs and reduce their bioavailability.

Changes in gastric-acid output Reduced production of chloride with a subsequent increase in gastric pH retards gastric emptying and may alter the balance between ionized and nonionized forms of therapeutic agents.

Reduction in the concentration of bile acids A reduction in the concentration of bile acids will affect the absorption of most fat-soluble compounds. Lower bile-acid concentration may result from increased binding and excretion or from decreased production. For example, the antibiotic neomycin binds to bile acids and increases their faecal excretion, thus reducing their luminal concentration and, in this fashion, decreasing the absorption of fat-soluble vitamins. This interaction, like many others, can be used therapeutically to reduce bile-acid turnover in patients with certain liver diseases and to lower cholesterol levels by reducing their reabsorption.

Alterations in the GI microflora Alterations in the GI microflora may affect the availability of nutrients produced by the normal gut flora, such as vitamin B_{12}′. Since many drugs are susceptible to bacterial metabolism, changes in the gut flora may also affect drug bioavailability. In certain cases, drug cleavage by intestinal microorganisms is an expected and necessary step for adequate drug action. For example, the anti-inflammatory agent 5-aminosalicylic acid is given as its precursor sulfasalazine, which is converted into the active compound by colonic bacteria. An altered colonic flora will affect the production of the active compound. Drugs can also affect nutrient absorption by directly inhibiting protein synthesis in the enterocyte. Since most transport systems require active protein synthesis and turnover, such inhibition results in a decreased rate of nutrient absorption. Furthermore, certain drugs undergo initial metabolism in the enterocyte, before reaching the bloodstream. Alterations in protein synthesis in the enterocyte, or an impaired turnover of the intestinal epithelia, will also affect this process.

Interactions Affecting Transport, Metabolism, and Excretion

Functional Synergism or Antagonism

The biological actions of nutrients and drugs can be synergistic or antagonistic, occur at different times after exposure, and affect a variety of target tissues. Some of the most common mechanisms are as follows.

Alterations in drug transport Drugs circulate in the bloodstream as free compounds or bound to other constituents, usually proteins. Drugs vary greatly in their propensity to bind to circulating proteins, covering virtually the entire spectrum from 0 to 100%. For a given drug, the bound fraction tends to be relatively constant under physiological conditions, but responds to changes in pH, electrolyte balance, and the presence of competing molecules. The major transport protein in plasma is albumin, and its concentration and the presence of other compounds with an affinity for albumin binding will affect the amount of drug that will ultimately be transported by this protein.

Increase in nutrient catabolism Certain drugs stimulate detoxifying systems, such as the cytochrome P-450 pathway. Activation of this system may result in increased catabolism of certain nutrients. In other cases, drugs directly affect nutrient catabolism, as in the case of anticonvulsant drugs, which stimulate vitamin D catabolism in the liver.

Changes in drug metabolism Certain nutrients (such as those found in grapefruit) can inhibit the activity of cytochrome P-450. Cytochrome P-450-3A is the only isoform affected in a clinically significant way. The mucosal cells of the small intestine are affected to a greater degree than the hepatic cytochrome P-450-3A. Certain HMG-CoA reductase

inhibitors (statins), simavastin, and lovastatin can have a significant interaction with grapefruit juice.

Biological antagonism Biological antagonism occurs when drug and nutrient have opposite biological actions, as is the case, for example, with vitamin K and salicylates in the coagulation process.

Increased nutrient losses Many drugs directly or indirectly enhance the urinary excretion of nutrients. Examples are the increased urinary losses of electrolytes caused by aminoglycoside antibiotics and amphotericin B antifungals, and the increase in urinary ascorbic acid excretion induced by barbiturates.

Host-related Functional Interactions

Nutrients and nutritional status can also affect drug action and disposition. Perhaps the most significant host-related factor affecting drug disposition is protein synthesis. Altered protein synthesis, usually resulting from insufficient dietary protein intake or severe diseases, will affect absorption, transport, metabolism, and excretion, as these are all protein-dependent processes. The role of plasma albumin in drug transport was discussed above and will certainly be affected by impaired albumin synthesis and/or sequestration in the extravascular space, as seen in protein-energy malnutrition. It should be noted, however, that malnutrition affects many aspects of drug metabolism, not all in the same direction. For example, drug delivery may be reduced by impaired albumin concentration, but the drug concentration in the bloodstream may be increased as a result of impaired clearance, which is also affected by malnutrition.

The plasma amino-acid profile may affect the efficacy of drug entry into the central nervous system. At the blood–brain barrier, certain drugs are transported into the brain by the same transport system that carries the large neutral amino-acids; thus they must compete with them for use of the carrier binding sites. Diet composition, by affecting the postprandial amino-acid profile, may significantly affect the clinical efficacy of drugs such as L-dopa, used in the treatment of Parkinson's disease.

Body composition is also a relevant determinant of drug disposition and action. Although most drug dosages are calculated by total body weight, most drugs act only in the fat-free body mass. Thus, at a given body weight, individuals with more body fat will tend to receive a higher effective dose than those with less body fat. The amount of body fat is also important for drugs that are stored in adipose tissue.

Major Drug–Nutrient Interactions of Clinical Relevance

Table 2 provides information on the major drug–nutrient interactions of clinical relevance. The list reflects well-known interactions of drugs that have been on the market for some time. The US Food and Drug Administration (FDA) maintains an on-line database of recently reported interactions and interactions of new drugs. The database can be assessed at http://www.fda.gov.

Table 2 Major drug–nutrient interactions of clinical relevance

Drug	*Class*	*Food/nutrient*	*Effect/mechanism*
Acarbose	Antidiabetic	Food	Delays carbohydrate breakdown and glucose absorption
		Iron	Decreased iron absorption
Acetaminophen	Analgesic	Food	May delay extended release; high-pectin food delays absorption
		Alcohol	Increased risk of hepatotoxicity
Acetohexamide	Antidiabetic	Glucose	Hypoglycaemia
		Alcohol	Flushing, headache, nausea, vomiting, sweating, tachycardia
		Sodium	Hyponatremia, SIADH
Acyclovir	Antiviral	Food	No effect; may take with meals
Aluminum hydroxide	Antacid	Thiamin	Affects bioavailability, owing to pH
		Iron	Decreased iron absorption
		Phosphorus	Inhibits phosphorus absorption
		Vitamin A	Inhibits vitamin A absorption
Amikacin	Antibiotic	Calcium, potassium, magnesium	Causes renal wasting of these nutrients
Amoxicillin	Antibiotic	Food	Decreased absorption owing to delayed gastric emptying
Amphotericin B	Antifungal	Potassium, magnesium	Causes renal wasting of potassium and magnesium
Ampicillin	Antibiotic	Food	Decreased absorption owing to delayed gastric emptying

Continued

Table 2 Continued

Drug	*Class*	*Food/nutrient*	*Effect/mechanism*
		Potassium	High doses increase urinary potassium losses
Antipyrine		Green vegetables, beef protein	Decreased absorption
Aspirin	Analgesic	Food	Decreased rate of absorption
		Folic acid	Increased excretion of folate
		Amino-acids	Decreased intestinal absorption of amino-acids, increased urinary excretion of tryptophan
		Iron	Chronic high dose 3–4 g day^{-1}, iron deficiency possible
		Alcohol	Gastric irritation, leading to possible gastric bleed
		Curry powder, liquorice, teas, raisins, paprika	Potential salicylate accumulation
		Ascorbic acid, fresh fruits, high vitamin	Increased urinary excretion; decreased concentration in serum and platelets
Astemizole	Antihistamine	Grapefruit juice	May result in cardiotoxicity
		Food	Decreased bioavailability
Atenolol	Antihypertensive	Food	Delayed absorption
Atovaquone	Antibiotic	Food	Bioavailability increased, especially in high-fat foods
Atropine	Anticholinergic	Iron	Delayed absorption
Azithromycin	Antibiotic	Food	Decreased rate and delayed absorption
Bacampicillin	Antibiotic (penicillins)	Food	Decreased absorption
Barbiturates	Anticonvulsants	Alcohol	Enhanced CNS depression
		Calcium, vitamin D	Increased vitamin D requirements, owing to increased metabolism
		Cyanocobalamin	Increased bone resorption
		Folic acid	Decreased serum levels, leading to megaloblastic anemia
		Serum lipids	Decreased CSF folate and erythrocyte concentration; may increase cholesterol, HDL triacylglycerols
Benzodiazepines	Anticonvulsants	Nutrient	Enhanced CNS depression
Clonazedam		Calcium	Increased vitamin D requirements secondary to increased metabolism
Clorazepate dipotassium		Vitamin D	Increased bone resorption
Lopazepam		Cyanocobalamin	Decreased serum levels, leading to megaloblastic anemia
Oxazepam		Folic acid	Decreased CSF folate and erythrocyte concentration
		Serum lipids	May increase cholesterol, HDL triacylglycerols
Buprenorphone HCL	Analgesic Narcotic Agonist–antagonist	Alcohol	Enhanced CNS depression
Butorphanol tartate	Analgesic Narcotic Agonist–antagonist	Alcohol	Enhanced CNS depression
Calcium carbonate	Antacid	Iron	Decreased iron absorption
		Fats	May cause steatorrhea
Captopril	Antihypertensive ACE inhibitor	Food	Reduced absorption
Carbamazpine	Anticonvulsants	Sodium	SIADH
		Food	Enhanced absorption, increased bile production
Carbenicillin iandanyl sodium	Antibiotic	Food	Decreased rate of absorption
Cephalosporins	Antibiotic	Alcohol	Flushing, headache, nausea, vomiting, tachycardia

Continued

Table 2 Continued

Drug	*Class*	*Food/nutrient*	*Effect/mechanism*
Cefadroxil		Food	No effect (may take with food)
Cefpodoxime proxetil		Food	Bioavailability increased with food
Cefuroxim axetil		Food	Bioavailability increased with food
Cefixime		Food	Decreased rate of absorption
Cefachlor		Food	Decreased rate of absorption
Cephalexin		Food	Absorption reduced for suspension, delayed for capsules
Cephradine		Food	Rate of absorption delayed
Ceftibuten		Food	Decreased absorption
Cefamandole		Vitamin K	Decreased vitamin K hypoprothrombinemia
Cefoperazone		Vitamin K	Decreased vitamin K hypoprothrombinemia
Cefotetan		Vitamin K	Decreased vitamin K hypoprothrombinemia
Cetrizine	Antihistamine	Food	Delays time to serum peak; no effect on overall absorption
Chlorambucil	Antineoplastic	Food	Reduced absorption
Chloramphenicol	Antibiotic	Iron	Increased serum level iron; increased total iron-binding capacity
		Folic acid	Antagonist to physiological action; increased requirements of folic acid
		Vitamin B_{12}	Increased requirements of vitamin B_{12} can cause peripheral neuropathy
Chlorothiazide	Diuretic	Food	Increased drug absorption owing to delayed gastric emptying
Chloroquine	Antimalarial	Food	Increased bioavailability
Chlorpromazine	Antiemetic	Food	Decreased absorption owing to delayed gastric emptying
Chlorpropamide	Antidiabetic	Glucose	Decreased blood glucose concentration
		Sodium	Hyponatremia, SIADH
		Alcohol	Flushing, headache, nausea, vomiting, tachycardia
Cholchicine	Antigout	Cyanocobalamin	Decreased absorption of cyanocobalamin
Cimetidine	Histamine 2 antagonist	Food	Delays absorption
Ciprofloxacin	Antibiotic (quinolone)	Caffeine	Decreased rate of absorption
		Food	Decreased elimination of caffeine
		Calcium	Calcium can bind quinolones
		Mineral supplement	Absorption of divalent and trivalent cations decreased by binding to quinolones
Cisapride	Motility	Food (grapefruit)	May result in increased cardiotoxicity
Clarithromycin	Antibiotic (macrolide)	Food	Decreased onset of absorption; no change in total amount absorbed
Clonazepam	Anticonvulsants (benzodiazepine)	Nutrient	Enhanced CNS depression
Clorazepate dipotassium		Calcium	Increased vitamin D requirements, owing to increased metabolism
		Vitamin D	Increased bone resorption
		Cyanocobalamin	Decreased serum levels, leading to megaloblastic anemia
		Folic acid	Decreased CSF folate and erythrocyte concentration
		Serum lipid	May increase cholesterol, HDL triacyclglycerols
Clorgyline	Antidepressant (MAO inhibitor)	Tyramine-rich foods (avacado, canned figs, aged cheese, cola beverage, coffee, chocolate, wine, soy sauce, fermented meats, yeast, yoghurts)	May increase blood pressure

Continued

Table 2 Continued

Drug	*Class*	*Food/nutrient*	*Effect/mechanism*
Cloxacillin	Antibiotic (penicillin)	Food	Decreased rate of absorption
Codeine	Narcotic agonist, analgesic	Alcohol	Enhanced CNS effect
		Glucose	Can cause hyperglycemia
Corticosteroids Prednisone Prednisolone Dexamethazone Methylprednisolone Hydrocortisone	Steroids	Calcium, phosphorus, vitamin D	Decreased absorption of calcium and phosphorus; increased urinary excretion; chronic high dose can cause osteomalacia
Corticosteroids	Steroids	Nitrogen	Increased urinary nitrogen losses
		Zinc	Increased urinary excretion and decreased serum levels
		Glucose	Impairs glucose tolerance; increases plasma levels
		Triacylglycerols, cholesterol	Increased serum levels
Co-trimaxazole	Antibiotic	Potassium	Decreased excretion hyperkalemia
		Sodium	Increased excretion hyponatremia
		Folic acid	Potential for folate deficiency
Cyclosporine	Antirejection	Milk, fat, pineapple juice	Increased absorption
Demeclocycline	Antibiotic	Food, calcium, iron	Decreased absorption of dairy products and divalent and trivalent cations
Diazepam Clonazedam Clorazepate dipotassium Lopazepam Oxazepam	Anticonvulsant	Food	Increased absorption with high-fat meals and delayed gastric emptying
Dicumarol	Anticoagulant	Food	Increased absorption with high-fat meals and delayed gastric emptying
Didanosine Tab	Antiviral	Food	Decreased rate and extent of absorption
Oral suspension		Fruit juice or acid liquid	Didanosine unstable in acid
Digoxin	Cardiac	Food	Delayed absorption; adsorbent to high-fiber high-pectin foods
Dirithromycin	Antibiotic (macrocide)	Food	Slightly increased absorption
Divalproex	Anticonvulsant	Food	Decreased rate of absorption; extent of absorption not affected
Doxycycline	Antibiotic	Food	Decreased absorption of food and milk
Erythromycin	Antibiotic (macrocide)	Food	Increased absorption by delayed gastric emptying
Erythromycin stearate		Food	Reduced absorption by delayed gastric emptying
Ethionamide	Antituberculosis	Pyridoxins	Reports of peripheral neuritis and paraesthesia
Etodolac	NSAID	Food (milk)	Decreased total bioavailability of tolmetin; decreased absorption of ibuprofen
		Sodium	Hyponatremia (indomethacin/ketorolac)
		Potassium	Hyperkalemia (indomethacin/ketorolac)
		Food	Increased rate of absorption
Felbamate	Anticonvulsant	Glucose	Hypoglycemia
		Magnesium	Hypomagnesemia
		Phosphorus	Hypophosphatemia
		Potassium	Hypokalemia
		Sodium	Hyponatremia
Fenoprofen	NSAID	Food (milk)	Decreased total bioavailability of tolmetin
Fenoprofen calcium		Sodium	Hyponatremia (indomethacin/ketorolac)
		Potassium	Hyperkalemia (indomethacin/ketorolac)
		Food (milk)	Decreased bioavailability of tolmetin; decreased absorption of ibuprofen

Continued

Table 2 Continued

Drug	*Class*	*Food/nutrient*	*Effect/mechanism*
		Food	Increased rate of absorption
Fluconazole	Antifungal	Potassium	Hypokalemia
Flucytosine	Antifungal	Food	Decreased rate of absorption; no change in extent of absorption
Foscarnet	Antiviral	Calcium	Hypocalcemia; drug chelates; divalent metal ions
		Magnesium	Hypomagnesemia
		Phosphorus	Hypophosphatemia and hyperphosphatemia
		Potassium	Hypokalemia
Furazolidone	Anti-infective	Tyramine-rich foods (avocados, canned figs, aged cheese, cola beverages, coffee, chocolate, wines, soy sauce, fermented meats, yeast preparation, yoghurts)	Prolonged large doses result in increased risk for hypertensive crisis
		Alcohol	Rushing, headache, nausea, vomiting, sweating, tachycardia
Furosemide	Diuretic	Food	Delayed absorption
Ganciclovir	Antiviral	Food	Increased area under curve plasma concentration
Glipizide	Antidiabetic	Food	Delayed absorption
		Alcohol	Flushing, headache, nausea, vomiting, sweating, tachycardia
		Sodium	Hyponatremia, SIADH
Griseofluvin	Antifungal	Alcohol	Can increase alcohol effect, flushing, tachycardia
		High-fat food	Increased drug absorption rate
Hydralazine	Diuretic	Food	Increased absorption
Hydrochlorothiazide	Diuretic	Food	Increased absorption by delayed gastric emptying
HMG-CoA Reductase inhibitors Simvastatin Lovastatin	Antihyperlipidemic	Food (grapefruit)	Increase drug serum concentration; increase area under curve concentration
Ibuprofen	NSAID	Food (milk)	Decreased total bioavailability of tolmetin; decreased absorption of ibuprofen
		Sodium	Hyponatremia (indomethacin/ketorolac)
		Potassium	Hyperkalemia (indomethacin/ketorolac)
		Food	Increased rate of absorption
Indinavir	Antiviral	Food	Decreased absorption of high-calorie, high-fat and protein-rich foods
		Grapefruit juice	Decreased area under curve concentration
Indomethicin	NSAID	Food (milk)	Decreased total bioavailability of tolmetin; decreased absorption of ibuprofen
		Sodium	Hyponatremia (indomethacin/ketorolac)
		Potassium	Hyperkalemia (indomethacin/ketorolac)
		Food	Increased rate of absorption
Iron	Mineral	Ascorbic acid	Increased absorption
		Amino-acids	Increased absorption
		Calcium phosphate	Decreased absorption
		Zinc	Inhibits absorption
		Vitamin A	Vitamin A deficiency inhibits iron utilization and accelerates the development of anemia
		Tea/coffee	Decreased absorption owing to formation of iron tannate

Continued

Table 2 Continued

Drug	*Class*	*Food/nutrient*	*Effect/mechanism*
		Vegetable polyphenols	Binds and insolubilizes iron
Isoniazid	Antituberculosis	Food	Decreased intestinal absorption
		Pyridoxine	Decreased metabolism, antagonism
		Food and histamine, tuna, liver, aubergine, parmesan cheese, tomato, spinach, tyramine-containing foods	Headache, redness, itching of eyes and face, chills, diarrhea, palpitation; potential hypertensive crisis due to monoamine oxidase inhibitor activity
Itraconazole	Antifungal	Food	Increased absorption, increased triacylglycerols
		Potassium	Hypokalemia
Ketoconazole	Antifungal	Alcohol	Flushing, headache, nausea, vomiting, sweating, tachycardia
Lansoprazole	H/K proton-pump inhibitor	Food	Delays absorption
Labetalol	Antihypertensive	Food	Increased absorption
Lamivudone (3TC)	Antiviral	Food	Decreased rate of absorption
Levodopa	Anti-Parkinson's	Food	Decreased absorption; with high-protein meals amino-acids compete for absorption
Lithium	Antimanic	Low-sodium diet	Increased lithium concentrations
		High-sodium diet	Increased lithium clearance
		Food	Increased absorption
Linezolid	Antibiotic	Tyramine-rich foods	May result in blood-pressure changes
Lomefloxacin	Antibiotic (quinolone)	Food	Decreased rate and extent of absorption
Loracarbef	Antibiotic	Food	Decreased rate of absorption
Lovastatin	Antihyperlipidemia	Food	Increased absorption
Mebendazole	Anthelmintic	Food	Increased absorption
Meclofenamate	NSAID	Alcohol	Additive CNS effects; increased prothrombin time
		Food	Decreased bioavailability
Melphalan	Antineoplastic	Food	Reduced absorption
Mercaptopurine	Antineoplastic	Food	Reduced absorption
Methacycline	Antibiotic	Food, calcium, iron	Decreased absorption of dairy products, cereals, divalent and trivalent cations
Methenamine mandelate	Urinary anti-infective	Milk products, citrus fruits	Excessive amounts inhibit drug conversion
Methosuximide	Anticonvulsant	Alcohol, calcium	Additive CNS effects; hypocalcemia
Methotreate	Antineoplastic	Food	Increased absorption
Methyldopa	Antihypertensive	Vitamin B_{12}, folate	In high doses methyldopa can increase vitamin B_{12} and folate losses
		Food	High-protein meals compete for absorption
Metoprolol	Antihypertensive	Food	Increased absorption
Metronidazole	Antibiotic	Alcohol	Flushing, headache, nausea, vomiting, sweating, tachycardia
		Food	Decreased peak serum concentration but total amount of drug absorbed is not affected
Minocycline	Antibiotic	Food, calcium	Decreased absorption
Nafcillin	Antibiotic	Food	Decreased absorption; decreased serum levels due to altered gastric pH
		Potassium	High doses can cause hypokalemia owing to increased urinary losses
Nifedipine	Antihypertensive	Food (grapefruit)	Increases pressor effect of drug
Nitrofurantoin	Antibiotic	Food	Increased absorption by delayed gastric emptying
NSAIDs	NSAID	Food (milk)	Decreased bioavailability of tolmetin
Diclofenl		Sodium	Hyponatremia (indomethacin/ketorolac)
Etodolac			

Continued

Table 2 Continued

Drug	*Class*	*Food/nutrient*	*Effect/mechanism*
Fenoprofen Ca Ibuprofen Ketoprofen Ketorolac Naproxn Oxapron Piroxican Sulindac Tolmetin NA		Potassium	Hyperkalemia (indomethacin/ketorolac)
		Food	Decreased absorption of ibuprofen
		Food	Increased rate of absorption
Norfloxacin	Antibiotic (quinolone)	Food, dairy products	Decreased rate of absorption
		Multivitamin and mineral supplements	Decreased absorption due to formation of divalent and trivalent cation complexes with quinolones
Nifedepine	Antihypertensive calcium-channel blocker	Grapefruit juice	Increased serum level of nifedepine flavonoids inhibits cytochrome P-450
		Food	Decreased bioavailability, formulation dependent
Ofloxacin	Antibiotic (quinolone)	Dairy products and mineral supplements	Decreased absorption by polyvalent cations
Ondansetron	Antiemetic	Food	Increased extent of absorption
		Potassium	Hypokalemia
Omeprazole	H/K proton-pump inhibitor	Food	Delays absorption
Oral contraceptives		Ascorbic acid	Decreased ascorbic-acid concentration in plasma, platelets, leucocytes
		Vitamin C, folic acid	Decrease in serum levels
		Vitamin B_{12}	Impairs tryptophan metabolism
		Amino-acids, vitamin A, vitamin E, copper	Increase in serum levels
Oxacillin	Antibiotic	Food	Decreased absorption and decreased serum concentration
		Fats	Oxacillin can cause steatorrhea
Paromomycin	Amoebicide	Food	Increased absorption by delayed gastric emptying
		Vitamins A, D, E, K	Malabsorption of fat-soluble vitamins owing to hypocholesterolemia
Penicillamine	Antidote (chelating agent)	Food	Decreased absorption
		Iron, zinc	Decreased absorption 30%–70% of increased zinc absorption; decreased penicillamine absorption
Penicillin G & VK	Antibiotic	Food	Decreased absorption by delayed gastric emptying
		Glucose	Hyperglycemia
Pentamidine	Antibiotic	Calcium, magnesium	Hypomagnesemia, hypocalemia
		Potassium	Hyperkalaemia due to nephrotoxicity
Phenacemide	Anticonvulsant	Fresh fruits and vitamin C	Increased urinary excretion of phenacemide
Phenobarbital	Anticonvulsant (see Barbiturates)	Food	Decreased absorption due to protein binding
		Protein	Low-protein diet increases duration of action of phenobarbitol
		Vitamin D, calcium	Decreased serum vitamin D by cytochrome P-450 hypocalcemia
		Fresh fruits and vitamin C	Increased urinary excretion of phenobarbitol
Phensuximide	Anticonvulsant (succinimides)	Calcium, vitamin D	Decreased serum vitamin D by P-450 cytochrome hypocalcemia

Continued

Table 2 Continued

Drug	*Class*	*Food/nutrient*	*Effect/mechanism*
		Vitamin B_{12}, folic acid	Decreased absorption and serum levels of folates; inhibits vitamin B_{12} transport
		Copper	Increased serum levels
Phenytoin	Anticonvulsant (hydantoins)	Fresh fruits and vitamin C	Increased urinary excretion
		Vitamin D, calcium	Decreased serum vitamin D by cytochrome P-450 hypocalcemia
		Enteral feeds	Decreased absorption
		Food	Increased absorption by delayed gastric emptying
Pimozide	Antineruoleptic	Food (grapefruit)	Increased risk of cardiotoxicity
Piroxicam	NSAID	Food	Delayed absorption
Praziquantel	Anthelmintic	Food	Decreased rate and extent of absorption
Primidone	Anticonvulsant	Fresh fruits and vitamin C	Increased urinary excretion of primidone
		Protein	Low-protein diet increases duration of action of primidone
Propantheline	Anticholinergic	Food	Decreased absorption
Propranolol	Antihypertensive	High-protein foods	Increased absorption
Proxyphene	Analgesic	Food	Increased absorption by delayed gastric emptying
Pyrimethamine	Antimalarial	Folic acid	Decreased serum folate concentrations
Quinidine	Antiarrhythmic	Food	Delayed absorption due to protein binding
Riboflavin	Vitamin	Food	Increased absorption by delayed gastric emptying
Rifampin	Antibiotic	Food	Decreased absorption
Ritonavir	Antiviral	Vitamins	Can cause vitamin deficiency
		Potassium	Hyperkalemia and hypokalemia
		Cholesterol	Hypercholesterolemia
		Triacylglycerols	Hypertriacylglycerolemia
		Food	Delayed absorption
Oral solution		Food	Increased extent of absorption
Capsules			
Salicylates	Analgesics	Iron	Long-term chronic use decreases serum iron
Magnesium salicylate			
Choline salicylate		Vitamin C	Decreases concentration in serum and platelets
Sodium salicylate		Amino-acids	Decreases their intestinal absorption and increases urinary secretion
Saquinavir mesylate	Antiviral	Food	Increased absorption of high-calorie, high-fat foods
		Calcium	Hypercalemia
		Glucose	Hyperglycemia and hypoglycemia
		Phosphorus	Changes in serum phosphorus
		Potassium	Hyperkalaemia and hypokalemia
Spironolactone	Diuretic	Food	Increased absorption by delayed gastric emptying
Sulfonamides	Antibiotic	Food	Delayed with no effect on extent of absorption
Sulfadiazine			
Sulfisoxazole		Folic acid	Decreased intestinal synthesis, absorption, and serum levels
Sulfamethoxazole			
Tetracycline	Antibiotic	Food	Decreased absorption
		Minerals	Inhibits absorption of iron, calcium, zinc, and magnesium; chelation by polyvalent cations
		Fats	Decreases absorption
		Vitamin K	Decreases bioavailability
		Vitamin C	Increases urinary losses; decreases

Continued

Table 2 Continued

Drug	*Class*	*Food/nutrient*	*Effect/mechanism*
Terfenadine	Antihistamine	Food (grapefruit)	Increased risk of cardiotoxicity
Theophylline	Broncodilator	Charbroiled beef	Increased metabolism of theophylline
		High-fat meals	Increased absorption dependent on formulation
Tolazamide	Antidiabetic	Sodium	Hyponatremia and SIADH
Tolbutamide	Antidiabetic	Ethanol	Prolonged hypoglycemia, disulfram reaction
Trimethoprim	Antibiotic	Folic acid	Decreased serum folate levels
Valproic acid Divalproex Sodium valproate Sodium oral solution	Anticonvulsant	Milk, food, carbonated drinks	Delayed absorption but no effect on extent of absorption
Warfarin	Anticoagulant	Alcohol, vitamin K	Inhibits warfarin metabolism; beef liver, pork liver, green tea, leafy green vegetables high in vitamin K inhibit anticoagulant effect
		Vitamin E	Can increase warfarin response
Zalcitabine	Antiviral	Food	Decreases rate and extent of absorption
Zafirlukast	Selective leukotiene antagonist	Food	Delayed absorption
Zidovudine	Antiviral	Food	Decreased rate of absorption

SIADH, Syndrome of inappropriate antidiuretic hormone excretion; CNS, central nervous system; CSF, cerebrospinal fluid; NSAID, nonsteroidal anti-inflammatory drug

Herb–Drug Interactions

Herbal botanicals have been used in many cultures throughout the world for hundreds of years. These products are usually seen as natural; they should not be synonymous with safe.

Possible interactions can involve hepatic cytochrome P-450 and changes in intestinal absorption, distribution, and renal excretion. Herbal interactions with certain drugs are listed in **Table 3**.

Table 3 Herbal–drug interactions

Herbal	*Drugs*	*Effect/mechanism*
Echinacea	Methotrexate, aminodarone, ketoconazole, steroids (anabolic)	Increased hepatotoxicity
Feverfew	NSAIDs	Decreased herbal effect
	Anticoagulants	Additive platelet inhibition
Garlic	Aspirin, anticoagulants	Reduced clotting time
Ginkgo biloba	Aspirin, anticoagulants, NSAIDs, tricyclic antidepressants, anticonvulsants	Decreased seizure threshold; increased risk of bleeding
Ginseng	Monoamine oxidase inhibitors	Headache, tremors, mania
	Corticosteroids	Increased steroid toxicity
	Warfarin	Decreased INR
	Digoxin	Increased digoxin levels
Kava kava	Benzodiazepines	Increased CNS depression
Epherda	Antidepressants, CNS stimulants	Increased herbal effect
St John's wort	Antidepressants, CNS stimulants	Additive effects
	Piroxicam, tetracycline	Increased photosensitivity
	Theophylline	Decreased theophylline levels
Saw palmetto	Oestrogen	Increased effect of herbal
Valerian	CNS depressants	Additive CNS depression

CNS, central nervous system; INR, International Normalization Ratio; NSAIDs, nonsteroidal anti-inflammatory drugs.

See also: **Amino Acids**: Metabolism. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management.

Further Reading

Caballero B (1988) Nutritional implications of dietary interactions: a review. *Food Nutrition Bulletin* **10**: 9–20.

Knapp HR (1996) Nutrient–drug interactions. In: Ziegler FF and Filer LJ (eds.) *Present Knowledge in Nutrition*, pp. 540–546. Washington, DC: ILSI Press.

Neuvonen P (1989) Clinical significance of food–drug interactions. *Medical Journal of Australia* **150**: 36–40.

Roberts J (1988) Age and diet effects on drug action. *Pharmacology and Therapeutics* **37**: 111–149.

Roe DA (1986) Drug–food and drug–nutrient interactions. *Journal of Environmental Pathology, Toxicology and Oncology* **5**: 115–135.

Roe DA (1994) Diet, nutrition and drug reactions. In: Shils ME, Olson JA, and Shike M (eds.) *Modern Nutrition in Health and Disease*, 8th edn, pp. 1399–1416. Philadelphia: Lea & Febiger.

Schmidt LE (2002) Food–drug interactions. *Drugs* **10**: 1481–1502.

Wurtman RJ, Caballero B, and Salzman E (1988) Facilitation of DOPA-induced dyskinesias by dietary carbohydrates. *New England Journal of Medicine* **390**: 1288–1289.

E

EARLY ORIGINS OF DISEASE

Contents
Fetal
Non-Fetal

Fetal

A J Buckley and S E Ozanne, University of Cambridge, Cambridge, UK

Introduction

The prevalence of metabolic diseases such as type 2 diabetes and cardiovascular disease is increasing at an alarming rate. Around one in ten people today suffer from type 2 diabetes and it is estimated that by 2010 over 250 million people worldwide will have this condition. Although the etiology of these metabolic diseases is considered to be multifactorial, there is now a substantial body of evidence suggesting that the pathogenic mechanisms underlying these adult-occurring diseases originate from disturbances experienced during *in utero* and early life.

The long-term effect of an insult during a critical period of development has been recognized for many years. As long as 70 years ago it was recognized that the early environment in which a child grows could have long-term effects on its health. This was based on the observations in England, Scotland, and Sweden that suggested that death rates in specific age groups at any time depended upon the year of birth, suggesting that the time of death was more related to the year the person was born in rather than the age of the person. Additional evidence supporting the importance of the early environment came from a study in Norway investigating the geographical variations in current death rates from arteriosclerotic heart disease. Forsdahl and coworkers demonstrated a significant positive correlation between the current death rates and geographical variation in past infant mortality rates.

In utero growth and development is an extremely critical period in one's life. This was first recognized by Barker and Osmond when they demonstrated a striking association between adult mortality from cardiovascular disease and past infant mortality rates earlier in the century in the same geographical regions of England and Wales. As infant mortality was greatest in regions where low birth weight was also present, it was hypothesized that adverse early life nutritional influences could result in low birth weight and lead to an increased predisposition to cardiovascular disease. Subsequent regional studies have also demonstrated associations between low weight both at birth and 1 year of age and high mortality rates from ischemic heart disease. Mounting support for the role that the early life environmental influences play in establishing the risk for disease came when this association was found not to disappear even with an improvement in diet during adult life or with moving to other regions of the country.

It is thus now well established that poor fetal growth confers an amplified risk for the development of diseases such as type 2 diabetes, cardiovascular disease, insulin resistance, and obesity. Metabolic programing, a concept defined as the process whereby exposure to a stimulus or insult during a crucial phase of growth and development results in permanent alterations in the structure or function of an organ or metabolic action, is involved in amplifying this risk. Maternal nutrition and the maternal metabolic milieu during gestation and lactation, as well as the functionality of the placenta, have been widely recognized as major influential factors of adverse fetal and early life metabolic programing.

Birth Weight and Adult Disease

Further evidence for the role that the early life environmental influences play in establishing the risk for disease came in the early 1990s when studies conducted by Barker and colleagues in a large cohort of men and women from Hertfordshire, UK revealed strong correlations between low birth weight and a high prevalence of metabolic diseases in later life (**Table 1**). Although not universally accepted, numerous epidemiological studies completed throughout the world, including other parts of the UK, Europe, US, and India, have found reproducible results in extensive population and ethnic groups and various age ranges.

The Hertfordshire study was a retrospective study that collected the birth records of 15 726 men and women born in Hertfordshire between 1911 and 1930. This study was the first to demonstrate that the incidence of death from coronary heart disease was highest in adults born with a low birth weight. Barker and colleagues replicated this finding in a cohort of 1586 men born in Sheffield, UK between 1907 and 1925. The link between low birth weight and coronary heart disease in adult life again received confirmation from studies involving over 70 000 nurses from the US born between 1921 and 1945. A study in South India demonstrated that in a cohort of 517 men and women born between 1934 and 1954, the prevalence of coronary heart disease rose to 11% when the recorded birth weight was less than 2.5 kg. In those people who had a birth weight of more than 3.1 kg the prevalence of coronary heart disease was only 3%. A Swedish study of over 14 000 men and women born in Uppsala, Sweden between 1915 and 1929 was able to clearly identify that the risk of death from cardiovascular disease was associated with being small for gestational age.

In the Hertfordshire study, Hales and Barker demonstrated a strong inverse relationship between birth weight and type 2 diabetes or impaired glucose tolerance in men aged 64 years. Similar findings have been reported in populations of men and women in other parts of Europe, Australia, and the US. Further support of the association between low birth weight and the development of metabolic disease comes from studies of monozygotic twins. Two studies have demonstrated that impaired glucose tolerance and type 2 diabetes is more prevalent in the twin with the lower birth weight.

Although initial studies focused on the relationship between cardiovascular disease and type 2 diabetes subsequent studies have reported associations with other conditions. From an early age, children born with a low birth weight demonstrate reduced endothelium-dependent dilation and increased arterial stiffness. It has been observed that this endothelial dysfunction persists into adult life. The mechanisms behind the association between low birth weight and impaired vascular function remain to be elucidated. Despite this, it has been suggested that endothelial dysfunction is an early feature and precedes the metabolic disorders that develop in low-birth-weight humans.

Whilst the association between low birth weight and high predisposition to disease has been reported quite substantially, it is important to note that a U-shape relationship does occur when investigating the association between birth weight and the development of metabolic diseases. Babies born large for gestational age are also at an elevated risk of developing diseases such as coronary heart disease and type 2 diabetes.

The etiology of breast cancer has now also been linked to prenatal influences. Investigations in Sweden, Norway, and the US have provided epidemiological data indicating that high birth weight potentially increases the risk of developing breast cancer. The specific biological mechanisms underlying this association still remain unclear. However, it has been suggested that prenatal exposure to the high level of estrogen that occurs during pregnancy may play a significant role. Other maternal hormones and growth factors may also be involved. A U-shape relationship also exists when investigating prenatal influences upon the development of breast cancer. Studies have reported birth weights below 2.5 kg and above 4 kg are significant risk factors for the development of breast cancer in women.

Table 1 Adult health characteristics associated with low birth weight

- Type 2 diabetes
- Coronary heart disease
- Hypertension
- Hypertriglyceridemia
- Impaired glucose tolerance
- Insulin resistance

Underlying Mechanisms

The Fetal Insulin Hypothesis

Hattersley and colleagues have suggested that the relationship between birth weight and type 2

diabetes could be mediated by mutations/polymorphisms in genes that influence insulin secretion or insulin sensitivity. Insulin is an important fetal growth factor, thus any defects in its secretion or action would result in both low birth weight and increased risk of diabetes. This is supported by studies of individuals with maturity onset diabetes of the young 2. These individuals who have mutations in the glucokinase gene have a lower birth weight compared to their unaffected siblings. The fetal insulin hypothesis therefore holds true in this rare monogenic form of diabetes.

The Thrifty Phenotype Hypothesis

An alternative hypothesis, termed the thrifty phenotype hypothesis, was proposed by Hales and Barker in 1992. This proposal focused on the role played by the fetal environment. It suggested that the mechanistic basis underlying the observed relationship between poor fetal growth and the future development of metabolic diseases was related to fetal nutrition.

Central to the thrifty phenotype hypothesis is the suggestion that during times of nutritional deprivation, the growing fetus undergoes metabolic adaptations that are beneficial to survival postnatally in similar conditions of poor nutrition. Such adaptations include the essential preservation of brain growth at the expense of the normal development of organs such as the liver, muscle, and the pancreas. This has no detrimental effect if the fetus is born into conditions of poor nutrition. Hence in sub-Saharan Africa where there is chronic malnutrition, rates of diabetes are very low. Detrimental consequences of fetal programing arise when the fetus is born into conditions that differ from those experienced *in utero*. The imbalance between the early and adult environments may then conflict with the programing that occurred during fetal life and predispose the offspring to the subsequent development of metabolic diseases in adulthood (**Figure 1**).

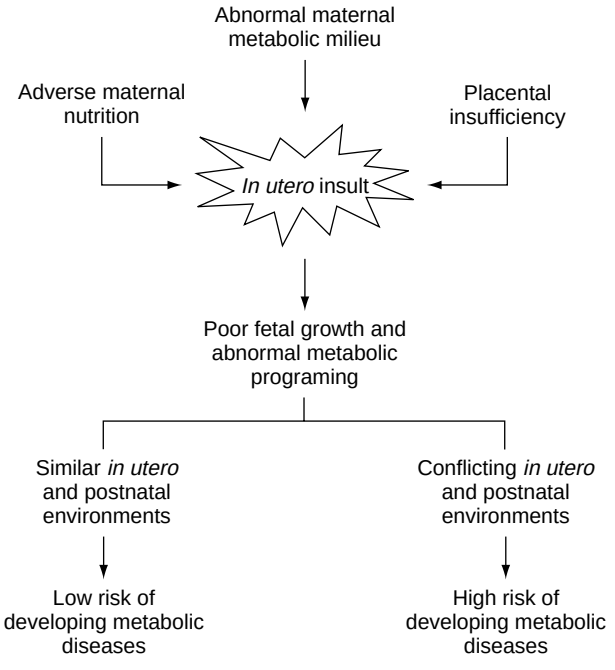


Figure 1 The interactions between *in utero* insults and the subsequent development of metabolic diseases.

Underlying Factors

Glucocorticoids

Maternal glucocorticoids can also influence birth weight of the offspring. Under normal conditions, fetal exposure to glucocorticoids is relatively low due to the presence of placental 11β-hydroxysteroid dehydrogenase 2 (11βHSD2), an enzyme that acts as a placental barrier by inactivating maternal glucocorticoids before they cross into the fetal environment. Maternal glucocorticoid treatment during pregnancy or inhibition of the placental 11βHSD2 can therefore increase the amount of active glucocorticoid crossing the placenta. Excess glucocorticoid exposure has also been implicated in disturbing the normal growth and development of the fetus with consequential effects on the overall health of the adult offspring. There does, however, appear to be a critical window of sensitivity where the developing fetus is particularly sensitive to glucocorticoids. Glucocorticoid overexposure in the 3rd trimester is known to cause reductions in birth weight. Studies in rats have established that glucocorticoid-exposed offspring undergo rapid postnatal catch-up growth, which proves deleterious to their adult health. These studies have demonstrated that excess exposure to glucocorticoids during fetal life is linked to low birth weight, altered functioning of the hypothalamic-pituitary-adrenal (HPA) axis, and the subsequent development of hypertension and impaired glucose tolerance in adulthood.

Increased HPA activity has been demonstrated in low-birth-weight human adult populations. Enhanced responsiveness of plasma cortisol to ACTH and increased urinary cortisol metabolite excretion has been reported in low-birth-weight adult males. Prenatal alteration of the HPA axis may therefore be involved in the subsequent development of cardiovascular disease in low-birth-weight adults.

Leptin

Cord plasma leptin levels have been shown to correlate positively with birth weight and neonatal adiposity. It has been suggested that leptin has a regulatory role in growth and development. Low levels of cord blood leptin have been reported in growth-restricted offspring. These low leptin levels may also predict significant weight gain and catch-up growth, both of which are evident in these growth-restricted offspring. Ong and colleagues hypothesized that there may be a link between *in utero* programing of leptin levels and the predisposition to the development of metabolic diseases.

Maternal Nutrition and Fetal Origins of Adult Metabolic Diseases

Assessing the impact of maternal nutrition on health of the offspring in humans is difficult. However, investigations involving offspring conceived during conditions of famine have provided direct evidence of the effects that maternal nutrition during gestation and lactation has on the overall health of the adult offspring. The Dutch famine, which occurred in the western part of the Netherlands at the end of World War II, only lasted around 5 months from late November 1944 to early May 1945, and was therefore defined as a short period of famine. Prior to the onset of the famine conditions, the affected area of the Netherlands consisted of a reasonably well-nourished population. The occurrence of this abrupt famine therefore granted researchers a unique opportunity to retrospectively study the effect of maternal nutrition during specific stages of gestation on insulin-glucose homeostasis and obesity risk in adult offspring (**Table 2**).

Investigators traced and studied individuals who were born immediately before the famine commenced, those born during the famine, and those born up to 21 months after the famine had ceased. Compared to the offspring not exposed to *in utero* famine conditions, individuals who were *in utero* during the famine had higher plasma glucose levels 2 h after a standard oral glucose tolerance test. These glucose levels were highest in those individuals who had been exposed to the famine during the final trimester of pregnancy and then become obese in adult life. In terms of obesity, individuals who were exposed to the famine during the first half of pregnancy were more obese at age 19 years. In contrast those who were exposed to the famine during the last trimester of pregnancy and in early postnatal life had reduced obesity. This suggests that the critical time windows for increased risk of obesity and type 2 diabetes differ. This study provided direct evidence that poor maternal nutrition leads to increased susceptibility of type 2 diabetes and obesity in offspring. It also supports the hypothesis that the greatest risk of developing metabolic diseases exists when there is a marked conflict between the environmental conditions experienced *in utero* and that experienced in adult life.

Table 2 Effects of famine conditions during the different stages of pregnancy

Maternal exposure to famine during pre-early gestation	*Maternal exposure to famine during mid-late gestation*
Increased birth weight	Reduced birth weight
Increased birth length	Reduced birth length
Increased obesity	Reduced glucose tolerance
Increased risk of coronary heart disease	Increased risk of type 2 diabetes

As nutritional studies in humans are complex and clouded by multiple confounding factors, a number of animal models of maternal nutritional insults have been developed. Investigations involving these models have significantly contributed to elucidating pathogenic mechanisms underlying the fetal origins of adult metabolic diseases.

Maternal Total Food Intake

Animal models replicating human famine conditions have been developed to enable a more in-depth investigation of maternal calorie restriction throughout pregnancy and the long-term health consequences that this nutritional insult imposes on the developing offspring. Various animal species have been utilized whilst studying the fetal origins of adult metabolic disease in response to maternal feed restriction using dietary insults of up to 70% *ad libitum* food restriction.

Maternal total food restriction (50% of normal food intake) in the rat throughout the gestation period can induce intrauterine growth restriction and result in significantly reduced birth weight. This poor fetal growth is then accompanied by numerous metabolic disturbances in later life. Compared to age-matched control rat offspring, the blood pressure of the maternal food-restriction offspring is significantly elevated and endothelial vascular dysfunction is evident. Insulin resistance, as defined by an elevated fasting plasma insulin level, has also been shown in adult rats exposed *in utero* to the adverse effects of severe maternal food restriction (to only 30% of *ad libitum* intake). In guinea pigs, mild to moderate maternal food restriction (70%–85% *ad libitum* intake) during

the pregnancy leads to perturbations in postnatal glucose-insulin homeostasis as well as alterations in the homeostasis of cholesterol metabolism in the male offspring.

Despite the observed significant reduction in birth weight, adult rats exposed to maternal undernutrition (30% of *ad libitum* intake) whilst *in utero* have been shown to develop obesity. Compared to the control offspring, the feed-restricted offspring appear to have been inappropriately programed and display hyperphagia and elevated food consumption as a consequence of *in utero* exposure to an adverse maternal diet. The underlying mechanisms leading to the hyperphagia in these offspring remains to be determined. However, the involvement of leptin resistance has been implied as these offspring also display hyperleptinemia and have significantly elevated fat pad mass as adults.

Even short-term maternal food restriction during the various stages of the gestational period has been demonstrated to provoke perturbations within the metabolic processes of the offspring. Exposure to maternal malnutrition, particularly in the final trimester and during lactation, impairs the programming of β-cell development and induces alterations in the fetal endocrine pancreas that persist into adulthood such that the offspring at 12 months of age display profound insulinopenia and marked glucose intolerance.

Maternal Protein Consumption

In addition to total calorie intake there is also evidence that composition of the maternal diet can have long-term consequences on the metabolism of the offspring. Experimental animal models involving maternal protein restriction have suggested that adequate protein intake is critical in both development of the fetus and its long-term health. They have also provided insight into the possible underlying mechanisms.

Offspring of rat dams fed a low (8%) protein diet throughout the gestational period are consistently smaller at birth than offspring of a control diet containing 20% protein. Initially, the low-protein offspring have significantly improved glucose tolerance than the control offspring. In humans, small-for-gestational-age infants also display increased insulin sensitivity with respect to glucose disposal in early postnatal life. Offspring of dams fed a low-protein diet appear to undergo a greater age-dependent loss of glucose tolerance. By 15 months of age, the male low-protein offspring are considered to have developed glucose intolerance that is associated with insulin resistance and by 17 months of age this has progressed to type 2 diabetes.

Consistent with the thrifty phenotype hypothesis, the growth restriction of the tissues and organs of the low-protein offspring is not uniform. In the growth-restricted rats, brain growth is spared at the expense of the growth of other developing tissues. In addition to the altered structure and growth patterns, the insulin-sensitive tissues (skeletal muscle and adipose tissue) and organs (liver and pancreas) of the low-protein offspring have been metabolically programed to have altered functionality.

Recently, evidence has been provided suggesting that taurine supplementation to the maternal low-protein diet may benefit the health outcomes of the rat offspring. Maternal taurine supplementation was found to restore and normalize the vascularization of the offspring's endocrine pancreas. Despite these findings, there is little evidence to suggest that a maternal high-protein intake has overall beneficial effects on the metabolic health of the offspring. Some human epidemiological studies and human trials involving high-protein dietary supplementation have in fact demonstrated that the consumption of a high-animal-protein, low-carbohydrate diet throughout late pregnancy can lead to metabolic disturbances in the offspring when they reach adulthood. It has been suggested that these high-protein diets stimulate the hypothalamic-pituitary-adrenal axis and cause maternal cortisol levels to increase. As a result, the developing fetus is presented with the metabolic stress of being exposed to excess cortisol levels. This inappropriate exposure to cortisol during fetal life appears to program lifelong hypercortisolemia and elevated blood pressure. It is likely that the type of protein is also important and this may in part explain some of these apparent discrepancies.

Adverse metabolic disturbances in offspring have therefore been demonstrated as consequential effects of both the *in utero* exposure to either maternal protein restriction or maternal high-protein consumption. The role of the carbohydrate level in these diets still needs to be ascertained. Nevertheless, it is essential that the optimal level of protein intake during pregnancy and lactation be clearly established so as to aid in the normal growth and development of the fetus.

Maternal Iron Restriction

Women, especially pregnant women, in today's society are often diagnosed as being anemic or iron deficient. A rodent model of maternal iron restriction has been developed to determine if metabolic health consequences are observed in the offspring of

iron-deficient women. Similar to other restricted maternal diets, iron restriction during pregnancy in rats can lead to disturbances within the events of early life metabolic programing and can induce permanent adaptations of the offspring's physiological and metabolic processes.

Although the mechanisms remain unclear, growth restriction of the offspring of rats fed an iron-deficient diet throughout gestation has been consistently reported. Elevated blood pressure appears to occur in response to maternal iron restriction and persists throughout the life of the offspring. Changes in cardiac size are evident prior to the initiation of hypertension in the offspring of iron-restricted rat dams. Elevated levels of cardiac hypertrophy may therefore contribute to the programed rise in blood pressure documented in these offspring. Renal development is also adversely altered in the offspring of iron-restricted dams. Maternal iron restriction during pregnancy can induce reductions in nephron and glomerular number in the adult rodent offspring. Significant inverse relationships between glomerular number and systolic blood pressure exist in the offspring of iron-restricted dams suggesting that abnormal renal development may also be involved in inducing the hypertensive state in these offspring. The principal mechanisms behind the association of maternal iron restriction during pregnancy with the altered cardiac and renal growth in the offspring and the subsequent induction of permanently elevated blood pressure still need to be further investigated.

Maternal High-Fat Consumption

In today's Western, more affluent society, the *in utero* environment is likely to be influenced by maternal nutritional insults such as excess fat consumption. There is little dispute regarding the deleterious effects of a high-fat diet. It is well documented that a diet high in fat has played a fundamental role in the prevalence of type 2 diabetes, reaching the epidemic proportions that is seen today. Both human epidemiological studies and experimental animal investigations have demonstrated clear associations between the consumption of a high-fat diet and the increasing prevalence of cardiovascular disease, insulin resistance, and type 2 diabetes. In light of this, a high-fat diet can increase the risk of a pregnant woman developing gestational diabetes. As will be discussed later, it is well established that offspring of diabetic mothers are themselves at an increased risk of developing the disease at an early age. Consumption of a high-fat diet may not only therefore cause deleterious effects to the current generation, but may ultimately have profound consequential effects on future generations. Considering this, the negative impact that a maternal high-fat diet has on the offspring is a topical area of research that requires more detailed attention.

Like any maternal nutritional insult, exposure to an abnormal *in utero* environment, induced by the maternal high-fat diet, can lead to subsequent disturbances in metabolic programing of the developing fetus. To date, investigations studying this nutritional insult have mainly concentrated on the effects of a maternal diet high in saturated fat. Such a diet has led to rat weanlings having increased amounts of body fat, increased liver weight, increased liver triglyceride content, higher blood glucose, and higher blood triglyceride levels. Permanent alterations in the structure and function of the pancreas, vascular dysfunction, and reduced insulin sensitivity have also been documented in rat weanlings and young adult offspring of high-saturated-fat-fed rat dams. Recently, it has emerged that feeding pregnant rats with diets containing a high proportion of animal lard can induce severe endothelial dysfunction in the offspring, along with the subsequent development of increased adiposity, hyperglycemia, insulin secretory deficiency, and insulin resistance. Taken together, these investigations demonstrate profound metabolic derangements in the offspring of fat-fed rats. The underlying mechanistic basis for these observations, however, still requires elucidation. It is also important to note that a high-fat diet is concurrently low in carbohydrate content and it cannot be ruled out that a carbohydrate deficiency in the female rats may account for the metabolic perturbations observed in the offspring of high-fat-fed dams.

Maternal Alcohol Consumption

Several laboratories have investigated the effects of sustained maternal alcohol consumption on the offspring's metabolic health. Alcohol consumption during pregnancy can lead to abnormal fetal development and a subsequent reduction in birth weight. Increased offspring morbidity may also be linked to gestational alcohol consumption. It has been previously documented that female rats fed a gestational diet supplemented with alcohol tended to have a higher number of pups die in early postnatal life. Of those alcohol-exposed offspring that survived, the reduced rate of prenatal growth and development has been linked to abnormalities in the offspring's glucose and insulin homeostasis. Both glucose intolerance and insulin resistance are evident in the rat offspring exposed during *in utero* life to maternal alcohol. Phenotypic abnormalities,

commonly associated with insulin resistance and other metabolic diseases, are also evident in the *in utero* alcohol-exposed offspring. The accumulation of triglycerides in nonadipocyte tissue, namely the skeletal muscle and the liver, is commonly observed in both insulin-resistant humans and experimental animal models of insulin resistance. Elevated levels of plasma and nonadipocyte tissue triglycerides have now also been documented in low-birth-weight rats that were exposed to maternal alcohol *in utero*.

Consumption of alcohol is quite common among breast-feeding mothers as studies have shown ethanol to aid in the promotion of lactation. Establishing the harmful effects of alcohol consumption during lactation is therefore important. Newborn rats exposed to maternal alcohol only during the lactation period have also been shown to develop reduced insulin sensitivity despite having normal prenatal growth and development. In early postnatal life some important metabolic processes are still undergoing development. Therefore, it must be considered that early postnatal life is still a vulnerable period of growth and the developing metabolic processes may still be particularly susceptible to adverse effects induced by alcohol consumption by breast-feeding mothers.

Maternal Metabolic Milieu and Fetal Origins of Metabolic Diseases

Over two decades ago it was hypothesized that perturbations in the metabolic milieu of pregnant women can disturb the intrauterine environment and influence long-term health consequences in the offspring. Abnormalities in the maternal metabolic milieu can evoke the adverse transfer of hormones and fuels from the mother to the growing fetus and thereby increase the predisposition to metabolic disease in the offspring's later life.

The effect of a diabetic pregnancy has been thoroughly examined in the Pima Indians of Arizona. The Pima Indian population has the world's highest prevalence and incidence of type 2 diabetes. Metabolic disorders are becoming increasingly common in the younger generation of this population. Consequently, it is possible that these disorders are present in Pima Indian women of childbearing age. In fact, 10–15% of Pima Indian pregnancies are complicated by type 2 diabetes.

Diabetic pregnant women are hyperglycemic, a characteristic of the general diabetic population. Whilst maternal insulin cannot cross the placental barrier, maternal glucose can freely do so. Elevated maternal glucose levels can therefore induce fetal hyperinsulinemia, subsequently promoting excessive growth and adipose tissue accumulation in the offspring. Compared to the offspring of either nondiabetic or prediabetic (those who developed type 2 diabetes after the birth of their offspring) women, the offspring of diabetic females generally have increased birth weights. This obese state then tracks with the offspring throughout life. Interestingly, excessive obesity can also develop in the subset of normal-birth-weight offspring of diabetic women. These findings suggest the offspring of diabetic women are detrimentally programed into developing altered metabolic processes regardless of birth weight.

Exposure to maternal diabetes whilst *in utero* is also largely responsible for the offspring having an increased risk of developing insulin resistance, impaired glucose tolerance, and type 2 diabetes at an early age. These metabolic disorders are significantly more prevalent in the offspring of diabetics than the offspring of nondiabetics and prediabetics. It has been established that more than one-third of Pima Indian children diagnosed with type 2 diabetes in the past decade seem to have developed the disease as a direct programed response to being exposed to an intrauterine diabetic environment.

Metabolic disorders such as type 2 diabetes have been suggested to have a strong genetic component. It has been postulated that diabetes-susceptibility genes are transmitted to the fetus and this then confers an increased risk for the offspring to develop the disease in adulthood. According to this explanation, offspring born either prior to or after the mother's type 2 diabetes diagnosis should carry the same risk of inheriting the diabetes-susceptibility genes. However, it has become evident that this does not appear to be the case. Sibship studies have compared the prevalence of type 2 diabetes and the degree of obesity in Pima Indian siblings either born before or after their mother was diagnosed with type 2 diabetes. Within the same family, offspring born prior to the mother having developed the disease remained relatively unaffected. Conversely, children born after the diagnosis developed obesity at an early age and were at an increased risk for type 2 diabetes. Therefore, these findings confirm that in addition to any genetic effect, direct exposure to the diabetic intrauterine environment is implicated in the offspring having an increased predisposition to the premature development of metabolic disorders.

Placental Insufficiency and Fetal Origins of Metabolic Diseases

Poor fetal growth and development can occur in the offspring of adequately nourished women or women

with normal metabolic milieu. It is believed that placental insufficiency may be largely responsible for the growth restriction observed in this subgroup of offspring. Placental transfer of nutrients and metabolites is pivotal to fetal growth and development. Interference within this transfer process can lead to placental insufficiencies and a disruption to fetal nutrition, hence disturbing the normal growth of the developing fetus.

Placental insufficiency has been artificially produced in both rats by uterine artery ligation and in sheep by placental embolization. Both models have demonstrated intrauterine growth restriction to be a direct consequence of functional disturbances within the placental nutrient transfer process. To date, these studies have mainly focused on the detrimental effects present during fetal and early postnatal life. Further investigations are indeed warranted to establish whether metabolic disturbances persist into adult life.

Summary and Conclusions

Studies investigating the fetal origins of metabolic disease have confirmed a pivotal role for the *in utero* environment mediating the relationship between poor fetal growth and the subsequent increased risk of developing metabolic diseases in adult life. Disturbances within the critical *in utero* environment may be induced by maternal nutritional insults, abnormalities within the maternal metabolic milieu, or by placental insufficiencies. Animal models have been developed in an attempt to elucidate the mechanistic basis of this adverse metabolic programing. However, there is still an urgent need to explore further the pathogenic mechanisms involved in order to allow suitable intervention studies to be initiated.

The escalating epidemic of obesity and type 2 diabetes may be a consequence of a vicious cycle (**Figure 2**). Exposure to an abnormal *in utero* environment may predispose the offspring to the premature development of metabolic diseases. Consequently, the female offspring that are programed to develop the metabolic disease at a young age may, when pregnant, perpetuate this cycle. Generation after generation then has the subsequent risk of also prematurely developing metabolic diseases such as obesity and type 2 diabetes.

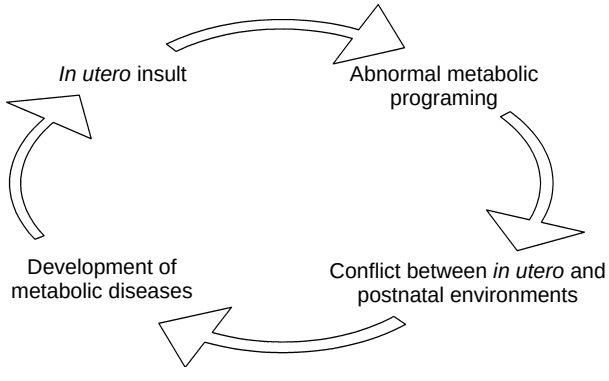


Figure 2 The vicious cycle that may be responsible for the increasing prevalences of metabolic diseases.

The ultimate aim in medical research is to prevent human disease. As maternal nutrition and their metabolic milieu status appears to have such a sizeable influence over the correct functioning of the metabolic processes in the offspring, there is an urgent need to establish ideal nutritional recommendations for pregnant and lactating women. Additionally, ways to treat the occurrence of placental insufficiencies successfully need to be identified. It is of utmost importance to optimize the growth, development, and metabolic programing of the offspring during the critical phase of *in utero* and early life. The development of possible prevention and treatment strategies may therefore aid in combating the epidemic prevalence of metabolic diseases such as obesity, coronary heart disease, and type 2 diabetes.

See also: **Alcohol**: Absorption, Metabolism and Physiological Effects; Disease Risk and Beneficial Effects. **Anemia**: Iron-Deficiency Anemia. **Cancer**: Effects on Nutritional Status. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Early Origins of Disease**: Non-Fetal. **Famine**. **Fats and Oils**. **Iron**. **Low Birthweight and Preterm Infants**: Causes, Prevalence and Prevention; Nutritional Management. **Pregnancy**: Nutrient Requirements; Energy Requirements and Metabolic Adaptations. **Protein**: Requirements and Role in Diet.

Further Reading

Barker DJP (1998) *Mothers, Babies and Disease in Later Life*, 2nd edn. Edinburgh, New York: Churchill Livingstone.

Bauer MK, Harding JE, Bassett NS, Breier BH, Oliver MH, Gallaher BH, Evans PC, Woodall SM, and Gluckman PD (1998) Fetal growth and placental function. *Molecular and Cellular Endocrinology* **140**: 115–120.

Bertram CE and Hanson MA (2001) Animal models and programming of the metabolic syndrome. *British Medical Bulletin* **60**: 103–121.

Dabelea D, Knowler WC, and Pettitt DJ (2000) Effect of diabetes in pregnancy on offspring: follow-up research in the Pima Indians. *Journal of Maternal-Fetal Medicine* **9**: 83–88.

Drake AJ and Walker BR (2004) The intergenerational effects of fetal programming: non-genomic mechanisms for the inheritance of low birth weight and cardiovascular risk. *Journal of Endocrinology* **180**: 1–16.

Godfrey KM and Barker DJP (2001) Fetal programming and adult health. *Public Health Nutrition* **4**(2B): 611–624.

Hales CN and Ozanne SE (2003) The dangerous road of catch-up growth. *Journal of Physiology* **547**(1): 5–10.
Harding JE (2001) The nutritional basis of fetal origins of adult disease. *International Journal of Epidemiology* **30**: 15–23.
Holness MJ, Langdown ML, and Sugden MC (2000) Early-life programming of susceptibility to dysregulation of glucose metabolism and the development of type 2 diabetes mellitus. *Biochemical Journal* **349**: 657–665.
Khan IY, Lakasing L, Poston L, and Nicolaides KH (2003) Fetal programming for adult disease: where next? *The Journal of Maternal-Fetal and Neonatal Medicine* **13**: 292–299.
Newnham JP, Moss TJM, Nitsos I, Sloboda DM, and Challis JRG (2002) Nutrition and the early origins of adult disease. *Asia Pacific Journal of Clinical Nutrition* **11**(supplement): S537–S542.
Ong KKL and Dunger DB (2001) Developmental aspects in the pathogenesis of type 2 diabetes. *Molecular and Cellular Endocrinology* **185**: 145–149.
Ozanne SE and Hales CN (2002) Early programming of glucose-insulin metabolism. *Trends in Endocrinology and Metabolism* **13**(9): 368–373.
Roseboom TJ, van der Meulen JHP, Ravelli ACJ, Osmond O, Barker DJP, and Bleker OP (2001) Effects of prenatal exposure to the Dutch famine on adult disease in later life: an overview. *Molecular and Cellular Endocrinology* **185**: 93–98.

Non-Fetal

L S Adair, University of North Carolina, Chapel Hill, NC, USA

Introduction

A substantial body of evidence supports the hypothesis that adult chronic diseases have origins in early life. The basic premise of research in this field is that nutritional insufficiency during sensitive developmental periods results in structural changes or programing of metabolic functions. In the short term, such changes may enhance survival and spare brain growth at the expense of other organs. In the long run, the cost of such adaptive responses may be an increased risk of chronic disease. The main focus of research has been on fetal origins of adult disease, but there remains substantial potential for nutritional programing of later disease risk during infancy and childhood. The young infant has high energy and nutrient needs to support rapid growth and development. Birth weight typically doubles in the first 4–6 months of life, and length increases by about 30% between birth and 6 months. Many organ systems continue to mature after birth, notably the immunologic, gastrointestinal, and renal systems. This combination of rapid growth and continued development make the infant highly susceptible to the effects of environmental exposures and suboptimal nutrition, which might affect the development of disease risk. Differentiating postnatal from fetal origins is challenging, however, owing to the inevitable link between pre- and postnatal growth.

Instances of purely postnatal effects relate primarily to infant feeding or exposure to pathogens or toxins. The potential effects of infant feeding relate to nutritional adequacy, and to exposure or lack of exposure to specific substances in human milk or human milk substitutes. Effects of feeding may occur independently of the infant's nutritional status at birth. This topic is discussed further in a separate section below.

There is also a continuum of fetal and postnatal effects. Intrauterine growth-restricted infants may experience optimal or even excess postnatal nutrition, or they may continue to be exposed to nutritional insufficiency. Their responses to postnatal challenges may be conditioned by their fetal nutritional history, such that there is an interaction or synergism of fetal and postnatal effects.

Prenatal nutritional insufficiency may be thought to result in 'downsizing.' It may produce smaller organs, for example, kidneys with a reduced nephron number, a pancreas with fewer islet cells, or a low skeletal muscle mass. Nutritional insufficiency may also alter metabolic or hormonal regulation, for example, hormone secretion or sensitivity of the hypothalamic-pituitary axis. In either case, the effects may be permanent, or subject to compensatory responses once nutritional or other insults are removed. For example, a permanently reduced nephron number is a hypothesized mechanism through which fetal growth restriction affects later blood pressure. Similarly, a reduced skeletal muscle mass may persist and affect insulin sensitivity in later life associated with a reduced number of insulin receptors. In such cases, the physiological capacity to respond to risk factors encountered later in life (e.g., diets high in sodium or excess calories relative to energy needs) may be compromised.

Alternatively, catch-up or compensatory postnatal growth may occur. Many infants who were underweight for length at birth typically undergo a period of rapid postnatal compensatory growth in weight, while those who are relatively short at birth have larger length increments (see **Figure 1** for an example from a Philippines infant cohort). A central finding in many studies is that chronic disease risk is most likely to be elevated in individuals who were growth restricted *in utero* and thus small at birth, but relatively large at the time health outcomes were measured, leading to the conclusion that excess

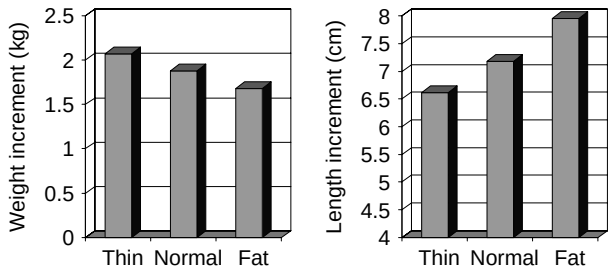


Figure 1 Early growth of Filipino infants is associated with relative weight at birth. Mean growth increments from birth to 2 months of age in children who were relatively thin (BMI $<$ 10th sample percentile) or fat (BMI $>$ 90th sample percentile).

postnatal growth contributes to disease risk. The extent to which rapid postnatal growth itself is a risk factor for the development of chronic disease has been the subject of extensive recent research. The relationship of early growth patterns to later disease risk is discussed in detail in a subsequent section.

Long-Term Effects of Infant Feeding

Much of the literature on the long-term effects of infant feeding is based on comparison of outcomes associated with human milk versus infant formula feeding. Postulated effects relate primarily to the different composition of human milk versus formula and different energy and nutrient intake by infants. The literature does not provide a clear and consistent picture of the long-term effects of feeding. When effects are found, they tend to be modest. Before discussing the results of these studies, it is important to raise several important methodological issues relevant to the interpretation of the literature.

First, breast feeding is a complex behavior chosen by mothers. Women who choose to breast-feed are likely to differ in systematic ways from those who do not. The choice to breast feed and the duration of breast feeding may be related to other short- and long-term health behaviors that affect the ultimate health outcomes of interest. To isolate the effect of infant feeding, it must be assumed that other concurrent and subsequent exposures are not systematically related to feeding history, or such exposures must be taken into account in multivariate analysis. Unfortunately, most studies have insufficient data to adequately control statistically for these other behaviors, particularly since they are often unmeasured or poorly measured.

Second, many studies use historical cohorts in which feeding method is recalled by the mother or based on limited records. While the decision to initiate breast feeding is likely to be accurately recalled, information about breast feeding duration and timing of introduction of other foods may be subject to recall bias.

Third, the composition of proprietary infant formulas has changed since their introduction in the 1920s. For example, sodium levels and fat sources have changed, and new ingredients such as n-3 fatty acids and nucleotides have been added recently. Therefore, results from older versus younger cohorts may differ either because true age-specific effects have emerged or because they were exposed to infant formula of different composition. Furthermore, the effects of breast and formula feeding on infant health are likely to differ depending on the environmental context.

The ideal study design for determining the long-term effects of infant feeding would require randomization to feeding regimens, and frequent follow-up of subjects up to the time when a disease risk factor or outcome is measured. Such designs are rarely ethical or feasible. An exception is a series of studies in the UK conducted by Alan Lucas and colleagues, which assessed long-term outcomes among preterm infants randomized to receive banked human milk or formula, and full-term infants whose mothers chose not to breast-feed randomized to different types of formula. While many of the studies have focused on neurodevelopment, some are now looking at other health outcomes.

Selected Outcomes Related to Infant Feeding

The following are examples of some chronic disease-related outcomes studied in relation to infant feeding. The selected outcomes are intended to be illustrative of a range of effects rather than a comprehensive treatment of all outcomes related to infant feeding.

Serum lipids Based on a systematic review of literature relating infant feeding to blood lipids in infants, adolescents, and adults, total cholesterol was found to be consistently higher in breast-fed infants compared to bottle-fed infants. No consistent differences related to feeding history were found in children and adolescents; and among adults, a majority of studies reported lower mean total cholesterol in those who had been breast-fed. The proposed but unproven mechanism for the protective effect of breast-feeding in adults is downregulation of endogenous cholesterol synthesis.

Blood pressure Differences in the sodium and fat content and composition of breast milk versus formula are thought to be the relevant determinants of

long-term effects of infant feeding on blood pressure. In a recent systematic review, data were compiled to compare exclusive breast feeding to formula feeding, with adjustment for current age, sex, height, and body mass index (BMI). The analysis was based on 26 studies of systolic blood pressure and 24 studies of diastolic blood pressure. On average, subjects who were breast-fed had a modestly lower systolic blood pressure than those who had been formula fed, with an average effect of −1.10 mmHg, and no marked differences by age. However, the analysis suggested publication bias since the effect was significantly larger in small studies than large studies. The studies showed no effects of feeding on diastolic blood pressure.

Taking advantage of a 1980 randomized trial to study the effect of a low or normal sodium diet in Dutch infants, a follow-up study at age 15 years found systolic blood pressure to be 3.6 mmHg lower and diastolic to be 2.2 mmHg lower in the low-sodium group. These results suggest that sodium intake in infancy may affect blood pressure later in life.

Further evidence of the effects of diet composition comes from a long-term follow-up of the Barry Caerphilly Growth study cohort. In this study, mothers and their offspring were randomly assigned to receive a milk supplement or usual care. In young adulthood (age 23–27 years), blood pressure was positively associated with dried formula milk supplement consumed in infancy. The effect was attenuated but remained significant after controlling for current BMI, suggesting an effect of diet composition independent of growth.

Reproductive function The relatively high levels of isoflavones in soy-based infant formula have raised concerns about potential effects on endocrine and reproductive function later in life. A recent retrospective cohort study of young adults who as infants had participated in controlled feeding studies during infancy found no differences associated with soy feeding across a large number of outcomes potentially susceptible to estrogenic or antiestrogenic activity of phytoestrogens, including timing of maturation, sexual development, or fertility in adolescents or adults. Another literature review reported no meaningful differences in child growth related to feeding of soy formula. However, data are limited and further randomized controlled trials are needed to provide definitive evidence.

Growth and body composition Mode of feeding may indirectly affect later disease risk through its effects on energy intake or aspects of metabolic regulation that affect growth and body composition. Numerous studies demonstrate different growth patterns in breast- and formula-fed infants that are hypothesized to reflect differences in nutrient intakes. In fact, evidence of systematic differences in breast- and formula-fed infants has led the World Health Organization to undertake the development of growth charts for breast-fed infants. In one careful study of body composition, total energy intakes and weight velocity from 3 to 6 months of age were higher in formula-fed compared to breast-fed infants. Estimates of fat and fat-free mass also indicate higher adiposity in formula-fed infants, however, none of these differences persisted into the second year of life. Similarly, in a study of nearly 5600 children who participated in the Third National Health and Nutrition Examination Survey, those who had been exclusively breast-fed for 4 months weighed less at 8-11 months than did infants who were fed in other ways, but few other meaningful differences in growth status through age 5 years were associated with early infant feeding.

Longer term effects of infant feeding have been assessed in studies that examined whether breast-feeding protects against later overweight or obesity. A recent review found inconsistent results, with some large cohort studies showing a moderate protective effect, and others showing no effect. The studies were also inconsistent in showing a dose response. An illustrative large study in 3–5-year-old children found that after adjusting for potential confounders, risk of having a BMI > 85th percentile was reduced among exclusively breast-fed children compared with those never breast fed, but there was no reduced risk of having a BMI > 95th percentile.

The findings are typically based on retrospective studies, in which breast-feeding data derive from maternal recall. This makes it difficult, if not impossible, to control for confounding, since a mother's decisions about breast-feeding may relate to subsequent child feeding and other factors associated with overweight. Thus, it is not clear based on the available data whether the effects of infant feeding are causal or whether breast-feeding serves as a marker for other health behaviors that may affect child and adolescent growth. Recent studies among siblings, which allow control for maternal characteristics, show no protective effects of breast-feeding on obesity in adolescents and young adults.

Exposure to antigens and development of autoimmune disease The infant's diet is the main source of exposure to antigens suspected to be related to the development of autoimmune diseases. A likely protective effect of exclusive breast-feeding relates to lack of exposure to food allergens, though some

other protective mechanisms related to specific substances in breast milk have been postulated. Exposure to bovine proteins by milk feeding, and to allergenic plant proteins such as those found in wheat is suspected to increase risk of developing diseases such as type 1 diabetes and celiac disease in genetically susceptible individuals.

Type 1 diabetes is one of the most prevalent chronic diseases with childhood onset. It is characterized by autoimmunity to pancreatic islet cells and is associated with a specific human leucocyte antigen (HLA) genotype. Not all individuals with the genotype develop the disease, suggesting an important role for gene–environment interactions. Hypothesized early exposures include infant feeding and enterovirus infections. Early introduction of cows' milk has received a great deal of attention as a potential risk factor. Numerous case–control studies associate increased risk with cows' milk, but a nearly equal number of studies show no effects. These retrospective studies have been criticized as suffering from recall bias and inappropriate control groups, for example, controls without the susceptible genotype. Recent prospective studies of at-risk infants in Australia and Germany found no association of type 1 diabetes with feeding of cows' milk. However, pilot study data from an international primary prevention trial suggests that eliminating cows' milk proteins in at-risk infants reduces risk of developing islet cell autoantibodies. This study also supports a role for early enterviral infections in the etiology of type 1 diabetes in genetically susceptible individuals. In fact, the research team has suggested that the effect of cows' milk may depend on viral exposures.

Recent studies suggest a role for other food antigens. A study of at-risk German children found that feeding of gluten-containing foods before 3 months of age was associated with risk of having pancreatic islet cell autoantibodies. Another study in the US also found an increased risk of islet cell autoimmunity among at-risk children given cereal before 3 months or after 7 months of age. Furthermore, they found that risks associated with cereal introduction were reduced by breast-feeding.

Other aspects of diet may have immunomodulatory effects. Vitamin D and the n-3 fatty acids EPA and DHA are suggested to be protective against immune-modulated diseases. For example, in a case–control study, Norwegian children given cod liver oil, a rich source of EPA and DHA, in the first year of life had significantly reduced risk of type 1 diabetes.

Type 2 diabetes Few studies have assessed the relationship of infant feeding to later development of Type 2 diabetes. Early feeding may affect patterns of insulin secretion in the newborn period, and thereby program subsequent development of metabolic control. Two studies in native American populations, one in Canada and one among Pima Indians, report a protective effect of breast-feeding on later development of Type 2 diabetes. In the Pima study, exclusive breast-feeding in the first 2 months of life was associated with a lower rate of Type 2 diabetes in children and adults. In the Canadian study, breast-feeding for more than 12 months was associated with decreased risk of Type 2 diabetes. Other studies have examined early risk factors related to subsequent development of Type 2 diabetes. For example, in a study of preterm infants randomized to human milk or formula of different composition, 32-33 split proinsulin, a marker of insulin resistance, was elevated in adolescents who had received a nutrient-enriched diet compared to those with a lower nutrient diet.

In sum, infant feeding, through nutritional adequacy, direct exposure to antigens, and protective substances provided in human milk, has the potential to alter response to subsequent exposures and to directly influence the beginning of disease processes.

Postnatal Growth and Later Risk of Disease

Small body size in childhood may reflect nutritional insufficiency that may program adult disease in ways similar to that observed in the fetal period. Independent of birth weight, low weight at 1 year of age has been associated with increased risk of cardiovascular disease in adult men. Similarly, poor childhood growth manifested as short stature has been linked with insulin resistance.

More attention has recently been paid to the effects of rapid childhood growth in height and weight. The observation in much of the fetal programing literature that effects of birth size emerge or are strengthened when current body size (typically represented as BMI) is taken into account suggests an important role for postnatal growth in the origins of adult disease. Individuals who are born small, but who end up relatively large (taller or heavier than their peers) have clearly experienced more rapid growth at some point between birth and when health outcomes and current size are assessed. Whether rapid growth is an independent risk factor or whether it confers increased risk only in individuals with a history of intrauterine growth restriction is a question requiring further research. Moreover, even when strong associations of growth rate and chronic disease risk are found, it is unclear whether the association is causal

or whether growth serves as a marker for other underlying causal processes.

Postnatal growth is clearly related to prenatal growth. Some metabolic changes associated with prenatal nutritional sufficiency may affect postnatal physiology and behavior that, in turn, affect growth. In addition, there is intriguing evidence from animal studies that prenatal nutritional restriction alters appetite and induces hyperphagia, and also reduces physical activity in adult animals (see **Figure 2**). If true in humans, this would be an important pathway by which disease risk is affected. Suggestive evidence comes from human infants whose cord blood leptin levels at birth were inversely related to weight gain in the first 4 months of life, independent of birth weight. Leptin may relate to subsequent growth by affecting appetite and energy intake.

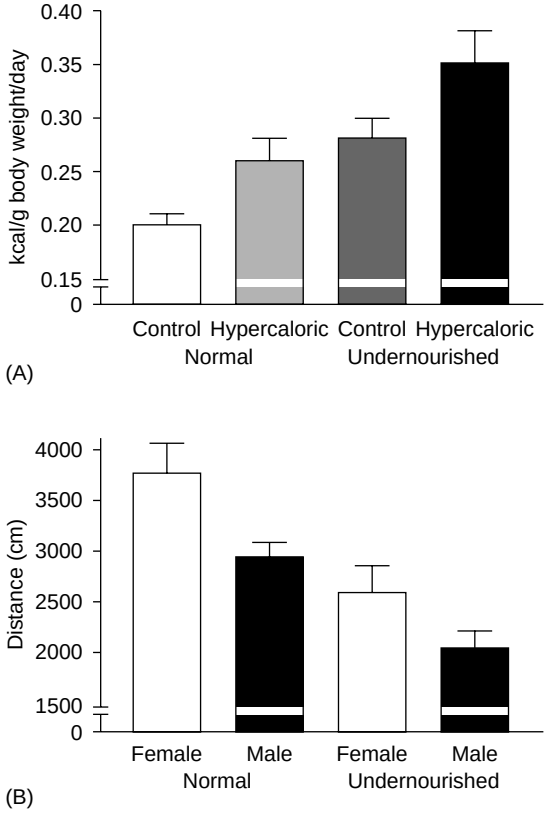


Figure 2 Locomotor behavior and food intake in Wistar rats as a consequence of a normal or adverse fetal environment ($n = 6$–8/group). (A) Food intake (kcal per gram body weight per day over a 5-day period) in females at day 145; $P < 0.005$ for effect of fetal programing, $P < 0.05$ for postnatal hypercaloric diet. (B) Locomotor activity at 14 months in males and females; $P < 0.005$ for effect of fetal programing and gender. Data analyzed by factorial ANOVA, and data are shown as means ±SE. (Reproduced from Vickers MH, Breier BH, McCarthy D, and Gluckman PD (2003) Sedentary behavior during postnatal life is determined by the prenatal environment and exacerbated by postnatal hypercaloric nutrition. *American Journal of Physiology. Regulatory Integrative and Comparative Physiology* **285**(1): R271-273 with permission from the American Physiological Society.)

Depending on the outcome under study, there are differences in whether linear growth or growth in weight, particularly weight relative to height, matters. Most often, more rapid weight gain is the risk factor, owing to the fact that excess adiposity is an important risk factor for many chronic diseases of adulthood. Another key issue concerns the timing of effects. There is controversy about whether early infancy compensatory growth following intrauterine growth restriction confers risk, or whether it is only later growth that matters.

Where many potential adverse outcomes might be affected by postnatal growth, the following sections focus on adiposity, blood pressure and coronary heart disease, insulin resistance and diabetes, and cancer.

Adiposity and Obesity

Early undernutrition followed by later overnutrition as well as early overfeeding independent of prior growth restriction are thought to increase risk of later obesity. Rapid postnatal weight gain occurs in a significant proportion of infants who are born small for gestational age. Prospective studies in US, South African, and British cohorts show that rapid growth in early infancy increases later risk of overweight. Longitudinal data from the US National Perinatal Collaborative study show that, independent of birth weight, one-third of obesity at age 20 is attributable to rapid weight gain in the first 4 months of life. In a Bristol, UK cohort, nearly one-third of children had an increased weight standard deviation (SD) score of more than 0.67 units from birth to age 2 years, and these children remained fatter, having more central fat distribution at age 5 years compared to children with lower early growth rates. Similarly, data from the South Africa Birth to Ten cohort showed that children with rapid weight gain in infancy were significantly lighter at birth and significantly taller, heavier, and fatter throughout childhood.

Early postnatal growth rates may program insulin-like growth factors, IGF-I and IGF-II. **Figure 3** illustrates this point with data on 5-year-old children from Bristol, UK in whom IGF levels were strongly related to current body size, but also that, independent of current size, children who had experienced catch up growth (change in Z-score >0.67 SD) from birth to age 2 had higher IGF levels. Childhood IGF levels are important as determinants of later linear growth and timing of puberty, and are associated with later risk of hormone-dependent cancers.

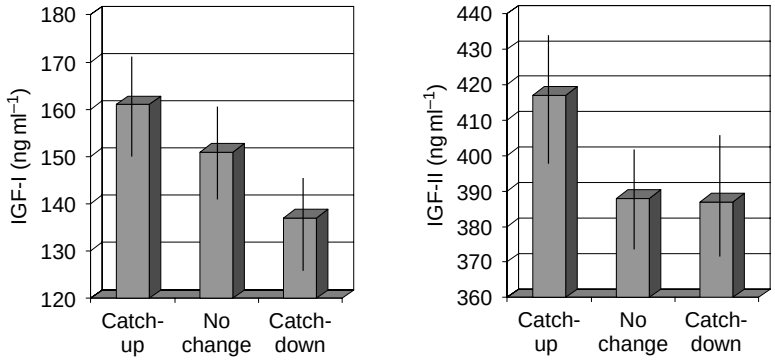


Figure 3 Hormone levels at age 5 years by change in weight Z-score from birth to 2 years of children in the ALSPAC cohort: means and 95% confidence intervals of IGF-I and IGF-II, adjusted for fat mass and fat-free mass. (Drawn from data from Ong K, Kratzsch J, Kiess W, Dunger D, and ALSPAC Study Team (2002) Circulating IGF-I levels in childhood are related to both current body composition and early postnatal growth rate. *Journal of Clinical Endocrinology and Metabolism* **87**(3): 1041–1044.)

Cancer

A large body of literature relates adult height to cancer risk, with the largest volume of evidence on breast, prostate, and colorectal cancers. In each case, risk of disease is increased with taller stature. A role for accelerated childhood growth is inferred, since taller individuals have experienced more linear growth. Possible mechanisms fall into two categories: childhood growth as a marker for other exposures that influence risk (fetal exposures, infections, timing of puberty, and energy intake) or growth as a mediator of risk (effects of growth promoting hormones such as IGF-I and IGF-II).

Few studies have directly addressed the effects of childhood growth, owing to lack of longitudinal data. Based on data from the UK Boyd Orr cohort, a one SD difference in height was associated with a 42% higher risk of overall cancer mortality in later life among males, but no effects were found in females. In another UK birth cohort, risk for breast cancer was elevated among women who were large at birth and tall at age 7. Based on data from the US Nurse's Health Study, rapid adolescent growth was associated with an increased risk of both pre- and postmenopausal breast cancer.

Blood Pressure and Coronary Heart Disease

Blood pressure is the one of the most well-studied outcomes in the context of fetal programing, with fairly consistent findings of a modest inverse relationship of birth weight to adult systolic blood pressure that increases with age. Substantial evidence demonstrates a synergistic relationship of fetal growth restriction with rapid postnatal growth. **Figure 4** presents the classic picture for systolic blood pressure: the highest pressure is found among adolescent males who were relatively thin at birth, but relatively heavy as adolescents. Current BMI is typically the strongest anthropometric predictor of blood pressure, but at the same BMI, those with a history of fetal growth restriction have higher mean blood systolic pressure and increased risk of having high blood pressure.

Owing to the existence of good longitudinal growth data in Scandinavia, child growth trajectories can be traced for individuals with and without hypertension or other adverse outcomes such as coronary heart disease. As shown in **Figure 5**, though initially smaller, adults with hypertension diverged in their BMI trajectory and were relatively heavier after age 7 compared to those without hypertension.

There remains controversy about the age at which higher growth rates pose risk of later disease. Some studies show elevated blood pressure in association with rapid weight gain in infancy, while other studies show no effect, or a protective effect (infants with larger weight increments have lower blood pressure

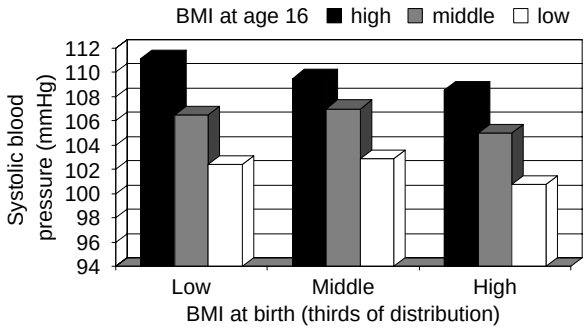


Figure 4 Synergistic effect of BMI at birth and age 16 on systolic blood pressure of Cebu (Philippines) boys: ■ high; ■ middle; and □ low BMI. Data from the Cebu Longitudinal Health and Nutrition Survey.

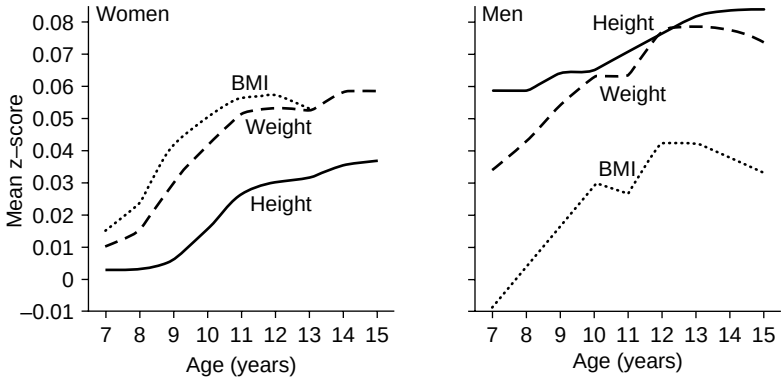


Figure 5 Z-scores for height, weight, and BMI from 7 to 15 years in 975 boys and 983 girls who later developed hypertension. Mean values for all 7086 subjects in cohort are zero. (Reproduced with permission from Eriksson J, Forsen T, Tuomilehto J, Osmond C, and Barker D (2000) Fetal and childhood growth and hypertension in adult life. *Hypertension* **36**(5): 790–794.)

as adults). The degree to which rapid infant growth represents risk may depend on whether it occurs in the context of recovery from fetal growth restriction and results in normalization of body weight versus excess growth leading to infant obesity.

There is more consistent evidence of increased risk associated with rapid weight gain in later childhood. In a Philippines cohort, larger weight increments from age 8 to 15 years increased risk of high blood pressure in boys who were relatively thin at birth. However, higher childhood weight gain in the absence of fetal growth restriction was not a risk factor in this population.

Fetal undernutrition may result in a reduced number of nephrons. Such deficits may not increase disease risk in individuals who remain small, but excess growth may challenge the ability of the kidneys to effectively regulate blood pressure. Catch-up linear growth has not been consistently implicated as a risk factor for later elevated blood pressure. In fact, continued poor linear growth, particularly in association with more rapid weight gain, increases risk of later elevated blood pressure.

Insulin Resistance and Diabetes

Most evidence relates to type 2 diabetes, but one large, population-based case–control study of type 1 diabetes in European populations found that height and weight were higher in cases starting at 1 month after birth, with maximum differences in cases and controls between 1 and 2 years of age. In the case of type 2 diabetes, both continued growth faltering in infancy and more rapid growth are associated with increased risk. Postnatal faltering in length is associated with impaired insulin metabolism.

As was the case for blood pressure, highest risk is associated with the combination of small size at birth and rapid postnatal growth gain. In a well-studied cohort in Finland, men and women who developed type 2 diabetes had lower birth weight, length, and ponderal index, and accelerated growth in weight and height from age 7 to 15 years. Precursors of diabetes such as insulin resistance have been studied. For example, in a follow-up study of British children who were born preterm, fasting split proinsulin and glucose concentration 30 min after a glucose load were highest in children with the greatest increase in weight centile between birth and time of measurement, regardless of early size.

Extensive studies of early origins of type 2 diabetes have been conducted in India, where rates are rising very rapidly. Indian babies who are small at birth have a deficit in skeletal muscle, but not body fat compared to normal size infants. These infants tend to grow into adults that retain a lower skeletal muscle mass, but have increased abdominal obesity. This body composition is strongly related to increased risk of type 2 diabetes. Prospective studies of Indian children show an interaction between birth weight and subsequent growth. For example, children who were born small but were relatively large at age 4 had higher plasma glucose and insulin concentrations 30 min after an oral glucose load, and greater insulin resistance at age 8.

Higher growth rates in previously growth-restricted individuals may pose excessive demands on systems initially adapted to function in the face of limited resources, leading to increased risk of diseases, particularly those associated with the metabolic syndrome. Rapid growth in weight during infancy and childhood, and in particular, rapid growth following prenatal growth restriction, increases risk of developing obesity, especially abdominal obesity. Factors that contribute to early onset of obesity are therefore important to control, since obesity tracks from early life to adulthood, and is a well-recognized risk factor for diseases such as type 2 diabetes, hypertension, and coronary heart disease.

In sum, the continued vulnerability and responsivity of the developing infant and child suggest the importance of a life course perspective on the development of diseases that are typically thought of as 'adult onset.'

See also: **Breast Feeding**. **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Coronary Heart Disease**: Hemostatic Factors; Lipid Theory. **Diabetes Mellitus**: Etiology and Epidemiology. **Hyperlipidemia**: Overview; Nutritional Management. **Hypertension**: Dietary Factors. **Infants**: Nutritional Requirements. **Lipids**: Chemistry and Classification.

Further Reading

Cameron N and Demerath EW (2002) Critical periods in human growth and their relationship to diseases of aging. *Yearbook of Physical Anthropology* **45**: 159–184.

Dewey KG (2003) Is breast feeding protective against child obesity? *Journal of Human Lactation* **19**: 9–18

Huxley RR, Shiell AW, and Law CM (2000) The role of size at birth and postnatal catch-up growth in determining systolic blood pressure: a systematic review of the literature. *Journal of Hypertension* **18**: 815–831.

Lucas A (1998) Programming by early nutrition: An experimental approach. *Journal of Nutrition* **128**: 401s–406s.

Martorell R, Stein AD, and Schroeder DG (2001) Early nutrition and later adiposity. *Journal of Nutrition* **131:** 874s–880s.

Okasha M, Gunnell D, Holly J, and Davey Smith G (2002) Childhood growth and adult cancer. *Best Practice and Research Clinical Endocrinology and Metabolism* **16**: 225–241.

Ong KK and Dunger DB (2002) Perinatal growth failure: the road to obesity, insulin resistance and cardiovascular disease in adults. *Best Practice in Research Clinical Endocrinology and Metabolism* **16**: 191–207.

Owen CG, Whincup PH, Gilg JA, and Cook DG (2003) Effect of breast feeding in infancy on blood pressure in later life: a systematic review and meta-analysis. *British Medical Journal* **327**: 1189–1195.

Owen CG, Whincup PH, Odoki K, Gilg JA, and Cook DG (2002) Infant feeding and blood cholesterol: A study in adolescents and a systematic review. *Pediatrics* **110**: 597–608.

Singhal A and Lucas A (2004) Early origins of cardiovascular disease: is there a unifying hypothesis? *Lancet* **15; 363(9421):** 1642–5.

Wasmuth HE and Kolb H (2000) Cow's milk and immune-mediated diabetes. *Proceedings of the Nutrition Society* **59**: 573–579.

Yajnik CS (2003) Nutrition, growth, and body size in relation to insulin resistance and type 2 diabetes. *Current Diabetes Reports* **2**: 108–114.

Eating Behavior *see* **Meal Size and Frequency**

EATING DISORDERS

Contents

Anorexia Nervosa

A R Rolla, Harvard Medical School, Boston, MA, USA

The relentless pursuit of thinness and the increasing prevalence of obesity in modern societies are the roots of the present higher prevalence of eating disorders. These eating disorders can be classified according to the interaction between the preoccupation with food and body weight and the self-control of hunger (**Figure 1**).

Classification of Eating Disorders

Obesity

Obesity can be classified as an eating disorder since, primarily or secondarily, obese patients eat

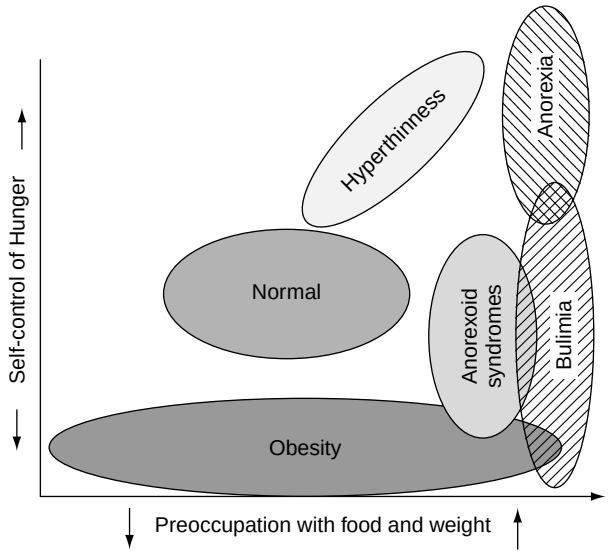


Figure 1 Classification of eating disorders based on the interaction between the preoccupation with food and body weight and the self-control of hunger. © 1999 Academic Press.

inappropriately for their increased weight and because obese individuals tend to suffer also from the other eating disorders.

Anorexia Nervosa

Anorexia nervosa is usually seen in younger women who restrict their food intake and increase exercise, causing a voluntary, stubborn malnutrition.

Bulimia

People who cannot control their hunger over a long period of time tend to have secret binging episodes. This is followed by an overwhelming feeling of guilt and depression, which frequently leads to self-induced vomiting. For this reason, the terms 'bulimia' (which means binge eating) and 'self-induced vomiting' are used interchangeably.

Anorexoid Syndromes

These abnormalities are seen in individuals who can no longer control their weight by dieting and exercising and have to resort to abnormal subterfuges, such as the following:

Self-induced vomiting
Ipecac abuse
Laxative abuse
Diuretic abuse
Anorexic agents abuse
Self-induced glycosuria in patients with insulin-dependent diabetes mellitus
Thyroid hormone abuse
Excessive, compulsive exercising

Professional Hyperthinness

This is a borderline condition, not necessarily pathological, in which individuals, usually with narcissistic tendencies, overvalue personal appearance and thinness as a way of obtaining professional success. It is commonly seen among models, figure skaters, ballerinas, artists, gymnasts, etc. They do not use the 'subterfuges' of the anorexoid patients; they are not socially isolated; their weight loss is not extreme; they have normal psychosexual activity; and they do not see themselves as overweight, unlike people with anorexia nervosa. For them, thinness is a means of obtaining success, not the final goal as in anorexia nervosa.

Anorexia Nervosa

Anorexia nervosa is a serious disease.

Psychological Disturbances

Psychological disturbances are most likely to be the initial event; they result in a complex obsession characterized by the following features:

1. An intrusive body image delusion makes the patients see themselves as being overweight when they are actually severely undernourished. This leads to a pathological fear of fatness (dysmorphophobia), a chronic voluntary starvation, and resistance to any external pressures to gain weight. Anorexic patients hide and dispose of food in the most ingenious ways to avoid eating.
2. An overwhelming sense of personal ineffectiveness makes anorexic patients believe that they cannot control the world around them. They continuously fear that they are going to lose their inner control. They therefore tightly control their world inside and slowly separate themselves from their social surroundings, with growing feelings of alienation and loneliness. There is no psychosexual development or interest, and no dating, unlike patients with bulimia.
3. Depression occurs which may be primary or secondary, obvious or atypical, and may or may not be amenable to treatment with psychotherapy and/or antidepressant medications.
4. Increased physical activity coexists with an apparent lack of hunger and fatigue and is inappropriate for the degree of malnutrition and depression.

Malnutrition

Anorexia nervosa is a self-imposed starvation. The term 'anorexia' is a misnomer since (at least initially) these patients are hungry. Anorexia nervosa is different from other forms of malnutrition because it is voluntary, resistant to nutritional treatments, accompanied by increased physical activity, without an initial organic cause (such as malignancy or surgery), and without associated infections. Because anorexia nervosa is a state of pure malnutrition without associated increased energy expenditure due to fever, immune response, or tissue reconstruction, these patients have a lowered metabolic rate and do not tend to develop opportunistic infections.

The exact mechanism by which these patients are able to control their hunger is unknown. Disturbed brain neurotransmitter activity may be implicated primarily or secondarily.

Endocrine and Metabolic Changes

Amenorrhea, decreased metabolic rate, hypothermia, hypotension, bradycardia, lanugo hair, carotinodermia, leucopenia, osteoporosis, etc. are mostly secondary to the severe malnutrition and are reversible with weight gain.

Other Clinical Characteristics

Anorexia nervosa is more frequent among daughters of white, affluent, achievement-oriented families in developed societies; it is extremely uncommon in areas of the world with poor nutrition. It tends to occur during the last years of high school or at the time of departure to a university or college.

The patients tend to be well-behaved perfectionists with good academic performance. Mothers of anorexic patients have a higher incidence of obsessive–compulsive personalities and preoccupation with diets and weight loss. Anorexic patients have done everything their mothers or families have trained them to do and, when faced with the increasing demands (and choices) of adult life, they exaggerate the only control left in their lives: *hunger*. There is an unconscious wish to revert to childhood, or to a prepubertal state, by means of undernourishing.

The onset of anorexia nervosa is usually subacute, over a period of weeks, not uncommonly after an episode of weight gain or after somebody has made a comment about the patient being overweight. Initially, it appears as an innocent attempt to lose weight, but soon thereafter it starts showing its rebellious and progressive nature. Anorexia nervosa appears in small epidemics in cities and countries, probably owing to social pressures and to imitation behaviors.

In contrast to their poor dietary intake, these patients have a paradoxical enhanced interest in nutrition and cooking. They collect recipes, read nutrition textbooks, plan a career in nutrition or cooking, or find a job in a restaurant (usually waitressing). Anorexic patients enjoy cooking and feeding the rest of the family. They know the precise energy content of all usual food and use their knowledge to select low-energy items.

When forced to eat, anorexic patients will dispose of or hide food. They use their above average intelligence to overcome all efforts to make them gain weight. They can be very resourceful in tampering with scales (adding weights to shoes or clothing, drinking large amounts of water just before weighing, etc.), and they have the most imaginative excuses as to why they are not gaining weight. They are extremely manipulative and master the art of confusing the different members of the treating team and family in their favor and against each other.

As they lose their natural insulation (subcutaneous tissue), it is difficult for anorexic patients to maintain their body temperature. They wear layer upon layer of clothing, which also helps them to hide their malnutrition. Lack of body fat is sensed by the hypothalamus as a sufficiency of stored energy, and therefore the cycling and amplitude of gonadotrophins decrease. This leads to hypothalamic amenorrhea, although approximately 30% of patients stop having menses before there is a significant weight loss. Depression may be another cause of hypothalamic amenorrhea in these patients. Some will remain amenorrheic for several months after regaining normal weight.

In primary or classical anorexia nervosa, patients lose weight by dieting (restrictive) and exercising. These patients tend to be younger, more naive, introverted, and obsessive, and they do not resort to subterfuges to lose weight. Their serum electrolytes, checked at frequent intervals, should be completely normal. Some patients find it impossible to control their hunger and start having binge episodes followed by forced vomiting (bulimia plus anorexia—'bulimarexia'). Others may start abusing laxatives or diuretics as they grow older.

There are patients in whom anorexia is secondary to an underlying, more serious psychiatric disorder, such as depression, schizophrenia, hysteria, or borderline personality disorder. In these cases, the course is longer and depends on the primary condition, as does its treatment. It is very uncommon for men to have anorexia nervosa, in men the condition tends to be associated more often with psychiatric problems and with homosexuality.

Physical Examination

The profound weight loss and cadaveric appearance contrast with the patient's increased physical activity. While hospitalized, if allowed, these patients try to perform some of the nursing chores or even to counsel other patients. Many patients exercise secretly in their rooms and jog or go for long walks when not supervised.

Pubic and axillary hair is preserved, and there is an increase in light, thin hair ('peach fuzz') on the face and neck, back, arms, and thighs (lanugo hair). Patients have low body temperature and poor tolerance to cold exposure because of their malnutrition-induced lowered metabolic rate and the loss of insulation of the diminished subcutaneous tissue. Layers of clothing tend to hide their cachectic appearance. Bradycardia and hypotension are common and secondary to decreased sympathetic drive due to malnutrition.

The skin is dry and has a peculiar bluish erythema over the knuckles and knees. Orange-yellow discoloration of the skin (carotinodermia), seen in palms and soles, is frequently found. It is caused, at least in part, by an increased intake of vegetables, since it may also be seen among vegetarians.

Symptoms

Symptoms are amazingly few. Usually these patients are forced to see a physician by their families. Spontaneous complaints may be amenorrhea, constipation, abdominal pain or distension after eating, 'fluid retention,' and inability to lose weight, for which they may ask to be placed on special diets.

Laboratory Investigations

The following findings are typical:

Mild normochromic, normocytic anemia.

Leucopenia with relative lymphocytosis.

Low sedimentation rates due to low fibrinogen levels.

Serum albumin and transferrin levels are normal, except in severe cases.

Serum carotene and cholesterol levels are normal or slightly elevated, which helps to rule out malabsorption.

Low normal blood glucose levels are found, with low levels of glycohemoglobin.

Electrolyte abnormalities, particularly low serum potassium values, occur only where there is self-induced vomiting or abuse of diuretics or laxatives.

Low serum levels of luteinizing hormone, follicle-stimulating hormone, and estradiol.

Increased growth hormone levels with decreased levels of insulin-like growth factor 1 (somatomedin C) in the serum.

Plasma renin activity and aldosterone levels may be very high in patients who abuse laxatives or diuretics (pseudo-Bartter's syndrome).

Electrocardiography shows sinus bradycardia, flat or inverted T waves, and prolonged QT_c.

Decreased bone density is due to decreased estrogen and progesterone secretion, decreased calcium and vitamin D intake, protein malnutrition, and increased marrow fat content of the bones. The conversion of hematopoietic to fatty marrow is related to the severity of the malnutrition and may be demonstrated with magnetic resonance imaging.

Endocrine Changes

Insulin

Low serum insulin levels occur with increased glucagon concentration. There is a tendency to asymptomatic low blood glucose levels and a low glycohemoglobin concentration. Fasting ketosis may be seen. The number and affinity of insulin receptors in target cells are increased, and abnormal glucose tolerance occurs due to prolonged fasting.

Adipose Tissue Hormones

The adipose tissue secretes different hormones called adipocytokines. Their secretion seems to vary in relation to the amount of adipose tissue accumulated, although the exact mechanism is not known. During profound weight loss, as in anorexia nervosa, there is a marked decrease in the adipose tissue mass with the typical changes in adipocytokines secretion that occur in these circumstances. One of the most studied adipocytokine changes is decreased leptin secretion. Increased fat mass stores are accompanied by an increased leptin secretion; decreased fat mass stores decrease leptin secretion. Low serum levels of leptin reaching the hypothalamus increase the activity of the 'hunger center,' in part by increasing the local activity of neuropeptide Y. Individuals with anorexia nervosa have very low levels of leptin in blood and cerebrospinal fluid, in relation to their decreased adipose tissue. This should cause an increase in hypothalamic neuropeptide Y content and hunger, but this compensatory mechanism to maintain a normal body weight does not seem to be effective in anorexic patients. Another important effect of the serum levels of

leptin on the hypothalamus is the modulation of the gonadal axis. Low levels of leptin are associated with decreased activity of the gonadal axis, and this explains the relationship between starvation and hypogonadism. After nutritional rescue and weight regain, the levels of leptin in the serum become normal.

Gonadal Axis

The female hypothalamus needs to 'sense' the presence of approximately 14–18 kg of body fat in order to allow fertility and menstrual cycles. With lesser amounts of fat, there is a progressive regression to the prepubertal state (low, nonspiking serum gonadotrophin levels). The signal from the fat stores to the gonadal hypothalamus seems to be the level of serum leptin. Very low levels of serum leptin, secondary to the decreased fat mass, seem to result in a decrease in luteinizing hormone-releasing hormone (LHRH) secretion. The hypothalamic, hypogonadal state of anorexia nervosa is due to the combined effects of malnutrition and the psychological disturbances on the hypothalamus. Secretion of LHRH and gonadotrophins improves as weight is regained and leptin levels increase; however, in up to one-third of these patients menses do not return immediately after nutritional rescue and weight restoration are accomplished. The decreased estrogen secretion from the ovaries causes a significant loss of bone mass at a critical time, which will subsequently aggravate postmenopausal osteoporosis.

In males, malnutrition seems to have a less important influence on the gonadal axis. Severe weight loss of long duration decreases serum testosterone and gonadotrophin levels, but to a lesser degree than in women.

Thyroid

Thyroid-stimulating hormone levels are normal but there is a delayed response to thyrotrophin-releasing hormone. Serum thyroxine, both total and free, is normal. The level of serum T_3 is low owing to decreased peripheral conversion (euthyroid sick syndrome), and there is a concomitant increase in reverse T_3. The basal metabolic rate is decreased by 20–30% and not fully corrected with T_3 replacement since it is also due to decreased sympathetic activity.

Sympathetic Nervous System

There is decreased peripheral sympathetic activity, with normal adrenomedullary function. This is due to decreased ingestion of energy, and it explains the tendency to bradycardia, postural hypotension, and low basal metabolic rate.

Adrenal Cortex

Serum cortisol levels are slightly raised, without diurnal variation, and may not suppress with dexamethasone overnight. Urinary 17-hydroxy and 17-keto steroids are decreased by 30–50%, but urinary-free cortisol may be increased. Corticotrophin-releasing factor (CRF) stimulation causes a subnormal corticotrophin rise but a normal or supernormal serum cortisol response. Levels of CRF in the cerebrospinal fluid are elevated. These changes in the hypothalamic–pituitary–adrenal axis are very similar to those seen in untreated depression.

Growth Hormone

Basal and pulsatile secretion of growth hormone (GH) is increased, with a peripheral resistance to its effects. Serum GH levels are elevated in 60% of patients, particularly in the most severe cases. This is due to decreased feedback from lowered serum concentrations of insulin-like growth factor 1 (IGF 1) and to increased serum levels of ghrelin. Growth hormone levels do not rise normally after L-dopa or insulin hypoglycemia, but there may be an unexpected rise in GH blood levels after stimulation with thyrotrophin-releasing hormone.

Ghrelin is a recently discovered polypeptide secreted by the stomach that increases in circulation with weight loss. Ghrelin is an activating ligand for the GH secretagogue receptor in the hypothalamus. With starvation and weight loss, the increased serum levels of ghrelin increase the release of GH and hunger. Individuals with anorexia nervosa have high serum levels of ghrelin that return to normal with normalization of body weight.

Vasopressin

There is decreased capacity to concentrate urine due at least in part to sluggish vasopressin secretion in response to osmotic stimuli. Levels of vasopressin in the cerebrospinal fluid are increased.

Inflammatory Cytokines

Inflammatory cytokines have important endocrinometabolic effects in people with infections and neoplasias, including anorexia and weight loss. In individuals with anorexia nervosa the serum levels of interleukin-1β, interleukin-6, tumor necrosis factor-α, and their soluble receptors are lower than normal, which may be due to decreased adipose production of these cytokines due to the decreased fat mass.

Hypothalamic Control of Hunger in Anorexia Nervosa

In normal individuals fasting and weight loss increase hunger by multiple mechanisms (decreased serum levels of leptin, insulin, and blood glucose and increased levels of ghrelin). At the level of the hypothalamus there is an increase in the potent orexigenic neuropeptide Y and other changes in neurotransmitters secondary to the fasting state. Some of these neurotransmitter changes may be the cause or a mechanism of anorexia nervosa, and for this reason they have received considerable attention in the past several years. It is important to understand that appetite control is a very complex hypothalamic function that involves many local and systemic neuropeptides, amines, and hormones.

Abnormal serotonin activity has been found in the brain of women with anorexia nervosa. An area in the chromosome 1 (p36.3–34.3) that contains genes for the serotonin 1D receptor and for the opioid delta receptor was associated with patients with anorexia nervosa by linkage analysis. One polymorphism in the Agouti related protein (Ala67Thr) has also been found associated with anorexia nervosa. Melanocortin system stimulants in the hypothalamus, such as Agouti related protein, are also involved in appetite and energy regulation. On the other hand, these genetic abnormalities may amount to only a biological tendency and do not explain the relatively short term of the illness during a life time or the changes in prevalence in the past decades.

Bone Density

Decreased estrogen and progesterone secretion; low serum levels of IGF-1; increased levels of serum cortisol; malnutrition with protein, calcium, and vitamin deficiencies; and fatty degeneration of the bone marrow lead to decreased bone density. Increased exercising does not counteract this-osteopenic tendency, which affects mostly young women during the years of skeletal growth. The osteopenia of anorexia nervosa is mostly asymptomatic, but some patients may present with stress fractures (diagnosed only with bone scans) related to their increased exercising. Many of these patients do not achieve their peak bone density even after their nutritional recovery and restoration of menses, and they are left with a propensity to fracture bones for the rest of their lives. Treatment with estrogens, calcium, and vitamin D is mildly effective. IGF-1 and DHEA-S have been used with partial success. Rapid restoration of nutrition seems to be the best management of anorexic osteopenia.

Differential Diagnosis

In the majority of cases the severe and voluntary malnutrition accompanied by the typical delusion of being fat and resistance to gain weight make the diagnosis very clear. Malnutrition due to organic causes in adolescents usually has an obvious reason and the patients want to improve their nutrition. Hypothalamic tumors may rarely present with severe loss of appetite.

The differential diagnosis should include the anorexoid syndromes. In pure anorexia nervosa the weight loss is due only to restrictive eating habits and exercise. Some anorexic patients may start binging and inducing vomiting, in which case their condition is called bulimarexia.

In some cases, anorexia nervosa is secondary to a serious, underlying psychiatric illness, with the weight loss being only an added problem. A particular diagnostic and therapeutic dilemma may occur for young women with personality disorder or chronic schizophrenia and anorexia nervosa.

Treatment

The multifaceted pathogenesis of anorexia nervosa requires an experienced team of psychiatrists, nutritionists, endocrinologists, internists or pediatricians, and nurses. *Each patient should be considered individually* since there are as many variations as there are patients. It is important to maintain communication between the different members of the team in order to present a unified front to the patient. Invariably, the patient will try to find and exploit the most minimal differences of opinion between the members of the team. Ideally, all important decisions should be made by one central team leader. Nurses, aides, and other paramedical personnel should be instructed about how to deal with the patient's behavior and charming search for allies.

There is no specific treatment and the methods reported are, at best, controversial. The etiology of this disorder is unknown, and etiological factors are probably different in each patient. It is therefore important to tailor the therapeutic approach to each patient.

Many cases of established and severe anorexia nervosa require prolonged hospitalization for psychological and nutritional rescue. Separation from parents and home environment is only part of what is to be gained from hospitalization. Administrators

and health insurance companies must understand this need.

Hospitalization is indicated for the following:

Severe and rapid weight loss
Serious metabolic or cardiovascular problems (hypokalemia less than 2.5 mmol l^{-1}) despite oral replacement, blood urea nitrogen more than 10.6 mmol l^{-1} of urea (30 mg dl^{-1}) in the presence of normal renal function, pulse less than 45 min^{-1}, systolic blood pressure less than 70 mmHg, or a body temperature less than 36 °C
Severe depression and suicide risk
Psychosis
Family crisis

Psychiatric Treatment

From the outset, the entire family should be interviewed to gain insight into the patient's previous behavior and to understand the family dynamics and enlist their help in therapy. Clear, simple contracts with the patient are a form of behavior modification that is simple to carry out. Initially, most daily activities and visits are curtailed and the patient is watched, particularly around mealtimes. As the patient improves, restrictions are lessened and privileges increased. Short-term goals are set from the beginning. Weight gains of 250 g daily or 1.3–1.8 kg per week are acceptable limits. Patients who accomplish these goals are rewarded by increasing levels of activity and autonomy within the hospital as a positive reinforcement.

The general attitude of the team should be one of understanding, concern, and firmness. One should try to build a trusting relationship in which the patients feel understood but without giving them a chance to deceive. The nature and course of the illness should be clearly explained to the patient and the family. This includes the serious complications of malnutrition and the fatal outcome of severe cases. Emphasize that the goal of treatment is not to make the patient fat but to make the patient feel better and to improve self-confidence and eating habits. Weight is only a by-product of the improvement, and 'muscle mass and protein recovery,' not fat, is what has to be gained.

This firm understanding should engage the patient in a *treatment alliance* with the team. Remember that many of these patients are very polite and 'out to please you' at least superficially, and many times their initial acceptance hides deeper feelings of isolation and resentment. Psychotherapy is of help for some patients, usually accompanied by behavior modification and family therapy.

Despite the common use of antidepressants, several double-blind trials have been inconclusive or only slightly favorable. Patients with clear manifestations of depression and the more severe cases seem to benefit more from these medications. Tricyclic antidepressants tend to increase appetite and are more suited for patients with pure anorexia nervosa. Selective serotonin reuptake inhibitors may help decrease binging in patients with associated bulimia. Olanzapine, an atypical antipsychotic medication associated with weight gain, has been shown to be useful in some patients with anorexia nervosa in uncontrolled studies.

Nutritional Treatment

The psychiatric treatment is beneficial only as long as the patient's nutrition is improved. The nutritional rescue breaks down the vicious circle of the psychological consequences of starvation and makes the patient more receptive to psychotherapy. The team should be prepared to deal with the most ingenious ways to deceive. The patient should be told that because of the tendency to deceive frequently found in her illness, close supervision will be necessary at least in the beginning of the treatment. Patients should be weighed fasting in the morning, in nightgown without shoes and with the same scale, daily or at regular intervals by a nurse.

Initially, oral intake should be monitored carefully with a nurse sitting through the eating period and for 30 min thereafter to prevent postprandial vomiting. The tray should be checked for any food not consumed. In this way, a careful energy count is obtained daily. If the energy intake is inadequate or if the patient is not gaining weight, the diet should be supplemented with low-residue, high-energy canned formulae dispensed by the nurse during medication rounds. These diet supplements should be consumed in front of the nurse. Many patients with anorexia nervosa have subclinical vitamin deficiencies and they should receive a multivitamin tablet every day.

It is not infrequent for these patients to complain of gastric distress after sudden increases in food intake; smaller and more frequent feedings and/or administration of metoclopramide or cisapride before meals may be of help. Tube feeding is poorly tolerated by most of these patients; it has connotations of a gastrointestinal 'rape.'

If severe malnutrition is present (low serum albumin and transferrin levels, anergic skin testing), parenteral hyperalimentation should be instituted from the beginning. It is recommended to start with small amounts of hyperalimentation fluid to avoid

excessive sodium and water retention (refeeding edema) which is very distressing to patients. The rate of hyperalimentation solution administration should be modified according to the improvement in oral intake and weight. Staff should be continuously aware of the possibility of tampering with the central lines by the patient, with the potential for air embolization, infection, and bleeding. In many patients it is important to curtail all physical activity initially, to the point of confining them to absolute bed rest with only bathroom privileges. As the patient improves, the activities are progressively increased.

Estrogen replacement is indicated to prevent the progressive decrease in bone density but it is poorly tolerated and accepted by these patients. Ideally, a birth control pill with good estrogen content should be administered.

Prognosis

The outcome for patients with anorexia nervosa is variable; a worse outcome is associated with older age of onset, severity and duration of the illness, male sex, and severe associated psychiatric disturbances. In general, 40–60% of patients achieve full nutritional and psychological recovery after 6–12 months. Approximately 20–40% attain a borderline normal weight and existence for the rest of their lives, but with the occurrence of significant stress they may revert to their previous anorexic behavior. There is a mortality rate of 5–30% in the most severe cases due to suicide, electrolyte imbalance, and starvation-induced myocardial damage causing intractable arrhythmias; it is rarely due to infection. Long-term follow-up of these patients has shown an increased later mortality due to alcoholism.

See also: **Adolescents**: Nutritional Requirements. **Eating Disorders**: Bulimia Nervosa. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **Obesity**: Definition, Etiology and Assessment. **Starvation and Fasting**.

Further Reading

Adan RAH, Hillebrand JJG, de Rijke C *et al.* (2003) Melanocortin system and eating disorders. *Annals of the New York Academy of Science* **994**: 267.

Ahima RS, Prabakaran D, Mantzoros C *et al.* (1996) Role of leptin in the neuroendocrine response to fasting. *Nature* **382**: 250.

Audenaert K, van Laere K, Dumont F *et al.* (2003) Decreased 5-HT2a receptor binding in patients with anorexia nervosa. *Journal of Nuclear Medicine* **44**: 163.

Becker AE, Grinspoon SK, Klibanski A *et al.* (1999) Eating disorders. *New England Journal of Medicine* **340**: 1092.

Bergen AW, van den Bree MB, Yeager M *et al.* (2003) Candidate genes for anorexia nervosa in the 1p33–36 linkage region: Serotonin 1 D and delta opioid receptor loci exhibit significant association with anorexia nervosa. *Molecular Psychiatry* **8**: 397.

Brambilla F *et al.* (2001) Plasma concentration of interleukin-1-beta, interleukin-6 and tumor necrosis factor-alpha, and their soluble receptors and antagonists in anorexia nervosa. *Psychiatry Research* **103**: 107.

Fairburn CG and Harrison PJ (2003) Eating disorders. *Lancet* **361**: 407.

Garfinkel PE and Gardner DM (1982) *Anorexia Nervosa, a Multidimensional Perspective*. New York: Brunner/Mazel.

Gordon CM, Grace E, Emans SJ *et al.* (2002) Effects of oral dehydroepiandrosterone on bone density in young women with anorexia nervosa: A randomized trial. *Journal of Clinical Endocrinology and Metabolism* **87**: 4935.

Kaye WH (1997) Persistent alterations in behavior and serotonin activity after recovery from anorexia and bulimia nerviosa. *Annals of the New York Academy of Science* **817**: 162.

Kennedy SH, Kaplan AS, Garfinkel PE *et al.* (1994) Depression in anorexia nervosa and bulimia nervosa: Discriminating depressive symptoms and episodes. *Journal of Psychosomatic Research* **38**: 773.

Klibanski A, Biller BM, Schoenfeld DA *et al.* (1995) The effects of estrogen administration on trabecular bone loss in young women with anorexia nervosa. *Journal of Clinical Endocrinology and Metabolism* **80**: 898.

Mantzoros C, Flier JS, Lesem MD *et al.* (1997) Cerebrospinal fluid leptin in anorexia nervosa: Correlation with nutritional status and potential role in resistance to weight gain. *Journal of Clinical Endocrinology and Metabolism* **82**: 1845.

Pannacciulli N, Vettor R, Milan G *et al.* (2003) Anorexia nervosa is characterized by increased adiponectin plasma levels and reduced nonoxidative glucose metabolsim. *Journal of Clinical Endocrinology and Metabolism* **88**: 1748.

Schwabe AD (moderator) (1981) Anorexia nervosa. *Annals of Internal Medicine* **94**: 371.

Soyka LA, Misra M, Frenchman A *et al.* (2002) Abnormal bone mineral accrual in adolescent girls with anorexia nervosa. *Journal of Clinical Endocrinology and Metabolism* **87**: 4177.

Tolle V, Kadem M, Bluet-Pajon MT *et al.* (2003) Balance in Ghrelin and leptin plasma levels in anorexia nervosa patients and constitutionally thin women. *Journal of Clinical Endocrinology and Metabolism* **88**: 109.

Treasurer J and Hollander AJ (1989) Genertic vulnerability to eating disorders. In: Remschmidt H and Schmidt MW (eds.) *Child and Youth Psychiatry: European Perspectives*. New York: Hogrefe & Hubert.

Vande Berg BC, Malghem J, Lecouvet FE *et al.* (1996) Distribution of serouslike bone marrow changes in the lower limbs of patients with anorexia nervosa: Predominant involvement of the distal extremities. *American Journal of Roentgenology* **166**: 621.

Bulimia Nervosa

A J Hill and S F L Kirk, University of Leeds, Leeds, UK

Episodes of ravenous overeating, referred to as compulsive eating or binge eating, have been recognized clinically since the 1950s. However, the disorder of bulimia nervosa was not formally described until 1979. This relatively recent recognition of the eating disorder has two important implications. First, the clinical picture and understanding of the psychopathology is changing with time. This has led both to a refinement in the diagnostic criteria used to characterize the disorder and to changes in reported prevalence. Second, the research base used to make judgments about development, prevalence, treatment, prognosis, and prevention is smaller than that for anorexia nervosa. Quite simply, there are still many unknowns in the area of bulimia nervosa.

This article focuses on the features used to make a diagnosis of bulimia nervosa, the psychopathology and developmental course of the disorder, and the groups at risk. Specific attention is paid to the nutritional consequences of bulimia nervosa and the ways in which dietary management is used in its treatment. Finally, long-term prognosis is considered.

Diagnostic Criteria

The behavior at the center of the disorder, binge eating, has been progressively redefined. A priority has been to separate binge eating from mere indulgence and everyday overeating. Accordingly, two features of a true binge have been identified: consumption of unusually large amounts of food and an aversive sense of lack of control over eating. The size of binges varies but is typically between 1000–2000 kcal.

Diagnostic schedules (such as DSM-IV and ICD-10) agree on three features that must be present in someone with bulimia nervosa. The first is the presence of binge episodes. Second, the person must use compensatory behavior to control body shape or weight. The most common is self-induced vomiting, but these strategies also include use of laxatives or diuretics, excessive exercise, and extreme dieting or fasting. Third, the person must show overconcern with body weight and shape. Importantly, the person should not be of low body weight, in which case a diagnosis of anorexia nervosa would be made.

The tightening of these formal diagnostic criteria has had the consequence of reducing misdiagnosis and prevalence but has increased the numbers of those with atypical eating disorders. Failing to exhibit one or more of the key diagnostic features, such as an insufficient frequency of binge eating, is classified variously as atypical, partial syndrome or 'eating disorders not otherwise specified' (EDNOS). It is also useful to note that the diagnostic criteria for another eating disorder, binge eating disorder (BED), are included in DSM-IV, albeit for research purposes. The key difference between BED and bulimia nervosa is the absence of the extreme compensatory behaviors that follow the binge. Those with BED are less likely to be restricting their eating but more likely to be overweight and to be older (most presenting between ages 30 and 50).

Psychopathology

The description of body image disturbance that is central to both anorexia and bulimia nervosa has undergone revision. A distinction has been argued for between dissatisfaction with body shape and overvalued ideas about weight and shape. Although body shape dissatisfactions are commonly found in these patients, it is their overvalued ideas about weight and shape that are the necessary diagnostic feature. In other words, concern should go beyond simply feeling fat to a point where a person's life is dominated by their feelings about body weight and shape.

If these overvalued ideas are accepted as the core psychopathology of bulimia nervosa, then the chaotic eating that typifies the condition can be seen as a behavioral consequence. Binges are often interspersed between periods of intense dieting, even fasting, themselves strategies to control weight. Purging always follows a binge and is a way of expelling the food ingested or compensating for the food energy intake. Binges are secretive, planned, often expensive, and emotionally self-destructive. Paired with purging, they are cyclical and self-perpetuating, although their frequency may wax and wane. In addition, this behavior may have a long history before treatment is considered and clinical attention sought.

Bulimic episodes may be triggered by a variety of factors, including anxiety, boredom, tension, or breaking the self-imposed dietary rules necessary to maintain rigid control over eating (see **Figure 1**). Only rarely is hunger identified as precipitating a binge, even though the person may not have eaten for 24 h or more.

Sustained depressive and anxiety symptoms are common and are part of a range of psychological

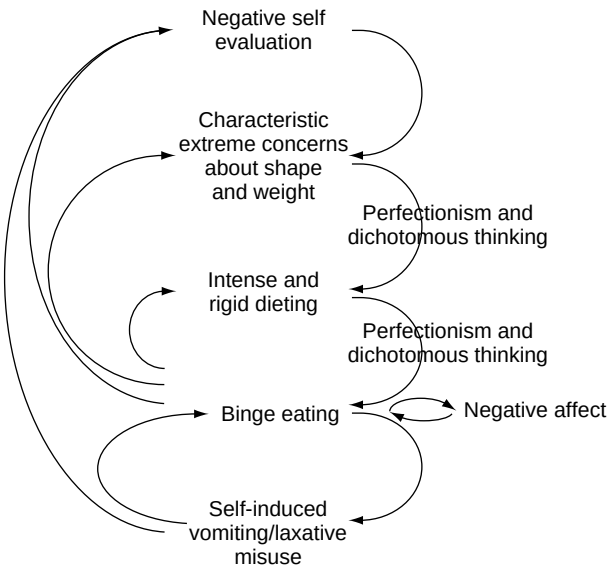


Figure 1 The cognitive behavioral view of the maintenance of bulimia nervosa (adapted from Fairburn & Brownell, 2002).

and social problems characteristic of bulimia nervosa. Impulsivity is also characteristic, with sexual promiscuity, self-harm, drug use, and stealing frequently noted. One suggestion is that impulsivity is a personality trait that favors bingeing over restriction, and so predisposes a person to bulimia nervosa rather than anorexia nervosa.

Etiology

As with anorexia nervosa, the picture of development is complex and multifactorial. There is no single cause of bulimia nervosa. Rather, a variety of psychological, biological, and social factors are involved in the emergence of the disorder. Although etiology is diverse, it has much in common with the forces responsible for anorexia nervosa and is clarified by looking at the groups of people most at risk. Overall, the balance of etiological factors is in favor of psychological and social causes, given that bulimia nervosa is a relatively new condition and has arisen at a time of profound social and cultural change, with little concurrent change in human biology.

The process of the development of eating disorders can be usefully divided into three stages. These conceptually separate the factors that predispose an individual to the disorder, precipitating events that lead to onset, and factors that perpetuate or maintain the disorder once initiated. Any framework drawn up for bulimia nervosa would be very similar to that for anorexia nervosa because the etiology of the two disorders appears to have a lot in common. Indeed, up to a third of patients with bulimia nervosa have a premorbid history of anorexia nervosa.

Although the etiological picture is very similar to that for anorexia nervosa, there are a few clues to differences. Genetic studies suggest the disorder is less heritable than anorexia nervosa, although heritability estimates of 46–71% have been calculated for the key behaviors, binge eating and self-induced vomiting. Evidence from case–control and cohort studies suggests two groupings of factors that contribute independently to the risk of developing bulimia nervosa. First, there is increased exposure to dieting and related risk factors, including parental and childhood obesity, and critical family comments about weight, shape, or eating. Second, a greater number of general risk factors for psychiatric disorder has been observed. These include parental psychiatric disorders such as depression, alcohol and substance abuse during childhood, low parental contact but high parental expectations, neglect, and abuse. Sexual abuse has been reported in 20–25% of patients with bulimia nervosa, a higher level than that found in restricting anorexia nervosa. Although the rate is increased compared to that of matched controls, it is no higher than the rate among young women with other psychiatric disorders. However, women with eating disorders in the context of sexual abuse appear to have higher rates of comorbid psychiatric conditions than other women with eating disorders.

The perspective on the perpetuation of bulimia nervosa is a cognitive one. **Figure 1** shows the vicious circles that maintain binge eating. Four points are emphasized in explaining this to patients. First, although dieting is a response to binge eating, it also maintains binge eating by both biological and psychological mechanisms. Second, compensatory purging encourages bingeing through a belief in its effectiveness at removing food for digestion. In other words, the barriers against overeating are removed since the food will not be absorbed. This is sometimes described as the reason why an individual initiated a binge–purge cycle of behavior. Third, extreme concern about body shape and weight promotes intense and rigid dieting and maintains the eating problem. Fourth, extreme concern about shape and weight is commonly associated with negative self-evaluation such as low self-esteem and longstanding feelings of worthlessness.

Groups at Risk

Like anorexia nervosa, bulimia nervosa is more common in women than in men, with a female:

male ratio around 10:1. The exact prevalence is notoriously difficult to establish for eating disorders, and for bulimia nervosa in particular. Problems in this regard include the recency of the disorder, changing diagnostic criteria, and the secrecy and non-life-threatening nature of the disorder preventing its routine appearance in clinical settings. Studies of US college students in the 1980s revealed up to 20% with bulimic symptoms. However, when such epidemiological studies investigate community samples and use interviews to follow-up questionnaire surveys, the average point prevalence among young women using strict diagnostic criteria is 1000 per 100,000 or 1.0%. In specific groups such as university students, there may be more than twice this level of the disorder. Bulimia nervosa is extremely rare in girls under 14 and the majority of cases are recognized between the ages of 18 and 25. Cases do present clinically in women in their late 20s and 30s, although they may have a long history of disordered eating.

The invisibility of bulimia nervosa is demonstrated by estimates of 1-year-period prevalence rates. These are calculated by adding figures for point prevalence and annual incidence. The 1-year-period prevalence rates for bulimia nervosa per 100,000 young women have been reported as follows:

In the community: 1500
In primary care: 170
In specialist mental health care: 87

These data indicate that only 11% of the community cases of bulimia nervosa are detected, and of these only half are received by specialist services. Since the first clinical description in 1979, there has been a dramatic upsurge in the number of bulimia nervosa cases seen, and much greater than that for anorexia nervosa. Although improved detection may account for some of this increase, there is general agreement that this represents a real increase in the number of women with bulimia nervosa.

Since women are most at risk, it is reasonable to ask why. One reason is that women are far more likely to diet than men. Dieting is itself a behavior that places individuals at risk of developing bulimia nervosa. But the motivations for dieting may be equally important. Women diet more than men for several reasons. These are bound together as a sociocultural perspective on bulimia, an approach that has become a powerful model for explaining who develops bulimia and why. At the heart of this perspective are three issues: the importance of a thin body shape for women, the centrality of appearance in women's gender role, and the importance of appearance for societal success. The arguments and evidence to support this analysis are compelling.

The above information on prevalence indicates that the average age of someone with bulimia nervosa is older than that for anorexia. This may reflect the observation that bulimia nervosa often follows a period of anorexia or at least low weight. Developmental challenges and age-dependent life events are also seen as important. The developmental task of achieving a sense of identity during mid- and late adolescence may be disrupted by relationship problems, peer or family difficulties, or events such as leaving home to go to college. The resultant erosion of self-esteem and perceived control can lead to problems with eating manifest as intensified dieting or periods of overeating and weight gain. The disrupted pattern of eating that follows may be the early stage of the disorder.

Nutritional Findings

A key feature of bulimia nervosa is the extreme dietary restraint that is exhibited in between episodes of binge eating. Such behavior has been described as all or nothing, so that on a good day the sufferer may describe consuming a very low-energy diet, whereas a bad day will consist of several episodes of uncontrolled eating. This will be accompanied by the purging behaviors previously described.

To sustain binge eating episodes, the person with bulimia nervosa may spend hundreds of pounds on food, selecting foods normally avoided during periods of dietary restraint which are easy to eat and subsequently remove from the body. To them, it is this overeating that is seen as the basic problem, not the dietary restraint that precedes it. Yet, it is this dietary restraint that drives the disorder. When not binge eating, it is common for patients to avoid eating for long periods, with 80% reporting consumption of one meal a day or less. While restricting their intake, they will consume reduced energy foods, with a strong tendency to avoid fat and choose energy-reduced foodstuffs.

It is often assumed that people with bulimia nervosa have good nutritional knowledge. Indeed, to the untrained eye, a diet history for a 'good day,' consisting of foods such as wholemeal bread, lots of vegetables and fruit, and skimmed milk, can be interpreted as conforming to healthy eating guidelines. However, this is not the case. Such restrictive behavior may fail to achieve even half the recommended energy intake and consequently may be deficient in micronutrients. The anxiety experienced

through consuming diet-breaking, 'unsafe' foods leads to the individual adopting extremely restricted diets between binges. Such intakes have been found to be lower in fat and higher in protein than the intakes of controls. People with bulimia nervosa also report feeling greater anxiety and guilt after eating foods they believed to be fattening.

Purging behaviors begin as a compensatory mechanism to offset episodes of binge eating. Consequently, it is a widely held belief that they are effective methods of weight control. However, the damage done to the body by these methods far exceeds any benefits in terms of weight. Any weight loss experienced is usually related to disruption of fluid balance rather than a loss of fat tissue. Furthermore, if self-induced vomiting is adopted, binges are likely to become more frequent and severe. If vomiting is prevented, the bulimic will consume significantly less food, thus maintaining the cycle previously described. Research has shown that vomiting fails to rid the body of all the food ingested. It has been estimated that only half the contents of the stomach are removed through vomiting, although this is variable and difficult to determine. Similarly, laxatives work on the system *after* food has been digested. One classic experiment looked at the amount of food energy lost through laxative abuse and found that, despite copious diarrhea, the amount of energy lost from the body was less than that found in the average chocolate bar.

What both laxative abuse and vomiting have in common is the depletion of fluid, leading to dehydration and electrolyte disturbances, particularly hypokalemia (low potassium). In some cases, hypoglycemia may develop as a response to fasting or binge eating and vomiting. In extreme cases, death may occur through cardiac arrest or gastrointestinal complications, such as oesophageal or gastric rupture. Vomiting also leads to erosion of dental enamel, resulting in periodontal disease and an increased incidence of dental caries. Other effects of bulimia nervosa include menstrual irregularities, swelling of the salivary glands secondary to vomiting, and reflex constipation, which occurs as a consequence of laxative abuse and dehydration. Laxative abuse has also been found to cause steatorrhea and protein-losing enteropathy in some cases.

People with bulimia nervosa may have lower energy requirements. Using indirect calorimetry, it has been found that patients have a measured resting energy expenditure below that predicted by standard formulas such as the Harris–Benedict equation. They also report consuming fewer kilocalories per kilogram body weight than control subjects. One explanation for this finding is that bingeing and purging may alter energy efficiency. These findings have implications for nutritional management, particularly in relation to the prescription of energy intakes.

Dietary Management

Dietitians and nutritionists are increasingly involved in the treatment of bulimia nervosa. While this is best utilized within a multidisciplinary team, ideally with some form of psychological intervention available, it is not uncommon for dietitians or nutritionisists to be the only professional involved. Any professional working with eating disorders should be clear about what they can address and be aware of when it is appropriate to enlist other forms of help. Thus, nutritional intervention should aim to separate food from underlying issues, leaving these to be addressed by professionals more experienced in psychological techniques. Research suggests that nutritional intervention, alongside other psychological therapies, most notably cognitive behavior therapy (CBT), is an important part of treatment. In addition, training in CBT techniques is advized for dietitians and nutritionists involved in bulimia nervosa management.

The aim of dietary management of bulimia nervosa is to break the binge–purge cycle previously described. The individual should be informed of the problems of maintaining this cycle through dieting and should be encouraged to stop dieting in an extreme way. They should also be educated about the damaging effects of vomiting and other purging behaviors. In some cases, this is enough to stop such behaviors. In others, this message should consistently be given to encourage them to work toward stopping these behaviors. An important part of breaking this cycle is to get the individual to monitor his or her intake through completing a food diary. In the example shown (**Table 1**), it can be seen how restricting intake earlier in the day can make the person more vulnerable to overeating later in the day. Food diaries are a powerful cognitive tool that enable the individual to understand his or her eating behavior more fully.

Education is essential to ensure that people understand why they are being asked to abandon what are some of the only coping mechanisms they have. They feel anxious that by giving up the pattern of dieting, binge eating, and purging they will gain excessive amounts of weight. These fears are very real and failure to address them with sensitivity can sabotage any attempt to control the disorder. This is

Table 1 Example of a food diary

Time	*Food/drink eaten and amount*	*Binge/vomit/laxatives*	*Comment/feelings*
Breakfast	Nothing	—	Not hungry.
Midmorning	Cup of black coffee × 2	—	Need something to fill my stomach. Really busy at work so no time to eat.
Midday (1.30 pm)	2 crispbreads, dry Small tub of diet cottage cheese 1 tomato, can of diet pop	—	Very hungry, feel as if I could eat more but mustn't.
Midafternoon	Chocolate eclair	Vomited	Someone's birthday in the office so couldn't refuse. Feel really guilty and had to be sick.
Evening meal	2 dishes of blackcurrant cheesecake, a choc ice, 4 bowls of ice cream, 6 snack-size chocolate bars, 5 cheese biscuits with butter and cheese, 5 slices of toast with butter and peanut butter, 2 packets of chocolate biscuits, 2 bowls of cereal, 1 packet of crisps and 1 chocolate and mint biscuit	Binge!! Vomited and took 10 laxatives	Couldn't decide what to have for tea, so started on cheesecake. Could not stop this binge at any cost. I feel terrible.
During evening	6 glasses of water		
Supper	—	—	Feel so terrible and bloated. Will have to cut back tomorrow.

particularly important when an individual has a history of overweight in the past. A detailed weight history should be carried out to include current, highest, lowest, and ideal weights and it should be stressed that recovery cannot be accomplished if the sufferer is trying to maintain a weight below normal. Thus, those with a premorbid history of obesity may have to accept that they will need to reach a weight that is higher than they would like it to be. Weight stabilisation should be an initial emphasis, particularly for those experiencing weight fluctuations. Initially, weight is likely to fluctuate through rehydration and repletion of glycogen stores. This effect should be explained to the individual to reduce unnecessary anxiety. They should also be discouraged from weighing themselves. If they must get weighed, this should be no more than once a week.

An important goal for nutritional management is to establish the individual on a regular pattern of eating. Often, normal cues for hunger and satiety are disrupted through repeated cycles of binge and restrictive eating so encouraging a regular meal pattern also helps the sufferer to begin to identify hunger and fullness again. They should be encouraged to eat regular meals and snacks and to maintain this pattern of eating even after a binge. Each meal or snack should be based around carbohydrate, with moderate amounts of protein foods and vegetables and fruit. They should be encouraged to include non-diet foods and to include foods containing fat. It is also worth getting them to compile a list of foods normally avoided or associated only with binges and to encourage them to include them within their meal pattern, when they feel able to do so. A system of food exchanges may also be useful (see the sample meal plan in **Table 2**). The amount of food required to meet energy needs is greater than that needed to consume sufficient nutrients. Thus, consumption of some energy dense, less nutritious food should be encouraged. A minimum intake of 6.0 MJ (1500 kcal) is usually an appropriate level to begin with, increasing to an intake corresponding to the estimated average requirement for women as recovery proceeds.

If the bulimic is used to keeping his or her stomach empty, even a normal amount of food may seem excessive and may trigger the urge to vomit. They should be informed that stomach distension is a normal consequence of eating and reassured that they will get used to the feeling in time. Similarly, if someone has been abusing laxatives, he or she may suffer from constipation and should be encouraged to have a reasonable fiber intake along with plenty of fluids.

Although it is important to give positive encouragement and feedback when working with individuals with bulimia nervosa, it should also be explained that relapse is a normal occurrence and should not be viewed so negatively that the individual feels a complete failure. Education on relapse prevention should be an important component of any treatment programme.

Table 2 Sample meal plan[a]

Breakfast:	Glass of fruit juice Bowl of cereal with milk Slice of toast, spread with butter/margarine and marmalade/jam if desired
Midmorning:	**Choose from exchange list**
Snack meal:	Sandwiches made with two large slices of bread, spread with butter/margarine, and filled with lean meat, egg, or cheese or beans on toast, etc.
Midafternoon:	**Choose from exchange list**
Main meal:	Average helping of meat, chicken, fish, or vegetarian alternative Potatoes, boiled rice, or pasta equivalent to two exchanges Vegetables or salad
	Option from exchange list
Supper:	1–2 slices of toast or option from exchange list

Exchange list
- 1 large slice of bread or a roll, teacake, or plain bun
- 1 small scone
- 2 plain biscuits or crackers
- 1 chocolate or cream biscuit
- Small bowl of breakfast cereal or porridge
- 2 spoons boiled rice or pasta
- 1 medium potato
- 4 spoons baked beans or tinned spaghetti
- 1 piece of fruit (apple, orange, pear, banana, etc.)
- 1 carton of yoghurt
- 1 glass of milk
- 6–8 tablespoons custard or rice pudding
- 1 packet crisps or nuts and raisins
- 1 average-sized chocolate bar

[a] This plan deliberately does not include specific portion sizes. However, some individuals may need the reassurance of a more detailed plan. The aim is to provide a minimum of 6.25 MJ (1500 kcal). In addition to using the above to exchange foods within the meal plan, some people benefit from having an additional number of exchanges (e.g., five extra) to allow for when they are feeling more hungry and to offset binges. This also ensures an energy intake that conforms to the current EAR for women.

Long-Term Prognosis

Once more, the relative recency of the disorder mitigates against definitive statements. However, the outcome studies conducted so far show that bulimia nervosa is far from being intractable, as was originally suggested. Of the studies with a follow-up of at least 5 years, between a third and a half of those with bulimia nervosa at outset still had an eating disorder, and between 10 and 25% still had bulimia nervosa. There is also considerable flux within samples. For example, in a community sample, each year about a third of patients remitted and a further third relapsed. Studies also report very low rates of spontaneous remission.

In terms of treatment, several psychotherapies have shown their effectiveness in improving the symptoms of bulimia nervosa. Most evidence is available on CBT, and the outcome is generally impressive and replicable. The treatment, usually provided by psychiatrists or clinical psychologists, aims to modify both the disturbed eating habits and the extreme concerns about shape and weight (the core psychopathology). Consequently, it combines psychological and dietetic approaches to patient management. Given a treatment program of around 20 sessions over 5 months, between a half and two-thirds of patients make a full and lasting recovery.

There is still uncertainty regarding prognostic indicators of treatment success. Patients with a less severe form of the disorder appear to do better in treatment. Self-help programs administered on their own or with modest support and encouragement from a nonspecialist therapist (guided self-help) may be of particular assistance to those in whom the disorder is less fully established. Conversely, those with childhood obesity, low self-esteem, or personality disturbance appear to do worse. Importantly, there is no evidence that bulimia nervosa evolves over time into other psychiatric disorders or of any persistent impairment in social functioning.

See also: **Appetite**: Physiological and Neurobiological Aspects; Psychobiological and Behavioral Aspects. **Eating Disorders**: Anorexia Nervosa; Binge Eating. **Hunger**.

Further Reading

Abraham S and Llewellyn-Jones D (2001) *Eating Disorders: The Facts*, 5th edn. Oxford: Oxford University Press.

American Academy of Pediatrics Committee on Adolescence (2003) Identifying and treating eating disorders. *Pediatrics* **111**: 204–211.

American Dietetic Association (2001) Nutrition intervention in the treatment of anorexia nervosa, bulimia nervosa, and eating disorders not otherwise specified (EDNOS). *Journal of the American Dietetic Association* **101**: 810–819.

American Psychiatric Association (2000) Practice guidelines for the treatment of patients with eating disorders. *American Journal of Psychiatry* **157**(supplement 1): 1–39. Available at http://www.psych.org/clin_res/guide.bk-2.cfm.

Fairburn CG (1995) *Overcoming Binge Eating*. New York: Guilford Press.

Fairburn CG and Brownell KD (eds.) (2002) *Eating Disorders and Obesity. A Comprehensive Handbook*, 2nd edn. New York: Guilford Press.

Fairburn CG, Cooper Z, Doll HA, Norman P, and O'Connor M (2000) The natural course of bulimia nervosa and binge eating disorder in young women. *Archives of General Psychiatry* **57**: 659–665.

Fairburn CG, Marcus MD, and Wilson GT (1993) Cognitive-behavioural therapy for binge eating and bulimia nervosa: A comprehensive treatment manual. In: Fairburn CG and

Wilson GT (eds.) *Binge Eating: Nature, Assessment, and Treatment*. New York: Guilford Press.
Fairburn CG, Norman PA, Welch SL *et al.* (1995) A prospective study of outcome in bulimia nervosa and the long-term effects of three psychological treatments. *Archives of General Psychiatry* **52**: 304–312.
Gordon RA (2000) *Eating Disorders: Anatomy of a Social Epidemic*, 2nd edn. Oxford: Blackwell.
Hay P and Bacaltchuk J (2002) Bulimia nervosa. *Clinical Evidence* **7**: 834–845.
Latner JD and Wilson GT (2000) Cognitive-behavioral therapy and nutritional counseling in the treatment of bulimia nervosa and binge eating. *Eating Behaviors* **1**: 3–21.
National Collaborating Centre for Mental Health (2004) Eating Disorders. Core interventions in the treatment and management of anorexia nervosa, bulimia nervosa, and related eating disorders. *National Clinical Practice Guideline CG9*. Leicester: British Psychological Society http://www.bps.org.uk/eatingdisorders/files/ED.pdf.
Polivy J and Herman CP (2002) Causes of eating disorders. *Annual Review of Psychology* **53**: 187–213.
Rome ES, Ammerman S, Rosen DS *et al.* (2003) Children and adolescents with eating disorders: The state of the art. *Pediatrics* **111**: e98–e108.
Russell GFM (1979) Bulimia nervosa: An ominous variant of anorexia nervosa. *Psychological Medicine* **9**: 429–448.
Schmidt U and Treasure J (1997) *Getting Better Bit(e) by Bit(e). A Survival Kit for Sufferers of Bulimia Nervosa and Binge Eating Disorders*. Hove: Psychology Press.
Stice E (1994) Review of the evidence for a sociocultural model of bulimia nervosa and an exploration of the mechanisms of action. *Clinical Psychology Review* **14**: 633–661.
Waller G (2002) Treatment of bulimia nervosa. *Psychiatry* **1**: 12–16.

Binge Eating

M D Marcus, M A Kalarchian and M D Levine, University of Pittsburgh, Pittsburgh, PA, USA

In 1959, Stunkard noted three patterns of eating behavior in obese patients: night eating, binge eating, and eating without satiation. It was not until the 1980s that binge eating began to receive attention as a distinct clinical syndrome. Spitzer and colleagues proposed diagnostic criteria for binge eating disorder (BED) and subsequently evaluated them in two field trials. These initial investigations led to the inclusion of BED in the *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition (*DSM-IV*), of the American Psychiatric Association as an example of Eating Disorder Not Otherwise Specified (EDNOS) and as a proposed diagnostic category requiring further study. BED is characterized by persistent and recurrent episodes of binge eating without the regular use of inappropriate compensatory behaviors seen in bulimia nervosa (BN). Research on BED is still in an early stage, and the key features of the disorder and its relationship to other eating disorders, especially BN, nonpurging type, continue to be debated in the field.

This article discusses the assessment of BED, prevalence and risk factors, and comorbid conditions associated with BED. In addition, empirically supported treatments are reviewed, including guidelines for choice of treatment approach. Throughout the article, a biopsychosocial framework for understanding aberrant eating behavior is emphasized.

Assessment of Binge Eating

A binge episode is defined as the consumption of a large amount of food within a discrete period of time, accompanied by a sense of loss of control over eating. Researchers and clinicians have agreed that loss of control involves the subjective feeling that one cannot stop eating or control what or how much is being eaten. However, there has been much less agreement about the size and duration of a binge eating episode. Specifically, there is no consensus as to what constitutes a large amount of food, and the duration of binge eating episodes can vary widely, sometimes continuing throughout an entire day. Many individuals have difficulty delineating binges into discrete episodes but can more readily recall whether a binge occurred or not on a given day. Thus, the BED diagnosis is made based on binge 'days' rather than 'episodes.' Similarly, many observers have concluded that the loss of control, rather than the amount of food ingested during a binge (i.e., a 'large' amount), is the hallmark of binge eating. See **Table 1** for the full research criteria for BED.

Several methods can be used to assess BED, including clinical interviews, self-reports such as questionnaires and food diaries, and observation of eating behavior in the laboratory. Currently, a clinical interview by a trained professional is the preferred assessment method. It provides the opportunity to standardize definitions of key concepts such as a 'large amount of food' and 'loss of control.' Although questionnaires are relatively easy to administer, there is high potential for misinterpreting these terms. Interview-based assessments tend to yield ratings of binge eating that are lower, but more precise, than questionnaire-based surveys.

Food diaries involve having individuals keep a daily record of the specifics of eating episodes, including how much food was consumed, whether or not there was loss of control over eating, any use of

Table 1 *DSM-IV-TR* criteria for binge eating disorder

1. Recurrent episodes of binge eating. An episode of binge eating is characterized by both of the following:
   - Eating, in a discrete period of time (e.g., within any 2-h period), an amount of food that is definitely larger than most people would eat in a similar period of time under similar circumstances
   - A sense of lack of control over eating during the episode (e.g., a feeling that one cannot stop eating or control what or how much one is eating)
2. The binge eating episodes are associated with at least three (or more) of the following:
   - Eating much more rapidly than normal
   - Eating until feeling uncomfortably full
   - Eating large amounts of food when not feeling physically hungry
   - Eating alone because of being embarrassed by how much one is eating
   - Feeling disgusted with oneself, depressed, or very guilty after overeating
3. Marked distress regarding binge eating is present.
4. The binge eating occurs, on average, at least 2 days a week for 6 months.
5. The binge eating is not associated with the regular use of inappropriate compensatory behaviors (e.g., purging, fasting, and excessive exercise) and does not occur exclusively during the course of anorexia nervosa or bulimia nervosa.

inappropriate compensatory behavior, and the associated context. Food diaries can provide detailed assessment information without introducing the bias of retrospective self-report; however, self-monitoring also has been shown to affect eating behavior and is frequently employed in clinical treatment. Findings from studies that have utilized food diaries indicate that BED patients report higher calorie intakes than non-binge eaters on both 'binge days' and 'non-binge days.'

Observation of binge eating in the laboratory is a specialized technique that is limited to use in research settings, providing the opportunity to document actual eating behavior and measure consumption. Laboratory studies with relatively small samples have shown that, compared to equally overweight patients who do not binge eat, BED patients ingest more calories, both during binges and at 'regular' meals. This difference in eating behavior in binge compared to non-binge eaters supports the validity of BED as a distinct diagnostic category.

Prevalence and Risk Factors

Available research suggests that the prevalence of BED among the general population is approximately 1 or 2% and thus more common than BN. In addition, preliminary findings suggest that the demographic profile of individuals with BED may be more diverse, affecting relatively more men and minority groups than BN or anorexia nervosa. Furthermore, binge eating is more prevalent among obese individuals in both clinical and community samples. It is estimated that up to one-third of individuals who present for treatment in university-based weight control clinics report significant binge eating.

The most comprehensive risk factor study to date suggests that the risk factors for BED may be weaker and more circumscribed than for BN. Fairburn and colleagues interviewed four groups of subjects matched for age and social class: individuals with BED, BN, another psychiatric disorder, and healthy controls. In comparing the BED group to the controls, negative self-evaluation, parental depression, adverse childhood experience, and exposure to repeated negative comments about shape, weight, or eating emerged as risk factors for BED. Further comparing BED patients to other groups with psychiatric diagnoses, childhood obesity and negative comments from family about eating, shape, and weight emerged as risk factors specific to BED. Thus, BED appears to be associated with two classes of risk factor—those that increase the risk of psychiatric disorder in general and factors that increase risk of obesity.

In order to improve our understanding of how multiple factors interact to determine the onset and maintenance of binge eating, prospective risk factor studies including males and females of different racial groups are needed. As suggested previously, biological (e.g., obesity), psychological (e.g., negative self-evaluation), and social (e.g., exposure to repeated negative comments about shape, weight, or eating) factors have been implicated in the pathogenesis of binge eating. Emergent research also has linked binge eating in a small proportion of individuals to a mutation in *MC4R*, a candidate gene for the control of eating behavior. Thus, future research may further elucidate the genetic influences on aberrant eating patterns.

Comorbidity

Binge eating is strongly associated with both obesity and psychiatric disorder. It is well documented that obesity is linked to adverse medical and psychosocial outcomes. Preliminary findings also suggest that BED may be associated with poor health, independent of the effects of comorbid psychopathology or comorbid obesity.

Severity of binge eating is positively associated with degree of overweight. Additionally, there are important differences between overweight individuals with

and without BED. BED patients report earlier onset of obesity, along with a history of more severe obesity, dieting, and weight fluctuations. When compared with equally overweight individuals without binge eating problems, BED patients report considerably less 'restraint' or control overeating, lower self-esteem, more fear of weight gain, more preoccupation with food, and higher body dissatisfaction.

Individuals with BED endorse high rates of psychiatric symptoms and disorders. For example, when compared to equally overweight individuals without binge eating problems, individuals with BED report significantly higher lifetime rates of major depressive disorder, substance abuse or dependence, and anxiety disorders. Some studies have shown that patients with BED report levels of eating disorder symptomatology, such as eating, shape, and weight concerns, that are comparable to those of normal weight patients with BN. Individuals with BED also have considerably higher rates of personality disorders than overweight individuals without an eating disorder. Thus, among individuals with BED, psychiatric symptomatology appears to be related to the binge eating rather than to the degree of obesity.

Treatment

Among those who seek treatment, BED tends to be a chronic and fluctuating disorder. The clinical picture in BED often involves onset in late adolescence or the early 20s, with numerous periods of relative control over eating, and weight loss during periods of successful calorie restriction, alternating with periods characterized by binge eating and weight gain. Individuals with BED often seek obesity treatment rather than treatment of disordered eating per se.

A variety of psychosocial and pharmacological interventions, as well as behavioral weight loss programs, can help individuals gain control over binge eating. **Figure 1** provides an overview of nonpharmacologic treatments and their postulated mechanisms of action; it should be noted that this list of treatments is not exhaustive. Furthermore, although the postulated treatment mechanisms reflect respective theoretical models of binge eating, the effects of psychosocial treatment usually lack specificity and the treatments often share common elements (e.g., self-monitoring is a central component of cognitive behavior therapy (CBT), dialectical behavior therapy (DBT), and behavioral weight control). Also, effects of psychosocial treatments often extend to areas that are not a focus of treatment (e.g., behavioral weight management programs that target changing eating and exercise also have been shown to improve mood). Thus, some of the treatments do overlap. Finally, it is important to note that a substantial number of patients are not abstinent from binge eating after treatment, suggesting

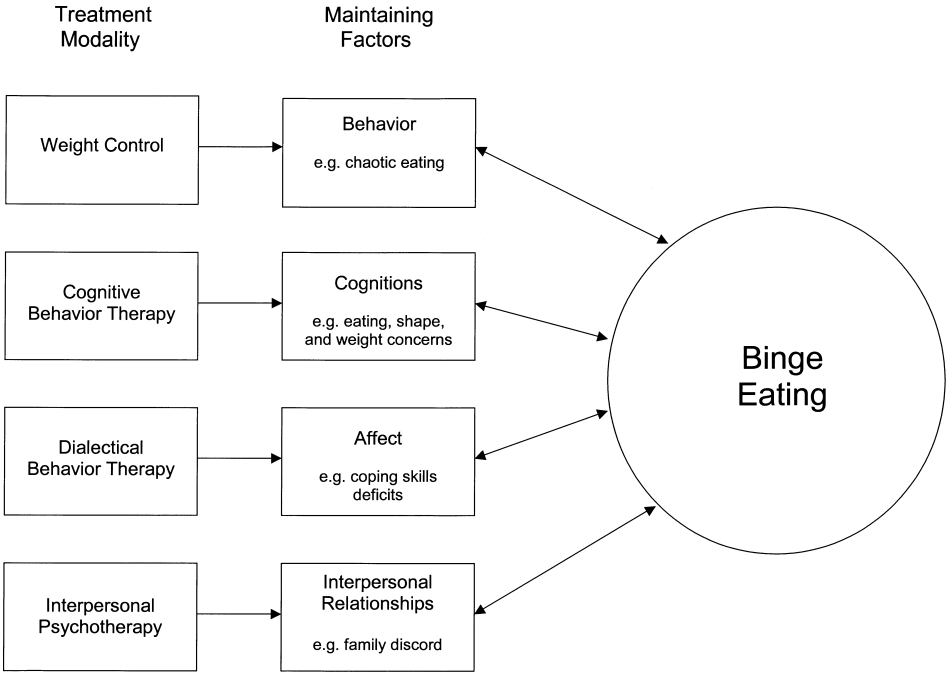


Figure 1 Overview of nonpharmacologic treatments for binge eating.

the need for clinical trials of novel therapeutic approaches as well as combinations and sequencing of treatments.

Psychosocial Treatments

Treatments for binge eating have been adapted from those that have been shown to be effective in reducing binge eating among individuals with BN. The majority of the research on psychosocial treatments has supported two structured, focused, short-term psychotherapies—CBT and interpersonal psychotherapy (IPT)—both of which have been shown to be more effective than no treatment in decreasing the frequency of binge eating and improving the psychopathology associated with binge eating. In addition, the use of DBT shows promise as an alternative treatment for BED.

Cognitive behavioral therapy CBT has been the most extensively studied treatment for individuals with binge eating. CBT for BED is based on the assumption that binge eating is maintained in the context of ongoing dietary restraint, weight concerns, negative emotions, and low self-esteem. Treatment focuses first on normalizing eating and then on the identification and restructuring of maladaptive thoughts and beliefs, particularly those related to eating, shape, and weight.

CBT for BED has been adapted to reflect important differences between individuals with BN and BED. Specifically, cognitions relating to having a large body size are directly targeted in treatment. Overweight individuals with BED may be helped to accept a larger than average body size and to change unrealistic expectations for weight loss. That is, for the majority of BED patients a 5- or 10-kg weight loss does not correspond with their desired weight loss, even though a modest weight loss may relate to improvements in binge eating and overall health. It is therefore important to help patients adopt realistic goals for the body weight and shape they are likely to achieve.

Another adaptation of CBT for BED relates to differences in the role of dieting between individuals with BED and those with BN. Although the treatment of BN stresses the role of dietary restraint in precipitating binge episodes, and treatment focuses on decreasing dietary restraint, patients with BED do not necessarily binge eat in response to restraint or hunger. Indeed, the preponderance of evidence suggests that increasing dietary restraint may help to ameliorate binge eating in obese individuals. Thus, CBT for BED does not stress decreased dietary restraint; rather, treatment encourages the development of a moderate, structured, healthy eating pattern.

Interpersonal psychotherapy Klerman and Weissman's IPT has also received empirical support in the treatment of individuals with BED. IPT for binge eating is based on the idea that dysfunctional eating behavior is maintained in the context of interpersonal difficulties. Treatment focuses on identifying and addressing specific, problematic interpersonal patterns in an effort to ameliorate binge eating. Treatment can focus on role disputes, such as marital or family discord; role transitions, such as the adjustment to motherhood or a new job; grief, such as the loss of a spouse or loved one; and interpersonal deficits, such as loneliness and social isolation.

IPT for BED does not directly target eating behaviors or attitudes about eating, shape, and weight. Although the ways in which CBT and IPT conceptualize and treat binge eating differ, both appear to be equally effective in reducing the frequency of binge eating.

Dialectical behavior therapy Developed by Linehan for the treatment of individuals with borderline personality disorder, DBT has shown promise in the treatment of BED. DBT is a comprehensive treatment program based on cognitive and behavioral principles and complemented by the use of mindfulness strategies derived primarily from Zen Buddhism. In addition to weekly individual outpatient treatment, traditional DBT prescribes a weekly group meeting in which the goal is to increase participants' behavioral skills. A group-only version of DBT for individuals with BED has also been shown to decrease binge eating and maladaptive attitudes about eating, shape, and weight. Additional research is needed to determine the efficacy of DBT relative to CBT and IPT.

Behavioral Weight Control

Because the majority of individuals with BED are also overweight and want to lose weight, and because obesity is associated with significant medical and psychosocial consequences, weight loss is a potentially important outcome in the treatment of BED. Numerous studies have documented that calorie restriction does not exacerbate binge eating in BED patients. Indeed, participation in behavioral weight control programs that focus on calorie restriction, provide education about sound nutritional principles, and promote physical activity may decrease binge eating and improve mood in BED patients. Therefore, concerns about the

potentially deleterious effects of dieting should not deter obese patients who binge eat from attempting behavioral weight management.

Weight lost through dieting is frequently regained, and sustained weight change involves a permanent modification of eating and exercise patterns. However, it is not necessary to achieve large weight losses to improve risk factors for heart disease, diabetes, and other obesity-related comorbidities. There is evidence that sustained weight losses of approximately 10% of initial body weight can lead to significant improvements in modifiable risk factors such as blood pressure, lipids, and blood sugar levels.

Pharmacotherapy

Pharmacologic approaches to the treatment of BED that have empirical support include antidepressant and weight loss medications. Additionally, one study has demonstrated the potential utility of the anticonvulsant agent, topiramate. However, studies to date have not shown that pharmacotherapy increases the effectiveness of psychotherapy for BED.

Antidepressant treatment Because of their efficacy in ameliorating binge eating and purge behaviors in BN, antidepressants have been used in the treatment of BED. Early research comparing tricyclic antidepressants, such as desipramine and imipramine, to placebo showed greater reductions in binge eating among obese binge eaters treated with the drug than with a placebo. Recently, several selective serotonin reuptake inhibitors (e.g., fluoxetine) have been shown to be associated with moderate reductions in binge eating in BED patients. Moreover, the effects of antidepressant treatment on binge eating are independent of any effects on mood.

Antidepressant treatment also may be useful in treating depression associated with BED and has been associated with weight loss among obese binge eaters. Antidepressant treatment also may enhance dietary restraint or improve compliance with a weight loss program. Thus, it seems possible that longer term antidepressant treatment may be useful in breaking the cycle of negative mood, binge eating, and weight gain that characterizes BED.

Anorectic agents The utility of anorectic agents in the treatment of BED has been investigated in a few studies. Early research found that obese binge eaters treated with the serotoneric agent, *d*-fenfluarmine, experienced a significantly greater short-term reduction in binge eating than did those given placebo. However, because of the association between the longer term use of *d*-fenfluramine and the combination of phentermine and fenfluramine with serious pulmonary and cardiac problems, these medications have since been withdrawn from the market. Currently, two anorectic agents, sibutramine and orlistat, have been approved for long-term use in the treatment of obesity. Sibutramine has been investigated in the treatment of BED, with initial findings suggesting decreases in binge eating and body weight among obese women with BED.

Selection of Treatment for Specific Patients

No single treatment approach is effective for all patients. Future research may provide data to guide the selection of treatment for individual patients as well as evaluate alternative treatments. Until such information becomes available, clinicians and patients must decide on a course of treatment based on a careful assessment and thorough consideration of the pros and cons of available options.

Eating disorder and obesity history A history of early onset of binge eating, binge eating in the absence of obesity, or obesity in combination with numerous bouts of weight loss and regain over time ('yo-yo' dieting) suggests a course of psychosocial treatment. Such patients can be reassured that significant improvements in the aberrant eating and eating disorders psychopathology associated with BED can be obtained without weight loss.

On the other hand, clinical experience suggests that patients who report adult onset of binge eating and obesity, and do not have a history of marked weight fluctuations, may be more likely to benefit from a behavioral weight control approach. Behavioral weight control may also be indicated for patients who remain overweight after a trial of eating disorders treatment. Although behavioral weight control appears to be beneficial on average, it is important for each individual to evaluate the likelihood that he or she will be able to sustain lifelong changes in eating and exercise.

Psychiatric status Given the high psychiatric comorbidity in BED, a thorough evaluation is important for all patients who seek treatment. Although mild to moderate depression or anxiety is likely to improve during treatment of binge eating, the presence of marked or severe current illness suggests primary treatment of the mood or anxiety disorder. In addition, the presence of personality

disorders characterized by emotional, dramatic, or impulsive behavior appears to be related to severity of binge eating but does not appear to predict treatment outcome.

Available resources Clinicians trained in the use of psychosocial treatments for eating disorders are likely to be found in most metropolitan areas but may not be available in smaller cities or rural areas. Insurance companies vary in coverage for treatment of eating disorders, and some insurance plans may pay for obesity treatment only if there is a clear medical indication (e.g., hypertension or other cardiovascular risk). Thus, treatment decisions may need to take into account pragmatic factors, such as clinician availability and training or patient insurance plan coverage. Self-help programs may be appropriate for carefully screened, motivated patients with mild to moderate binge eating. However, comorbid psychopathology and high-frequency binge eating may require more intensive clinical intervention.

Conclusion

BED is a chronic and fluctuating disorder that is common among obese individuals who seek treatment, and it is associated with elevated rates of psychopathology. A biopsychosocial model shows most promise in understanding the etiology of aberrant eating. Once established, binge eating is maintained by a complex interplay of eating behaviors, cognitions, affect, and interpersonal factors. Nevertheless, available research indicates that most who binge eat can be helped with either a behavioral weight control program or an eating disorders treatment. Pharmacotherapy may also reduce binge eating but generally does not add to the effectiveness of psychosocial treatment. A careful assessment, review of the benefits and disadvantages of the different therapies, and consideration of the availability of trained clinicians should guide the choice of treatment for an individual with BED. More research is necessary to fully understand this problematic eating pattern and to improve strategies for management and treatment.

See also: **Eating Disorders**: Anorexia Nervosa; Bulimia Nervosa. **Obesity**: Fat Distribution; Complications; Treatment. **Weight Management**: Weight Maintenance.

Further Reading

American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed.–text revision. Washington, DC: American Psychiatric Association.

Cooper P (1995) *Bulimia Nervosa & Binge-Eating: A Guide to Recovery*. New York: New York University Press.

De Zwaan M (2001) Binge eating disorder and obesity. *International Journal of Obesity* **25**(supplement 1): S51–S55.

Fairburn C (1995). *Overcoming Binge Eating*. New York: Guilford Press.

Fairburn CG and Brownell KD (eds.) (2002) *Eating Disorders and Obesity: A Comprehensive Handbook*. New York: Guilford Press.

Fairburn CG, Doll HA, Welch SL *et al.* (1998) Risk factors for binge eating disorder. *Archives of General Psychiatry* **55**: 425–432.

Fairburn GG and Wilson GT (eds.) (1993) *Binge Eating: Nature, Assessment and Treatment*. New York: Guilford Press.

Levine MD and Marcus MD (1998). The treatment of binge eating disorder. In: Hoek H, Treasure J, and Katzman M (eds.) *The Integration of Neurobiology in the Treatment of Eating Disorders*, pp. 363–381. London: John Wiley & Sons.

Marcus MD and Kalarchian MA (2003) Binge eating in children and adolescents. *Int J Eat Disord*. **34**(suppl.): S47–S57.

Marcus MD and Levine MD (2004) Obese patients with binge eating disorder. In: Goldstein DJ (ed.) *The Management of Eating Disorders*. Clifton, NJ: Humana Press.

Marcus MD and Levine MD (2004) Dialectical behavior therapy in the treatment of eating disorders: Brewerton T (ed.) *Eating Disorders*. New York: Marcel Dekker.

Striegel-Moore RH and Smolak L (2001). *Eating Disorders: Innovative Directions in Research and Practice*. New York: Guilford Press.

Stunkard AJ (1959) Eating patterns and obesity. *Psychiatric Quarterly* **33**: 284–295.

Wadden TA and Stunkard AJ (eds.) (1994) *Handbook of Obesity Treatment*. New York: Guilford Press.

Walsh BT (ed.) (2003) The current status of binge eating disorder. *International Journal of Eating Disorders* **34**.

Walsh BT and Devlin MJ (1998) Eating disorders: Progress and problems. *Science* **280**: 1387–1390.

EGGS

D J McNamara and H S Thesmar, Egg Nutrition Center, Washington, DC, USA

Introduction

Eggs have been a staple in the human diet for thousands of years. From hunter–gatherers collecting eggs from the nests of wild birds, to the domestication of fowl for more reliable access to a supply of eggs, to today's genetically selected birds and modern production facilities, eggs have long been recognized as a source of high-quality protein and other important nutrients. Over the years, eggs have become an essential ingredient in many cuisines, owing to their many functional properties, such as water holding, emulsifying, and foaming.

An egg is a self-contained and self-sufficient embryonic development chamber. At adequate temperature, the developing embryo uses the extensive range of essential nutrients in the egg for its growth and development. The necessary proteins, lipids, carbohydrates, vitamins, minerals, and functional nutrients are all present in sufficient quantities for the transition from fertilized cell to newborn chick, and the nutrient needs of an avian species are similar enough to human needs to make eggs an ideal source of nutrients for us. (The one essential human nutrient that eggs do not contain is ascorbic acid (vitamin C), because non-passerine birds have active gulonolactone oxidase and synthesize ascorbic acid as needed.) This article summarizes the varied nutrient contributions eggs make to the human diet.

Egg Types

While the majority of eggs consumed today are chicken eggs, a variety of eggs from different species of bird are commercially available in different parts of the world, from the petite quail egg to the very large ostrich egg. The data presented in **Table 1** compare the caloric, protein, lipid, and cholesterol contents per 100 g for eggs from various species. Eggs from commercial chickens differ from those from wild breeds in that they have lower cholesterol and lipid contents. This difference could be the result of many years of genetic selection of breeds with increased feed-to-egg conversion ratios and faster rates of lay.

The commercial hen used in today's egg production has been selected for optimal feed conversion and egg production along with overall health, disease resistance, livability, and temperament. The majority of egg production is carried out using a battery cage system, which offers a high degree of control over environment, feed, water, hygiene, biosecurity, and egg collection. This system also facilitates mechanization. Other production systems include barn and free-range, which offer more freedom to the birds but often lead to higher disease and mortality rates and potentially to increased susceptibility to bacterial contamination of the eggs.

Shifting dietary patterns in the population have resulted in compensatory changes in the egg industry. A major change has been the increased use of eggs in egg products for the pre-prepared packaged-food industry. In the USA over 30% of the total egg production is used to make egg products, and

Table 1 Macronutrient composition of various raw eggs (per 100 g)

Nutrient	*Species (average egg weight)*				
	Quail (9 g)	*Chicken (50 g)*	*Duck (70 g)*	*Turkey (79 g)*	*Goose (144 g)*
Water (g)	74.35	75.84	70.83	72.50	70.43
Energy					
kJ	663	617	776	716	775
kcal	158	147	185	171	185
Protein (g)	13.05	12.58	12.81	13.68	13.87
Lipid (g)	11.09	9.94	13.77	11.88	13.27
SFA (g)	3.56	3.10	3.68	3.63	3.60
MUFA (g)	4.32	3.81	6.53	4.57	5.75
PUFA (g)	1.32	1.36	1.22	1.66	1.67
Cholesterol (mg)	844	423	884	933	852

USDA National Nutrient Database for Standard Reference, Release 16 (July 2003). Nutrient Data Laboratory Home Page, http://www.nal.usda.gov/fnic/foodcomp

SFA, saturated fatty acids; MUFA, monounsaturated fatty acids; PUFA, polyunsaturated fatty acids.

egg-product usage has been the most rapidly growing part of the industry, accounting for the majority of the increased per capita egg consumption over the past decade. Another area of growth has been the speciality egg market. As consumers become more health conscious, there has been an emphasis on functional components of foods that contribute to health and well-being. Eggs with enhanced nutrient benefits, especially with increased content of omega-3 fatty acids, are available worldwide.

Egg Macronutrient and Micronutrient Content and Distribution

The levels of many nutrients in an egg are influenced by the age and breed or strain of hen as well as the season of the year and the composition of the feed provided to the hen. While most variations in nutrients are relatively minor, the fatty acid composition of egg lipids can be significantly altered by changes in the hen's diet. The exact quantities of many vitamins and minerals in an egg are determined, in part, by the nutrients provided in the hen's diet.

Hen eggs contain 75.8% water, 12.6% protein, 9.9% lipid, and 1.7% vitamins, minerals, and a small amount of carbohydrates (**Table 2**). Eggs are classified in the protein food group, and egg protein is one of the highest quality proteins available. Virtually all lipids found in eggs are contained in the yolk, along with most of the vitamins and minerals. Of the small amount of carbohydrate (less than 1% by weight), half is found in the form of glycoprotein and the remainder as free glucose.

Egg Protein

Egg proteins, which are distributed in both yolk and white (albumen), are nutritionally complete proteins containing all the essential amino-acids (EAA). Egg protein has a chemical score (EAA level in a protein food divided by the level found in an 'ideal' protein food) of 100, a biological value (a measure of how efficiently dietary protein is turned into body tissue) of 94, and the highest protein efficiency ratio (ratio of weight gain to protein ingested in young rats) of any dietary protein.

The major proteins found in egg yolk include low density lipoprotein (LDL), which constitutes 65%, high density lipoprotein (HDL), phosvitin, and livetin. These proteins exist in a homogeneously emulsified fluid. Egg white is made up of some 40 different kinds of proteins. Ovalbumin is the major protein (54%) along with ovotransferrin (12%) and ovomucoid (11%). Other proteins of interest include flavoprotein, which binds riboflavin, avidin, which can bind and inactivate biotin, and lysozyme, which has lytic action against bacteria.

As shown in **Table 3**, egg protein contains substantial amounts of EAAs and nonessential amino-acids. The first column shows the amount of each amino-acid in one large egg. The second column indicates the amount of each amino-acid per 100 g of whole egg. The third column shows the dietary reference intake (DRI) for all of the EAAs per 50 g of total dietary protein, and the last column indicates the percentage of the DRI for each EAA provided by one large egg. While a large egg provides only 3% of the energy in a 2000 kcal (8394 kJ) diet, it provides 11% of the

Table 2 Macronutrient distribution in raw chicken egg (per 50 g large egg)

	Whole egg	*Egg albumin*	*Egg yolk*
Weight (%)	100	66	34
Water (g)	37.9	28.9	8.9
Energy			
kJ	308.5	71.3	228.8
kcal	73.5	17.2	54.7
Protein (g)	6.29	3.60	2.70
Lipid (g)	4.97	0.06	4.51
Sugars (g)	0.39	0.24	0.10

USDA National Nutrient Database for Standard Reference, Release 16 (July 2003). Nutrient Data Laboratory Home Page, http://www.nal.usda.gov/fnic/foodcomp

Table 3 Amino-acid content of a large egg

Amino-acid	*Grams per large egg*	*Grams per 100 g whole egg*	*DRI (g) EAA per 50 g protein day^{-1}*	*Percentage EAA DRI per large egg*
Alanine	0.38	0.69		
Arginine	0.42	0.77		
Aspartic acid	0.65	1.18		
Cystine[a]	0.15	0.28	1.25	12
Glutamic acid	0.85	1.54		
Glycine	0.22	0.40		
Histidine[a]	0.16	0.29	0.9	18
Isoleucine[a]	0.36	0.66	1.25	29
Leucine[a]	0.57	1.04	2.75	21
Lysine[a]	0.45	0.82	2.55	18
Methionine[a]	0.21	0.39	1.25	17
Phenylalanine[a]	0.35	0.64	2.35	15
Proline	0.26	0.48		
Serine	0.50	0.91		
Threonine[a]	0.32	0.59	1.35	24
Tryptophan[a]	0.11	0.19	0.35	31
Tyrosine[a]	0.28	0.51	2.35	12
Valine[a]	0.43	0.79	1.6	27

[a]Essential amino-acids (EAA) are not synthesized by the body and must be consumed in foods; therefore, only EAA have a dietary reference intake (DRI) value.

USDA National Nutrient Database for Standard Reference, Release 16 (July 2003). Nutrient Data Laboratory Home Page, http://www.nal.usda.gov/fnic/foodcomp

protein needs. The EAAs in an egg contribute between 12% and 31% of the DRI for the various EAAs.

Egg Lipids

A large egg yolk contains 4.5 g of lipid, consisting of triacylglycerides (65%), phospholipids (31%), and cholesterol (4%). Of the total phospholipids, phosphatidylcholine (lecithin) is the largest fraction and accounts for 26%. Phosphatidylethanolamine contributes another 4%. The fatty-acid composition of egg-yolk lipids depends on the fatty-acid profile of the diet. The reported fatty-acid profile of commercial eggs indicates that a large egg contains 1.55 g of saturated fatty acids, 1.91 g of monounsaturated fat, and 0.68 g of polyunsaturated fatty acids. (Total fatty acids (4.14 g) does not equal total lipid (4.5 g) because of the glycerol moiety of triacylglycerides and phospholipids and the phosphorylated moieties of the phospholipids). It has been reported that eggs contain less than 0.05 g of trans-fatty acids. Egg yolks also contain cholesterol (211 mg per large egg) and the xanthophylls lutein and zeaxanthin. The lipid profile of a large egg is presented in **Table 4**.

Egg Vitamins

Eggs contain all the essential vitamins except vitamin C, because the developing chick does not have a dietary requirement for this vitamin. As shown in **Table 5**, the yolk contains the majority of the water-soluble vitamins and 100% of the fat-soluble vitamins. Riboflavin and niacin are concentrated in the albumen. The riboflavin in the egg albumin is bound to flavoprotein in a 1:1 molar ratio. Eggs are one of the few natural sources of vitamins D and B_{12}. Egg vitamin E levels can be increased up to tenfold through dietary changes. While no single vitamin is found in very high quantity relative to its DRI value, it is the wide spectrum of vitamins present that makes eggs nutritionally rich.

Table 4 Egg yolk lipid profile per large egg

Lipids	*Amount*
Fatty acids, total saturated (g)	1.55
8:0–14:0	0.02
16:0	1.16
18:0	0.41
20:0–24:0	0.01
Fatty acids, total monounsaturated (g)	1.99
16:1	0.16
18:1	1.82
20:1	0.02
Fatty acids, total polyunsaturated (g)	0.72
18:2	0.60
18:3	0.02
20:4	0.07
20:5–22:6 *n*-3	0.02
Cholesterol (mg)	211
Carotene, β (μg)	15
Carotene, α (μg)	6.5
Cryptoxanthin, β (μg)	5.6
Lutein + zeaxanthin (μg)	186

USDA National Nutrient Database for Standard Reference, Release 16 (July 2003). Nutrient Data Laboratory Home Page, http://www.nal.usda.gov/fnic/foodcomp

Table 5 Egg vitamin content per large egg

Vitamin	*Whole*	*Albumen*	*Yolk*
Thiamin (mg)	0.04	<0.01	0.03
Riboflavin (mg)	0.24	0.15	0.09
Niacin (mg)	0.04	0.04	<0.01
Pantothenic acid (mg)	0.72	0.06	0.51
Vitamin B_6 (mg)	0.07	<0.01	0.06
Folate, total (μg)	23.5	0	24.8
Vitamin B_{12} (μg)	0.65	0.03	0.33
Vitamin A (IU)	243.5	0	245.1
Choline (mg)	125.5	0	125.5
Retinol (μg)	70	0	63.1
Vitamin E (mg)	0.49	0	0.44
Vitamin D (IU)	17.3	0	18.3
Vitamin K (μg)	0.15	0	0.12

USDA National Nutrient Database for Standard Reference, Release 16 (July 2003). Nutrient Data Laboratory Home Page, http://www.nal.usda.gov/fnic/foodcomp

Egg Minerals

Eggs contain small amounts of all the minerals essential for life. Of particular importance is the iron found in egg yolks. Research evaluating the plasma iron and transferrin saturation in 6–12-month-old children indicated that infants who ate egg yolks had a better iron status than infants who did not. The study indicated that egg yolks can be a source of iron in a weaning diet for breast-fed and formula-fed infants without increasing blood antibodies to egg-yolk proteins. Dietary iron absorption from a specific food is determined by iron status, heme- and nonheme-iron contents, and amounts of various dietary factors that influence iron absorption present in the whole meal. Limited information is available about the net effect of these factors as related to egg iron bioavailability.

In addition to iron, eggs contain calcium, phosphorus, sodium, potassium, magnesium, zinc, copper, and manganese (**Table 6**). Egg yolks also contain iodine (25 μg per large egg), and this can be increased twofold to threefold by the inclusion of an iodine source in the feed. Egg selenium content can also be increased up to ninefold by dietary manipulations.

Egg Choline

Choline was established as an essential nutrient in 1999 with recommended daily intakes (RDIs) of 550 mg for

Table 6 Egg mineral content per large egg

Mineral	*Whole*	*Albumen*	*Yolk*
Calcium (mg)	26.5	2.3	21.9
Iron (mg)	0.92	0.03	0.46
Magnesium (mg)	6.0	3.63	0.85
Phosphorus (mg)	95.5	4.95	66.3
Potassium (mg)	67.0	53.79	18.53
Sodium (mg)	70.0	54.78	8.16
Zinc (mg)	0.56	0.01	0.39
Copper (mg)	0.05	0.01	0.01
Manganese (mg)	0.02	<0.01	0.01
Selenium (μg)	15.85	6.60	9.52

USDA National Nutrient Database for Standard Reference, Release 16 (July 2003). Nutrient Data Laboratory Home Page, http://www.nal.usda.gov/fnic/foodcomp

men and 450 mg for women. The RDI for choline increases during pregnancy and lactation owing to the high rate of choline transfer from the mother to the fetus and into breast milk. Animal studies indicate that choline plays an essential role in brain development, especially in the development of the memory centers of the fetus and newborn. Egg-yolk lecithin (phosphatidylcholine) is an excellent source of dietary choline, providing 125 mg of choline per large egg.

Egg Carotenes

Egg yolk contains two xanthophylls (carotenes that contain an alcohol group) that have important health benefits – lutein and zeaxanthin. It is estimated that a large egg contains 0.33 mg of lutein and zeaxanthin; however, the content of these xanthophylls is totally dependent on the type of feed provided to the hens. Egg-yolk lutein levels can be increased up to tenfold through modification of the feed with marigold extract or purified lutein. An indicator of the lutein + zeaxanthin content is the color of the yolk; the darker yellow-orange the yolk, the higher the xanthophyll content. Studies have shown that egg-yolk xanthophylls have a higher bioavailablity than those from plant sources, probably because the lipid matrix of the egg yolk facilitates greater absorption. This increased bioavailability results in significant increases in plasma levels of lutein and zeaxanthin as well as increased macular pigment densities with egg feeding.

Egg Cholesterol

Eggs are one of the richest sources of dietary cholesterol, providing 215 mg per large egg. In the 1960s and 1970s the simplistic view that dietary cholesterol equals blood cholesterol resulted in the belief that eggs were a major contributor to hypercholesterolemia and the associated risk of cardiovascular disease. While there remains some controversy regarding the role of dietary cholesterol in determining blood cholesterol levels, the majority of studies have shown that saturated fat, not dietary cholesterol, is the major dietary determinant of plasma cholesterol levels (and eggs contain 1.5 g of saturated fat) and that neither dietary cholesterol nor egg consumption are significantly related to the incidence of cardiovascular disease. Across cultures, those countries with the highest egg consumption actually have the lowest rates of mortality from cardiovascular disease, and within-population studies have not shown a correlation between egg intake and either plasma cholesterol levels or the incidence of heart disease. A 1999 study of over 117 000 men and women followed for 8–14 years showed that the risk of coronary heart disease was the same whether the study subjects consumed less than one egg a week or more than one egg a day.

Clinical studies show that dietary cholesterol does have a small influence on plasma cholesterol levels. Adding one egg per day to the diet would, on average, increase plasma total cholesterol levels by approximately $5\,\mathrm{mg\,dl^{-1}}$(0.13 mmol/L). It is important to note, however, that the increase occurs in both the atherogenic LDL cholesterol fraction ($4\,\mathrm{mg\,dl^{-1}}$(0.10 mmol/L)) and the antiatherogenic HDL cholesterol fraction ($1\,\mathrm{mg\,dl^{-1}}$(0.03 mmol/L)), resulting in virtually no change in the LDL:HDL ratio, a major determinant of cardiovascular disease risk. The plasma lipoprotein cholesterol response to egg feeding, especially any changes in the LDL:HDL ratio, vary according to the individual and the baseline plasma lipoprotein cholesterol profile. As shown in **Table 7**, adding one egg a day to the diets of three hypothetical patients with different plasma lipid profiles results in very different effects on the LDL:HDL ratio. For the individual at low risk there is a greater effect than for the person at high risk, yet in all cases the effect is quantitatively minor and would have little impact on their heart-disease risk profile. Overall, results from clinical studies indicate that egg

Table 7 Changes in plasma lipoprotein cholesterol levels with addition of one large egg per day to the diet

	Cholesterol ($mg\,dl^{-1}$)		*LDL:HDL ratio*
	LDL	*HDL*	*(% change)*
Baseline	125	50	2.50
$+1\,\mathrm{egg\,day^{-1}}$	129	51	2.53 (+1.2%)
Baseline	150	50	3.00
$+1\,\mathrm{egg\,day^{-1}}$	154	51	3.02 (+0.6%)
Baseline	175	50	3.50
$+1\,\mathrm{egg\,day^{-1}}$	179	51	3.51 (+0.3%)

feeding has little if any effect on cardiovascular disease risk. This is consistent with the results from a number of epidemiological studies.

A common consumer misperception is that eggs from some breeds of bird have low or no cholesterol. For example, eggs from Araucana chickens, a South American breed that lays a blue-green egg, have been promoted as low-cholesterol eggs when, in fact, the cholesterol content of these eggs is 25% higher than that of commercial eggs. The amount of cholesterol in an egg is set by the developmental needs of the embryo and has proven very difficult to change substantially without resorting to hypocholesterolemic drug usage.

Undue concerns regarding egg cholesterol content resulted in a steady decline in egg consumption during the 1970s, 1980s, and early 1990s, and restriction of this important and affordable source of high-quality protein and other nutrients could have had negative effects on the well-being of many nutritionally 'at risk' populations. Per capita egg consumption has been increasing over the past decade in North America, Central America, and Asia, has remained relatively steady in South America and Africa, and has been falling in Europe and Oceania. Overall, world per capita egg consumption has been slowly increasing over the past decade, in part owing to the change in attitude regarding dietary cholesterol health concerns.

Allergenic Aspects of Egg Proteins

Eggs are one of the most common causes of food allergies in infants and young children. Although the majority of egg allergies are caused by egg-white protein, proteins in both the egg white and the yolk are associated with allergies. The egg white contains 50% ovalbumin, which is the major allergen. Other egg-white allergenic proteins are ovomucoid, ovotransferrin, and lysozyme. Most egg allergies in young children are outgrown by the age of 5 years following an elimination diet.

Owing to the allergenicity of egg proteins, it is advised not to feed egg yolks to infants younger than 6 months of age and to wait until children are 12 months old to feed them egg whites. When feeding egg yolks to children between the ages of 6 months and 12 months, the eggs should be prepared in such a way that the egg white can be completely removed, as in hardboiled eggs.

Speciality Eggs

There is an increasing interest worldwide in the production and marketing of speciality eggs with enhanced nutrient benefits. The nutrient composition of an egg can be significantly modified by altering the composition of the feed. Commercially available nutrient-enhanced eggs contain increased amounts of omega-3 fatty acids, vitamin E, selenium, and lutein. Other enhancements include increased contents of vitamin D and the B vitamins as well as incorporation of conjugated linoleic acid.

Omega-3 Fatty Acids

The fatty-acid content of eggs is easily and significantly affected by the fatty-acid profile of the hen's feed. The omega-3 fatty-acid content of eggs can be increased by feeding hens a source of omega-3 fatty acids. In some countries, fish meal is used as a source of omega-3 fatty acids, but this can result in eggs with a fishy odour and taste. Marine algae are another source of omega-3 fatty acid and result in higher concentrations of eicosapentanoic acid and docosahexanoic acid (DHA) in egg yolks. Flaxseed oil is also used as a source of omega-3 fatty acids and results in increased levels of α-linolenic acid in egg yolks. The relative proportion of DHA to α-linolenic acid can be controlled by feeding a mixture of flaxseed oil and marine algae. It is possible to attain levels as high as 200 mg of omega-3 fatty acids per large egg.

Although omega-3 fatty-acid levels in eggs are well below levels found in fishes such as salmon and tuna, eggs can still be an important source of omega-3 fatty acids in the diet. For people who cannot eat fish, eggs with higher levels of omega-3 fatty acids can be an important way of including these beneficial fatty acids in the diet.

Other Nutrients in Speciality Eggs

By altering the content in the feed, other nutrients in eggs can be enhanced, for example lutein, vitamin E, and selenium. Vitamin E is usually added to the feed to serve as an antioxidant when the polyunsaturated fatty acids are increased. Vitamin E levels in eggs have been increased as much as 25-fold. The vitamin E in these eggs can provide an additional natural source of this important fat-soluble vitamin. Lutein (a xanthophyll) content can also be increased in eggs by increasing the amount in feed, usually in the form of marigold extract. Lutein is deposited in the egg yolk at levels as high as 2 mg per large egg, and the human body readily absorbs lutein from the egg phospholipid matrix. Nutritional needs for selenium vary widely owing to differences in the selenium content of regional soils. Egg selenium levels can be increased between 5-fold and 8-fold by the addition of an organo-selenium source to the feed.

Egg Food Safety

Eggs pose a unique food-safety problem because they can be contaminated internally with the pathogenic bacteria *Salmonella enterica* Serovar Enteritidis (SE). If SE infects the reproductive tracts of laying hens, it can be deposited in the eggs during formation. In addition to internal egg contamination by SE, eggshells can be contaminated with a number of microorganisms. Caution is required when selecting eggs for consumption. Only clean eggs should be consumed. Vaccinating hens against salmonella, together with temperature control, proper handling, and cooking are important control measures to reduce the incidence of SE illness.

When SE internally contaminates an egg, it is thought to be deposited at the yolk membrane in the egg white. The integrity of the vitelline membrane is very important in preventing SE from entering the yolk, where it could grow very rapidly in the nutrient-rich environment. The egg white has natural antimicrobial compounds, such as lysozyme, that help prevent SE from growing.

In naturally contaminated eggs, scientists have documented that between 10 and 100 cells of SE may be deposited in an egg. The bacterial cell count will remain low unless the egg is exposed to temperatures that would allow rapid growth of SE or the vitelline membrane breaks down. Even when flocks are infected with SE, only a small percentage of the eggs produced will contain SE. Properly cooking eggs to a temperature of 63 °C for 3 min, 65 °C for 1 min, or 70 °C for 1 s will destroy SE if it is present in an egg.

The Role of Eggs in the Diet

The nutritional contribution of eggs to a diet is determined by the per capita consumption profile of a given country. In countries such as Japan, with the highest per capita egg consumption, eggs play an important role as a source of nutrients, while in countries such as India, with very low per capita consumption, their role is minor. Worldwide there are many misperceptions and myths regarding eggs, which influence consumption patterns (**Table 8**).

Eggs are a nutrient-dense source of many EAA, vitamins, and minerals, and, as shown in **Figure 1**, eggs contribute a number of nutrients to the American diet in amounts proportionally greater than their caloric contributions. While providing only 1.3% of the calories, they provide nine different nutrients in amounts ranging from 2% to 6% of the DRI. Such nutrient-dense foods can play an

Table 8 Common myths and misperceptions about eggs

Myth	*Fact*
Brown eggs are healthier than white eggs; fertile eggs have less or no cholesterol; free-range eggs have more nutritional value than commercial eggs	There are no substantive nutritional differences between white eggs, brown eggs, fertile eggs, and free-range eggs; nutritional content is determined by the hen's diet
Eggs contain the hormones they give the hen to force her to lay eggs when there isn't a rooster around	Hens are not given hormones to produce eggs in the absence of a rooster; hens lay eggs with or without a rooster; there are no harmful hormones in eggs
Eggs contain the antibiotics they give hens to increase the number of eggs they'll lay	Antibiotics have no effect on egg production and there is no value in using them unless needed for therapeutic reasons
Eggs in the store are a mixture of fertile and non-fertile eggs; that stringy stuff is the embryo	Commercial eggs are not fertile (can be included in a lactoovo- or ovo-vegetarian diet); that stringy stuff (chalaza) is an egg protein that anchors the yolk in the centre of the egg
Eating eggs can cause liver problems	No study has ever shown that eggs cause liver problems
Eggs with blood or meat spots are fertilized or are bad	The tiny meat or blood spot is caused by the rupture of a blood vessel during egg formation; it has no adverse effect on the egg and can be either removed or eaten
If an egg floats in water, it is bad	As an egg ages the air sac expands and an egg will stand on end in water; this is not a sign that the egg is bad

Additional information and facts can be obtained from the American Egg Board *Eggcyclopedia*. (http://www.aeb.org/).

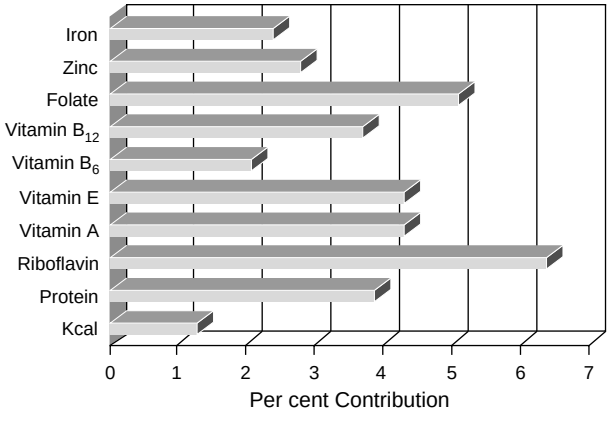


Figure 1 Nutrient contributions of eggs to the American diet.

important role in the diets of seniors who have decreased caloric intakes as well as in weight-reduction and weight-maintenance diets. Studies have shown that egg intake has a significant effect on satiety beyond what would be predicted from its protein and fat contents. Egg intake slows the rate of gastric emptying, resulting in a flatter blood glucose response and a lower insulin response. The effects on gastric emptying appear to be related to the effects of egg yolk (not white) intake on the secretion of cholecystokinin and gastric inhibitory peptide.

Summary

For nutritionally vulnerable populations, including the poor, the very young, the very old, pregnant women, and those suffering from chronic diseases, eggs are an affordable nutrient-dense source of high-quality protein important for maintaining health and facilitating recovery. Pregnancy is an especially important time to optimize the intake of high-quality protein and other essential nutrients to reduce the risk of low birth weight and the associated development of chronic diseases and other health problems during the infant's adult life. Eggs also serve as an important dietary source of choline during pregnancy and lactation, providing the fetus and newborn with choline for brain development. In addition, eggs provide a satiety effect, which, in view of the global problem of obesity, can be a valuable addition to weight-loss and weight-maintenance programs. For various populations, from infants to the aged, there are a multitude of health reasons to include nutrient-dense eggs as part of the diet, and for many of these groups it is economically feasible.

The high-quality protein, many nutritional components, low caloric content, affordability, blandness, ease of digestibility, and satiety response all make eggs ideal for inclusion in the diet at all ages, from very young to very old, and in times of both health and convalescence.

See also: **Antioxidants**: Diet and Antioxidant Defense. **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Choline and Phosphatidylcholine**. **Coronary Heart Disease**: Lipid Theory. **Fatty Acids**: Omega-3 Polyunsaturated. **Food Allergies**: Etiology. **Food Safety**: Bacterial Contamination. **Phytochemicals**: Classification and Occurrence; Epidemiological Factors. **Pregnancy**: Nutrient Requirements. **Protein**: Requirements and Role in Diet; Digestion and Bioavailability; Quality and Sources.

Further Reading

American Egg Board. *Eggcyclopedia*, Chicago: American Egg Board. http://www.aeb.org/eggcyclopedia/main_frame_page.html

Egg Nutrition Center, Washington, DC. http://www.enc-online.org

Handelman GJ, Nightingale ZD, Lichtenstein AH, Schaefer EJ, and Blumberg JB (1999) Lutein and zeaxanthin concentrations in plasma after dietary supplementation with egg yolk. *American Journal of Clinical Nutrition* **70**: 247–251.

Herron KL and Fernandez ML (2004) Are the current dietary guidelines regarding egg consumption appropriate? *Journal of Nutrition* **134**: 187–190.

Hu FB, Stampfer MJ, Rimm EB *et al.* (1999) A prospective study of egg consumption and risk of cardiovascular disease in men and women. *JAMA* **281**: 1387–1394.

Humphrey TJ (1994) Contamination of egg shell and contents with *Salmonella enteritidis*: a review. *International Journal of Food Microbiology* **21**: 31–40.

McNamara DJ (2000) Dietary cholesterol and atherosclerosis. *Biochimica et Biophysica Acta* **1529**: 310–320.

McNamara DJ (ed.) (2000) Where would we be without the egg? A conference about nature's original functional food *Journal of the American College of Nutrition* **19**: 495S–562S.

Stadelman WJ and Cotterill OJ (1995) *Egg Science and Technology*, 4th edn. New York: Food Products Press.

Watson RR (ed.) (2002) *Eggs and Health Promotion*. Ames: Iowa State Press.

Yamamoto T, Juneja LR, Hatta H, and Kim M (1997) In *Hen Eggs: Their Basic and Applied Science*. Boca Raton: CRC Press.

Eicosanoids *see* **Prostaglandins and Leukotrienes**

ELECTROLYTES

Contents
Acid-Base Balance
Water–Electrolyte Balance

Acid-Base Balance

A G Jardine and P B Mark, University of Glasgow, Glasgow, UK

Introduction

Maintenance of cellular and extracellular pH (hydrogen ion concentration) is essential to life, in view of the exquisite pH dependence of processes such as enzyme function. Hydrogen ions (H^+) are generated by cellular metabolism and, to a lesser extent by the ingestion of acids in the diet. Acid-base homeostasis regulates pH between 7.36 and 7.44 (corresponding to a [H+] of 36–44 nmol l^{-1}) in extracellular fluids, such as blood, whereas intracellular pH is more acidic (pH 6.3–7.4) depending on individual organs and circumstances. The pH of subcellular organelles may be more acidic, reflecting their physiological function (e.g., lysosomes). Blood and extracellular fluid pH are tightly regulated by the presence of buffer systems, which attenuate changes as a consequence of acid load. These buffer systems, both extracellular and intracellular, include hemoglobin, other proteins, phosphate, and bicarbonate – the latter being of greatest importance. However, the acid load must ultimately be eliminated by the subsequent excretion of volatile acid by the lungs and fixed acids by the kidney.

Definitions, Acids, Bases, and Buffers

pH

The term pH is an expression of hydrogen ion (H^+) concentration (such that pH and H^+ are inversely related):

$$pH = -\log_{10}[H^+] \qquad [1]$$

Acids and Bases

Acids are substances that dissociate to donate H^+ (eqn [2]); the stronger the acid, the more readily it dissociates. The dissociation constant (p*K*a) is the pH at which 50% of the acid is dissociated. At pH values greater than p*K*a more H^+ will dissociate; the lower the p*K*a, the stronger the acid. A base is a substance that accepts hydrogen ions. In the following text the term 'fixed acid' is used to describe formed acid, and 'volatile acid' is used to describe the potential acid load imposed by carbon dioxide (CO_2). Where 'A' represents an acid, the following applies:

$$AH \leftrightarrow A^- + H^+ \qquad [2]$$

The importance of this relationship in physiological terms is that since the p*K*a of most organic acids is much lower than the pH of extracellular fluids, most organic acids exist in a dissociated state (as acid anion salts) the free H^+ being buffered. In urine, where the minimum achievable pH is around 5, most strong acids (with a p*K*a below this value) will be in a dissociated state, necessitating the excretion of H^+ together with urinary buffers.

Acidosis is the term used to describe conditions where pH is low and those where pH would be low were it not appropriately buffered; similarly, alkalosis is the term used for a high pH and for a potentially elevated pH that has been appropriately buffered. Acidemia and alkalemia reflect low or elevated blood pH. It is common to describe acidosis/alkalosis as respiratory or metabolic depending on their causation.

Buffers

Buffering is the ability of weak acids, present in excess, to accept H^+ donated from strong acids, thus limiting the changes in free H^+ concentrations and pH changes (equation [3]):

$$AH + \text{Buffer} \leftrightarrow \text{Buffer-}H^+ + A^- \qquad [3]$$

The principal buffer system in blood (and other extracellular fluids) is based on bicarbonate (HCO_3^-), accounting for approximately 70% of the buffering capacity of the blood. In blood, CO_2 (the major product of oxidative metabolism) reacts with water in the presence of the enzyme carbonic anhydrase (CA) to form carbonic acid (H_2CO_3). This compound is relatively unstable and tends to dissociate (eqn [4]). The rate of formation of

carbonic acid is dependent on the concentration of carbon dioxide and the rate constant of reaction [i]; the dissociation of carbonic acid to generate H^+ and HCO_3^- is governed by the rate constant of reaction [ii]. In practice, these two reactions can be combined, and the relationship between pH ($[H^+]$), carbon dioxide, and bicarbonate is described by a single equation – the Henderson-Hasselbalch equation [5]:

$$\underset{[i]}{CO_2 + H_2O} \leftrightarrow \underset{[ii]}{H_2CO_3} \leftrightarrow H^+ + HCO_3^- \quad [4]$$

$$pH = 6.1 + \log_{10}([HCO_3^-]/K.S.PCO_2) \quad [5]$$

pH reflects $-\log [H^+]$; 6.1 is the value of $-\log (1/K)$, K being the equilibrium constant describing the overall equation [4]; P_{CO_2} is the partial pressure of carbon dioxide; S is the solubility constant for carbon dioxide. $K.S.$ is constant and equal to 0.225 when P_{CO_2} is measured in kPa, 0.03 when P_{CO_2} is measured in mmHg). **Table 1** shows the normal range for these parameters in humans.

From eqn [5] the principles of acid-base balance can be appreciated. Acidification may occur in two ways: either by the production of CO_2 or by the consumption of bicarbonate (as part of the buffering of fixed acid). The excretion of CO_2 (see below) is controlled by the lungs, and the excretion of fixed acid takes place in the kidney.

The Henderson-Hasselbalch equation allows basic understanding of acid-base physiology, in health and disease, but has limitations. In the presence of either metabolic or respiratory derangement of acid-base homeostasis it does not allow assessment of the severity of the metabolic derangement, analogous to the respiratory component. It also does not assess the influence of other acids other than carbonic acid. For this reason some authors propose analysis of acid-base physiology using a more complex method based on the principles of physical chemistry. This method proposes that all changes pH in plasma can be explained in terms of relative concentrations of CO_2, relative electrolyte, and weak acid. This concept allows more rigorous interrogation of acid-base disorders and may permit greater insight into their pathophysiology and management.

Table 1 Normal ranges

Variable	*Normal range*
pH	7.36–7.44
Hydrogen ion (H^+)	37–44 nmol l^{-1}
Partial pressure CO_2 (P_{CO_2})	34–46 mmHg; 4.5–6.1 kPa
Bicarbonate (HCO_3^-)	24–30 mmol l^{-1}

Maintenance of the pH of the Blood and Extracellular Fluids

Acid and Alkali Load

The sources of acids (and alkalis) are from the diet and metabolism. The major potential source of acid is CO_2 ('volatile acid'; eqn [4]) generated by oxidative metabolism; a total of 12–20 mol of CO_2 are produced daily. Other metabolic products include lactic acid, other organic acids, and urea, the synthesis of which produces H^+. Because of its role in the metabolism of lactic acid and in the synthesis of urea, the liver plays a major role in acid-base homeostasis that is often not appreciated.

Volatile acid (CO_2) is excreted by the lungs, whereas the breakdown of sulfur and phosphorus-containing compounds are 'fixed' acids, requiring excretion by the kidney. For example, cysteine or methionine metabolism leads to the production of sulfuric and phosphoric acid (H_2SO_4, H_3PO_4), while the metabolism of other amino acids (lysine, arginine, and histidine) leads to the production of hydrochloric acid (HCl). In contrast, organic acids (e.g., lactate, fatty acids) may be completely metabolized to CO_2 and H_2O and thus excreted by the lungs. In addition, the absorption of dietary phosphate and the fecal loss of bicarbonate represent an additional acid load. In total, the net acid load of fixed acid is approximately 1 mmol kg^{-1} day^{-1} and may be increased by a high protein intake or reduced by a strict vegetarian diet.

There is surprisingly little information on the direct contributions of individual foods to the acid burden. However, this source of dietary acid is of increasing importance in view of current popular weight reduction diets (e.g., the Atkins diet). The major acids contained in food are citric acid (in fruit, fruit juices), acetic acid (as a preservative, pickles, vinegar), lactic acid (yogurt, fermented foods), malic acid (fruit), oxalic acid (vegetables that contain smaller amounts of citric and malic acids), and tartaric acid (wine). Oxalic acid precipitates in the gut to form calcium salts, which are excreted in the stool and little is absorbed. The other acids are absorbed but quickly metabolized and present an acid burden in the form of CO_2. The largest source of fixed acid comes from the metabolism of amino acids (particularly those from animal proteins – see above). The significance of this source of acid is readily demonstrated in patients consuming a high-protein diet (particularly one rich in animal protein) who have increased urinary acid excretion. Based on studies on the relationship between diet, renal excretion of acid, and urine pH it is theoretically feasible to quantify urinary acid

excretion for individual foods. However, because of daily variation in diet (and therefore absence of a metabolic steady state) and inherent variation in the composition of foodstuffs, it has not been possible to date to estimate accurately the effects of diet on renal acid-base metabolism in circumstances reflective of normal dietary intake.

Alkalis are often prescribed to compensate for metabolic acidosis (see below) and in the past were often used to neutralize gastric acidity. Milk and milk products are also alkaline but seldom cause any disturbance, unless consumed in great excess. Excessive consumption of milk or alkali is now rarely seen.

Regulation

Blood and extracellular fluid pH is regulated at three levels: (1) buffering within the blood and tissues; (2) excretion of volatile acids by the lungs; and (3) excretion of fixed acids by the kidney. Whilst buffering is immediate, respiratory compensation occurs over minutes to hours and renal excretion takes many hours to days (**Table 2**).

Blood/Extracellular Fluid

Immediate buffering of an acid load, for example by the release of lactic acid and CO_2 by anaerobic and aerobic metabolism in exercising muscle, occurs in the blood and other extracellular fluids, which together contain approximately 350 mmol of bicarbonate buffer. Sixty to seventy per cent of the buffering capacity of blood is accounted for by the bicarbonate buffer system; 20–30% is dependent on direct binding to hemoglobin and to other proteins, including plasma proteins. Blood is in equilibrium with extracellular fluid H^+. H^+ ions move across cell membranes depending on concentration and charge; thus, H^+ ions may move into cells in exchange for K^+ (and to a lesser extent Na^+ ions) when extracellular H^+ is increased. Hence, acidosis is often accompanied by increased serum K^+, and alkalosis by low K^+. Large amounts of H^+ may be 'buffered' by direct binding to proteins within cells and tissues, particularly bone where H^+ ions are also buffered by calcium salts, such as apatite.

Table 2 Buffering and acid-base regulation

Mechanism	*Site*	*Role (time)*
Protein (e.g., Hb)	Cell	Rapid binding of H^+ (seconds)
Bicarbonate buffer	ECF	Buffering of H^+ (seconds)
Ventilation	Lungs	Excretion of CO_2, respiratory compensation (hours)
Fixed acid excretion	Kidney	Excretion of H^+, reabsorption and regeneration of bicarbonate, renal compensation (hours to days)

Lungs

The lungs excrete volatile acid (CO_2) by changes in the rate and volume of respiration. This is regulated by respiratory centers in the brainstem that respond to changes in the pH of the cerebrospinal fluid (which is in equilibrium with extracellular fluids elsewhere in the body), and signals from chemoreceptors in the carotid and aortic bodies that are responsive to changes in pH and P_{CO_2} of the arterial blood (increased P_{CO_2} or reduced pH cause an increase in respiration). Thus, acidosis leads to an increase in respiratory rate and ventilatory volume (the pattern in severe acidosis being described as Kussmaul breathing) and alkalosis leads to the opposite effect.

Kidneys

The kidneys have two major roles in acid-base homeostasis: the recovery of filtered bicarbonate and generation of new bicarbonate; and the excretion of fixed acid (**Figures 1** and **2**; eqn [5]). Blood is filtered in the glomeruli and the glomerular filtrate is subsequently modified in the renal tubules so that the final urine volume is less than 1% of the glomerular filtrate volume. Plasma bicarbonate concentration is approximately 25 mmol l^{-1} and glomerular filtration rate (GFR) is 100 ml min^{-1}, thus 3600 mmol of bicarbonate must be reabsorbed daily.

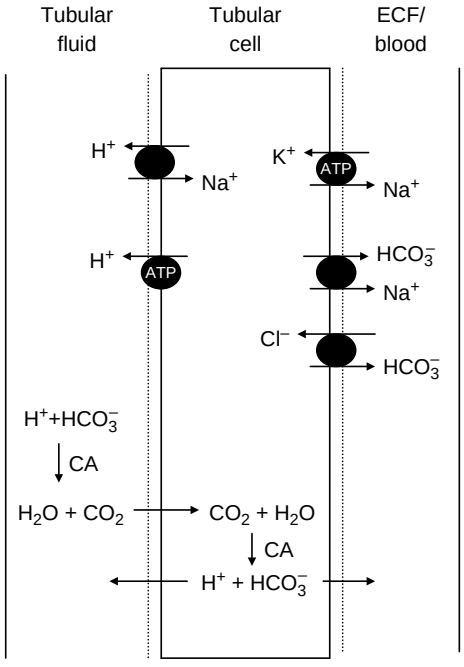


Figure 1 Recovery of filtered bicarbonate in the proximal convoluted tubule. CA, carbonic anhydrase.

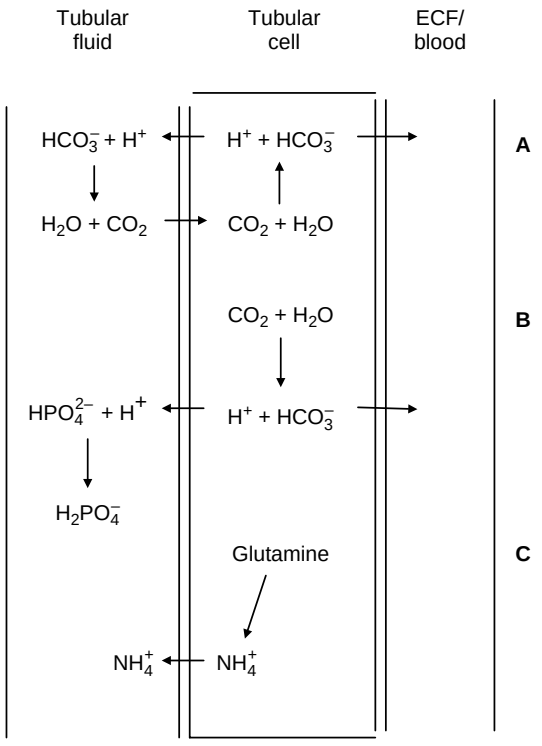


Figure 2 Excretion of acid in the collecting duct.

Bicarbonate reabsorption mainly occurs in the proximal convoluted tubule (PCT, **Figure 1**). Eighty-five per cent of filtered bicarbonate is reabsorbed at this site: 10% in the thick ascending limb of the loop of henle, the remainder being titrated to regulate total acid excretion in the collecting duct (**Figure 2**). As shown in the Figure, different mechanisms are involved at each tubular site. The enzyme carbonic anhydrase, on the luminal brush border of tubular cells, catalyzes the combination of filtered bicarbonate with H^+, secreted by the apical H^+-ATPase and Na^+/H^+ exchangers on tubular cells, to generate CO_2. CO_2 then diffuses into the tubular cells down its concentration gradient. Within the cell, carbonic anhydrase catalyzes the reverse reaction generating the production of H^+ and HCO_3^-. Hydrogen ions are then recycled to the tubular lumen and bicarbonate is secreted into the extracellular fluid (by basolateral anion exchangers or Na^+–HCO_3^- cotransporters) passing into the extracellular fluid and blood. The tubule cells are also exposed to CO_2 in the extracellular fluid and will continue to generate H^+ even in the absence of filtered bicarbonate. This H^+ is then buffered by other buffers in the glomerular filtrate including HPO_4^{2-} and, to a lesser extent, creatinine. Strong acids (e.g., H_2SO_4) with low p*K*a values will dissociate in the urine (pH range 5–8) and are buffered, whereas weaker acids may be excreted intact. In the presence of alkalosis, the function of cellular transporters may be reversed so that H^+ secretion occurs on the basolateral membrane and HCO_3^- secretion on the brush border of tubular cells resulting in alkaline urine.

Classically, the final mechanism by which the kidney can excrete H^+ is by the generation of ammonium (NH_4^+) from the metabolism of glutamine by glutaminase (**Figure 2**), a process that is stimulated by low pH and increased P_{CO_2}. The excretion of H^+ as part of ammonium accounts for around $70\,\text{mmol}\,\text{day}^{-1}$, increasing several-fold (albeit over a period of days) in the face of an acid load. Whether this is truly a urinary buffer is the subject of some debate as ammonium (NH_4^+) is generated directly from glutamine rather than accepting additional protons. There are alternative mechanisms for the role of NH_4^+ in overall acid-base homeostasis that involve the liver. After being pumped into the glomerular filtrate, NH_4^+ may be reabsorbed by the tubule and used by the liver to synthesize urea, generating free H^+ ions. Thus, there is no net loss of H^+ and the overall role of NH_4^+ in acid-base balance is dependent on the balance between tubular reabsorption of NH_4^+ and the hepatic synthesis of urea. The latter function may also be directly influenced by extracellular pH.

Liver and Bone

The liver plays additional roles in acid-base balance that may be underestimated. For example, the liver metabolizes lactate and keto acids; the rate of metabolism is dependent on pH (e.g., ketogenesis is suppressed at low pH) and may be exceeded at higher concentrations of lactate or in liver disease. The synthesis of urea from ammonium and carbon dioxide (above) results in the genesis of two protons, and is reduced in the presence of acidosis. Some buffering also occurs in bone due to the slow exchange of bone calcium carbonate for extracellular phosphate.

Measurement of Urinary Acid Excretion

Urinary pH can be measured by commercially available 'dipstix' or by using a pH meter on a fresh sample of urine. The loss of CO_2 or the production of NH_4^+ from urea-splitting organisms in infected urine will alter the pH with time. The excretion of fixed acid can be determined by the chemical titration of urine to pH 7.4, and is commonly termed 'titratable' acidity. The amount of NH_4^+ is usually estimated from the difference between the most abundant cation (Na^+, K^+) and anion (Cl^-) concentrations in the urine.

Effects of Acid-Base Disturbance

In addition to the adaptive changes occurring in acidosis, a range of metabolic and pathophysiological changes occur; alkalosis tends to produce opposite but milder effects. The metabolism of carbohydrate is altered: both glycolysis and gluconeogenesis are inhibited in the liver. Delivery of oxygen to the tissues is increased by the reduced ability of hemoglobin to retain oxygen in an acid environment (the Bohr effect). Consciousness is impaired, leading to coma in severe cases. However, the most important effects from a clinical perspective are cardiovascular: vasodilatation occurs in peripheral tissues, cardiac contractility is impaired resulting in reduced blood pressure, and, when severe, in reduced tissue perfusion. It is these effects that contribute to the adverse effects of acidosis in, for example, septic shock and contribute to the high mortality in these conditions.

Abnormalities in Acid-Base Balance

Disturbances in acid-base balance are classified as either 'acidosis,' indicating an excess of H^+ ions in the blood (reduced pH) or alkalosis, indicating the opposite. In practice, acidosis is the more common, varied, and serious problem. Disturbances in acid-base balance are usually labeled according to their origin. For example, respiratory acidosis reflects a primary problem in gas exchange with impaired excretion of CO_2, whereas metabolic acidosis reflects the over-production of fixed acid or loss of bicarbonate. Compensation refers to the body's ability to offset the primary problem. Thus, the response to a primary metabolic acidosis is to increase the excretion of CO_2 – respiratory compensation. If the pH returns to normal the problem is said to be 'fully compensated' whereas most disturbances tend to be only partially compensated (**Table 3**).

Table 3 Changes in blood and ECF during acid-base disturbance, the mechanism and degree of compensation

Problem	*$[H^+]$*	*$[HCO_3^-]$*	*P_{CO_2}*	*Compensation*
Metabolic				
Acidosis	↑	1°↓	2°↓	Partial respiratory
Alkalosis	↓	1°↑	2°↑	Partial respiratory
Respiratory				
Acidosis	↑	2°↑	1°↑	Complete renal
Alkalosis	↓	2°↓	1°↓	Complete renal

↑, Increase; ↓, decrease; 1°, primary; 2°, secondary.

Metabolic Acidosis

The main causes of metabolic acidosis are excessive acid production, inappropriate urinary loss of bicarbonate, or failure of the kidney to excrete fixed acid. Although the Henderson-Hasselbalch equation provides mathematical information concerning the equilibrium of bicarbonate species, in practice it provides little information regarding the nature of the underlying cause of the acid-base disorder and the concept of 'anion gap' is useful in assessing the cause of metabolic acidosis. This is derived from the principle of electroneutrality and is calculated thus:

$$([NA^+] + [K^+]) - ([Cl^-] + [HCO_3^-]) \qquad [6]$$

The anion gap represents an artificial disparity between the concentrations of these cations and anions routinely measured in clinical practice, therefore signifying the concentration of unmeasured anions such as proteins (the most important in healthy subjects), sulfate, phosphate, and others. The normal anion gap is 10–18 mmol l^{-1} although recent calculations using more sensitive measurements estimate this to be 6–12 mmol l^{-1}. This concept has limitations but is useful for dividing metabolic acidoses into those characterized by an increased anion gap as a marker of excess generation of organic acids and those with a normal anion gap due to decreased excretion of acid or external losses of bicarbonate. There are exceptions to this rule, e.g., the acidosis of chronic renal failure, but, nonetheless, it remains a useful concept in clinical practice. Classification of the causes of metabolic acidoses according to the presence of an increased or normal anion gap is shown in **Table 4**.

Diabetic Ketoacidosis

The absence of pancreatic insulin secretion in insulin-dependent diabetes results in increased plasma glucose and reduced tissue uptake and utilization of glucose. In the place of glucose, there is increased utilization of nonesterified fatty acid (NEFA) as an alternative source of energy that is metabolized to acetyl coenzyme A (acetyl-CoA). Under normal circumstances this substance is further metabolized in the liver via the tricarboxylic acid (TCA) cycle to CO_2 and water. In diabetic crises this cycle cannot accommodate the excess acetyl-CoA that is, instead, converted to acetoacetic acid, which can be further reduced to β-hydroxybutyric acid or decarboxylated to acetone. These three metabolites are known as 'ketone bodies' and their accumulation results in metabolic acidosis. In diabetic ketoacidosis, the homeostatic compensation is to increase ventilation

Table 4 Causes of metabolic acidoses according to the presence of an increased or normal anion gap

Increased anion gap	*Normal anion gap*
Ketoacidosis	Decreased renal acid excretion
Diabetic	Distal renal tubular acidosis
Starvation	
Alcoholic	
Inborn enzyme defects of metabolism	
Lactic acidosis	Loss of alkali
	Diarrhea
	Ureterosigmoidostomy (urinary conduit)
Renal failure	Increased renal bicarbonate loss
	Proximal renal tubular acidosis
	Azetazolamide
	Renal tubular damage
Intoxication	Increased HCl production
Salicylates	Ammonium chloride ingestion
Methanol	Increased catabolism of lysine, arginine
Ethylene glycol	
Paraldehyde	

and CO_2 excretion, leading to the characteristic pattern of ventilation known as Kussmaul respiration.

Lactic Acidosis

Reduced tissue perfusion, or perfusion that is inadequate to meet the metabolic demands of the tissues (such as exercising muscle), results in an inadequate supply of oxygen and a change from oxidative metabolism (the end products of which are CO_2 and H_2O) to anaerobic metabolism. The end product of anaerobic glycolysis is lactic acid, which is normally metabolized (to CO_2 and H_2O) by the liver or used in the synthesis of glucose (gluconeogenesis). The normal plasma [lactate] is less than $1\,mmol\,l^{-1}$ but may increase 10-fold in extreme exercise. When the ability to metabolize lactate is exceeded, either by increased production, or reduced delivery to the liver (in, for example, circulatory shock) or in the presence of impaired liver function, accumulation results in metabolic acidosis. Thus, lactic acidosis may occur in a variety of conditions, including circulatory shock, severe diabetic ketoacidosis, as a consequence of drugs (for example, the oral hypoglycemic agent metformin that inhibits gluconeogenesis and lactate transport), chronic liver disease, and poisoning (including ethanol and methanol).

Excess Bicarbonate Loss

The secretion of acid into the stomach is neutralized by alkaline secretions in the intestine. It follows that excessive loss of pure intestinal secretions (for example, in the presence of an enteric fistula) may lead to acidosis. A more common circumstance is the presence of an ileal conduit where the ureters are implanted into an isolated loop of intestine, which is then externalized (a 'urinary conduit'). The delivery of urine rich in chloride to the isolated intestine leads to exchange of Cl^- for HCO_3^-, and thence to excessive loss of HCO_3^- in the conduit, resulting in metabolic acidosis.

There are also a group of conditions known as renal tubular acidosis (RTA). These are mostly inherited but may be acquired, for example, as a consequence of recurrent infection. There are two major forms – proximal and distal – reflecting the site of the tubular defect in the nephron. In distal tubular RTA (type I) H^+ secretion is impaired resulting in impaired H^+ excretion, whereas in proximal RTA (type I) HCO_3^- reabsorption is impaired (usually as part of multiple tubular abnormalities) leading to net loss of bicarbonate. Both cause acidosis, the features of which are low pH and hypokalemia as a result of increased distal tubular H^+/K^+ exchange. The precise causes of these conditions is not known but is likely to reflect genetic defects on individual transporter subtypes, for example, those of the Na^+/H^+ exchanger (**Figure 2**).

Renal Failure

In progressive renal failure, renal clearance of all substances is impaired, reflecting the progressive loss of individual nephron function. Reduced excretion of fixed acid leads to bicarbonate consumption in the extracellular fluids and to acidosis. Tubular recovery of HCO_3^- may also be impaired (see RTA), as may the production of tubular NH_4^+, and may be associated with overproduction of urea in the liver.

Drugs and Other Causes

Many drugs can cause metabolic acidosis, generally in overdose. A classic example is aspirin (acetylacetic acid). Lactic acidosis is also associated with oral hypoglycemic agents (specifically metformin, used in the treatment of noninsulin-dependent diabetes), paracetamol, alcohol, and ethylene glycol (antifreeze) poisoning.

Compensation

The body's response to metabolic acidosis is a compensatory increase in ventilation to excrete excessive CO_2, restoring the equilibrium in the Henderson-Hasselbalch equation (eqn [5]). This respiratory compensation is usually incomplete, resulting in pH values or H^+ concentrations at, or marginally outside, the limits of 'normal' (**Table 3**). Complete

compensation depends on renal excretion of excess H^+, or resolution of the underlying condition.

Treatment

Treatment of metabolic acidosis is essentially that of the underlying condition: correction of tissue hypoxia in lactic acidosis; correction of fluid depletion and insulin therapy in diabetic ketoacidosis; and dialysis in renal failure. Rapid correction of pH can be achieved by the administration of intravenous sodium bicarbonate if necessary; treatment of chronic metabolic acidosis (e.g., in chronic renal failure or RTA) may be achieved by the administration of oral sodium bicarbonate. In uremia the prescription of a low-protein diet will also reduce acid load.

Metabolic Alkalosis

Metabolic alkalosis may be caused either by the excessive loss of acid or intake of alkali. The latter may be iatrogenic or factitious, with the excessive intake of prescribed antacids (such as sodium bicarbonate for heartburn or peptic ulcer disease) – the 'milk-alkali' syndrome. The loss of acid-rich gastric secretions in severe vomiting, for example, in cases of gastric outlet obstruction (due to pyloric stenosis, or a consequence of peptic ulcer disease), also leads to alkalosis. Compensation is by reducing ventilation to promote retention of CO_2 and thus balance the Henderson-Hasselbalch equation. Treatment is of the underlying condition rather than by administration of acid.

Respiratory Acidosis

Impaired ventilation reduces CO_2 excretion, increases Pa_{CO_2}, and thus lowers pH. This may occur acutely or chronically. Causes of respiratory acidosis include factors that interfere with the neurological 'drive' for respiration (e.g., head injury, cardiac arrest, opiate and anesthetic drugs), diseases of the respiratory muscles (e.g., poliomyelitis, Guillain-Barré syndrome), or primary lung diseases (acute pulmonary edema or pneumonia, chronic bronchitis or emphysema). In acute conditions, pH may fall dramatically, whereas in chronic conditions, such as chronic lung disease, the pH is generally nearer normal. In chronic conditions complete compensation occurs in the kidney where elevated $PaCO_2$ levels are offset by the increased generation of bicarbonate and excretion of fixed acid by the kidney, to balance the Henderson-Hasselbalch equation.

Respiratory Alkalosis

Respiratory alkalosis occurs as a result of inappropriately increased ventilation and increased excretion of CO_2. This may occur as a transient response to pain or hysteria. Such stimuli tend to be short lived and can be offset by analgesia, sedation, or short-term re-breathing of expired air. Additional causes include the early phases of aspirin poisoning (where the respiratory centers are activated), hypoxia, stroke, and other conditions affecting the brainstem respiratory control centers. Most causes of respiratory alkalosis are short term and, although adaptive responses would be expected to require excretion of bicarbonate to balance the Henderson-Hasselbalch equation, resolution usually occurs by resolution of the underlying condition.

Transporter Mechanisms: Physiology and Pathophysiology

Developments in molecular biology have led to major improvements in our understanding of the physiology and pathophysiology of renal tubular function. It is now possible to subdivide the various types of renal tubular acidosis, for example, by the precise biochemical defect rather than simply the tubular location. Thus, distal (or type 1) RTA may be a consequence of impaired distal tubular H^+ excretion, either due to increased permeability to H^+ or to impaired secretion, the latter, in turn, being a consequence of a variety of defects that include carbonic anhydrase type 2 deficiency, mutations in anion transport protein AE1, or deficiency of collecting duct proton transport ATPase. Whilst specific knowledge of the molecular defect is not necessary to either diagnose or manage these disorders, it is likely that future classification of acid-base disorders will change to recognize the underlying defect.

See also: **Brain and Nervous System**. **Electrolytes**: Water–Electrolyte Balance.

Further Reading

Adrogue HJ and Madias NE (1998) Management of life-threatening acid-base disorders. Second of two parts. *New England Journal of Medicine* **338**(2): 107–111.

Adrogue HJ and Madias NE (1998) Management of life-threatening acid-base disorders. First of two parts. *New England Journal of Medicine* **338**(1): 26–34.

Corey HE (2003) Stewart and beyond: New models of acid-base balance. *Kidney International* **64**: 777–787.

Galla JH (2000) Metabolic alkalosis. *Journal of the American Society of Nephrology* **11**: 369–375.

Gluck SL (1998) Acid-base. *The Lancet* **352**: 474–479.
Gunnerson KJ and Kellum JA (2003) Acid-base and electrolyte analysis in critically ill patients: are we ready for the new millennium? *Current Opinion in Critical Care* **9**: 468–473.
Halperin ML and Goldstein MB (1999) *Fluid, Electrolyte and Acid-Base Physiology: A Problem-Based Approach*, 3rd edn. London: W.B. Saunders.
Kellum JA (2000) Determinants of blood pH in health and disease. *Critical Care* **4**(1): 6–14.
Palmer BF, Narins RG, and Yee J (2005) Clinical acid–base disorders. In: Davison AM, Cameron JS, Grünfeld J-P, Ponticelli C, Ritz E, Winearls CG, and van Ypersele C (eds.) *Oxford Textbook of Clinical Nephrology*, 3rd edn, vol. 1, Ch 2.6, pp. 321–346. Oxford: Oxford University Press.
Remer T (2000) Influence of diet on acid-base balance. *Seminars in Dialysis* **13**(4): 221–226.
Sirker AA, Rhodes A, Grounds RM, and Bennett ED (2002) Acid-base physiology: the 'traditional' and the 'modern' approaches. *Anaesthesia* **57**: 348–356.
Stewart PA (1983) Modern quatitative acid-base chemistry. *Canadian Journal of Physiology and Pharmacology* **61**: 1444–1461.
Williams AJ (1998) ABC of oxygen: Assessing and interpreting arterial blood gases and acid-base balance. *British Medical Journal* **317**: 1213–1216.

Water–Electrolyte Balance

S M Shirreffs and R J Maughan, Loughborough University, Loughborough, UK

Body Water and Electrolytes

Man is dependent on ready access to water for survival. Water is the largest component of the human body and the total body water content varies from approximately 45 to 70% of the total body mass; this therefore corresponds to about 33–53 l for a 75 kg man. The water content of the various tissues is maintained relatively constant; as adipose tissue has a low water content (**Table 1**) the fraction of water in the body is determined largely by the fat content. The body's water can be divided into two components—the intracellular fluid and the extracellular fluid. The intracellular fluid is the major component and comprises approximately two-thirds of total body water. The extracellular fluid can be further divided into the interstitial fluid (that between the cells) and the plasma; the plasma volume represents approximately one-quarter of the extracellular fluid volume (**Table 2**).

Numerous electrolytes and solutes are dissolved within the body water compartments: an electrolyte can be defined as a compound which dissociates into ions when in solution. The major cations (positively charged electrolytes) in the body water are sodium and potassium, with smaller amounts of calcium and magnesium; the major anion (negatively charged electrolytes) is chloride, with smaller amounts of bicarbonate and protein. Sodium is the major electrolyte present in the extracellular fluid, while potassium is present in a much lower concentration (**Table 3**). Within the intracellular fluid the situation is reversed, and the major electrolyte present is potassium, while sodium is found in much lower concentrations. Maintenance of the transmembrane electrical and chemical gradients is of

Table 1 Water content of various body tissues for an average 75 kg man

Tissue	*% water*	*% of body mass*	*Water per 75 kg (l)*	*% of total body water*
Skin	72	18	9.7	22
Organs	76	7	4.0	9
Skeleton	22	15	2.5	5
Blood	83	5	3.1	7
Adipose	10	12	0.9	2
Muscle	76	43	24.5	55

From Sawka (1990).

Table 2 Body water distribution between the body fluid compartments in an adult male

	% of body mass	*% of lean body mass*	*% of body water*
Total body water	60	72	100
Extracellular water	20	24	33
Plasma	5	6	8
Interstitium	15	18	25
Intracellular water	40	48	67

From Sawka (1990).

Table 3 Ionic concentrations ($mmol\,l^{-1}$) of body water compartments[a].

Ion	*Plasma*	*Intracellular fluid*	*Sweat*
Sodium	140 (135–145)	12	20–80
Potassium	4.0 (3.5–4.6)	150	4–8
Calcium	2.4 (2.1–2.7)	4.0	0–1
Magnesium	0.8 (0.6–1.0)	34	<0.2
Chloride	104 (98–107)	4	20–60
Bicarbonate	29 (21–38)	12	0–35
Inorganic phosphate	1.0 (0.7–1.6)	40	0.1–0.2

[a]The normal ranges of the plasma electrolyte concentrations are shown.

paramount importance for maintaining the integrity of the body's cells and allowing electrical communication throughout the body.

Daily Regulation of Body Water

The body's total body water content is normally maintained within a small window of fluctuation on a daily basis by intake of food and drink to balance the excretion of urine and other losses. Hyperhydration is corrected by an increase in urine production and hypohydration by an increase in water intake via food or drink consumption initiated by thirst. Most of our water intake is related to habit rather than thirst, but the thirst mechanism is effective at driving intake after periods of deprivation. There are also water losses via the respiratory tract, the gastrointestinal tract, and the skin. The extent of these losses will vary from individual to individual and will be strongly influenced by environmental conditions and physical activity levels, but for a sedentary individual in a cool environment these generally represent only a small proportion of the total body water loss.

All the major textbooks of nutrition and physiology include data on the various components of water intake and output, although it is difficult to find the original data on which the various mean values and ranges are based. The Geigy Scientific Tables suggest that the minimum daily water intake for adults is on the order of 1.5 l, but others indicate that the minimum intake should be 2 l per day. Body size has a major influence on water turnover, but the total body water content will also be markedly affected by the body composition. Water turnover should therefore be more closely related to lean body mass than to body mass itself. It is expected, therefore, that there will be differences between men and women and between adults and children.

Environmental conditions will affect the basal water requirement by altering the losses that occur by the various routes (i.e., respiration, sweat, and urine). Water requirements for sedentary individuals living in the heat may be two or threefold higher than the requirement when living in a temperate climate, even when not accompanied by pronounced sweating. Transcutaneous and respiratory losses will be markedly influenced by the humidity of the ambient air, and this may be a more important factor than the ambient temperature. Respiratory water losses are incurred because of the humidification of the inspired air with fluid from the lungs. These losses are relatively small in the resting individual in a warm, moist environment (amounting to about 200 ml per day) but will be increased approximately 2-fold in regions of low humidity, and may be as high as 1500 ml per day during periods of hard work in the cold, dry air at altitude. To these losses must be added insensible water loss through the skin (about 600 ml per day) and urine loss, which will not usually be less than about 800 ml per day.

Variations in the amount and type of food eaten have some effect on water requirements because of the resulting demand for excretion of excess electrolytes and the nonvolatile products of metabolism. An intake of electrolytes in excess of the amounts lost (primarily in sweat and feces) must be corrected by excretion in the urine, with a corresponding increase in the volume and osmolality of urine formed. The daily intake of electrolytes is subject to wide variation among individuals, with strong trends for differences among different geographical regions. Daily dietary sodium chloride intakes for 95% of the young male UK population fall between 3.8 and 14.3 g, with a mean of 8.4 g; the corresponding values for young women are 2.8–9.4 g, with a mean value of 6.0 g. For the same population, mean urinary sodium losses were reported to account for about 175 mmol per day, which is equivalent to about 10.2 g of sodium chloride.

There are also large differences among countries in the recommended intake of salt. The British health authorities advise a maximum of 6 g per day, but in Germany a maximum of 10 g per day is recommended. In contrast, Sweden recommends a maximum of 2 g per day, and Poland recommends a minimum of 1.4 g per day. The differences among countries reflect in part different interpretations with regard to the evidence linking salt intake and health, but also reflect regional consumption patterns dictated by food choices.

A high-protein diet requires a greater urine output to allow for excretion of water-soluble nitrogenous waste; this effect is relatively small compared with other routes of water loss but becomes meaningful when water availability is limited. The water content of the food ingested will also be influenced greatly by the nature of the diet, and water associated with food may make a major contribution to the total fluid intake in some individuals. Some water is also obtained from the oxidation of nutrients, and the total amount of water produced will depend on the total metabolic rate and is also influenced by the substrate being oxidized. An energy expenditure of 3000 kcal (12.6 MJ) per day, based upon a diet composed of 50% carbohydrate, 35% fat, and 15% protein, will yield about 400 ml of water per day. Reducing the daily energy

expenditure to 2000 kcal (8.4 MJ), but keeping the same diet composition, will yield about 275 ml of water. The contribution of this water-of-oxidation to total water requirements is thus appreciable when water turnover is low but becomes rather insignificant when water losses are high.

Thirst and the Control of Intake

In man, daily fluid intake in the form of food and drink (plus that formed from substrate oxidation) is usually in excess of the obligatory water loss (transcutaneous, pulmonary, and renal output), with renal excretion being the main mechanism regulating body water content. However, conservation of water or electrolytes by the kidneys can only reduce the rate of loss; it cannot restore a deficit. The sensation of thirst, which underpins drinking behavior, indicates the need to drink and hence is critical in the control of fluid intake and water balance. While thirst appears to be a poor indicator of acute hydration status in man, the overall stability of the total water volume of an individual indicates that the desire to drink is a powerful regulatory factor over the long term.

The act of drinking may not be a direct consequence of a physiological need for water intake but can be initiated by habit, ritual, taste, or desire for nutrients, stimulants, or a warm or cooling effect. A number of the sensations associated with thirst are learned, with signals such as dryness of the mouth or throat inducing drinking, while distention of the stomach can stop ingestion before a fluid deficit has been restored. However, the underlying regulation of thirst is controlled separately by the osmotic pressure and volume of the body fluids and as such is governed by the same mechanisms that affect water and solute reabsorption in the kidneys and control central blood pressure.

Regulatory Mechanisms

Areas of the hypothalamus and forebrain, that are collectively termed the thirst control centers, appear to be central to the regulation of both thirst and diuresis. Receptors in the thirst control centers respond directly to changes in osmolality, volume, and blood pressure, while others are stimulated by the fluid-balance hormones that also regulate renal excretion. These regions of the brain also receive afferent input from systemic receptors monitoring osmolality, circulating sodium concentration, and alterations in blood volume and pressure. Changes in the balance of neural activity in the thirst control centers regulated by the different monitoring inputs determine the relative sensations of thirst and satiety, and influence the degree of diuresis. Input from the higher centers of the brain, however, can override the basic biological need for water to some extent and cause inappropriate drinking responses. Cases of water intoxication (hyponatremia) during endurance sports events lasting more than about 6–8 h have been reported in which the major cause of the illness is due to overhydration as a result of overdrinking.

A rise of between 2 and 3% in circulating osmolality (i.e., about 6–8 mosm kg^{-1} H_2O) is sufficient to evoke a profound sensation of thirst coupled with an increase in the circulating concentration of antidiuretic hormone, also known as vasopressin. The mechanisms that respond to changes in intravascular volume and pressure appear to be less sensitive than those that monitor plasma osmolality, with hypovolemic thirst being evident only following a 10% decrease in blood volume. As fairly large variations in blood volume and pressure occur during normal daily activity, primarily in response to postural changes, this lack of sensitivity presumably prevents overactivity of the volume-control mechanisms. Prolonged exercise, especially in the heat, is associated with a decrease in plasma volume and a tendency for an increase in osmolality, but fluid intake during and immediately following exercise is often less than that required to restore normal hydration status. This appears to be due to a premature termination of the drinking response rather than to a lack of initiation of that response. Also, the composition of the beverage consumed has an effect on the volume of fluid ingested, with water prematurely abolishing the osmotic drive to drink, while sodium-containing drinks help maintain the osmotic drive to drink and increase voluntary intake.

When a water deficit is present and free access to fluid is allowed, the drinking response in man usually consists of a period of rapid ingestion, during which more than 50% of the total intake is consumed, followed by intermittent consumption of relatively small volumes of drink over a longer period. The initial alleviation of thirst occurs before significant amounts of the beverage have been absorbed and entered the body water. Therefore, although decreasing osmolality and increasing extracellular volume promote a reduction in the perception of thirst, other preabsorptive factors also affect the volume of fluid ingested. Receptors in the mouth, oesophagus, and stomach are thought to meter the volume of fluid ingested, while distension of the stomach tends to reduce the perception of thirst. These preabsorptive signals appear to be

behavioural, learned responses and may be subject to disruption in situations which are novel to the individual. This may partly explain the inappropriate voluntary fluid intake in individuals exposed to an acute increase in environmental temperature or to exercise-induced dehydration.

Renal Function

As well as acting to regulate body water levels by an increase or decrease in the amount of urine produced, the kidneys are also responsible for the elimination of waste products from the body. This, therefore, also affects the daily urine volume. For example, a healthy individual eating a normal diet excretes approximately 600–800 mosmol of solute per square metre of body surface area per day, amounting to a total of about 1000–1500 mosmol day^{-1}. The kidneys can dilute urine to at least as low as 100 mosmol kg^{-1} and can concentrate it to 1200 mosmol kg^{-1}. Therefore, the daily solute load to be excreted can be accommodated in a volume ranging between approximately 500 ml and more than 13 l. To allow for waste product excretion, an obligatory minimum amount of urine must always be excreted, and this is generally in the region of 20–50 ml per hour. However, in the majority of healthy individuals in most situations, the volume of urine produced and excreted is in excess of these basal levels.

Hormonal Control of Urine Production

The volume of urine produced in a healthy individual is largely determined by circulating hormone levels, and in particular by levels of vasopressin. Vasopressin is a cyclic, nine-amino acid peptide. It is released from the posterior pituitary after having been transported there along the axons of neurons whose cell bodies are located in the paraventricular and supraoptic nuclei of the hypothalamus, the site of vasopressin synthesis. An increase in the rate of secretion of vasopressin results in a reduced urine production. Vasopressin acts on the renal distal tubules and collecting ducts to cause an increased permeability to water and hence an increased reabsorption of water from the filtrate. Therefore, a hyperosmotic urine can be formed and the solute load to be excreted can be accommodated in a small volume of water. A decrease in vasopressin secretion results in an increase in the volume of urine produced by causing a reduction in the permeability of the renal distal tubule and collecting ducts to water. Vasopressin secretion is largely influenced by changes in plasma osmolality. An increase in plasma osmolality results in a increased vasopressin secretion and vice versa. The vasopressin is released rapidly in response to the stimuli and begins to act within minutes. When the secretion is inhibited, the half-life of clearance from the circulation is approximately 10 minutes. Therefore, changes in body fluid tonicity are rapidly translated into changes in water excretion by this tightly regulated feedback system.

In addition to the influence of plasma osmolality on vasopressin secretion, other (nonosmotic) factors with an influence are baroregulation, nausea, and pharyngeal stimuli. A fall in blood pressure or blood volume will stimulate vasopressin release; vasopressin secretion is, however, less sensitive to these changes than to changes in osmolality. Nausea is an extremely potent stimulus to vasopressin secretion in man; vasopressin levels can increase 100- to 1000-fold in response to nausea induced by various chemical agents. After a period of water deprivation followed by access to drink, vasopressin levels fall before there is any change in plasma tonicity, suggesting activation of neuronal pathways from the oropharynx.

Aldosterone, a steroid hormone, is released into the circulation after synthesis by the zona glomerulosa cells of the adrenal cortex. Its primary role, in terms of renal function, is to increase renal tubular reabsorption of sodium and in doing so will bring about an increased excretion of potassium and, in association with vasopressin, increase water reabsorption in the distal segments of the nephron. Aldosterone causes this response by increasing the activity of the peritubular sodium/potassium pump and by increasing the permeability of the luminal membrane to both sodium and potassium. The increased luminal permeability allows potassium to move down its concentration gradient from the inside of the membrane cells into the tubule lumen. The majority of the sodium present is reabsorbed into the cell down the concentration gradient. The sodium absorption and potassium excretion are closely correlated with a 3 sodium:2 potassium ratio. Chloride follows the sodium to maintain the electrical neutrality of the urine.

The release of aldosterone is determined by a number of factors including the renin–angiotensin system: A fall in blood or extracellular fluid volume increases renin production and, via angiotensin II, results in an increase in aldosterone secretion.

The presence in the renal filtrate of ions such as bicarbonate and sulphate, which are not reabsorbed, promotes secretion of potassium into the distal tubule of the nephron and will also result in an increased urinary loss of potassium.

Sweat

Exercise, particularly in a warm environment, and diarrheal illness are two situations which will increase the requirement for salt to substantially greater levels. Sweating, therefore, is an important consideration in the area of water and electrolyte balance as this is the one route where there can be extensive losses of water and electrolytes from the body in a healthy individual. If these losses are not replaced, serious consequences can ensue.

Eccrine sweat is a clear, watery, odourless substance whose primary function is to promote heat loss by evaporation from the skin surface. When sweat is produced, the daily water losses increase and the intake must be increased or urine production decreased accordingly if euhydration is to be maintained.

Sweat Evaporation

There is a daily loss on the order of approximately 500 ml of water through the skin. However, when the body is exposed to a heat stress and behavioral and vasomotor mechanisms are insufficient to prevent a rise in temperature, the physiological responses generally include an increase in sweat production in an attempt to prevent hyperthermia; the high latent heat of vaporization of water makes the evaporation of sweat an effective heat loss mechanism (evaporation of 1 l of water from the skin surface will remove 2.4 MJ (580 kcal) of heat from the body). The heat stress may be of external origin (i.e., from the environmental conditions), of internal origin due to muscular work or fever, or from a combination of these factors.

In many individuals sweat rates can be in excess of 2 l per hour, especially in situations of exercise undertaken in a warm, humid environment, and these high sweat rates can be maintained for a number of hours. For example, body mass losses in marathon runners have been reported to range from about 1–6% (0.7–4.2 kg) at low (10 °C) ambient temperatures to more than 8% (5.6 kg) in warmer conditions. However, when sweat rates are high, a significant fraction of the sweat secreted onto the skin may drip from the body and is therefore ineffective at removing heat.

Mechanism of Sweat Secretion

The human body has approximately 2 million sweat glands. The eccrine sweat gland consists of a single tubule, opening onto the epidermis at one end and closed at the other. The proximal half of the tubule is the secretory coil and the distal half the reabsorptive duct. The sweat secreted onto the skin is the original tubular secretion minus the substances which are, further up the tubule, reabsorbed; from the isoosmotic fluid secreted by the coil most of the major electrolytes (Na^+, Cl^-, HCO_3^+, and lactate) are transported out of the duct back into the extracellular fluid in excess of water. The final sweat secreted onto the skin is therefore hypotonic with respect to body fluids.

Sweat Composition

The composition of human sweat is highly variable, both between individuals and within an individual over time. However, sodium and chloride are the major electrolytes lost in sweat, with other ions being present in smaller amounts relative to the whole body status. The sweat electrolyte composition of an individual seems to be related primarily to sweat rate but can be influenced by training status, extent of heat acclimation, and diet. However, the range of values for sweat electrolyte composition reported in the literature probably reflects not only the interindividual differences but also differences in the methodology used for collection of the sweat. This last factor may be the result of errors caused by contamination or incomplete collection of the sample, or it may reflect a real difference induced by the collection procedure.

Due to the secretion and reabsorption process involved in sweat production within the sweat gland and duct, sweat composition is influenced by sweat rate, at least within single ducts, such that a reduction in rate allows for greater reabsorption of certain electrolytes (Na^+, Cl^-, but not K^+) from the duct resulting in a lower concentration in the final sweat produced. There also appear to be regional variations in sweat composition, as evidenced by the different values obtained when the composition of sweat obtained from different parts of the body is compared, and the values obtained by regional collection also differ from those obtained by the whole body washdown technique. Reported values for sweat electrolyte composition are summarized in **Table 3**.

Restitution of Water and Electrolyte Balance

When substantial sweat losses have been incurred, restitution of water balance requires both volume repletion and replacement of electrolyte losses. This can be achieved by ingestion of electrolyte-containing drinks or by ingestion of water and consumption of electrolyte-containing foods. Problems may occur

when large sweat losses are replaced with electrolyte-free drinks. Hyponatremia and blood volume expansion ensue and will promote a diuresis that will prevent effective recovery.

Conclusions

In healthy individuals, water is the largest single component of the body. Although water balance is regulated around a range of volumes rather than a finite set point, its homeostasis is critical for virtually all physiological functions. To further ensure proper regulation of physiological and metabolic functions, the composition of the individual body water compartments must also be regulated.

Humans continually lose water through the renal system, gastrointestinal system, skin, and respiratory tract, and this water must be replaced. Thirst is implicated in our water intake, but behavioral habits also have an important influence on drinking.

When exercise is undertaken or when an individual is exposed to a warm environment, the additional heat load is lost largely due to sweating and this can increase greatly the individual's daily water loss and therefore the amount that must be consumed. Sweat rates on the order of 2 to 3 l per hour can be reached and maintained by some individuals for a number of hours and it is not impossible for total losses to be as much as 10 l in a day. These losses must of course be replaced and when they are so extreme, the majority must be met from fluid consumption rather than food ingestion. A variety of drink types and flavors are likely to be favored by individuals who have extreme losses to replace. Sports drinks have an importance role in this recovery when no food is ingested by their contribution to sweat electrolyte loss replacement which is crucial for retention of the ingested water.

See also: **Calcium**. **Exercise**: Beneficial Effects. **Magnesium**. **Potassium**. **Sodium**: Physiology.

Further Reading

Briggs JP, Sawaya BE, and Schnermann J (1990) Disorders of salt balance. In: Kokko JP and Tannen RL (eds.) *Fluids and Electrolytes*, 2nd edn., pp. 70–138. Philadelphia: WB Saunders.

Costill DL, Coté R, and Fink W (1976) Muscle water and electrolytes following varied levels of dehydration in man. *Journal of Applied Physiology* **40**: 6–11.

Engell DB, Maller O, Sawka MN *et al.* (1987) Thirst and fluid intake following graded hypohydration levels in humans. *Physiology and Behaviour* **40**: 229–236.

Grandjean AC, Reimers KJ, Bannick KE, and Haven MC (2000) The effect of caffeinated, non-caffeinated, caloric and non-caloric beverages on hydration. *Journal of the American College of Nutrition* **19**: 591–600.

Kirby CR and Convertino VA (1986) Plasma aldosterone and sweat sodium concentrations after exercise and heat acclimation. *Journal of Applied Physiology* **61**: 967–970.

Lentner C (ed.) (1981) *Geigy Scientific Tables*, 8th edn. Basel: Ciba-Geigy Limited.

Maughan RJ (2001) Water, hydration status and human wellbeing. In: Berk Z *et al.* (ed.) *Water Science for Food, Health, Agriculture and Environment,* ISOPOW 8, pp. 43–57. Lancaster: Technomic.

Maughan RJ and Murray R (eds.) (2000) *Sports Drinks: Basic Science and Practical Aspects*. Boca Raton, FL: CRC Press.

Rose BD (1984) *Clinical Physiology of Acid–Base and Electrolyte Disorders*, 2nd edn. New York: McGraw-Hill.

Sawka MN (1990) Body fluid responses and hypohydration during exercise-heat stress. In: Pandolf KB, Sawka MN, and Gonzalez RR (eds.) *Human Performance Physiology and Environmental Medicine at Terrestrial Extremes*, pp. 227–266. Carmel: Cooper.

Shirreffs SM and Maughan RJ (1997) Whole body sweat collection in man: An improved method with some preliminary data on electrolyte composition. *Journal of Applied Physiology* **82**: 336–341.

Sterns RH and Spital A (1990) Disorders of water balance. In: Kokko JP and Tannen RL (eds.) *Fluids and Electrolytes*, 2nd edn., pp. 139–194. Philadelphia: WB Saunders.

Taylor NAS (1986) Eccrine sweat glands. Adaptations to physical training and heat acclimation. *Sports Medicine* **3**: 387–397.

Valtin H (2002) "Drink at least eight glasses of water a day." Really? Is there scientific evidence for "8 × 8"? *American Journal of Physiology* **283**: R993–R1004.

ENERGY

Contents
Metabolism
Balance
Requirements
Adaptation

Metabolism

S Cox, London School of Hygiene and Tropical Medicine, London, UK

The energy required for the growth and maintenance of living organisms is acquired through the cellular respiration of organic compounds. In man and other nonphotosynthetic organisms, these metabolic fuels are obtained from food and consist of carbohydrates, fats, and proteins.

Metabolism

Metabolism is defined as the sum of anabolic or synthetic chemical reactions that require energy and the catabolic chemical reactions that break down large organic molecules into smaller molecules, thereby releasing energy for anabolic reactions.

Cellular Respiration and Adenosine Triphosphate

Cellular respiration can be defined generally as the process by which chemical energy is released during the oxidation of organic molecules. If it requires oxygen, it is called aerobic respiration, whereas if it takes place in the absence of oxygen it is anaerobic respiration.

Organic molecules, usually carbohydrate or fat, are broken down by a series of enzyme-catalyzed reactions. Many of these reactions release a small amount of energy that is channeled into molecules of a chemical nucleotide called adenosine triphosphate (ATP) (**Figure 1**).

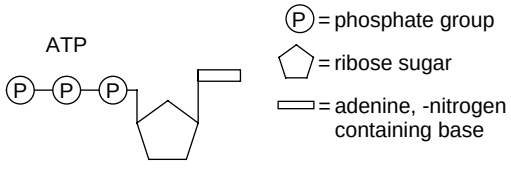


Figure 1 Structure of adenosine triphosphate (ATP).

ATP is the standard unit in which the energy released during respiration is stored. ATP is an instant source of energy within the cell. It is mobile and transports energy to wherever energy-consuming processes are occurring within the cell. The energy is released by the dephosphorylation of ATP to ADP, which can then be rephosphorylated to ATP by being coupled to the processes of respiration. ATP is found in all living cells and can be thought of as a universal energy transducer.

The principal metabolic fuel is glucose, and there are three stages in its oxidation to carbon dioxide, water, and energy, captured as ATP. This process can be summarized very simply by the following equation:

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O + \text{energy (ATP)}$$

In the first stages, glucose and other metabolic fuels are oxidized, linked to the chemical reduction of coenzymes (nicotinamide adenine dinucelotide (NAD^+), flavin adenine dinucleotide (FAD), and flavin mononucleotide (FMN)). In the final stage, ATP is synthesized from ADP and phosphate via a common pathway using energy released from the oxidation and recycling of the reduced coenzymes (**Table 1**). Thus, the oxidation of metabolic fuels is tightly coupled to energy consumption and the production of ADP from ATP in energy-consuming processes (**Figure 2**).

Glycolysis

The main substrate for glycolysis is glucose. Glycolysis does not require oxygen and is important for the direct production of ATP when oxygen is limiting (i.e., in rapidly contracting muscle). Glycolysis results in the splitting of glucose, a six-carbon (6C) compound, into two molecules of pyruvic acid (3C), which in the cytoplasmic solution becomes pyruvate (**Figure 3**). Pyruvate can enter the mitochondrion and be metabolized by oxidative decarboxylation to CO_2, or if oxygen is unavailable it can be further metabolized to lactic acid resulting in the

Table 1 The three principal stages in the production of ATP from one molecule of glucose

Metabolic pathway	*Location*	*O_2 required?*	*Net ATP or reduced coenzymes/glucose*	*Products*
Glycolysis	Cytoplasm	Anaerobic	Net gain 2 ATP 2 $NADH + H^+$	Glucose → 2 pyruvate
Pyruvate → acetyl-CoA	Mitochondrial matrix	Aerobic	2 $NADH + H^+$	
TCA cycle	Mitochondrial matrix	Aerobic	2 GTP → 2 ATP 8 $NADH + H^+$ 2 $FADH_2$	2 Pyruvate → $6CO_2$
Electron transfer chain (oxidative phosphorylation)	Mitochondrial crista and primary particles	Aerobic	12 $NAD^+ + 2$ FAD → 38 ATP[a]	$12H_2 + 6O_2 \rightarrow 6H_2O$

[a]The exact net gain in the number of ATP produced from the oxidation of the reduced coenzymes $NADH + H^+$ and $FADH_2$ varies depending on the mechanism used to transport them across the crista membrane in the mitochondria, the site of oxidative phosphorylation.

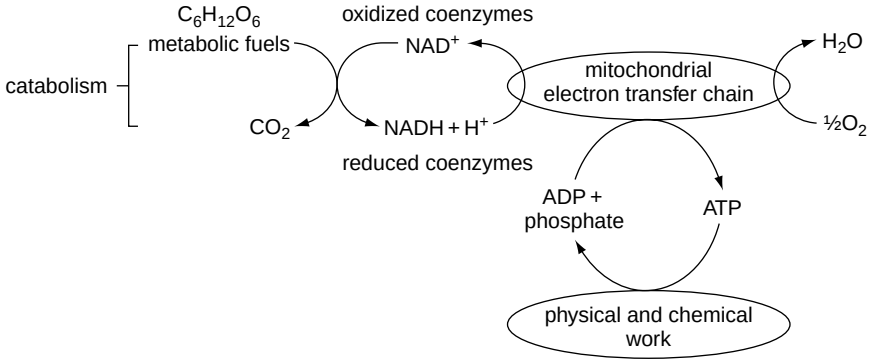


Figure 2 Linkage between ATP utilization in physical and chemical work and the oxidation of metabolic fuels. (Adapted with permission from Bender DA (2002) *Introduction to Nutrition and Metabolism*, 3rd edn. London: Taylor & Francis.)

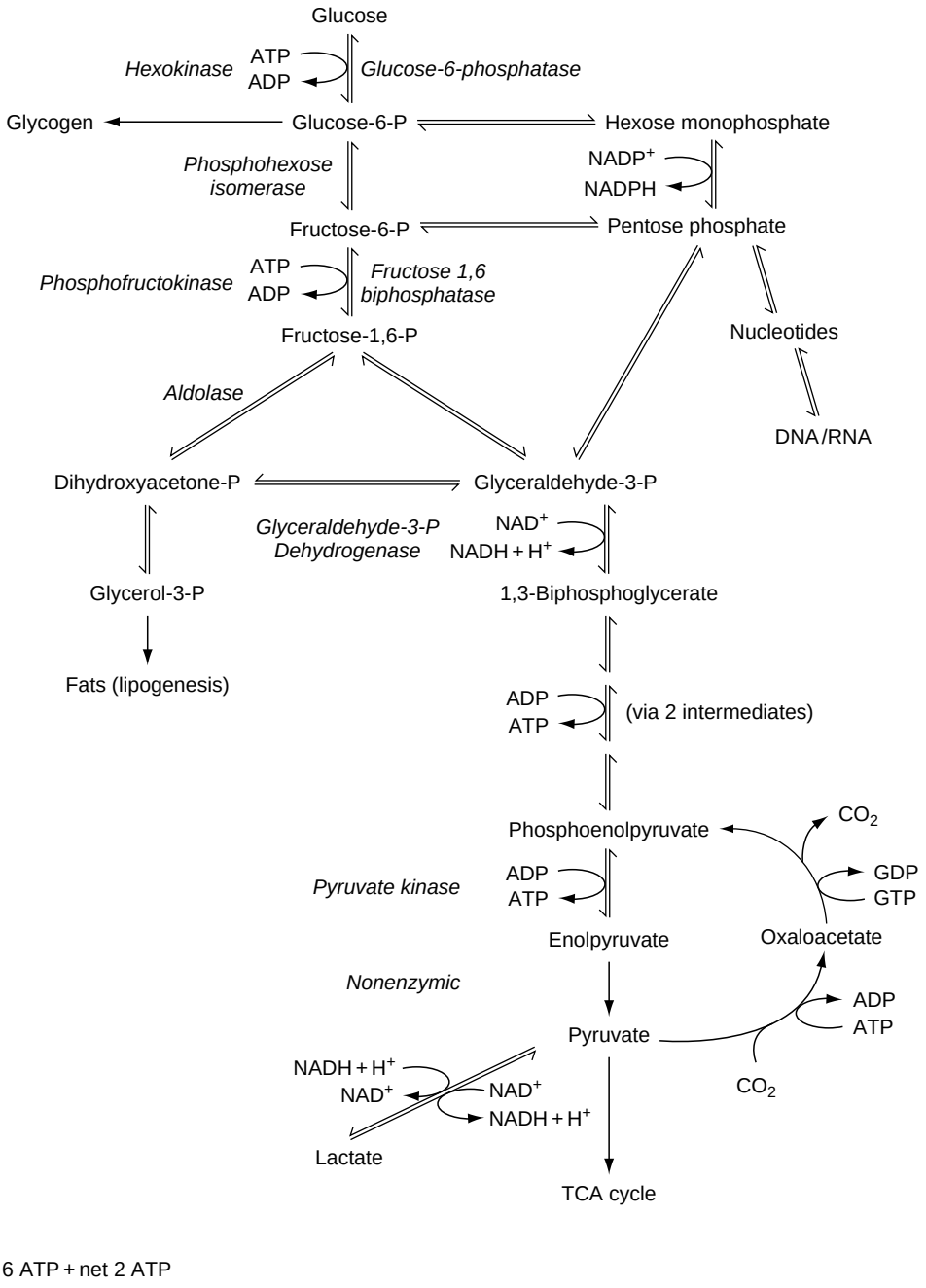
regeneration of NAD^+ from $NADH + H^+$, thus allowing glycolysis to continue in the absence of oxygen. Red blood cells lack mitochondria and therefore glycolysis is the only source of energy metabolism. Thus, red cells can only metabolize glucose or other simple sugars and not fats or proteins. Red cells produce lactate that is excreted into the blood. Lactate is primarily metabolized back to pyruvate in the liver, where it is mostly used for the synthesis of glucose (gluconeogenesis), which is essentially the reverse of glycolysis except for the irreversible reaction of phosphoenolpyruvate (PEP) to pyruvate. Hence, in the liver pyruvate is converted back to PEP via oxaloacetate (**Figure 3**). This cycling of lactate and pyruvate is known as the Cori cycle.

Other sugars, such as fructose and galactose, can be fed into glycolysis at different points and then metabolized in the same way as glucose to pyruvic acid.

The pentose–phosphate shunt (sometimes also known as the hexose–monophosphate pathway) occurs when glucose-6-phosphate is metabolized via an alternative route to glycolysis to generate pentose phosphates to be used as components in DNA and RNA nucleotides. Alternatively, pentose phosphates can be returned into the glycolytic pathway by conversion back to fructose-6-phosphate or glyceraldehyde-3-phosphate. Another purpose of this shunt is the production of $NADPH + H^+$ from $NADP^+$, which is the required coenzyme for fat synthesis (lipogenesis). Reduced $NADP^+$ is also required for the reduction and recycling of oxidized glutathione, an important intermediate in antioxidant defence and in the generation of the respiratory burst, used to kill parasites ingested by macrophages, a type of white blood cell.

Tricarboxylic Acid Cycle

The tricarboxylic acid (TCA) cycle (also known as the citric acid cycle) is located in the mitochondrial matrix and is a common metabolic pathway for all fuels. It is responsible for the production of the majority of the reduced coenzymes used for the generation of ATP in the electron transfer chain. It also plays a central role in the interconversion of



net = 2 NADH = 6 ATP + net 2 ATP

Figure 3 Glycolysis and its interactions with other metabolic pathways.

fuels and metabolites. The TCA cycle participates in gluconeogenesis from amino acids and lactate during fasting between meals and longer term in starvation. TCA cycle intermediates are the source of most of the nonessential amino acids, such as aspartate and glutamate. It is also involved in the conversion of carbohydrates to fat for storage after a carbohydrate-rich meal.

Pyruvate (3C) from glycolysis is oxidatively decarboxylated to acetyl-CoA (2C) in the mitochondria, catalyzed by the multienzyme complex pyruvate dehydrogenase and the coenzyme A (Co-ASH):

$$\text{Pyruvate} + \text{Co-ASH} + \text{NAD}^+ \rightarrow \text{acetyl-CoA} + \text{Co}_2 + \text{NADH} + \text{H}^+$$

Pyruvate dehydrogenase requires several coenzymes derived from vitamins, including vitamin B_1 or thiamine, niacin (NAD), riboflavin (FAD), and pantothenic acid (a component of CoA). Deficiencies in any of these vitamins can affect energy metabolism, as

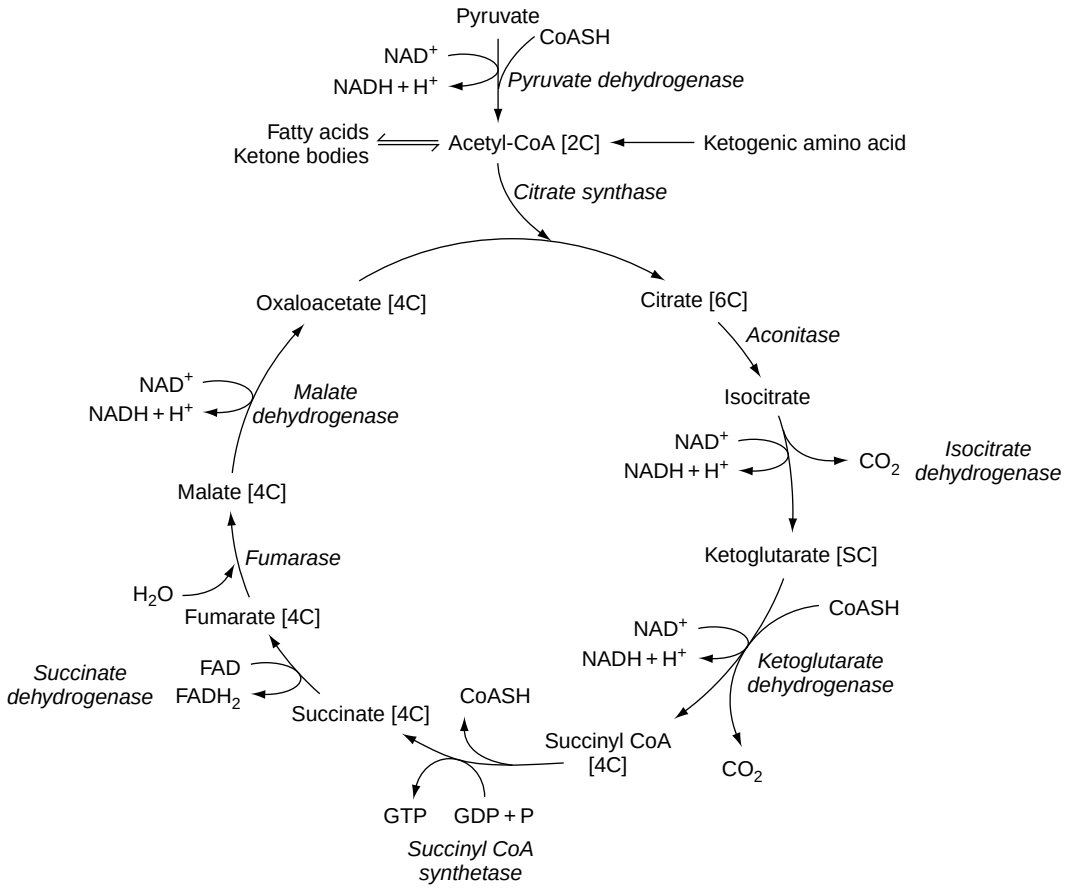


Figure 4 The oxidative decarboxylation of pyruvate and the tricarboxylic acid cycle.

evidenced by the increased cellular pyruvate and cardiac and skeletal muscle weakness in beriberi caused by thiamine deficiency. Pyruvate dehydrogenase catalyzes a central reaction in carbohydrate metabolism, and therefore its activity is regulated by both allosteric and covalent mechanisms.

Acetyl-CoA can be produced from pyruvate but also from fatty acids released from fat stores and from amino acids released from proteolysis of protein tissue, which can be converted to acetyl-CoA or cycle intermediates. In the first of the eight enzymatic reactions, acetyl-CoA (2C) combines with oxaloacetate (4C), forming citrate (6C) and releasing the CoA for further reactions with pyruvate to acetyl-CoA. A cycle of reactions follows in which two molecules of CO_2 are released and three molecules of $NADH + H^+$ and one of $FADH_2$ are produced along with one molecule of GTP (equivalent to ATP). At the end of the cycle, oxaloacetate is regenerated and able to react again with another molecule of acetyl-CoA, and so the cycle continues (**Figure 4**).

In the electron transfer chain, each $NADH + H^+$ yields approximately 3 ATP and $FADH_2$ yields 2 ATP. Thus, each rotation of the TCA cycle produces approximately 12 ATP ($3\ NADH + H^+ \approx 9\ ATP + 1\ FADH_2 \approx 2\ ATP + 1\ GTP$). Because two molecules of acetyl-CoA are formed from one glucose molecule, the TCA cycle rotates twice for each molecule of glucose respired, producing a net of 24 ATP (**Table 1**).

Electron Transfer Chain (Oxidative Phosphorylation)

Oxidative phosphorylation occurs in the crista of mitochondria, formed by invaginations of the inner mitochondrial membrane (**Figure 5**). The hydrogen accepted by NAD^+ and FAD during glycolysis and the TCA cycle is oxidized to water by molecular oxygen with accompanying phosphorylation of $ADP \rightarrow ATP$. This is achieved by phosphorylation of ADP coupled with a series of redox reactions whereby the hydrogen ions (H^+) and their electrons (e^-) are passed along a chain of intermediate carriers in the crista membrane of the mitochondria

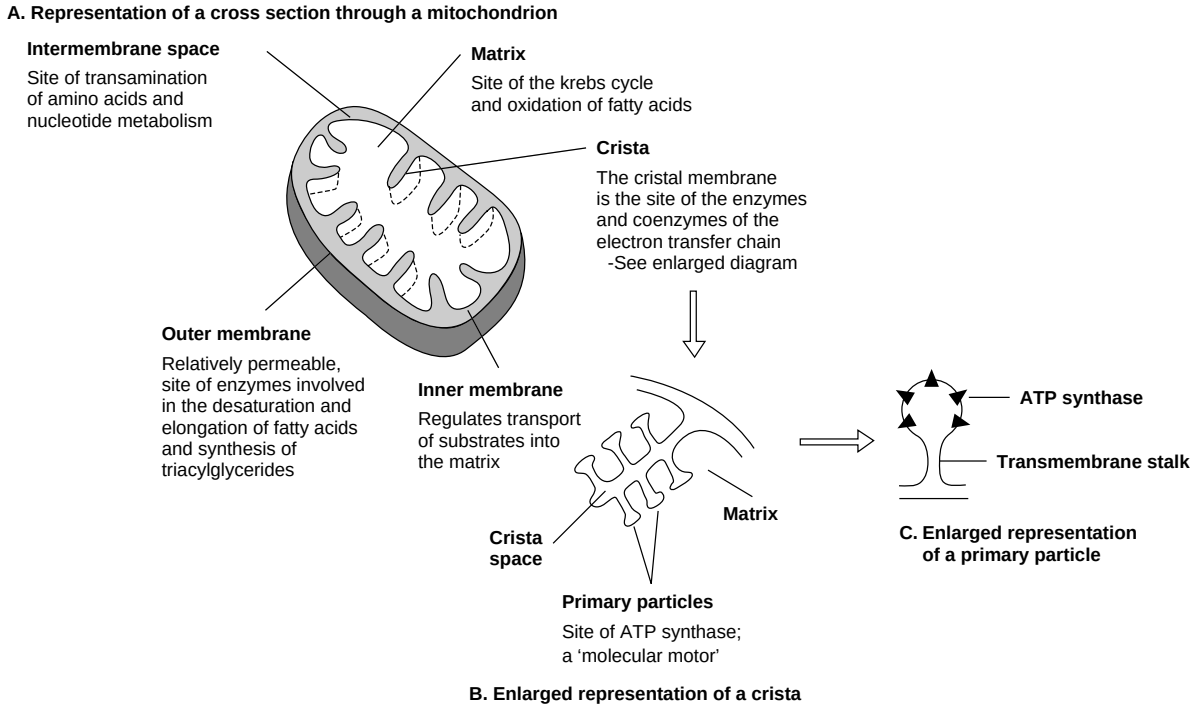


Figure 5 Structure and related functions of a mitochondrion.

(**Figure 6**), with each intermediate being reduced by the proceeding one and in turn reducing the next one and hence itself being oxidized. The chain consists of a flavoprotein and ubiquinone (coenzyme Q), both hydrogen carriers, followed by a series of cytochromes that are carriers of electrons only. Finally, at the end of the chain is cytochrome oxidase, which catalyzes the formation of water from hydrogen ions, electrons, and molecular oxygen. Unlike the other cytochromes, cytochrome oxidase contains copper (Cu^{2+}) in a prosthetic group instead of Fe^{3+} (in the form of a heme molecule), and this final stage can be inhibited by the irreversible binding of cyanide to Cu^{2+} preventing it from accepting electrons and therefore terminating the entire hydrogen electron transfer chain and hence all aerobic respiration. This is the basis of the toxicity of cyanide and also several other substances.

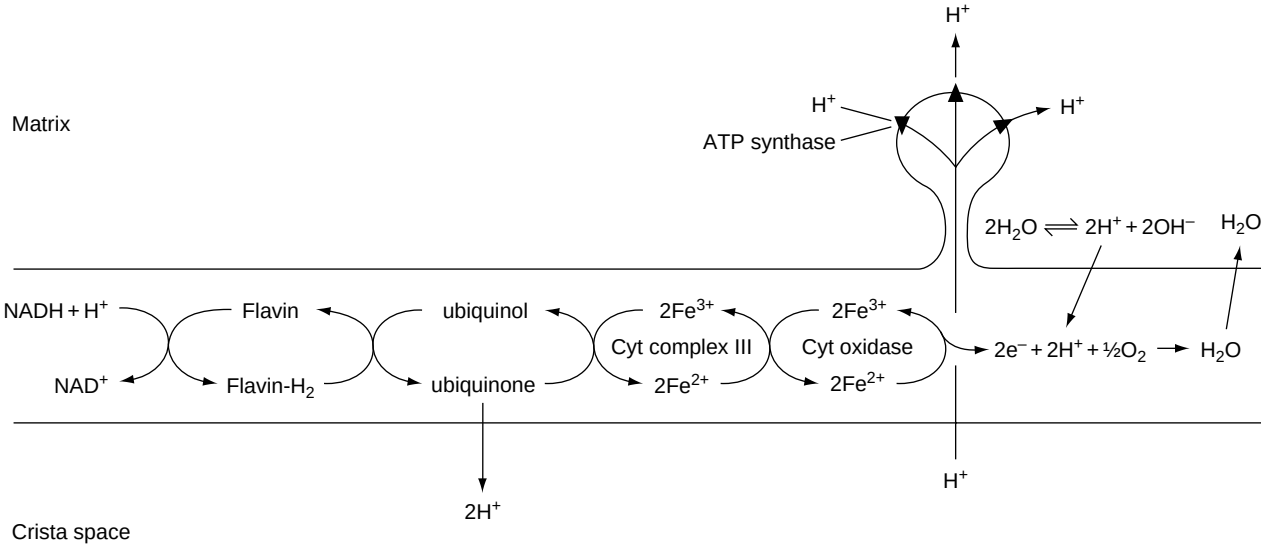


Figure 6 Overview of the electron transfer chain.

When the hydrogen from $NADH + H^+$ or $FADH_2$ is passed from ubiquinone to the first cytochrome, the hydrogen dissociates into a hydrogen ion (proton) and an electron. The proton is excreted into the crista space, whereas the electron carries on down the chain of cytochromes. This creates a proton gradient across the crista membrane. In the last step, the reduction of molecular oxygen to water, the hydrogen protons are obtained not from the hydrogen excreted into the crista space but from the mitochondrial matrix from the dissociation of water ($H_2O \leftrightarrow H^+ + OH^-$), thus maintaining a proton gradient across the crista membrane. The resulting movement of protons from the crista space to the matrix through the transmembrane stalk of the primary particle drives the multienzyme complex of ATP synthase. It is the energy of the flow of the protons that provides the energy required for the synthesis of ATP in a manner analogous to a water mill, where the flow of water can be used to turn a motor, grind wheat, or generate electricity.

The oxidation of $FADH_2$ and $NADH + H^+$ is normally tightly coupled to the phosphorylation of $ADP \rightarrow ATP$ because the phosphorylation of ADP cannot occur unless there is a proton gradient across the crista membrane of the mitochondria, resulting from the processes in the hydrogen electron transfer chain. Metabolic fuels can only be oxidized if there is NAD^+ and FAD available to be reduced. If there is no ADP because it has all been phosphorylated to ATP, protons cannot move down the concentration gradient across the crista membrane and through the 'water mill' of ATP synthase because it will not turn if there is no ADP to bind to it. Once a critical concentration of protons is reached in the crista space, it is no longer possible to extrude any more protons from the oxidation of FADH or $NAD + H^+$ during the transfer of hydrogen from ubiquinone to the first cytochrome. Hence, the whole chain stops and no more oxidized NAD^+/FAD is available for glycolysis or the TCA cycle.

The uncoupling of the electron transfer chain from the production of ATP can only occur if protons move down the concentration gradient across the crista membrane and through routes independent of ATP synthase. In some circumstances tissues, uncoupling proteins can be expressed in the crista membrane, which allow protons to flow down the concentration gradient into the mitochondrial matrix without passing through the ATP synthase and therefore not generating ATP but resulting in the generation of a large amount of heat energy. This is the basis of nonshivering thermogenesis, which is thought to occur mostly in brown adipose tissue in infants by the uncoupling protein thermogenin. In adults, the importance of brown adipose tissue compared with other uncoupling proteins in muscle and other tissues is unclear. In addition to maintenance of body temperature, uncoupling proteins may also be important in overall energy balance and body weight. Leptin, a hormone secreted from white adipose tissue, increases the expression of uncoupling proteins in muscle and adipose tissue, thus increasing energy expenditure and utilisation of adipose tissue fat reserves.

Energy Metabolism of Other Nutrients

The TCA cycle and pyruvate are central in the metabolism of carbohydrate, fat, and protein in the fed and fasting state. Pyruvate can have three main fates depending on the metabolic circumstances. It can be a substrate for gluconeogenesis, or it can undergo oxidative decarboxylation to acetyl-CoA and either enter the TCA cycle or be used for fatty acid synthesis.

Acetyl-CoA can be made from carbohydrates via pyruvate, from fatty acids via β-oxidation in the mitochondrial matrix, or from the proteolysis of proteins to amino acids, some of which are converted to acetyl-CoA.

There are three main stores of metabolic fuels: triacylglycerols in adipose tissue, glycogen as a carbohydrate reserve in liver and muscle, and protein as a source of amino acids that can be oxidized via the TCA cycle or used as a substrate for gluconeogenesis.

Overview of Fat Metabolism

β-Oxidation of Fatty Acids and Ketogenesis

Fats (triacylglycerides) are stored mainly in adipose tissue. Lipolysis breaks down fats into the constituent fatty acids and glycerol. Fatty acids can be oxidized via the β-oxidation pathway in the mitochondrial matrix. In this process, a cyclical series of reactions removes the last two carbon atoms from the carboxyl end of the fatty acyl-CoA, with the addition of another CoA to form a new fatty acyl-CoA that is two carbon atoms shorter plus acetyl-CoA. In muscle, the acetyl-CoA is metabolized via the TCA cycle to produce reduced coenzymes for the production of ATP. In the liver, it is shunted largely to the synthesis of ketone bodies (ketogenesis), which, like glucose, are exported for use in other tissues.

On the outer face of the mitochondrial membrane, fatty acids are esterified to CoA to form

fatty acyl-CoA, which cannot enter the matrix of the mitochondria, the site of the enzymes for β-oxidation. This function is performed by the carnitine shuttle. On the outer mitochondria membrane, fatty acyl is transferred onto carnitine to form fatty acyl-carnitine that is transported across the inner and outer mitochondrial membranes on a countercurrent transporter system, in exchange for transporting free carnitine into the intermembrane space. Once in the matrix, the fatty acyl is esterified to CoA, thus releasing free carnitine. There is no dietary requirement for carnitine because it is readily synthesized from the amino acids lysine and methionine.

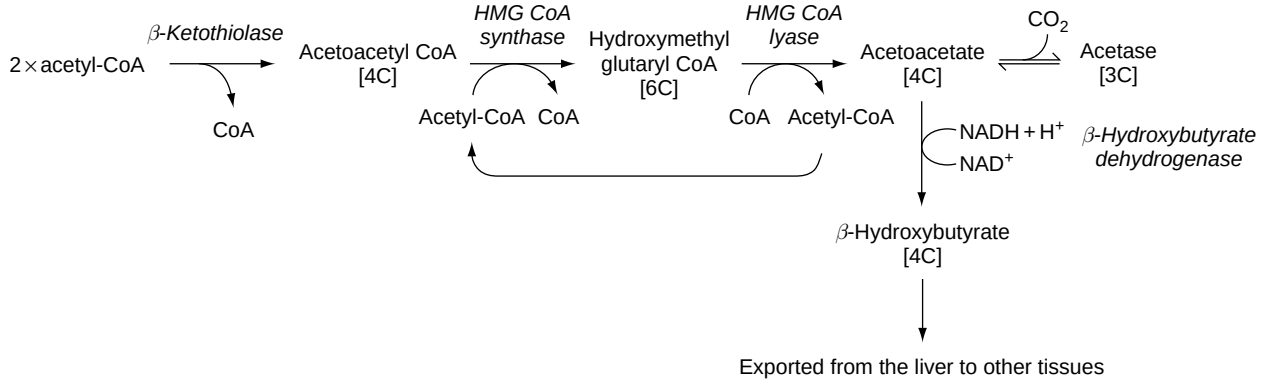
Most tissues have a limited capacity for β-oxidation. However, the liver can produce large amounts of acetyl-CoA by β-oxidation and can then convert some of this into four-carbon ketone bodies that can be easily transported to other tissues for use as a metabolic fuel. Acetoacetate is formed by the combination of 2 acetyl-CoA and the removal of the CoA molecules. This is unstable and undergoes a nonenzymic reaction to acetone, which is poorly metabolized, most of it being excreted in urine and exhaled air. Hence, most of the acetoacetate is reduced to β-hydroxybutyrate before being released from the liver. β-Hydroxybutyrate is metabolized by extrahepatic tissues by adding a CoA via succinate-CoA to form succinate and acetoacetyl-CoA that is then broken down into 2 acetyl-CoA by β-ketothiolase and CoA (**Figure 7**).

Synthesis of Fatty Acids and Triacylglycerides

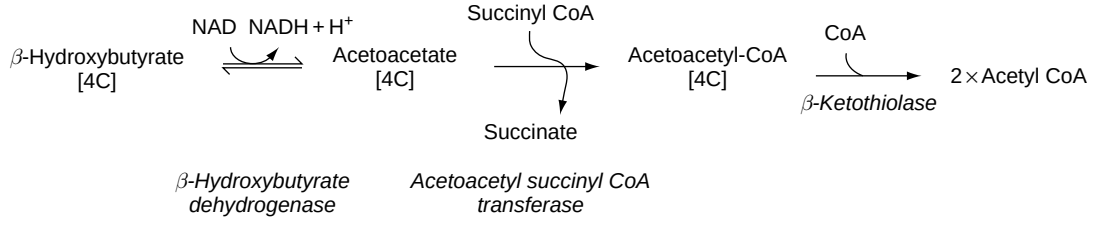
The majority of fatty acids are supplied by the diet, but many tissues are capable of *de novo* synthesis, including the liver, brain, kidney, mammary glands, and adipose tissue. The *de novo* synthesis of fatty acids occurs in conditions of excess energy intake.

Fatty acid synthesis occurs in the cytosol but is essentially the reverse of β-oxidation of fatty acids (although it employs a separate set of enzymes), whereby fatty acids are synthesized from the successive additions of 2C acetyl-CoA followed by reduction.

Acetyl-CoA is formed in the mitochondrial matrix, but it cannot pass across the mitochondrial inner membrane. Hence, the source of acetyl-CoA for fatty acid synthesis is citrate, which can pass out of the mitochondria, where, with CoA, it is cleaved to produce acetyl-CoA and oxaloacetate. The oxaloacetate is returned indirectly to the mitochondrial matrix via its oxidation to pyruvate, which is linked to the generation of reduced NADP, required for fatty acid synthesis. Once in the mitochondrial matrix, the pyruvate is converted back to oxaloacetate and thus returned into the TCA cycle.



(A) Ketogenesis in the liver from acetyl-CoA



(B) Extrahepatic metabolism of β-hydroxybutyrate (Ketone bodies) to acetyl-CoA

Figure 7 The production (A) and metabolism (B) of ketones produced from acetyl-CoA.

The first step in fatty acid synthesis is the carboxylation of acetyl-CoA to malonyl-CoA, followed by the addition of a series of malonyl-CoA units by a complex series of reactions via the multienzyme complex fatty acid synthase. The carboxylation of acetyl-CoA to malonyl-CoA is regulated by the hormones insulin and glucagon, which affect the activity of the enzyme catalyzing this reaction. Hence, in the fed state insulin increases the activity of the enzyme, whereas in the fasting state glucagon decreases its activity. Also, malonyl-CoA inhibits the uptake of fatty acids into the mitochondria via acetyl carnitine-CoA so that when fatty acid synthesis is occurring, β-oxidation is inhibited by limiting the supply of substrate into the mitochondria.

Storage of fatty acids in adipose tissue can only occur when glycolysis is activated in the fed state because the source of glycerol in adipose tissue is blood glucose entering glycolysis and DHA-P removed to be converted to glycerol-3-P.

During fatty acid synthesis, desaturase enzymes can introduce double bonds to make mono- and polyunsaturated fatty acids. However, these enzymes cannot introduce double bonds after C10; this is why there is a requirement for the essential fatty acids linoleic acid (n-6) and linolenic acid (n-3), which can then be converted to the long-chain polyunsaturated fatty acids, arachidonic acid and eicosapentaenoic acid, which are important metabolic precursors.

Protein Metabolism

After a meal there is an increase in the synthesis of tissue protein from absorbed amino acids and the increased availability of metabolic fuel to provide ATP for protein synthesis. During fasting some of the relatively labile protein laid down in response to a meal can be mobilized and the amino acids used both as a metabolic fuel and as a source of TCA cycle intermediates for gluconeogenesis.

After removal of the nitrogen-containing amino group of amino acids, their carbon skeletons can undergo gluconeogenesis (gluconeogenic amino acids only), be converted into ketone bodies via acetyl-CoA (ketogenic amino acids), be fully oxidized to CO_2 and H_2O, be converted into fat or glycogen for storage, or be used as a precursor for a wide range of important biomolecules (**Figure 8**).

Gluconeogenesis

The brain can normally only metabolize glucose as an energy source. Therefore, it is very important to maintain relatively constant levels of circulating glucose. In normal circumstances, glycogen serves as a source of blood glucose as free fatty acids from adipose tissue and ketone bodies from the liver are used preferentially as metabolic fuels by muscle and some other tissues. However, this still leads to the exhaustion of glycogen reserves within 12–18 h.

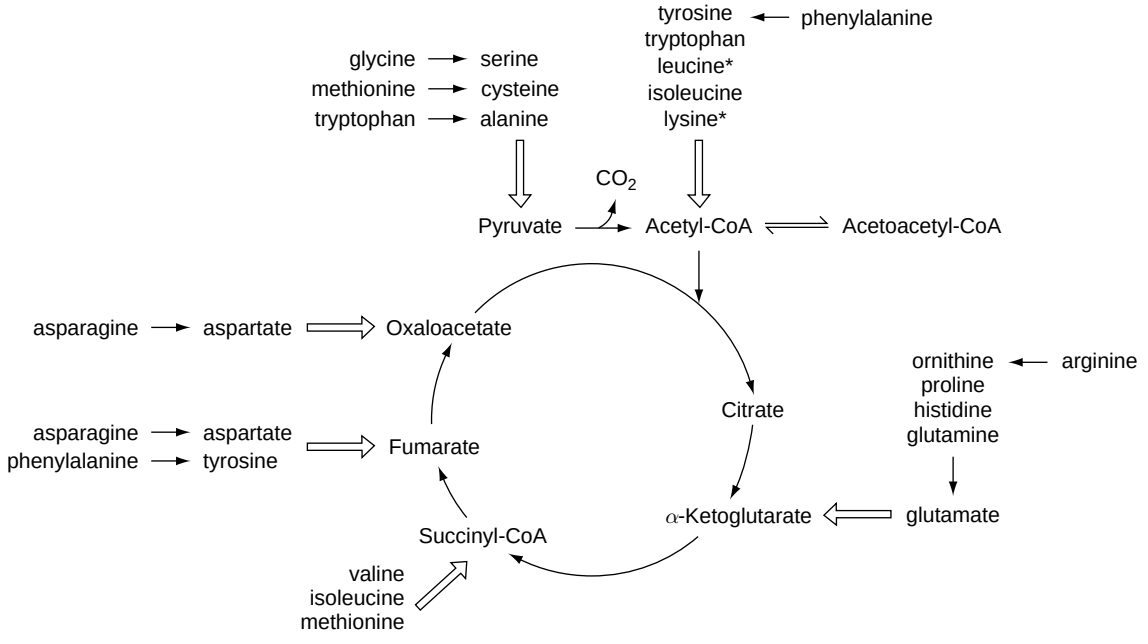


Figure 8 Entry of amino acid carbon skeletons into the tricarboxylic acid cycle. *These amino acids are ketogenic only. (Adapted with permission from Bender DA (2002) *Introduction to Nutrition and Metabolism*, 3rd edn. London: Taylor & Francis.)

Hence, the formation of glucose from noncarbohydrate sources (gluconeogenesis) is important.

Glucose can be formed from the gluconeogenic amino acids and from glycerol released from the lipolysis of TAGs in adipose tissue. Amino acids can enter the TCA cycle as intermediates and be converted to oxaloacetate, the excess of which can then be removed and metabolized to phophoenolpyruvate and then, by a process that is the reverse of glycolysis, be converted to glucose. Glycerol can be converted to an intermediate on the glycolytic pathway and therefore undergo gluconeogenesis if required.

Amino acids that can only be metabolized to acetyl-CoA cannot undergo gluconeogenesis because acetyl-CoA cannot be converted back to pyruvate and the inclusion of more acetyl-CoA will not generate a net increase in oxaloacetate, which can be removed from the cycle. Hence, fatty acids and ketones that are broken down into acetyl-CoA also cannot be used for gluconeogenesis.

There are three enzymes in gluconeogenesis that are different from those in glycolysis, and the relative activity of these compared to the equivalent glycolytic enzymes is tightly controlled by hormones, hence controlling whether glycolysis or gluconeogenesis is the dominant pathway.

Table 2 Summary of relative importance of different metabolic pathways in intermediary metabolism in different tissues

Tissue	*Principal catabolic and anabolic pathways*
Brain	25% basal O_2 consumption
	Metabolizes glucose only, except after prolonged starvation when it can adapt to uptake and metabolize ketones
Blood	Mature red blood cells have no mitochondria ∴ energy from anaerobic glycolysis: glucose → lactate
Muscle	Preferentially metabolize fatty acids and ketones produced from the liver
	Anaerobic glycolysis of glucose from glycogen stores
	Aerobic respiration of glucose from glycogen or fatty acids/ketones
Liver	Mostly amino acid oxidation for generation of ATP
	Most important tissue for maintaining blood glucose by gluconeogenesis from amino acids and lactate (via Cori cycle) and glycerol and also from breakdown of glycogen stores
	Fatty acid synthesis and synthesis of lipoproteins for transport
	Production of ketones into circulation
	Site of the pentose–phosphate pathway for generation of $NADPH + H^+$
Adipose tissue	Designed for the storage of fat
	Can synthesize fat from glucose
Kidneys	Gluconeogenesis
	Amino acid oxidation for ATP generation

Glycogen Metabolism

The red blood cells and the brain have an absolute requirement for glucose for energy metabolism. Glucose is absorbed from the intestines only for 2 or 3 h after a meal; therefore, there must be another source of glucose to maintain a constant blood glucose level. When blood glucose levels increase after a meal, the liver can uptake large amounts of glucose, where it is converted to glucose-6-phosphate that can be used to synthesize glycogen (glycogenesis). When glycogen stores are full, glucose-6-phosphate can enter glycolysis or be used to synthesize glycerol for the formation of fat. When blood glucose levels decrease, during fasting between meals, glycogen is broken down in the liver and glucose is released (glycogenolysis). In the fasting state, glycogen is broken down by the removal of glucose units as glucose-1-phosphate from the many ends of the molecule. This is then isomerised to glucose-6-phosphate. Only the liver can release free glucose because muscle tissue lacks glucose-6-phosphatase. The free glucose released by the liver is used by the brain and red blood cells.

Glucose-6-phosphate released in the muscle tissue from glycogen can enter directly into glycolysis for energy production by the muscle. Alternatively, it can be metabolized to pyruvate and then transaminated to alanine that is exported from the muscle to the liver, where it can be used as a substrate for gluconeogenesis. **Table 2** shows the relative importance of energy metabolic pathways in different tissues of the body.

See also: **Amino Acids**: Metabolism. **Energy**: Balance. **Fatty Acids**: Metabolism. **Glucose**: Metabolism and Maintenance of Blood Glucose Level. **Protein**: Requirements and Role in Diet. **Sports Nutrition**.

Further Reading

Bender DA (2002) *Introduction to Nutrition and Metabolism*, 3rd edn. London: Taylor & Francis.

Brody T (1999) Regulation of energy metabolism. In: *Nutritional Biochemistry*, 2nd edn, pp. 157–262. London: Academic Press.

Stillway W (1999) Bioenergetics and oxidative metabolism. In: Baynes J and Dominiczak M (eds.) *Medical Biochemistry*, pp. 83–94. St. Louis: Mosby.

Stillway W (1999) The tricarboxylic acid cycle. In: Baynes J and Dominiczak M (eds.) *Medical Biochemistry*, pp. 157–167. St. Louis: Mosby.

Wildman REC and Medeiros DM (2000) Energy metabolism. *Advanced Human Nutrition*, pp. 283–316 Boca Raton, FL: CRC Press.

Balance

Y Schutz, University of Lausanne, Lausanne, Switzerland

Introduction

To maintain physiologic functions, the human body continuously expends energy by oxidative metabolism. This energy is used to maintain chemical and electrochemical gradients across cellular membranes for the biosynthesis of macromolecules such as proteins, glycogen, and triglycerides, and for muscular contraction. Another part of the energy is lost as heat because of the inefficiency of metabolic transformations. Ultimately all the energy produced by the organism is dissipated as heat.

The energy expended by an individual can be assessed by two different techniques: indirect and direct calorimetry. The term indirect calorimetry stems from the fact that the heat released by chemical processes within the body can be indirectly calculated from the rate of oxygen consumption ($\dot{V}O_2$). The main reason for the close relation between energy metabolism and $\dot{V}O_2$ is that the oxidative phosphorylation at the respiratory chain level allows a continuous synthesis of adenosine triphosphate (ATP). The energy expended within the body to maintain electrochemical gradients, support biosynthetic processes, and generate muscular contraction cannot be directly provided from nutrient oxidation. Almost all chemical processes requiring energy depend on ATP hydrolysis. It is the rate of ATP utilization that determines the overall rate of substrate oxidation and therefore $\dot{V}O_2$. With the exception of anaerobic glycolysis, ATP synthesis is coupled with substrate oxidation. The biochemical theory of oxidative phosphorylation considers that 3 mol of ATP are generated per gram-atom of oxygen consumed (i.e., a P:O ratio of 3:1). The energy expenditure per mole of ATP formed should be calculated from the heat of combustion of 1 mol of substrate, divided by the total number of moles of ATP generated in its oxidation. It is interesting to note that each mole of ATP generated is accompanied by the release of about the same amount of heat (~75 kJ/mol ATP) during the oxidation of carbohydrates, fats, or proteins. Because there is a proportionality between $\dot{V}O_2$ and ATP synthesis, and because each mole of ATP synthetized is accompanied by the production of a given amount of heat, one understands the rationale of using $\dot{V}O_2$ measurement to calculate heat production within the body.

Since indirect calorimetry measures the heat released by the oxidative processes and direct calorimetry assesses the heat dissipated by the body, a relationship exists between the two: for a subject in resting conditions, the difference between metabolic heat production and heat dissipation represents the body heat balance (**Figure 1**). The heat production from oxidative processes is equal to the sum of

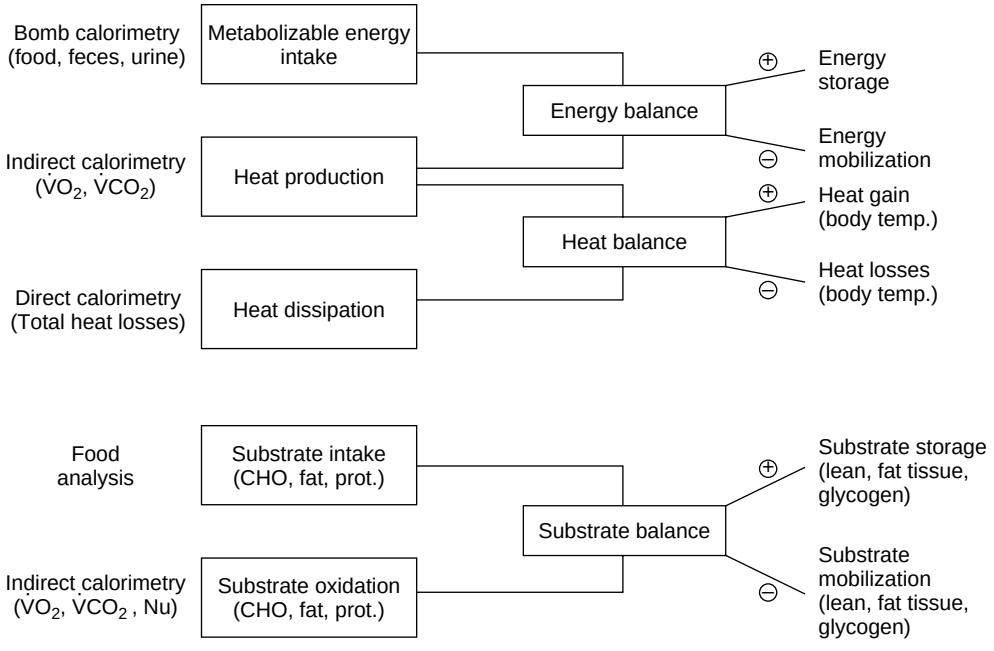


Figure 1 Heat balance, energy balance, and substrate balance: three different concepts.

the nonevaporative components (radiant heat exchange + convective + conductive heat transfer) plus the evaporative heat transfer. In order to assure the equality of the equation an additional term representing the rate of storage of body heat must be included:

$$\text{Heat production} = \text{Heat losses} \pm \text{Heat storage}$$

Heat storage can be positive when excess heat is gained, resulting in a rise in internal body temperature. Heat storage can be negative when excess heat is lost, resulting in a cooling of the body. The rate of heat storage can be estimated from the body weight, the specific heat capacity of the body (which depends upon body composition), and the rate of change in internal body temperature. In practice, this calculation remains somewhat uncertain since the changes in temperature within the body are not uniformly distributed within each tissue.

Under most environmental conditions, heat is lost by all channels (i.e., radiative + convective + conductive + evaporative). However, except during immersion in water, the rate of heat gain or loss by conduction constitutes a small proportion of the total heat loss (typically 3%). Heat can be lost by convection (air currents) but it can also be gained in very hot conditions such as in a desert characterized by high movement of hot air.

Energy Balance: Definition

Overall energy balance is given by the following equation:

$$\text{energy intake} = \text{energy expenditure} + \Delta \text{ energy stores}$$

Thus, if the total energy contained in the body (as fat, protein, and glycogen) is not altered (i.e., Δ energy stores = 0), then energy expenditure must be equal to energy intake. In this case, the individual is said to be in a state of energy balance.

If the intake and expenditure of energy are not equal, then a change in body energy content will occur, with negative energy balance resulting in the utilization of the body's energy stores (glycogen, fat, and protein) or positive energy balance resulting in an increase in body energy stores, primarily as fat.

The difference between the concepts of energy balance and heat balance is presented in **Figure 1**.

Model of Energy Balance: A Dynamic State

There are multiple reciprocal direct and indirect influences of energy intake on energy expenditure and vice versa: for example, energy intake influences resting energy expenditure by increasing postprandial and dietary-induced thermogenesis, whereas changes in energy expenditure via a modification of physical activity is susceptible to influence energy intake to maintain energy balance. In order to assure an accurate regulation of body stores, a double control is essential (see below). Body weight and body composition are not invariant with time but small corrections of both input and output from day-to-day or week-to-week, assure energy homeostasis (**Figure 2**). When attempting to explain the actual responses in energy balance and weight regulation in real life, we need to recognize that several factors may be operating at once on both sides of the energy balance equation. Compensatory adjustments occur in both intake and output, so unravelling the importance of one or other adjustment is not easy in man.

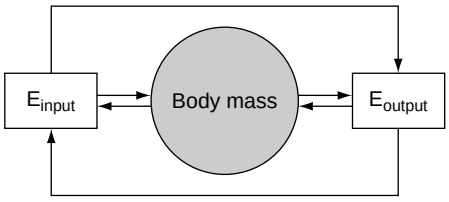


Figure 2 Simple model of energy (E) balance. Long-term constancy of body weight through the regulation of energy balance is achieved through a highly complex network of regulatory systems operating through changes in food intake, in energy expenditure, and body energy content (i.e., change in body composition).

Gross and Metabolizable Energy

The traditional way of measuring the energy content of foostuffs is to use a 'bomb calorimeter' in which the heat produced when a sample of food is combusted (under high pressure of oxygen) is measured. When the food is combusted, it is completely oxidized to water, carbon dioxide, and other incompletely burned elements. The total heat liberated (expressed in kilocalories or kilojoules) represents the gross energy value or heat of combustion of the food. The heat of combustion differs between carbohydrates, proteins, and fats. There are also important differences within each category of macronutrient. The gross energy yield of sucrose, for example, is $16.5\,\text{kJ}\,\text{g}^{-1}$, whereas starch yields $17.5\,\text{kJ}\,\text{g}^{-1}$. The energy yield of butterfat is $38.5\,\text{kJ}\,\text{g}^{-1}$ and of lard $39.6\,\text{kJ}\,\text{g}^{-1}$. These values have been rounded off to give $17.3\,\text{kJ}\,\text{g}^{-1}$ for carbohydrates rich in starch and poor in sugar, $39.3\,\text{kJ}\,\text{g}^{-1}$ for average fat, and $23.6\,\text{kJ}\,\text{g}^{-1}$ for mixtures of animal and vegetable proteins.

The gross energy value of foodstuffs (**Table 1**), however, does not represent the energy actually

Table 1 Metabolizable energy (ME) and Atwater's factors

Nutrient	*Gross energy in kJ g^{-1}(kcal g^{-1})*	*% Absorbed (Atwater's values)*	*Digestible energy in kJ g^{-1}(kcal g^{-1})*	*Urinary loss in kJ g^{-1}(kcal g^{-1})*	*Metabolizable energy in kJ g^{-1}(kcal g^{-1})*	*Atwater's factor[1] (kcal g^{-1})*
Starch	17.5 (4.2)	99	17.3 (4.15)	–	17.3 (4.15)	4
Glucose	15.6 (3.75)	99	15.4 (3.7)	–	15.4 (3.7)	4
Fat	39.1 (9.35)	95	37.1 (8.88)	–	37.1 (8.88)	9
Protein	22.9 (5.47)	92	21.1 (5.04)	5.2 (1.25)	15.9 (3.8)	4
Alcohol	29.8 (7.1)	100	29.8 (7.1)	Trace	29.8	7

[1]Values are rounded off.

available to the body, since no potentially oxidizable substrate can be considered available until it is presented to the cell for oxidation. None of the foodstuffs are completely absorbed; therefore, some energy never enters the body and is excreted in feces. Digestibility of the major foodstuffs, however, is high; on the average, 97% of ingested carbohydrates, 95% of fats, and 92% of proteins are absorbed from the intestinal lumen.

In the body, the tissues are able to oxidize carbohydrate and fat completely to carbon dioxide and water, but the oxidation of protein is incomplete, and results in the formation of urea and other nitrogenous compounds, which are excreted in the urine. Determination of both the heat of combustion and the nitrogen content of urine indicates that approximately 33.0 kJ g^{-1} of urine nitrogen is equivalent to 5.2 kJ g^{-1} of protein since 1 g urinary nitrogen arises from ~6.25 g protein. This energy represents metabolic loss and must be subtracted from the 'digestible' energy of protein to obtain metabolizable energy.

With a mixed diet, rich in carbohydrates and fibers, the metabolizable energy of food is approximately 90% of the gross energy (heat of combustion) (**Figure 3**). The remaining 10% is mainly due to unabsorbed energy (fecal energy losses) and urinary excretion of metabolites.

The collection of representative samples of all food eaten combined with the collection of urine and feces for a week (i.e., complete nutritional balance) is technically difficult and cumbersome in practice. The pioneer investigator Atwater developed in the early 20th century a practical approach for calculating, rather than measuring, the metabolizable energy (ME) of a diet based on its composition of carbohydrates, fat, and proteins. A specific calorimetric factor was developed according to the digestibility and absorption of each macronutrient and the loss of energy in urine (measured by nutrition balance technique). These are the so-called 'Atwater factors.' The ME values for the three substrates and their derivation are shown in **Table 1**.

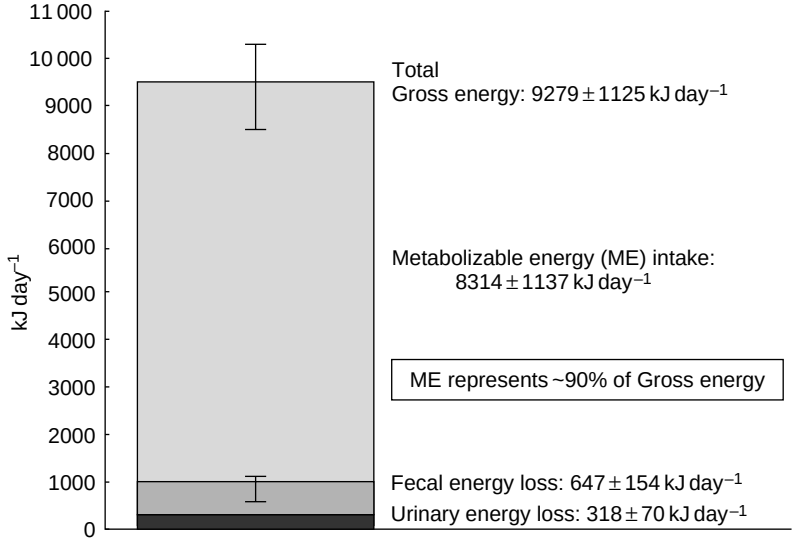


Figure 3 Gross and metabolizable energy intakes of a mixed, high-carbohydrate diet (data from McNeil (2000)). Calculation of metabolizable energy for 10 women on a high-carbohydrate, high-nonstarch polysaccharide (NSP) diet for a period of 7 days.

Assessment of Energy Expenditure

The study of energy metabolism in humans has recently raised a great interest in the regulation of these processes thanks to advances in the construction of open-circuit ventilated hood indirect calorimeters and comfortable respiration chambers.

With the only measurement of $\dot{V}O_2$ (in liters of O_2 STPD (standard temperature (0 °C), pressure (760 mm Hg), and dry) per min), metabolic rate (M), which corresponds to energy expenditure, can be calculated (in kilojoules per min) as follows:

$$\dot{M} = 20.3 \times \dot{V}O_2 \qquad [1]$$

The number 20.3 is a mean value (in $kJ\,l^{-1}$) of the energy equivalent for the consumption of 1 l (STPD) of oxygen. The value of the energy equivalent of oxygen depends on the composition of the fuel mixture oxidized. However, the error in using equation 1 instead of an equation that takes into account the type of fuels oxidized (equations 2 and 3, see below) is no greater than ±1 to 2%.

To take into account the heat released by the oxidation of the three macronutrients (carbohydrates, fats, and proteins), three measurements must be carried out: oxygen consumption ($\dot{V}O_2$), carbon dioxide production ($\dot{V}CO_2$), and urinary nitrogen excretion (N).

Simple equations for computing metabolic rate (or energy expenditure) from these three determinations are written under the following form:

$$\dot{M} = a\dot{V}O_2 + b\dot{V}CO_2 - cN \qquad [2]$$

The factors a, b, and c depend on the respective constants for the amount of O_2 used and the amount of CO_2 produced during oxidation of the three classes of nutrients. An example of such a formula is given below:

$$\dot{M} = 16.18\dot{V}O_2 + 5.02\dot{V}CO_2 - 5.99N \qquad [3]$$

where $\dot{M}$ is in kilojoules per unit of time, $\dot{V}O_2$ and $\dot{V}CO_2$ are in liters STPD per unit of time, and N is in grams per unit of time. As an example, if $\dot{V}O_2 = 600\,l\,day^{-1}$, $\dot{V}CO_2 = 500\,l\,day^{-1}$ (respiratory quotient, or RQ = 0.83), and $N = 25\,g\,day^{-1}$, then $\dot{M} = 12\,068\,kJ\,day^{-1}$. With the simpler equation (1) the results give a value of 12 180 kJ per day.

Slightly different factors for the amounts of O_2 used and of CO_2 produced during oxidation of the nutrients are used by other authors, and the values for the factors a, b, and c are modified accordingly. The difference in energy expenditure calculated by the various formulae is not greater than 3%.

Total Energy Expenditure and its Components

It is customary to consider energy expenditure as being made up of three components: the energy spent for basal metabolism (or basal metabolic rate), the energy spent on physical activity, and the increase in resting energy expenditure in response to a variety of stimuli (in particular food, cold, stress, and drugs). These three components are depicted in **Figure 4**.

Basal Metabolic Rate (BMR) or Resting Metabolic Rate (RMR)

This is the largest component of energy expenditure accounting for between half to three-quarters of daily energy expenditure. It is measured under standardized conditions, i.e., in an awake subject lying in the supine position, in a state of physical and mental rest in a comfortable warm environment, and in the morning in the postabsorptive state, usually 10–12 h after the last meal. There is an arbitrary distinction between RMR and BMR in the literature. RMR may be considered equivalent to BMR if the measurements are made in postabsorptive conditions. It seems difficult to partition RMR into various subcomponents since the metabolic rates of individual organs and tissues are hard to assess in humans under noninvasive experimental conditions. BMR can vary up to ±10% between individuals of the same age, gender, body weight, and fat-free mass (FFM), suggesting that genetic factors are also important. Day-to-day intraindividual variability in BMR is low in men (coefficient of variation of 1–3%) but is greater in women because the menstrual cycle affects BMR. In both women and men, sleeping metabolic rate is lower than BMR by

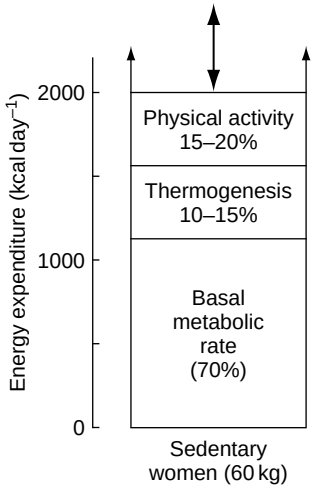


Figure 4 The three classical components of total energy expenditure. (Inactive person).

5–10%, the difference being explained by the effect of arousal. BMR is known to be depressed during starvation.

The major part of the whole-body RMR stems from organs with high metabolic activity such as the liver, kidneys, brain, and heart, although these account for a small proportion of the total body weight (5%). Per unit body weight, the kidneys and heart have a metabolic rate more than twice as high as the liver and the brain. In contrast, the metabolic rate of muscle per unit body weight is nearly 35 times lower than that of the heart and kidneys. Since the proportion of muscle to nonmuscle changes with age from birth to adulthood, the RMR per unit body weight is not constant with age. The tissue with the lowest metabolic activity per unit body weight is adipose tissue, which accounts for only 4% of the whole-body RMR in nonobese subjects. Calculations show that this value can increase up to 10% or more in obese subjects with a large excess in body fat. Skin and intestines (which have a relatively large protein mass and protein turnover), as well as bones and lungs, also contribute significantly to RMR.

Numerous studies have demonstrated that major factor explaining the variation in RMR between individuals is FFM. FFM is a heterogeneous component that can be partitioned into muscle mass and nonmuscle mass. Unfortunately, there is no simple and accurate way to assess these two subcomponents. Owing to the larger variation between individuals in fat mass, as compared to FFM, and because in grossly obese women fat mass can represent a nonnegligible component of total RMR, the prediction models for RMR that include both FFM and fat mass explain significantly more variance in RMR than FFM alone. In addition, age, sex, and family membership are additional factors that should be taken into account.

The effects of gender on resting metabolic rate are explained by differences in body composition. Caution should be used when comparing resting metabolic rate expressed per kilogram FFM in men and women, because the composition of FFM is influenced by gender. The muscle mass of men is greater than that is greater of women and this tends to give a lower value of RMR per kilogram FFM in men when compared to that of women. This is explained by a greater component of tissue with a low metabolic rate (resting muscle) in men than in women.

In clinical work, where body composition is difficult to assess, body weight, gender, and age can be used to estimate BMR and RMR (**Table 2**), bearing in mind that many important determinant of RMR, in addition to body size, have been tracked (**Table 3**).

Table 2 Simple formulae for the prediction of resting metabolic rate in men and women of different ages (equations for predicting basal metabolic rate from body weight alone)

Age range (years)	*kcal per day*	*MJ per day*
Males		
0–3	60.9 W − 54	0.255 W − 0.226
3–10	22.7 W + 495	0.0949 W + 2.07
10–18	17.5 W + 651	0.0732 W + 2.72
18–30	15.3 W + 679	0.0640 W + 2.84
30–60	11.6 W + 879	0.0485 W + 3.67
>60	13.5 W + 487	0.0565 W + 2.04
Females		
0–3	61.0 W − 51	0.255 W − 0.214
3–10	22.5 W + 499	0.0941 W + 2.09
10–18	12.2 W + 746	0.0510 W + 3.12
18–30	14.7 W + 496	0.0615 W + 2.08
30–60	8.7 W + 829	0.0364 W + 3.47
>60	10.5 W + 596	0.0439 W + 2.49

W, body weight expressed in kilograms; MJ, megajoules.
(Data from WHO (1986) *Energy and Protein Requirements*. Report of a Joint FAO/WHO/UNU Expert Consultation. Technical Report Series 724. Geneva: World Health Organization.)

Thermic Effect of Food or Postprandial Thermogenesis

The energy expenditure increases significantly after a meal. The thermic effect of food is mainly due to the energy cost of nutrient absorption and storage. The total thermic effect of food over 24 h represents ~10% of the total energy expenditure in sedentary subjects. The thermic effect of nutrients mainly depends on the energy costs of processing and/or storing the nutrient. Expressed in per cent of the energy content of the nutrient, values of 8%, 2%, 20–30%, and 22% have been reported for glucose, fat, protein, and ethanol, respectively.

Glucose-induced thermogenesis mainly results from the cost of glycogen synthesis and substrate cycling. Glucose storage as glycogen requires 2 mol ATP/mol. In comparison with the 38 mol ATP produced on complete oxidation of glucose, the energy cost of glucose storage as glycogen corresponds to

Table 3 Determinants of resting (basal) metabolic rate

- Body size
- Body composition (lean vs. obese)
- Gender
- Age
- Physiological status (growth, pregnancy, and lactation)
- Genetic make-up
- Hormonal status (e.g., Follicular ve luteal phase)
  - Temperature (body internal and environment)
  - Pharmacological agents (e.g., nicotine and caffeine)
  - Disease (fever, tumors, burns, etc.)

5% (or 2/38) of the energy content of glucose stored. Cycling of glucose to glucose-6-phosphate and back to glucose, to fructose-1,6-diphosphate and back to glucose-6-phosphate, or to lactate and back to glucose, is occurring at variable rates and is an energy-requiring process that may increase the thermic effect of carbohydrates.

The thermic effect of dietary fat is very small; an increase of 2% of its energy content has been described during infusion of an emulsion of triglyceride. This slight increase in energy expenditure is explained by the ATP consumption in the process of free fatty acid reesterification to triglyceride. As a consequence, the dietary energy of fat is used very efficiently.

The thermic effect of proteins is the highest of all nutrients (20–30% of the energy content of proteins). Ingested proteins are degraded in the gut into amino acids. After absorption, amino acids are deaminated, their amino group transferred to urea, and their carbon skeleton converted to glucose. These biochemical processes require the consumption of energy amounting to ~25% of the energy content of amino acids. The second pathway of amino acid metabolism is protein synthesis. The energy expended for the synthesis of the peptide bonds also represents ~25% of the energy content of amino acids. Therefore, irrespective of their metabolic pathway, the thermogenesis induced after absorption of amino acids represents ~25% of their energy content.

Energy Expenditure Due to Physical Activity

The energy spent on physical activity depends on the type and intensity of the physical activity and on the time spent in different activities. Physical activity is often considered to be synonymous with 'muscular work', which has a strict definition in physics (force × distance) when external work is performed in the environment. During muscular work (muscle contraction), the muscle produces 3–4 times more heat than mechanical energy, so that useful work costs more than muscle work. There is a wide variation in the energy cost of any activity both within and between individuals. The latter variation is due to differences in body size and in the speed and dexterity with which an activity is performed. In order to adjust for differences in body size, the energy cost of physical activities are expressed as multiples of BMR. These generally range from 1 to 5 for most activities, but can reach values between 10 and 14 during intense exercise. In terms of daily energy expenditure, physical activity accounts for 15–40% of total energy expenditure but it can represent up to 70% of daily energy expenditure in an individual involved in heavy manual work or competition athletics. For most people in industrialized societies, however, the contribution of physical activity to daily energy expenditure is relatively small. The numerous factors influencing the 3 components of energy expenditure are outlined in **Table 4**.

Table 4 Exogenous and endogenous factors influencing the three components of energy expenditure

Components	*Endogenous*	*Exogenous*
• Basal metabolic rate	• Fat-free mass • Thyroid hormones • Protein turnover	
• Thermogenesis	• Nutritional status • Sympathetic nervous system activity • Insulin resistance (obesity)	• Macronutrient intake (+alcohol) • Cold exposure • Stress • Thermogenic stimuli (coffee, tobacco) • Thermogenic drugs
• Physical activity	• 'Fidgeting' • Muscular mass • Work efficiency • Fitness level ($\dot{V}O_2$max)	• Duration intensity, and frequency of physical activity

The effect of body weight in average women (~60 kg) on energy expenditure is illustrated in **Figure 5**. The relationship is slightly curvilinear because of differences in body composition in terms of leanness and fatness. Resting metabolic rate is shown as a baseline value.

Just as described above for a specific activity, it has been customary to express total energy expenditure

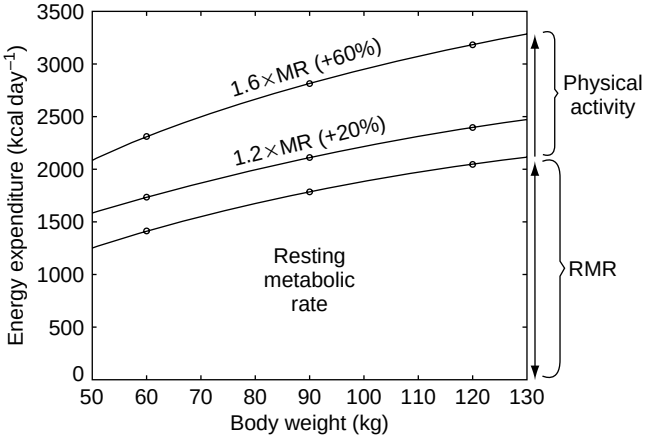


Figure 5 Effect of body weight on total energy expenditure at two levels of physical activity in young women. A physical activity level (PAL) of 1.2 represents minimal physical activity compatible with health, whereas a value of 1.6 represents a 'medium' level of physical activity.

(TEE) relative to RMR (TEE/RMR or TEE/BMR) to offset the large variation in RMR among subjects of difference body weight & body composition. This quotient is called physical activity level (PAL) and reflects multiples of RMR. A PAL of 1.5 indicates that TEE is 50% greater than RMR over 24 h.

Macronutrient Balance, Energy Balance, and Storage

Since macronutrients (carbohydrate, fat, protein, and alcohol) are the sources of energy, it is logical to consider energy balance and macronutrient balance together as the opposite side of the same coin.

There is a direct relationship between energy balance and macronutrient balance, and the sum of individual substrate balance (expressed as energy) must be equivalent to the overall energy balance. Thus:

$$\text{carbohydrate balance} = \text{exogenous carbohydrate} - \text{carbohydrate oxidation}$$

$$\text{protein balance} = \text{exogenous protein} - \text{protein oxidation}$$

$$\text{lipid balance} = \text{exogenous lipid} - \text{lipid oxidation}$$

It follows that Δ substrate balance $\equiv \Delta$E balance. Fat balance is closely related to energy balance (**Figure 6**).

Indirect calorimetry also allows computation of the nutrient oxidation rates in the whole body. An index of protein oxidation is obtained from the total amount of nitrogen excreted in the urine during the test period. One approach to calculate the nutrient oxidation rate is based on the oxygen consumption and CO_2 production due to the oxidation rates of the three nutrients carbohydrate, fat, and protein **Figure 6** respectively. In a subject oxidizing c grams per min of carbohydrate (as glucose), f grams per min of fat, and excreting n grams per min of urinary nitrogen, the following equations, can be used:

$$\dot{V}O_2 = 0.746c + 2.02f + 6.31n \quad [4]$$

and

$$\dot{V}CO_2 = 0.746c + 1.43f + 5.27n \quad [5]$$

We can solve equations 4 and 5 for the unknown c and f this way:

$$c = 4.59\dot{V}CO_2 - 3.25\dot{V}O_2 - 3.68n \quad [6]$$

$$f = 1.69\dot{V}O_2 - 1.69\dot{V}CO_2 - 1.72n \quad [7]$$

Because 1 g urinary nitrogen arises from approximately 6.25 g protein, the protein oxidation rate (p in grams per min) is given by the equation

$$p = 6.25n \quad [8]$$

Energy stores (constituted mainly of fat stores) are big as a proportion of the food intake (2000 kcal day^{-1}, mixed diet in a 60-kg nonobese woman with 25% body fat). The total energy stored is about 90 times total daily energy intake: typically fat stores are 175 times daily fat intake, protein 133 times daily protein intake, and carbohydrate only 1.3 times daily carbohydrate intake (**Figure 7**).

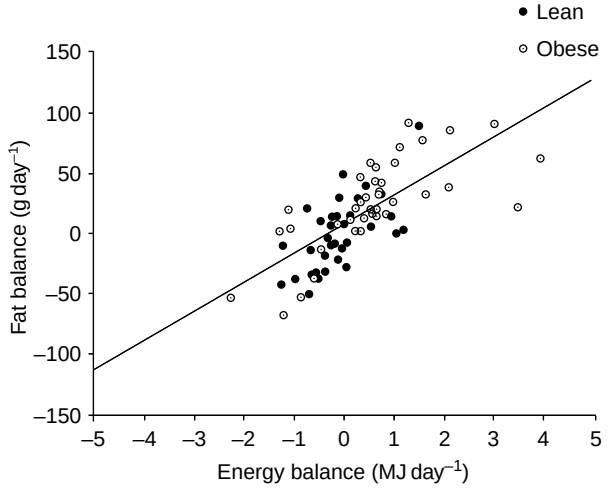


Figure 6 Relationship between energy balance and fat balance in lean and obese individuals. Note that at zero energy balance fat balance is zero. At an excess or deficit of 4.2 MJ day^{-1} (1000 kcal) the fat imbalance (about 100 g day^{-1} = 900 kcal day^{-1}) accounts for more than 90% of the magnitude of energy balance. (Reproduced from Schrauwen P, Lichtenbelt WD, Saris WH, Westerterp KR (1998) Fat balance in obese subjects: role of glycogen stores. *AM J Physiol.* **274**: E1027–33.)

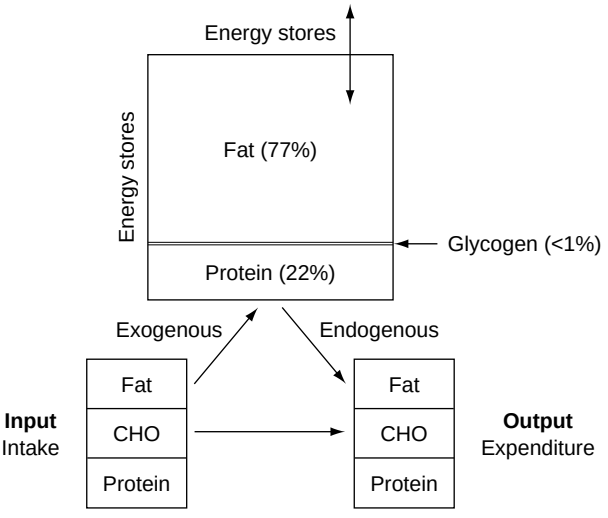


Figure 7 Macronutrient (substrate) stores versus macronutrient intake.

Energy Imbalance and Body Weight

Positive energy balance leads to body weight gain and negative energy leads to body weight loss. There is no fixed relationship between these two variables so that relatively small energy retention can be accompanied by large body weight gain and vice versa. The confounding factor is the associated water storage.

Long-term fluctuations in fat stores will be reflected in body weight. There is a difference in the energy value of fat mass and fat-free mass, the latter including the glycogen–water pool and the protein–water pool (**Table 5**).

Energy density of the tissue stored (or the substrate pool stored) represents an indicator of the composition of tissue stored or mobilized. It is defined as the total calorie per gram of substance. It is about $8\,kcal\,g^{-1}$ for adipose tissue compared to the fat value (triglyceride) of $9\,kcal\,g^{-1}$. This lower former value is due to the fact that fat is diluted out by the small amount of water (5–10%) and proteins the adipose tissue contains. As explained previously, the energy density of the glycogen–water pool is low, about $1\,kcal\,g^{-1}$, since glycogen ($4.2\,kcal\,g^{-1}$) is associated with approximately 3 times its weight of water.

Let us take an energy imbalance of say 1000 kcal. The body weight change will be approximately 8 times lower (i.e., ≈125 g) if fat is stored in adipose tissue, as compared to glycogen stored (under the form of glycogen-water pool) in liver and muscles (≈1000 g). In other words, rapid weight gain (or weight loss) means little fat storage despite what the layman thinks. Day-to-day energy imbalance is generally accommodated by water retention due to changes in carbohydrate storage and sodium intake.

In real life, it is more reasonable to consider that the reserve is composed of a mixture in different proportions of fat and glycogen. If about half of the energy imbalance is accounted for by fat and half by glycogen storage, the energy density will be $4–5\,kcal\,g^{-1}$. With the imbalance value described above, it will generate a body weight change of 400–500 g.

The energy balance varies from day to day. The changes in daily energy intake and expenditure are not necessarily synchronized. Positive energy balance on one day may not be spontaneously compensated by negative energy balance on the subsequent day, so that it is important to consider the overall energy balance regulation over a prolonged period of time. Short-term day-to-day energy imbalance is mostly accommodated by rapid changes in carbohydrate balance, whereas over a prolonged period of time, positive energy balance is mostly expressed as fat storage since carbohydrate stores are small (**Figure 6**).

To what extent do alterations in energy output contribute to the regulation of energy balance and stability of body weight? To understand the regulation of a system, it must be subjected to perturbation. Excess food intake during overfeeding or deficit in food intake during underfeeding disrupts the balance system.

Overfeeding Studies (Figure 8)

In a perfectly regulated system, any increase in energy intake should be offset by a change in energy expenditure of the same magnitude and direction. However, a 100% efficient adaptive process would obviously be counter productive, since this would signify that an increase in energy storage (required during nutritional rehabilitation) or an increase in energy mobilization (required for decreasing body weight) would be very limited. Adaptation to energy imbalance only occurs at the cost of increasing (or decreasing) body weight. In fact, excess energy intakes result in an increased metabolic turnover and energy flux through the mechanism of adaptive thermogenesis. The efficiency of energy storage is not constant and depends upon several factors including the magnitude of energy imbalance and the composition of the surfeit energy fed, as well as endogenous factors. As shown in **Figure 8**, the energy expenditure increases during acute overfeeding, an evidence of the 'flexibility' of the metabolism.

Table 5 Body stores of energy as different macronutrients

Substrate	*Form of storage*	*Pool size*	*Tissues*	*% Water in tissues*	*Daily imbalance*	*Energy density (kcal/g^{-1})*	*Postprandial thermogenesis*
Carbohydrate	Glycogen	Small (limited)	Liver + muscles	70–80	Large	≃4	Average
Fat	Triglycerides	Moderate–large (unlimited)	Adipose tissue	5–10	Small	≃8	Low
Protein	Protein	Moderate (limited)	Lean tissue (muscle)	70–75	Small	≃4	High

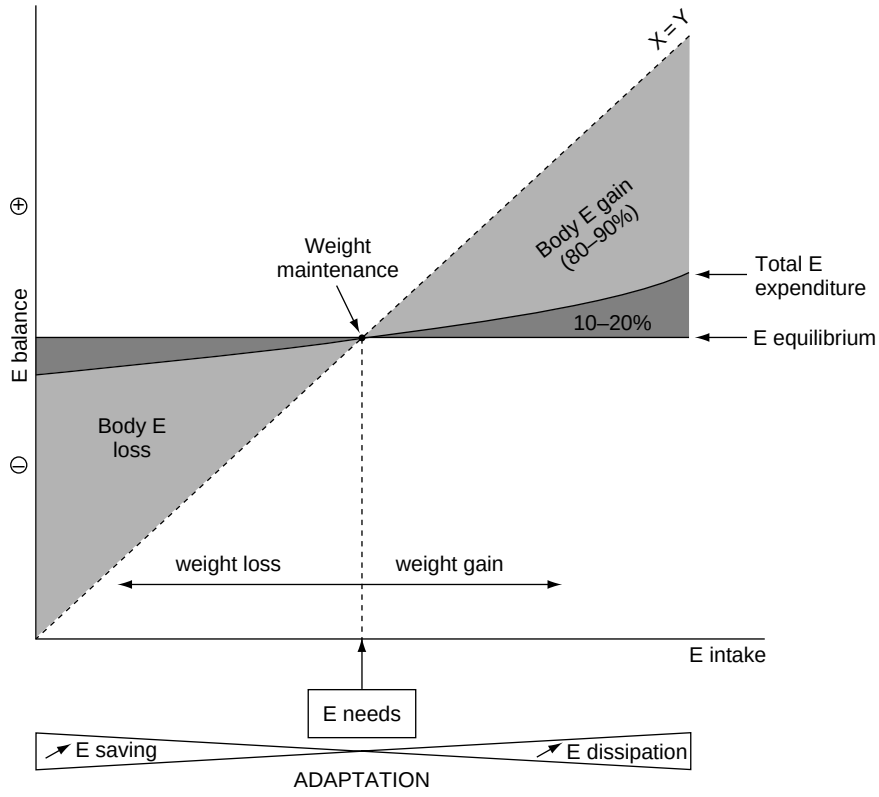


Figure 8 Energy balance in underfeeding (below maintenance) and overfeeding (above maintenance) conditions. E, energy.

Underfeeding Studies (Figure 8)

Analysis of underfeeding experiments shows that the decrease in energy expenditure has three components. First, if energy intake is decreased the thermic effect of feeding (about 10% of energy intake) is similarly decreased. Second, there is an adaptive decrease in metabolic rate during the first week, related in part to a decrease in sympathetic activity. The magnitude of this decrease is significantly related to the initial metabolic rate, and is usually about 5–8%. Third, there is a decrease in metabolic rate related to the weight lost: most investigators find a decrease of 10–12 kcal per day per kg weight loss. The effect of all three processes is that a person who lost weight from, say, 100 kg to 70 kg (a 30% reduction in weight) would experience about a 15% reduction in energy requirements for weight maintenance. Thus, a decrease in energy intake causes a reduction in body weight but, provided the decrease is not too great, a new equilibrium will be reached at which the reduced requirement will be satisfied by the reduced intake, and body weight will stabilize. Taken together we can conclude that the efficiency of energy utilization is lower in overfeeding than in underfeeding conditions because, substrate storage in tissues is energetically costly (ATP needs), whereas the process of energy mobilization requires little energy. In the former situation excess energy must be dissipated.

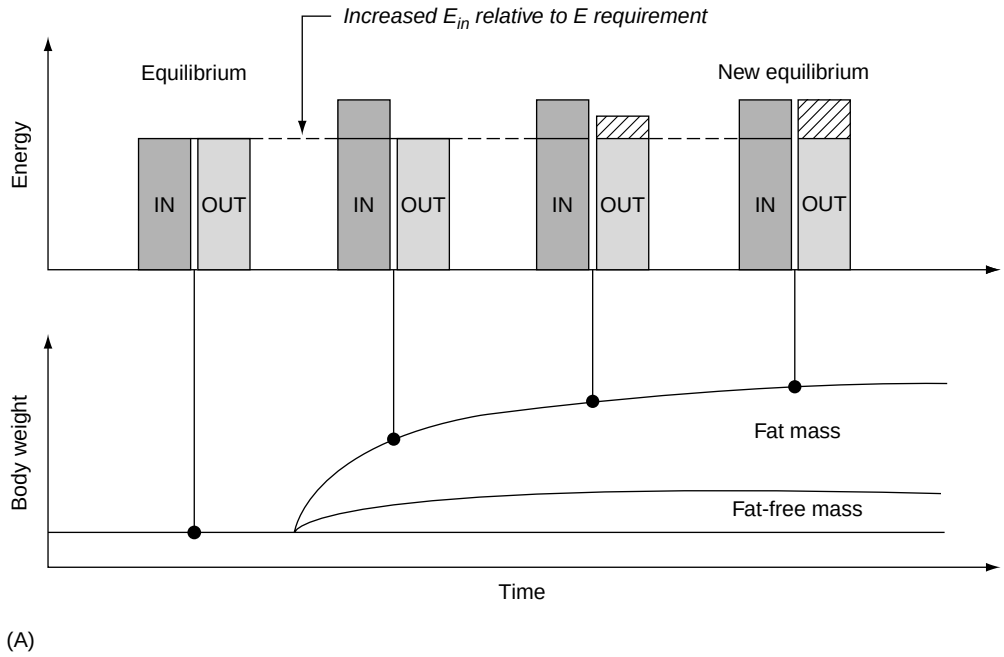
Adaptive changes in thermogenesis do attenuate the impact on energy balance of excessive or insufficient food consumption (as compared to requirement). The magnitude of adaptive thermogeneis varies as a function of the nature of excess substrates fed (protein is higher than carbohydrate and fat).

Energy Expenditure is Less Effective than Food Intake as a Control Mechanism of Energy Balance

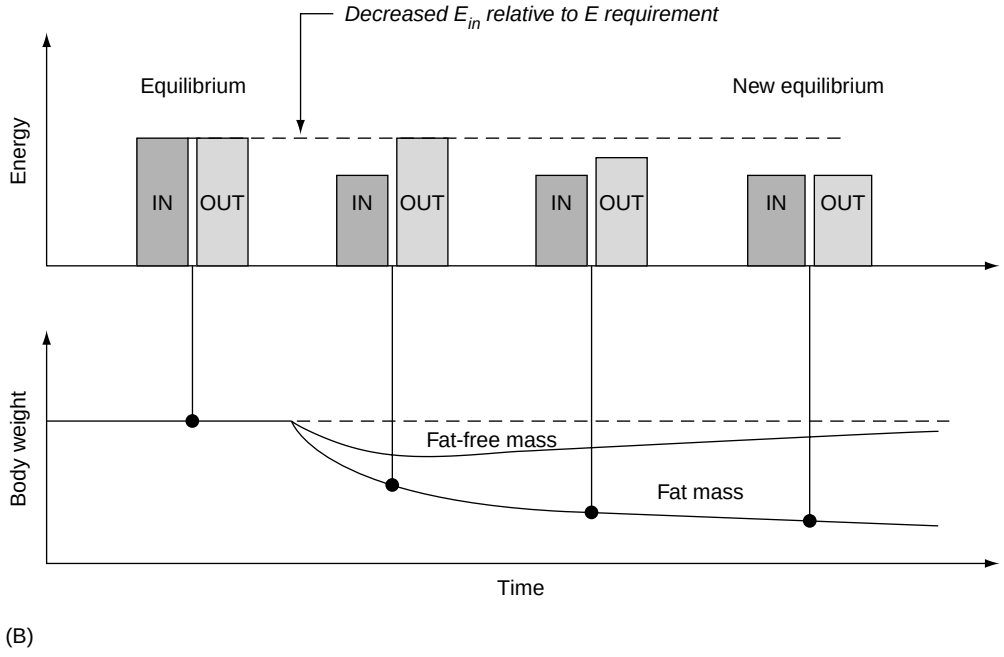
It should be stressed that the relationship between the change in energy intake below and above energy equilibrium and energy storage is not quite linear, indicating an increased net efficiency of energy utilization below energy maintenance and a decreased net efficiency of energy utilization above energy equilibrium (**Figure 8**).

Dynamics of Energy Balance with Overfeeding and Underfeeding (Figure 9)

To understand the dynamic aspect of energy balance while overfeeding is of the utmost important, since as mentioned previously the system is not invariant.



(A)



(B)

Figure 9 Dynamic change in energy balance following a step steady increase (A) overfeeding (or decrease) (B) underfeeding in energy intake. The time required to reach a new equilibrium in energy balance is very long (years) and depends upon the initial energy imbalance, the magnitude of adaptation of energy expenditure in response to change in energy intake, and on the factors related to the subject (obesity versus leanness). The figure shows that the static energy balance as such tells us nothing about the absolute level of energy intake and expenditure (see initial and final balance).

Continuous increase in energy intake above energy requirement will lead to a gradual gain in body weight. The size of the energy imbalance will progressively diminish with time as weight is gained. The reason for this is that the expansion of fat-free mass and fat mass (adipose tissue) will be accompanied by a rise in energy metabolism. A new equilibrium in weight is eventually reached after adaptation of each component of total energy expenditure, i.e., resting metabolic rate, diet-induced thermogenesis, and the increasing energy cost of supporting a heavier body weight. Note that each

kg of excess body weight increases total energy expenditure by about 20–25 kcal day^{-1}, and 10–12 kg day^{-1} when RMR is considered (**Figure 5**).

Let us take the following practical example: small increase in daily intake, e.g., of 100–200 kcal, will induce small increases in body weight with its associated rise in energy expenditure as the mass of lean tissue increases. If these changes occur on a daily basis, month by month, and if after 3–5 years the adult is still eating 200 kcal day^{-1} more than at baseline, they will now be heavier, and have a higher energy expenditure, and will come into energy balance; therefore, they cease to gain more weight.

Summary

- Energy balance is the difference between metabolizable energy intake and total energy expenditure. It is strongly related to macronutrient balances, and the sum of the individual substrate balances, expressed as energy, must be equivalent to the overall energy balance.
- Energy in foods is furnished by carbohydrate, proteins, fats, and alcohol; only 5–10% is lost through the feces and urine.
- The energy available to the body, called 'metabolizable energy', is on average 17 kJ g^{-1} of carbohydrate, 17 kJ g^{-1} of protein, 37 kJ g^{-1} of fat, and 29 kJ g^{-1} of alcohol. These figures vary slightly according to the type of carbohydrate, protein, or fat in the diet.
- The energy used in the body, or energy expenditure, is classically assessed by indirect calorimetry. It involves measuring the oxygen consumption and carbon dioxide production by an individual.
- Short-term regulation of energy balance is poor, but (in most people) long-term regulation is accurate. The mechanism is unknown, but must include conscious alterations in lifestyle to correct unwanted changes in body weight.
- During long periods of energy imbalance the weight gained (or lost) is initially glycogen plus water with an energy density of ~1.0 kcal g^{-1}. If the imbalance continues, after a week the tissue gained (or lost) is a mixture of mostly fat, water, and protein with an energy density of ~7 kcal g^{-1}.
- Undernutrition leads to a decrease in energy expenditure. Part of the decrease in metabolic rate is related to weight loss.
- In overfeeding, although some of the excess energy intake will be stored in adipose tissue, there are compensatory increases in energy expenditure.
- The changes in modern affluent society can be considered in terms of energy balance. The increasing prevalence of overweight in industrialized society can be attributed to a profound change in the pattern of physical activity due to increased mechanization, robotics, and computerization, which have substantially reduced the need for even modest physical activity. Today, the demand for heavy labor is rare. In developed countries increased car ownership and heavy road traffic result in fewer opportunities to travel on foot. Television-watching now fills a large proportion of leisure time and numerous gadgets minimize housework.

See also: **Amino Acids**: Metabolism. **Energy**: Metabolism; Requirements; Adaptation. **Energy Expenditure**: Indirect Calorimetry; Doubly Labeled Water. **Fats and Oils**.

Further Reading

Blaxter K (1989) *Energy metabolism in animals and man*. Cambridge: Cambridge University Press.

Ravussin E and Bogardus C (2000) Energy balance and weight regulation: genetics versus environment. *Br J Nutr* **83**(Suppl 1): S17–20.

Ravussin E and Bogardus C (1989) Relationship of genetics, age, and physical fitness to daily energy expenditure and fuel utilization. *Am J Clin Nutr* **49**(Suppl): 968–75.

Schutz Y (1995) Macronutrients and energy balance in obesity. *Metabolism* **44**(Suppl 3): 7–11.

Requirements

W P T James, International Association for the Study of Obesity/International Obesity Task Force Offices, London, UK

The metabolic rate of the body is the overall rate of tissue oxidation of fuels by all the body's organs. The dietary fuels are the carbohydrate, fat, protein, alcohol, and minor dietary components that are oxidized in the tissues, with oxygen being taken up by the lungs and the combusted end products (carbon dioxide, water, and urea) being excreted by the lungs, urine, and skin. The total rate of body metabolism is assessed by monitoring the rate of oxygen uptake by the lungs. The sources of fuel can then be estimated from the proportion of carbon dioxide produced and the rate of urea production.

The energy equivalence of oxygen varies depending on the precise nature of the fuels being oxidized, but a value of 20 kJ per liter of oxygen is taken as an appropriate average.

Factors Affecting Metabolic Rate

The process of oxidation involves a series of enzymatically controlled biochemical reactions leading eventually to the combination of oxygen with the carbon and hydrogen components of the body's fuels to yield the carbon dioxide and metabolically derived water. The incompletely oxidized nitrogen is excreted as urea, which is synthesized by the liver and excreted by the kidneys. The intermediate steps in the metabolism of the body's fuels are linked biochemically to drive the generation of phosphate-containing organic molecules, such as adenosine triphosphate (ATP), which in turn serve as the direct energy sources for all the body's cell activities, including the synthesis of complex molecules, the maintenance of tightly controlled ionic gradients in the cell, and the excretion of ions and molecules outside the cell. Thus, the oxygen being taken up by the lungs reflects the tissue metabolism of the fuels needed to regenerate the ATP used up in either biochemical 'internal' work or mechanical external work undertaken by the body's muscles. The rate at which the body burns its own stored fuels in the fasted, resting, and relaxed state (i.e., in the basal state in a warm room) is called the basal metabolic rate (BMR). This varies with the age of the individual, mainly because of the varying sizes of metabolically very different organs at different ages. Thus, a child has a relatively large brain, liver, and intestine, with a higher metabolic rate per kilogram of body weight than a more muscular adult. Body fat cells are metabolically active but contain a substantial amount of inert fat so that the larger fat mass of a woman means she has a lower BMR per unit body weight than a man. However, if the oxygen uptake is calculated in terms of the metabolically active fat-free mass, then her metabolic rate is the same. As men and women age, they tend to lose lean tissue and store extra fat, so the BMR on a weight basis decreases with age.

Equations can be used to estimate a group's BMR from their sex, age, and body weight (**Table 1**), but there is a range of BMR amounting to $\pm 20\%$ of the mean value at each weight. Thus, in a 25-year-old woman who weighs 55 kg, the anticipated mean BMR is 5460 kJ (1305 kcal) per day but may vary under normal conditions from 4448 to 6473 kJ per day. The kilojoule is the standard measure, and 4.184 kJ corresponds to 1 kcal, which was originally defined in energy terms as that required to increase the temperature of 1 g of pure water by 1 °C from 14.5 to 15.5 °C. The variation in BMR at a constant weight in part reflects differences in the fat content of individuals of the same weight. Thus, the BMR per unit fat-free mass varies by 12–15% rather than by $\pm 20\%$ as for weight. Approximately 40% of the BMR variation may be explained by differences in the size of the body's organs (e.g., liver, intestine, and muscle), but there is a residual difference between individuals that seems to be explicable only in terms of differences in the rate at which every organ of the body metabolizes its fuel. This in turn is controlled principally by the circulating concentration of thyroid hormones. Adults with a normal but above average level of thyroid hormones in their blood tend to have a BMR in the upper normal range. Smokers also have a BMR that is approximately 5% above normal, but whether this relates to changes in thyroid hormones is unknown. Young women who have normal menstrual cycles show a change in BMR that is at its lowest in the late follicular phase, just before ovulation. On ovulation, the basal body temperature rises rapidly by approximately 0.5 °C. The BMR is also increased but rises further to a peak in the later luteal phase, immediately before menstruation. This metabolic cycle with changes of $\pm 5\%$ of the mean is independent of changes in food intake, but the recognized 5–10% decrease in intake during the follicular phase with a similar increase in the luteal phase may accentuate the hormonally dependent change in metabolism. The effects of contraceptives that inhibit ovulation and the subsequent increase in basal temperature are unknown. The previous day's food intake does not affect the BMR unless there has been substantial overeating. However, the mixture of fuels combusted during fasting is influenced by the proportion of the previous 3 or 4 days' intake, which is derived from carbohydrate, with much of the glucose from glycogen being metabolized in the fasting state if carbohydrate intake was previously high. When the carbohydrate store of glycogen in the liver is nearing exhaustion, the body's output of carbon dioxide declines as the

Table 1 Equations for estimating basal metabolic rate (BMR) from body weight (kg)[a]

Age (years)	*Males (MJ/day)*	*Females (MJ/day)*
<3	BMR = 0.255 kg − 0.226	BMR = 0.244 kg − 0.130
3–9.9	BMR = 0.0949 kg + 2.07	BMR = 0.085 kg + 2.033
10–17.9	BMR = 0.74 kg + 2.754	BMR = 0.056 kg + 2.898
18–29.9	BMR = 0.063 kg + 2.896	BMR = 0.062 kg + 2.036
30–59.9	BMR = 0.048 kg + 3.653	BMR = 0.034 kg + 3.538
60–74	BMR = 0.0499 kg + 2.930	BMR = 0.0386 kg + 2.875
75+	BMR = 0.035 kg + 3.434	BMR = 0.0410 kg + 2.610

[a]The BMR values for infants and children are no longer used to calculate energy requirements but provide an indication of the likely values. The adult data are those from the original Schofield *et al.* analyses, with the data for 60-year-old and older adults derived from both the Schofield data and information provided to the UK Department of Health (1991).

body switches to using body stores of fat. The oxygen uptake for combusting the fatty acids continues since the demand for regenerating ATP is unaffected by the change in fuel supply, but a carbohydrate-rich diet tends to induce a slightly higher fasting metabolic rate than an energy-equivalent, fat-enriched diet, probably because of a slight induction of thyroid metabolism by dietary carbohydrates.

The BMR decreases 2–5% when individuals move to a tropical, warm environment, and in uninsulated houses seasonal changes in the BMR have been readily seen, with a 5–10% increase from summer to winter observed in the Japanese before World War II. The BMR formulas shown in **Table 1** ignore any temperature effects. The BMR of some people living in the tropics may be as much as 10% below the values shown, but these studies have been conducted on children and adults who are, or were, undernourished. Poor nutrition may have both an immediate and a long-term effect in lowering the BMR. Semistarvation leads to a decrease in BMR beginning on approximately day 4, and within 2 weeks the BMR can decrease by 15% as thyroid metabolism changes and the body's organs become more efficient. More prolonged or severe semistarvation leads to a progressive loss of the body lean tissues as well as fat, and the BMR therefore continues to decline in proportion to the loss of lean tissues. Body weight can eventually stabilize at a new low level, and if physical activity is also reduced, semistarved volunteers can return to energy balance on 50% of their initial intake. However, this requires a 40% loss of weight and marked lethargy if energy balance is to be preserved on such a low intake.

Components of Metabolic Rate

Traditionally, the metabolic rate is divided into three components: BMR, postprandial thermogenesis, and physical activity. The BMR usually comprises 50–60% of an individual's total energy expenditure and postprandial thermogenesis comprises 10%, which is used for the metabolic cost of processing (i.e., eating, absorbing, transporting, and storing food). The remaining energy is used for physical activity.

Postprandial Thermogenesis

The surge in oxygen uptake after a meal, known as postprandial thermogenesis, has been variously described as the specific dynamic action of food, dietary-induced thermogenesis, or the thermic effect of feeding. The last term is particularly favored by animal nutritionists. It is difficult to measure it accurately because after ingesting food with minimum physical effort, an individual has to lie at complete rest while oxygen uptake and carbon dioxide production are monitored for many hours until the metabolic rate has returned to the basal rate. This may take more than 10 h, which explains why BMR is measured after a 14-h fast. Separate feeding of different fuels shows that the maximum effect on oxygen uptake occurs after protein intake. This response is equivalent to approximately 30% of the protein's energy: Glucose induces a 5–10% effect; fat only a 2–5% effect, consistent with its slow absorption by the lymphatic tissue; and alcohol a 0–8% effect. Certain dietary components also increase metabolic rate; for example, a caffeine equivalent to two cups of tea increases metabolic rate by 1–3%, and spices, such as those found in an Indian curry, may increase it by 25% compared with a nonspiced meal. Moderate exercise amplifies the metabolic response to a standard meal so that the combined effect of exercise and food is greater than the sum of the response to each stimulus given separately. The effect, however, is small, amounting to 2% of the total energy expenditure.

Differences in postprandial energy expenditure have been sought as an explanation for the propensity of some individuals and animals to obesity. Results are often conflicting because in any person, the response tends to vary from day to day and is readily influenced by changes in gastric emptying. A proportion of obese subjects have a reduced metabolic response to a meal; this effect may depend on the degree of abdominal insulation since the response is reduced if volunteers are swathed in insulation to reduce abdominal heat loss, thereby increasing the temperature of the blood entering and leaving the liver. This seems to reduce the stimulus to body metabolism. Lactating mothers (and pregnant women) have a lower postprandial thermogenesis that returns to normal after they have stopped breast-feeding. Smoking and postprandial thermogenesis interact synergistically so the thermic output after a meal is enhanced. The small postprandial response during lactation is consistent with that observed in many species of animals in which brown adipose tissue is used as the organ for modulating heat production as a mechanism to maintain body temperature. However, this organ is not very active in humans.

Prolonged underfeeding and overfeeding lead to changes not only in postprandial thermogenesis but also in BMR. The effects of semistarvation when expressed per unit of tissue mass are modest, but overfeeding can produce a much greater response, provided that the intensity of overfeeding (especially

with carbohydrate) is high. Thus, progressive overconsumption of 6.3 MJ (1500 cal) per day leads to a 33% increase in daily energy expenditure. This composite response to more prolonged overfeeding is usually classified as dietary-induced thermogenesis. Nevertheless, this apparent mechanism for dissipating excess energy is limited because, at most, 27% of the excess intake is metabolised, and the remaining 73% is stored—two-thirds as fat and approximately one-third as lean tissue. The majority of the excess response in metabolism is accounted for by the theoretical cost of fat synthesis from carbohydrate, although the human capacity to transform carbohydrate into fat is limited, with preference being given to the selective storage of the fat component of the ingested energy.

Physical Activity

The energy cost of physical activity is predictable, but in order to obtain a reasonable estimate of energy expenditure on a daily basis, an analysis of activity patterns on a minute-by-minute basis is required. Children can be very active, making a detailed analysis difficult because the type of activity needs to be specified if an energy cost is to be assigned to each type of activity. Weight-bearing movement and antigravitational moves (e.g., walking up a hill with a load) are particularly expensive. The simplest way to estimate individual costs is to use the extensive tables on the energy cost of different movements in children and adults. For simplicity, these can be expressed as a ratio of the BMR since, in this way, differences between the sexes and individuals of varying size are removed. **Table 2** illustrates how this is achieved for adults with a moderately active lifestyle. The physical activity ratios (PARs) (i.e., energy costs in relation to BMR) can be assigned on the basis of extensive measures of similar activities. Those involved in sports medicine also call this unit cost a metabolic equivalent (MET), where 1 MET is equal to the resting metabolic rate of approximately 4.2 kJ (1 kcal) per minute. If the average energy requirement of the individual is to be estimated, account must be taken of the different types of work involved throughout the year. Each day is compartmentalized on a minute-by-minute basis, with a division often being made for convenience between occupational and other work. Maintenance of the individual's household varies depending on the day of the week and the season. Integrating all these activities, the ratio of the total daily energy expenditure to the BMR is designated the physical activity level (PAL). Physical activity is of general health benefit, so it is desirable that the overall PAL of individuals should be 1.75 or higher, which requires at least 60 minutes of moderately vigorous activity daily. In sedentary societies, however, 30 minutes of moderately vigorous exercise three times a week benefits muscular tone and physical fitness. This improves cardiovascular health and insulin sensitivity and therefore is likely to limit the development of type 2 diabetes in susceptible individuals. **Table 3** provides a listing of PAL values for adults of all ages according to their activity patterns. Thus, by knowing the sex, age, and body weight of individuals, it is possible to estimate their BMR (see **Table 1**). Given this BMR figure in millijoules per day, multiply by the PAL value shown in **Table 3**, and the energy needs can be estimated. Note that individuals vary in their energy needs by 10% at equivalent weights and activity levels, so an individual's needs cannot be predicted very accurately

Table 2 Calculating the appropriate energy requirements as physical activity levels (PALs) for adults with an active or moderately active lifestyle[a]

Main daily activities	*Time allocation (h)*	*Energy cost (physical activity ratio)*	*Time (h) × Energy cost*	*Mean PAL (multiple of 24-h BMR)*
Sleeping	8	1	8.0	
Personal care (dressing, showering)	1	2.3	2.3	
Eating	1	1.5	1.5	
Standing, carrying light loads[b] (waiting on tables, arranging merchandise)	8	2.2	17.6	
Commuting to/from work on the bus	1	1.2	1.2	
Walking at varying paces without a load	1	3.2	3.2	
Low-intensity aerobic exercise	1	4.2	4.2	
Light leisure activities (watching TV, chatting)	3	1.4	4.2	
Total	**24**		**42.2**	**42.2/24 = 1.76**

[a]If this PAL was from a female population, 20–25 years old, with mean weight of 57 kg and mean BMR of 5.60 MJ/day (1338 kcal/day), TEE = 1.76 × 5.60 = 9.86 MJ (2355 kcal), or 173 kJ (41 kcal)/kg/day.
[b]Composite of the energy cost of standing, walking slowly, and serving meals or carrying a light load.

Table 3 Classification of lifestyles in relation to the intensity of habitual physical activity or physical activity level (PAL)

Category	*PAL value*[a]
Sedentary or light activity lifestyle	1.40–1.69
Active or moderately active lifestyle	1.70–1.99
Vigorous or vigorously active lifestyle	2.00–2.40

[a]PAL ranges apply to both men and women.
Reproduced with permission from FAO Human Energy Requirements. Food and Nutrition Technical Report Series No. 1. (2004) FAO, Rome.

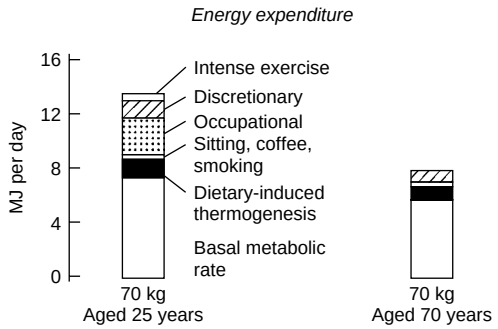


Figure 1 Different components of energy expenditure in a young man and the impact of ageing. Note that the energy expenditure decreases mainly because of a decline in physical activity. There is also a decrease in the basal metabolic rate, but this predominantly reflects a reduction in the lean tissues (i.e., fat-free mass) in the elderly.

unless his or her BMR is measured and account taken of his or her personal lifestyle (i.e., activity pattern). Nevertheless, the total energy expenditure of an individual child or adult is remarkably consistent from day to day, varying by only 1 or 2% provided that food intake and physical activity are meticulously standardized and account is taken in women of the stage of the menstrual cycle. The factors modulating BMR or the metabolic response to food are many, but their effect is small, so it is not surprising that energy expenditure is very predictable; the human body is thus a finely tuned and well-regulated machine. Abnormalities of regulation (e.g., in obesity) only arise because of a consistent discrepancy between the physiological controlled intake and expenditure, which, if discrepant by 1 or 2%, produces a 2- to 5-kg weight change in a year.

New measures of total energy expenditure estimated over 2 or 3 weeks can be obtained by using the double-labeled water technique, which relies on the difference in labeling of urine or saliva with the two heavy isotopes of water, deuterium and ^{18}O. The differential dilution of ^{2}H and ^{18}O in urinary water is monitored over a 2- or 3-week period following a single oral dose of $^{2}H_2\,^{18}O$. The ^{18}O content is diluted more rapidly than the ^{2}H because the oxygen in water exchanges rapidly with the body's bicarbonate pool, which turns over rapidly as carbon dioxide is produced by tissue metabolism. Thus, the difference in dilution rates of ^{18}O and deuterium provides a measure of the rate of carbon dioxide production. The technique is expensive and difficult to perform analytically but very convenient for the subject being studied since only single daily or occasional urine or saliva specimens are needed during the period of observation.

This method is increasingly used, and together in children with a heart rate monitoring system, has been used to revise the estimates of population energy requirements. These are accurate for groups of people but not for individuals, who vary not only in their physical activity patterns but also in their metabolic rate at rest. PAL estimates are less useful in infants because their energy requirement includes not only their total energy expenditure but also their cost of growth; BMRs are also difficult to measure at this age, so the $^{2}H_2\,^{18}O$ method together with estimates of growth costs have been used to compile new lower estimates of energy needs than were originally estimated by the old factorial method.

Age-related changes in energy needs are important not only in children but also in adults for different reasons. **Figure 1** gives an indication of the decline in energy needs during adult life. This results from the atrophy of the lean tissues, which may be related to a decline in physical activity. Lack of exercise is therefore a handicap because it directly reduces energy expenditure, and it may also lead to a slow shrinkage of tissues, such as muscle, thereby producing a long-term decline in metabolism at rest. There may also be up to an additional 5% decrease in the rate of tissue metabolism. Thus, unless people adapt their intake extraordinarily well to this progressive decline in energy output, energy storage, weight gain, and obesity are inevitable.

Extra Energy Costs

The cost of growth amounts to 10–25 kJ/g of new tissue deposited; the value is higher if fat with little lean tissue is laid down. A newborn has a high energy requirement of approximately 460 kJ/kg, with a cost of weight gain amounting to 26 kJ/g; however, by 1 year of age the total daily requirement decreases to approximately 335 kJ/kg as growth slows, with growth now costing 10 kJ/g. Breast-fed infants have an energy requirement approximately 10% lower than that of bottle-fed infants (**Table 4**).

Table 4 Average energy requirements of infants (breast and bottle fed) and of children up to age 18 years

Infants (months)	*Energy requirement (kJ/kg/day)*			*Children (years)*	*Weight (kg)*	*PAL*	*Energy requirement (kJ/kg/day)*	*Weight (kg)*	*PAL*	*Energy requirement (kJ/kg/day)*
	Boys	*Girls*	*Mean*		*Boys*			*Girls*		
1	475	445	460	1.1–2	11.5	1.43	345	10.8	1.42	335
2	435	420	430	2.1–3	13.5	1.45	350	13.0	1.42	337
3	395	395	395	3.1–4	15.7	1.44	334	15.1	1.44	320
4	345	350	345	4.1–5	17.7	1.49	322	16.8	1.49	309
5	340	345	345	5.1–6	19.7	1.53	312	18.6	1.53	299
6	335	340	340	6.1–7	21.7	1.57	303	20.6	1.56	290
7	330	330	330	7.1–8	24.0	1.60	295	23.3	1.60	279
8	330	330	330	8.1–9	26.7	1.63	287	26.6	1.63	267
9	330	330	330	9.1–10	29.7	1.66	279	30.5	1.66	254
10	335	330	335	10.1–11	33.3	1.71	270	34.7	1.71	242
11	335	330	335	11.1–12	37.5	1.75	261	39.2	1.74	229
12	335	330	335	12.1–13	42.3	1.79	252	43.8	1.76	217
				13.1–14	47.8	1.82	242	48.3	1.76	206
				14.1–15	53.8	1.84	233	52.1	1.75	197
				15.1–16	59.5	1.84	224	55.0	1.73	189
				16.1–17	64.4	1.84	216	56.4	1.73	186
				17.1–18	67.8	1.83	210	56.7	1.72	185

The energy requirements of infants were derived from double-labeled water measurements of total energy expenditure to which was added the age-specific energy deposited during growth, taking into account the different proportions of lean and fat tissue laid down during infancy. The children's requirements were estimated from quadratic equations relating body weight to total energy expenditure of girls and boys measured separately or from estimates of total energy expenditure based on calibrated heart rate recordings. Again, the energy deposited as growth was added to give a requirement expressed on a weight basis to allow adjustments for children of different weights at each age.PAL, physical activity level.

Without sufficient energy, a child will fail to grow, but the causes of growth failure usually relate to a deficiency of other nutrients or to infection rather than to a lack of dietary energy. Adolescents, particularly boys, who are physically very active may have a high demand for energy. However, the actual cost of even rapid growth rates at this age is modest.

Traditionally, pregnancy is considered, incorrectly, a time of great demand for food. Good nutrition is extremely important, and a weight gain in pregnancy of approximately 12 kg is desirable for reducing the risk of maternal and fetal complications and preterm and low-birth-weight infants and increasing the probability of delivering a 3.3-kg infant. With a weight gain of 12 kg, increases in maternal BMR are 5, 10, and 25% in the first, second, and third trimester of pregnancy, respectively. In practice, the intensity of physical activity often declines, particularly in late pregnancy, and some enhanced metabolic efficiency seems to occur. Thus, the increase in total energy expenditure amounts to only 1, 6, and 17% in the three trimesters, respectively. Therefore, the need for additional energy is small, amounting to 85, 350, and 1300 kJ/day (20, 85, and 310 kcal/day) for sequential trimesters, and in practice this means that a pregnant women needs to increase her food intake by 1.5 MJ/day (360 kcal/day) in the second trimester and 2.0 MJ/day (475 kcal/day) in the third trimester.

Lactation imposes a greater demand on mothers since their milk contains 1.9 MJ/day after birth, increasing to approximately 2.3 MJ/day on exclusive breast-feeding at 3 months. Extra energy is involved in making this milk, and the total extra energy demand is 2.6 MJ/day. Part of this additional energy derives from the extra fat stored by the mother during pregnancy, with the average, well-nourished women losing 0.8 kg/month, so the mother needs to eat approximately 1.9 MJ/day (450 kcal/day). This explains why mothers are more hungry when nursing their child than when pregnant. During lactation, there are no significant changes in BMR, efficiency of work performance, or total energy expenditure, and in most societies women resume their usual level of physical activity in the first month postpartum or soon thereafter.

Convalescent patients who need to gain weight require extra food, but the cost of this weight gain is 20–40 MJ/kg. If 1 kg is gained per month, the extra food needed amounts to approximately 1 MJ/day.

See also: **Adolescents**: Nutritional Requirements. **Breast Feeding**. **Children**: Nutritional Requirements. **Energy**: Balance. **Energy Expenditure**: Indirect Calorimetry. **Exercise**: Diet and Exercise. **Infants**: Nutritional Requirements. **Lactation**: Dietary Requirements. **Obesity**: Definition, Etiology and Assessment; Childhood Obesity; Complications. **Pregnancy**: Nutrient Requirements; Energy Requirements and Metabolic Adaptations. **Protein**: Requirements and Role in Diet.

Further Reading

Department of Health (1991) *Dietary Reference Values for Food Energy and Nutrients for the United Kingdom*, No. 41. London: HMSO.

FAO Human Energy Requirements (2004) Food and Nutrition Technical Report Series No. 1. FAO, Rome.

James WPT and Schofield EC (1990) *Human Energy Requirements*. Oxford: Oxford Medical.

Schofield WN, Schofield C, and James WPT (1985) Basal metabolic rate: Review and prediction. *Human Nutrition: Clinical Nutrition* **39C**: 1–96.

World Health Organization (1985) *Energy and Protein Requirements. Report of a Joint FAO/WHO/UNU Expert Consultation*, WHO Technical Report Series 724. Geneva: World Health Organization.

Adaptation

A G Dulloo, University of Fribourg, Fribourg, Switzerland
J Jacquet, University of Geneva, Geneva, Switzerland

Throughout much of evolutionary history, the mammalian species have been faced with periodic food shortages, specific nutrient deficiencies, and, sometimes, food abundance. Within such a lifestyle of famine and feast, it is conceivable that adaptive mechanisms—operating through adjustments in energy expenditure and in management of the body's main energy-containing compartments (fat and protein)—have evolved to the extent that they constitute key control systems in the regulation of body weight and body composition. These control systems and how they operate to enable the human body to adapt to nutritional stresses and to achieve body weight homeostasis are the focus of this article.

Beyond Adaptation through Mass Action

There is in fact a built-in stabilizing mechanism in the overall homeostatic system for body weight. Any imbalance between energy intake and energy requirements will result in a change in body weight that, in turn, will alter the maintenance energy requirements in a direction that will tend to counter the original imbalance and hence be stabilizing. The system thus exhibits 'dynamic equilibrium.' For example, an increase in body weight will be predicted to increase metabolic rate (on the basis of the extra energy cost for synthesis and subsequent maintenance of extra lean and fat tissues), which will tend to produce a negative energy balance and hence a subsequent decline in body weight toward its set or preferred value. Similarly, a reduction in body weight will result in a reduction in metabolic rate due to the loss in lean and fat tissues, which will tend to produce a positive balance and hence a subsequent return toward the 'set' or 'preferred' weight. In reality, however, the homeostatic system is much more complex than this simple effect of mass action since the efficiency of metabolism (or metabolic efficiency) may also alter in response to the alterations in body weight. Indeed, subjects forced to maintain body weight at a level 10% above their initial body weight showed an increase in daily energy expenditure even after adjusting for changes in body weight and body composition. Conversely, in subjects maintaining weight at a level 10% below the initial body weight, daily energy expenditure was also lower after adjusting for losses in weight and lean tissues. These compensatory changes in energy expenditure (~15% above or below predicted values) reflect changes in metabolic efficiency that oppose the maintenance of a body weight that is above or below the set or preferred body weight.

Interindividual Variability in Metabolic Adaptation

The experiments on forced changes in body weight have also revealed that there is a large interindividual variability in the ability to readjust energy expenditure, with some individuals showing little or no evidence for altered metabolic efficiency and others a marked capacity to decrease or increase energy expenditure through alterations in metabolic efficiency. Indeed, the most striking feature of virtually all experiments of human overfeeding (lasting from a few weeks to a few months) is the wide range of individual variability in the amount of weight gain per unit of excess energy consumed. Some of

these differences in the efficiency of weight gain could be attributed to interindividual variability in the gain of lean tissue relative to fat tissue (i.e., variability in the composition of weight gain), but most are in the ability to convert excess calories to heat—that is, in the large interindividual capacity for diet-induced thermogenesis (DIT). A detailed reanalysis of data from approximately 150 humans participating in the various overeating experiments conducted between 1965 and 1999 suggested that at least 40% of these overfed subjects must have exhibited an increase in DIT, albeit to varying degrees. Part of this variation in DIT could be explained by differences in the dietary protein content of the diet, with DIT being more pronounced in unbalanced diets that are low or high in percentage protein. As shown in **Figure 1**, in a subgroup of volunteers who were overfed on two occasions—once on a normal protein diet and once on a low-protein diet—

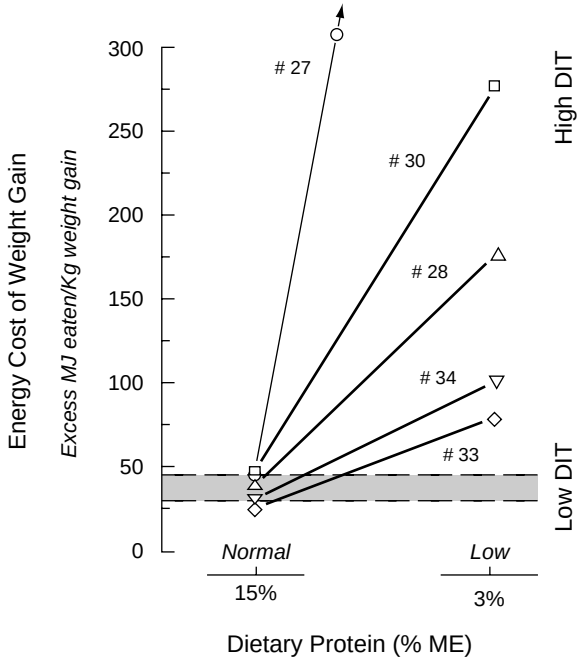


Figure 1 Interindividual variability in thermogenesis by low-protein overfeeding in humans. The data represent the energy cost of weight gain (excess MJ consumed per kilogram weight gained) during 3 or 4 weeks of overfeeding in the five human volunteers who participated in both the normal protein and low-protein overfeeding in the gluttony experiments of Miller and Mumford (1967). The horizontal broken lines (enclosing the shaded area) correspond to predicted energy cost of weight gain on the assumption that weight gain is either 100% fat (45 MJ/kg) or 60% fat (30 MJ/kg), the latter value including the cost of fat-free mass gain. The greater the deviation from the predicted values, the greater the likelihood that the excess calories were dissipated via enhanced diet-induced thermogenesis (DIT). Adapted from Dulloo AG and Jacquet J (1999) Low-protein overfeeding: A tool to unmask susceptibility to obesity in humans. *International Journal of Obesity* **23**: 1118–1121.

relatively small individual differences in DIT on balanced normal protein diet were amplified on the protein-deficient diet. That genes play an important role in variability in metabolism that underlies such susceptibility to weight gain and obesity has in fact been established from overfeeding experiments in identical twins. Conversely, a role for genotype in human variability in both the composition of weight loss (i.e., ratio of lean to fat tissue) and the enhanced metabolic efficiency (i.e., adaptive reduction in thermogenesis) during weight loss has been suggested from studies in which identical twins underwent slimming therapy on a very low-calorie diet. Taken together, it is evident that in addition to the control of food intake, changes in the composition of weight changes (via partitioning between lean and fat tissues) and in metabolic efficiency (via adaptive thermogenesis) also play an important role in the regulation of body weight and body composition, and the magnitude of these adaptive changes is strongly influenced by the genetic makeup of the individual.

What Constitutes Adaptive Thermogenesis?

The quantitative assessment of adaptive thermogenesis in the regulation of body weight and body composition is hampered by difficulties in determining which component(s) of energy expenditure may be contributing importantly to the changes in metabolic efficiency. As depicted in **Figure 2**, energy expenditure in the resting state is measured as basal metabolic rate (BMR) or as thermic effect of food (classically known as the specific dynamic action). Changes in the thermic effect of food (as percentage of calories ingested) or resting energy expenditure (after adjusting for changes in fat-free mass and fat mass) can be quantified, and they reflect changes in metabolic efficiency and hence in adaptive changes in thermogenesis. In contrast, any change in heat production from what is generally labeled nonresting energy expenditure is more difficult to quantify. The efficiency of muscular contraction during exercise is low (~25%), but that of spontaneous physical activity (SPA)—including fidgeting, muscle tone and posture maintenance, and other low-level physical activities of everyday life—is even lower since these essentially involuntary activities comprise a larger proportion of isometric work that is simply thermogenic. Since actual work done on the environment during SPA is very small compared to the total energy spent on such activities, the energy cost associated with SPA

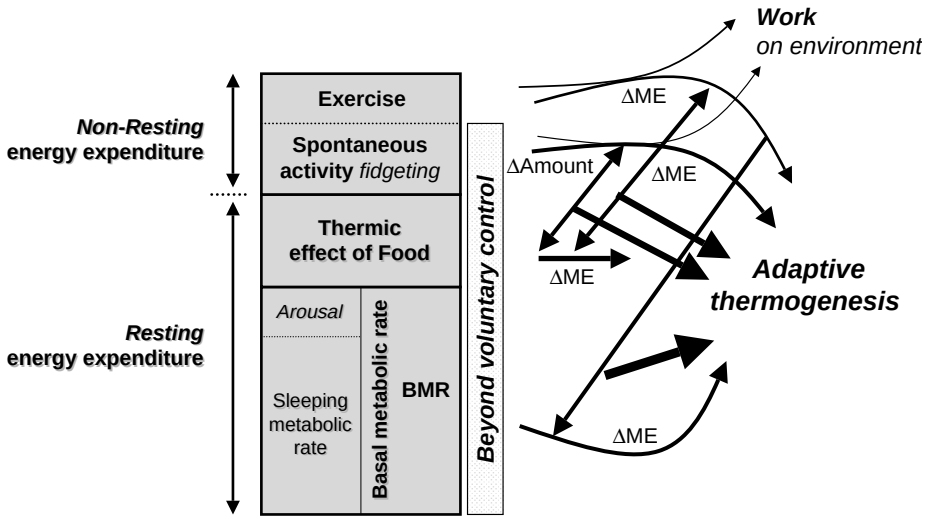


Figure 2 Schematic diagram showing the various compartments of human energy expenditure and how changes in metabolic efficiency (ΔME) both within and across these compartments can lead to adaptive changes in thermogenesis. The diagonal arrows depict possible *interactions* – between compartments of energy expenditure – that also constitute adaptive changes in thermogenesis; see text for details.

has been referred to as movement-associated thermogenesis or SPA-associated thermogenesis. It has also been argued that since SPA is essentially subconscious and hence beyond voluntary control, a change in the level or amount of SPA in a direction that defends body weight also constitutes autoregulatory changes in energy expenditure. In this context, an increase in the amount of SPA in response to overfeeding or a decrease during starvation also constitute adaptive changes in thermogenesis.

Spontaneous Physical Activity

The most direct evidence that changes in SPA contribute to autoregulatory changes in energy expenditure in humans derives from data obtained from the eight men and women who participated in the Biosphere 2 experiment. Biosphere 2 was a self-contained ecologic 'miniworld' and prototype planetary habitat built in Arizona. As a result of an unexpected shortage of food, their loss in body weight (8–25%) over a 2-year period was accompanied by a major reduction in SPA, which, like their reduced energy expenditure, persisted several months after the onset of weight recovery and disproportionate recovery of fat mass. Whether interindividual variability in the amount of SPA during overfeeding contributes to variability in resistance or susceptibility to obesity has also been the focus of a few human studies of energy expenditure. The importance of SPA-associated thermogenesis in human weight regulation was in fact underscored by the finding that even under conditions in which subjects were confined to a metabolic chamber, the 24-h energy expenditure attributed to SPA (as assessed by radar systems) was found to vary between 100 and 700 kcal/day and to be a predictor of subsequent weight gain. In fact, a main conclusion of early overfeeding experiments conducted in the late 1960s was that most of the extra heat dissipation in some individuals resisting obesity by increased DIT could not be accounted for by an increase in resting metabolic rate but could be due to an increased energy expenditure associated with simple (low-level) activities of everyday life. This notion has recently gained support from the findings that more than 60% of the increase in total daily energy expenditure in response to an 8-week overfeeding period could be attributed to SPA, and that interindividual variability in energy expenditure associated with SPA, referred to as nonexercise activity thermogenesis (NEAT), was the most significant predictor of resistance or susceptibility to obesity.

Efficiency of Muscle Work

In addition to changes in SPA, there is also evidence that changes in the efficiency of muscle work also contribute to adaptive changes in energy expenditure. Indeed in experiments of forced changes in weight in which subjects maintained body weight at 10% above or 10% below their habitual body weight, changes in muscle work efficiency could account for a-third of the change in daily energy expended in physical activity. These findings are

consistent with other reports of an increase in skeletal muscle work efficiency (i.e., decreased thermogenesis) after experimentally-induced weight reduction or in chronically undernourished subjects. Instead, changes in muscle work efficiency could account for one-third of the change in daily energy expended performing physical activity. These findings are consistent with other reports of an increase in skeletal muscle work efficiency (i.e., decreased thermogenesis) after experimentally induced weight reduction or in chronically undernourished subjects.

Interactions between Resting and Nonresting Energy Expenditure

It must be emphasized that the separation of adaptive thermogenesis between resting and nonresting is artificial, given the possibilities of their interactions illustrated in **Figure 2**. For example, energy expenditure during sleep, which is generally nested under resting energy expenditure, also comprises a nonresting component due to spontaneous movement (or SPA) occurring during sleep, the frequency of which seems to be highly variable between individuals. Furthermore, nonresting energy expenditure or NEAT may also include heat production resulting from the impact of physical activity (exercise or SPA) on postabsorptive metabolic rate or postprandial thermogenesis. There is evidence that relatively low-intensity exercise can lead to potentiation of the thermic effect of food, and that the effect of physical activity on energy expenditure can persist well after the period of physical activity (postexercise or post-SPA stimulation of thermogenesis). Reduction in postexercise stimulation of metabolic rate has also been proposed as a mechanism for energy conservation in individuals who are considered to be chronically energy deficient since childhood. Thus, any changes in metabolic efficiency in the resting or nonresting state that would tend to attenuate energy imbalance or to restore body weight and body composition toward its set or preferred value constitute adaptive changes in thermogenesis

Autoregulation of Body Weight and Body Composition

From a system physiology standpoint, the available evidence, based on classic longitudinal studies of starvation, refeeding, and overfeeding, suggests the adaptive mechanisms for optimal survival in an environment of famine and feast are embodied in three distinct autoregulatory control systems: the control of partitioning between protein and fat (the two main energy-containing compartments in the body) and two distinct control systems underlying adaptive changes in thermogenesis, as depicted in **Figure 3**. One control system is a direct function of changes in the food energy supply and responds relatively rapidly to the energy deficit. Its effector mechanisms are suppressed early during the course of starvation, and upon refeeding they are restored relatively rapidly as a function of energy reavailability and are activated further if hyperphagia occurs during refeeding, which may account for increased DIT. Because the efferent limb of this control system, which is primarily under the control of the sympathetic nervous system (SNS), is dictated not only by the dietary energy supply but also by a variety of other environmental factors, such as diet composition, specific nutrient deficiencies, ambient temperature, and psychological stress, it is referred to as the nonspecific control of thermogenesis. In contrast, the other control system has a much slower time constant by virtue of its response only to signals arising from the state of depletion/repletion of body fat stores; it is therefore referred to as the control system operating through an adipose-specific control of thermogenesis. The definitions of these two control systems underlying adaptive

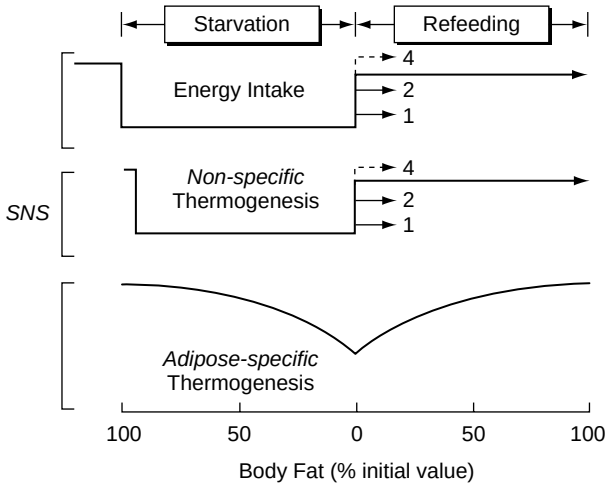


Figure 3 Schematic representation of the concept of two distinct control systems underlying adaptive thermogenesis during prolonged starvation and subsequent refeeding, showing *(i)* the *non-specific* control of thermogenesis, which is mediated primarily by the sympathetic nervous system (SNS) and which is a direct function of food energy supply, and *(ii)* the *adipose-specific* control of thermogenesis which is independent of the functional state of the SNS and which is a direct function of the state of depletion/repletion of the fat stores. Note that the different caloric loads during refeeding correspond to levels 1–4 of energy reavailability, with energy intake at level 1–2 below and level 4 above that prior to starvation, respectively. Adapted from Dulloo AG and Jacquet J (2001) An adipose-specific control of thermogenesis in body weight regulation. *International Journal of Obesity* **25**: (Supplement 5) S22–29.

thermogenesis are thus made on the basis of their differential commands—either deriving solely from the state of adipose tissue fat stores or not.

A Compartmental Model

An overall integration of these autoregulatory control systems in the regulation of body weight and body composition during a cycle of weight loss and weight recovery is discussed with the help of a schematic diagram presented in **Figure 4**. This diagram embodies the finding that the control of body energy partitioning between protein and fat is an individual characteristic—that is, individuals vary in their partitioning characteristic (Pc) during weight loss and weight recovery—and takes into account the two distinct control systems for adaptive thermogenesis that can operate independently of each other.

During starvation, the control of partitioning determines the relative proportion of protein and fat to be mobilized from the body as fuel—the individual's Pc being dictated primarily by the initial body composition. The functional role of the control of partitioning is to meet the fuel needs of the individual in such a way that the energy reserve component in both the fat and the protein compartments (i.e., the part that can be lost without death or irreversible damage) would reach complete depletion simultaneously—a strategy that ensures the maximum duration of survival for a given individual during long-term food scarcity. Furthermore, the energy conserved resulting from suppressed thermogenesis is directed at reducing the energy imbalance, with the net result that there is a slowing down in the rate of protein and fat mobilization in the same proportion as defined by the Pc of the individual. Indeed, the fact that the fraction of fuel energy derived from protein (i.e., the *P* ratio) remains relatively constant during the course of prolonged starvation, albeit in normal-weight humans, implies that neither control system underlying suppressed thermogenesis is directed at sparing specifically protein nor specifically fat but, rather, at sparing both protein and fat compartments simultaneously. Therefore, during starvation the functional role of both control systems underlying suppressed thermogenesis is to reduce the overall rate of fuel utilization (i.e., for energy conservation directed at sparing both lean and fat tissues).

During refeeding, the control of partitioning operates in such a way that protein and fat are deposited in the same relative proportion as determined by the Pc of the individual during starvation, and this serves to reestablish the individual's prestarvation capacity for survival during long-term food scarcity. Furthermore, the increased availability of food leads to the rapid removal of suppression upon the nonspecific (SNS-mediated) control of thermogenesis. In contrast, the suppression of thermogenesis under

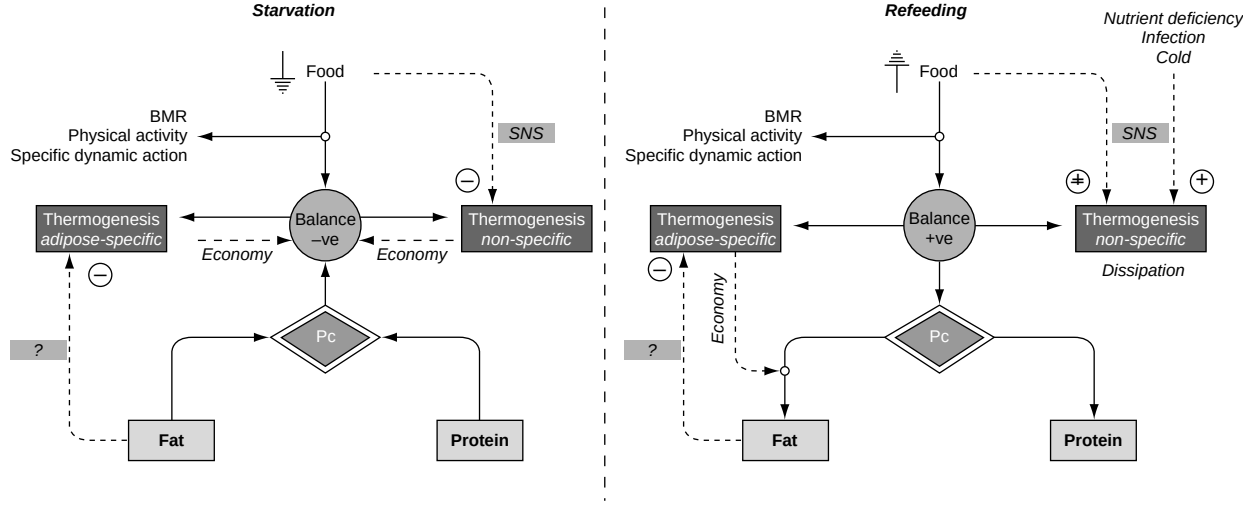


Figure 4 Schematic representation of a compartmental model for the regulation of body weight and body composition during a cycle of weight loss (prolonged starvation) and weight recovery (refeeding). In this model, the two distinct control systems underlying adaptive thermogenesis—the nonspecific control and the adipose-specific control—are integrated with the more 'basal' control of partitioning between the body fat and protein compartments as determined by the partitioning characteristic (Pc) of the individual. Adapted from Dulloo AG and Jacquet J (2001) An adipose-specific control of thermogenesis in body weight regulation. *International Journal of Obesity* **25**(supplement 5): S22–S29.

adipose-specific control is only slowly relieved as a function of fat recovery, such that the energy that continues to be spared is directed specifically at the replenishment of the fat stores. The net effect is that fat is deposited in excess of that determined by the Pc of the individual, thereby contributing to the disproportionate rate of fat relative to lean tissue recovery. This phenomenon of catch-up of fat (rather than catch-up of lean tissue) is often observed, both in adults after severe weight loss due to food unavailability and disease and in infants and children recovering from protein energy malnutrition and growth arrest.

Biological Significance

Such an adaptive phenomenon that accelerates the restitution of fat stores rather than diverting the energy saved toward compensatory increases in body protein synthesis (an energetically costly process) would have survival value in ancestral famine-and-feast lifestyle. This is because by virtue of the fact that body fat has a greater energy density and a lower energy cost of synthesis/maintenance than protein, it would provide the organism with a greater capacity to rapidly rebuild an efficient energy reserve and hence to cope with recurrent food shortage. Thus, the functional role of the adipose-specific control of thermogenesis during weight recovery is to accelerate specifically the replenishment of the fat stores whenever food availability is increased after a long period of food deficit and severe depletion of body fat stores. It provides an alternative mechanism to recover survival capacity in the absence of hyperphagia. However, equally important for the survival of mammals during weight loss and weight recovery is the need to retain the capacity to increase heat production (i.e., to activate thermogenesis) in response to a number of other environmental stresses, namely (i) for increased thermoregulatory needs in cold environments, (ii) for the generation of fever during exposure to infections, or (iii) for increased heat production as an adaptation to nutrient-deficient diets. The necessity to increase DIT in the face of nutrient-deficient diets probably had evolutionary survival advantage of 'homeostatic waste' because it enables individuals to overeat relatively large quantities of poor-quality food in order to obtain essential nutrients without the deposition of excess, nonessential energy as fat. Excessive weight gain would be a hindrance to optimal locomotion, hunting capabilities, and the ability to fight or flee. It has been proposed that DIT may have evolved as a means of regulating the metabolic supply of essential nutrients (protein, minerals, and vitamins) with only a secondary role in regulating energy balance and body weight. Whatever the exact functional significance of DIT, however, it is clear that in the context of weight recovery an elevated efficiency for catch-up fat can be shown to persist even under conditions of hypermetabolism (a net increase in thermogenesis) induced by hyperphagia or nutrient-deficient diets. To explain this apparent paradox, the model presented in **Figure 4** provides a structural framework that illustrates how suppressed adipose-specific thermogenesis that results in enhanced fat deposition during refeeding, and that is postulated to occur in the skeletal muscle, persists under conditions when the nonspecific control of thermogenesis is activated in organs/tissues recruited by the SNS (liver, kidneys, heart, and brown adipose tissue). Such differentially regulated control systems for thermogenesis may thus have arisen during the course of mammalian evolution as dual-adaptive processes that can satisfy the need for energy conservation during weight loss or for catch-up fat during weight recovery, even under environmental stresses when SNS-mediated activation of heat production has equally important survival values.

Energy Adaptation during a Longitudinal Human Study of Weight Fluctuations

The existence and operation of this dual-control system for adaptive thermogenesis are consistent with the temporal changes of BMR and body composition during the unique longitudinal study of semistarvation, refeeding, and subsequent overfeeding in men participating in a Minnesota experiment. The pattern of changes in food intake and body weight, together with kinetics of altered thermogenesis (assessed as changes in BMR adjusted for fat-free mass (FFM) and fat mass and expressed as a percentage of baseline BMR value) are presented in **Figure 5**.

During the phase of weight loss, the operation of the two control systems for adaptive thermogenesis is suggested by the fact that reduction in thermogenesis is biphasic in nature, with an initial rapid reduction in adjusted BMR at week 4, corresponding to 10% of baseline BMR, followed by a slower reduction in adjusted BMR, corresponding to 20 and 25% of baseline BMR at weeks 20 and 24, respectively. At the latter time points during starvation (at S20 and S24), the magnitude of reduced adjusted BMR was found to be associated with the reduction in fat mass (i.e., the greater the degree of depletion of the fat stores, the greater the suppression of thermogenesis).

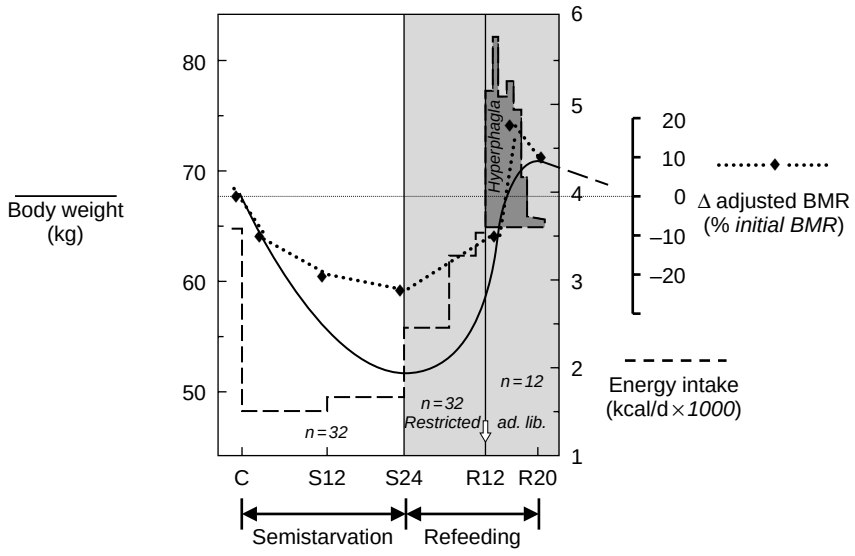


Figure 5 Pattern of changes in body weight, food intake, and adaptive thermogenesis during the various phases of the longitudinal 'Minnesota experiment' of human semistarvation and refeeding. The changes in adaptive thermogenesis at the various time points are assessed as changes in basal metabolic rate (BMR) after adjusting for changes in fat-free mass and fat mass and are expressed as a percentage of the baseline (control) BMR level. C, end of control (baseline) period; S12 and S24, week 12 and week 24 of semistarvation, respectively; R12 and R20, week 12 and week 20 after onset of refeeding, respectively. Data from Keys A, Brozek J, Henschel A, Mickelson O and Taylor HL (1950) *The Biology of Human Starvation*. Minneapolis: University of Minnesota Press; and Dulloo AG and Jacquet J (1998) Adaptive reduction in basal metabolic rate in response to food deprivation in humans: A role for feedback signals from fat stores. *American Journal of Clinical Nutrition* **68**: 599–606.

During the phase of weight recovery, the operation of the two control systems for thermogenesis is also suggested by the following:

1. The relation between the degree of depletion of fat stores and suppressed (adipose-specific) thermogenesis persists at week 12 of restricted refeeding, at which time point (R12) the mean adjusted BMR is still approximately 10% below baseline BMR level, the body fat is 80% recovered, and body weight and FFM recoveries are less than 50%.
2. After withdrawal of the dietary restriction during the subsequent period of ad libitum refeeding, the development of hyperphagia is accompanied by a prompt (perhaps SNS-mediated nonspecific) increase in thermogenesis, as judged by increases in adjusted BMR corresponding to approximately 20% of baseline BMR at week 14 of refeeding.

It is also noticeable that by week 20 after the onset of refeeding (R20), when FFM has been almost 100% recovered and body fat has overshot baseline (prestarvation) level by >70%—a phenomenon referred to as 'poststarvation obesity'—the adjusted BMR remains significantly higher (by approximately 10%) above baseline BMR despite the fact that hyperphagia is no longer present. This post-overfeeding sustained elevation of thermogenesis is consistent with a feedback mechanism existing between thermogenesis and body fat (i.e., the result of an activated adipose-specific control of thermogenesis), which may well have contributed to the subsequent slow return of body weight toward the baseline level after the phase of fat overshooting. Body fat was still higher than pre-starvation values when examined 33 weeks after the end of starvation, but was no longer significantly higher than pre-starvation values when examined 58 weeks after the end of starvation.

It should be noted that this study only enabled analysis of adaptive changes in thermogenesis in the BMR compartment since no measurements were performed pertaining to the thermic effect of food or to the energy cost of physical activity. The authors nonetheless observed that there was a profound decrease in SPA of the subjects, particularly during weeks S12 and S24 of semistarvation, thereby suggesting that adjustments in energy expenditure occurred in both resting and nonresting energy expenditure.

Energy Adaptation and Susceptibility to Leanness and Fatness

In addressing the issue of energy adaptation in human susceptibility to leanness and fatness, it must be noted that even in individuals who maintain a relatively stable lean body weight over decades, there is no 'absolute' constancy of body weight over

days, weeks, and years. Instead, body weight tends to fluctuate or oscillate around a mean constant value, with deviations from a set or preferred value being triggered by events that are cultural (e.g., weekend parties and holiday seasons), psychological (e.g., stress, anxiety, or emotions), and pathophysiological (ranging from minor health perturbations to more serious disease states). Very short-term day-to-day changes in body weight have a standard deviation of approximately 0.5% of body weight, whereas longitudinal observations over periods of between 10 and 30 years indicate that individuals experience slow trends and reversal of body weight amounting to between 7 and 20% of mean weight. In such a dynamic state within which weight homeostasis occurs, it is likely that long-term constancy of body weight is achieved through a network of regulatory systems and subsystems through which autoregulatory changes in food intake, body composition, and energy expenditure are interlinked.

The autoregulatory control systems—operating through adjustments in energy partitioning and through the two distinct control systems underlying thermogenesis—can play a crucial role in attenuating and correcting deviations of body weight from its set or preferred value. The extent to which these adjustments are brought about is dependent on the environment (e.g., diet composition) and is highly variable from one individual to another, largely because of the previous nutritional status of the individual and because of genetic variations. They probably conferred varying capacities to defend the body's protein and fat stores in an ancestral hunter–gatherer lifestyle of famine and feast, but they now underlie varying metabolic susceptibilities to fatness in societies in which palatable foods are abundant year-round. The resultant subtle variations between individuals in energy partitioning and in adaptive thermogenesis can, over the long term, be important in determining constancy of body weight in some and in provoking drift toward obesity in others.

Furthermore, the adaptive responses to starvation, so far discussed in the context of experimentation in normal weight (lean) individuals, also persist in individuals in whom obesity has developed spontaneously and contribute the defense of the obese state once acquired. In fact, longitudinal studies in obese humans losing weight in response to therapeutic slimming also indicate that they show a reduction in BMR (even after adjustments for losses of lean and fat tissues) as well as in SPA during both the dynamic phase of weight loss and subsequent weight maintenance. These findings support the notion that suppressed thermogenesis in response to food deprivation is a factor that reduces the efficacy of therapeutic regimens and contributes to obesity relapse. Furthermore, since the initial body composition (percentage of fat) is the most important determinant of energy partitioning between lean and fat tissue (i.e., Pc of the individual) during weight loss and weight recovery, the higher percentage body fat (i.e., the more obese the individual), the lower the fraction of energy mobilized as protein and hence the greater the propensity to mobilize fat during weight loss and to subsequently deposit fat during recovery. The low partitioning characteristic of the obese, coupled with sustained (adipose-specific) suppression of thermogenesis in response to their relative state of body fat depletion, will contribute to the relapse of obesity.

Acknowledgments

This work was supported by Grant 3200B0-102156 from the Swiss National Science Foundation.

See also: **Body Composition**. **Energy**: Metabolism; Balance. **Obesity**: Treatment. **Starvation and Fasting**. **Weight Management**: Weight Cycling.

Further Reading

Dulloo AG (2002) A sympathetic defense against obesity. *Science* **297**: 780–781.

Dulloo AG and Jacquet J (1999) The control of partitioning between protein and fat during human starvation: Its internal determinants and biological significance. *British Journal of Nutrition* **82**: 339–356.

Dulloo AG, Jacquet J, and Montani J-P (2002) Pathways from weight fluctuations to metabolic diseases: Focus on maladaptive thermogenesis during catch-up fat. *International Journal of Obesity* **26**(supplement 2): S46–S57.

Elia M, Stubbs RJ, and Henry CJK (1999) Differences in fat, carbohydrate, and protein metabolism between lean and obese subjects undergoing total starvation. *Obesity Research* **7**: 597–604.

Henry CJK, Rivers J, and Payne RR (1998) Protein and energy metabolism in starvation reconsidered. *European Journal of Clinical Nutrition* **42**: 543–549.

Hirsch J, Hudgins LC, Leibel RL, and Rosenbaum M (1998) Diet composition and energy balance in humans. *American Journal of Clinical Nutrition* **67**(supplement): 551S–555S.

Levine JA, Eberhardt NL, and Jensen MD (1999) Role of non-exercise activity thermogenesis in resistance to fat gain in humans. *Science* **283**: 212–214.

Luke A and Schoeller D (1992) Basal metabolic rate, fat-free mass, and body cell mass during energy restriction. *Metabolism* **41**: 450–456.

Miller DS and Mumford P (1967) Gluttony 1. An experimental study of overeating low- or high-protein diets. *American Journal of Clinical Nutrition* **20**: 1212–1222.

Shetty PS (1999) Adaptation to low energy intakes: The responses and limits to low intakes in infant, children and adults. *European Journal of Clinical Nutrition* **53**(supplement 1): S14–S23.

Stock MJ (1999) Gluttony and thermogenesis revisited. *International Journal of Obesity* **23**: 1105–1117.

Weyer C, Walford RL, Harper IT *et al.* (2000) Energy metabolism after 2 y of energy restriction: The Biosphere 2 experiment. *American Journal of Clinical Nutrition* **72**: 946–953.

ENERGY EXPENDITURE

Contents
Indirect Calorimetry
Doubly Labeled Water

Indirect Calorimetry

A Raman and D A Schoeller, University of Wisconsin–Madison, Madison, WI, USA

All living organisms require a source of energy for survival. Among animals, this energy is provided in the form of chemical energy in the nutrients they consume, which are converted to other forms of energy through respiration. This conversion is subject to the same laws of thermodynamics that govern all energy systems. The first law of thermodynamics states that energy can neither be created nor destroyed; it can only be exchanged from one system to another. Hence, the chemical energy consumed in the form of food is converted into mechanical energy for work performed by the body, thermic energy for maintenance of body temperature, or stored as chemical energy in tissues as fat, protein, or a small fraction as carbohydrates. This conservation of energy can be stated mathematically as

$$\text{Energy}_{\text{in}} = \text{Energy}_{\text{work}} + \text{Energy}_{\text{heat}} \pm \text{Energy}_{\text{stored}}$$

The sum of energy converted to work and heat is defined as metabolism. Although metabolism constitutes thousands of chemical reactions occurring at the same time throughout the body that cannot be individually measured, their sum can be measured as either the sum of work and heat energy or, in the absence of any measurable work, the rate of heat production by the body. This is based on the assumption that all the cellular events ultimately result in heat.

The process of measuring heat produced by the body during combustion of substances or nutrients in animals or humans is called calorimetry. The term 'direct calorimetry' is used when the rate of heat production is directly measured by placing a person in a thermally isolated chamber. The term 'indirect calorimetry' is used when heat production is not measured directly but is instead calculated from the measurement of the rates of oxygen consumption (V_{O_2}) and carbon dioxide production (V_{CO_2}). In both measurements, the rate of metabolism is commonly referred to as the rate of energy expenditure, which in the absence of work output is the rate at which chemical energy in food is converted to heat. The nutrients in food that provide this chemical energy are the macronutrients: carbohydrates, fat, protein, and alcohol. The chemical process that releases the chemical energy is respiration, in which each of these macronutrients is combined with oxygen to produce carbon dioxide and water. These chemical reactions are chemically equivalent to those that would be observed if the nutrient were combusted in a flame, except the reaction in the body is an enzymatic process that does not produce a flame.

For example, one molecule of sugar (glucose) breaks down as follows:

$$C_6H_{10}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O$$

It should be noted that during this reaction, six molecules of CO_2 are produced and six molecules of O_2 are consumed. Thus, the ratio of CO_2 to O_2 has a value of 1.0. This ratio is commonly called the respiratory quotient (RQ), although many investigators prefer the term respiratory exchange ratio (RER) when it is applied to a whole body measurement. Similarly, when one molecule of fat (tripalmitin) is broken down completely, the chemical reaction is

$$C_{57}H_{104}O_6 + 80O_2 \rightarrow 57CO_2 + 52H_2O$$

In the instance of fat oxidation, 57 molecules of CO_2 are produced while 80 molecules of O_2 are consumed when 1 molecule of fat is oxidized. This yields an RER of 0.71. When only carbohydrate and fat are being used to support energy expenditure, this difference in RER makes it possible to calculate what percentage of energy expenditure is being supported by each of the two energy substrates.

There is, however, a third macronutrient that is oxidized to produce energy. The third macronutrient, protein, is more difficult to describe on a chemical basis because a protein is made from a mixture of amino acids, and for each dietary protein the number and composition of amino acids differ.

The breakdown of the average dietary protein, however, can be described by the chemical reaction

$$C_{100}H_{159}N_{26}O_{32}S_{0.7} + 105.3O_2 \rightarrow 13CON_2H_4 + 87CO_2 + 52.8H_2SO_4$$

In the instance of protein oxidation, 87 molecules of CO_2 are produced while 105.3 molecules of O_2 are consumed when 1 molecule of protein is oxidized. This yields an RER of 0.83. Although this RER value is intermediate between carbohydrate and fat, protein is unique among the three energy substrates because it is the only one to contain nitrogen. As such, urinary nitrogen can be assayed to obtain an estimate of protein oxidized by an individual. Combining this with the knowledge that the average protein is 16% nitrogen by weight, it is possible to use the previous chemical relationship to calculate the O_2 consumption and CO_2 production that result from the oxidation of the protein represented by the urinary nitrogen. Subtracting these from the total respiratory gas exchange yields a nonprotein O_2 consumption, CO_2 production, and nonprotein RER. This is used to calculate the nonprotein metabolic rate and eventually the carbohydrate and fat oxidation rates. Because urinary nitrogen is often not measured, results from indirect calorimetry often use the Weir equation to calculate the energy expenditure. This equation was derived assuming that protein oxidation supports 12% of total energy expenditure (**Table 1**).

Over the years, different instrumental methods of indirect calorimetry have been developed to accurately measure V_{O_2} and V_{CO_2} rates. Despite being a precise and accurate method of measurement of macronutrient oxidation and hence energy expenditure, constraints such as expense, portability, gas collection issues, samplers, and applicability of measurements to habitual expenditure prevented it from being available to different types of research. Hence, the quest to develop new instrumental techniques is driven by the desire to make it a more generally applicable and easier to use technique.

Table 1 Formulas for calculation of energy expenditure

Variable	*Formula*
Oxygen consumption (ml/min)	= (Volume of inspired air per minute × fraction of inspired O_2) − (volume of expired air per minute × fraction of expired O_2)
Carbon dioxide production (ml/min)	= (Volume of expired air per minute × fraction of expired CO_2) − (volume of inspired air per minute × fraction of inspired CO_2)
Respiratory exchange ratio (RER or RQ)	$= V_{CO_2}/V_{O_2}$
Weir Equation (TEE, kcal/min)	$= (0.0039 \times V_{O_2}) + (0.0011 \times V_{CO_2})$ − (2.2 × urinary nitrogen, g/min)

Laboratory Methods

Whole Body Indirect Calorimetry

The advent of indirect calorimetry was a significant event in the history of animal and human nutrition. In whole body indirect calorimetry, the subject is kept in a sealed room or chamber, which is ventilated with a constant, measured supply of air. 'It is a setting similar to his habitual living and hence a more applicable measurement of energy expenditure.' The respiratory gas exchange of the subject is measured by the change in composition of the air going into the chamber and that of the air expelled from the chamber. Well-mixed samples of the chamber air are drawn to be analyzed for chamber air composition. The difference in O_2 and CO_2 composition of the incoming and outgoing chamber air is used to calculate the energy expenditure and macronutrient oxidation of the subject. Two main types of indirect calorimetry systems exist.

Closed-circuit indirect calorimetry involves the recirculation of the same air through the chamber. This can be performed by placing the subject in a sealed chamber. The recirculated air is kept breathable by removing the CO_2 produced by the subject and replacing the O_2 consumed by the subject. The replacement of O_2 is controlled by continuously monitoring the change in the volume of the gas in the closed breathing circuit. As the subject consumes O_2, a sensor detects the decrease in volume and a signal is sent to an external source to release constant calibrated pulses of O_2 back into the system to restore the original values. The rate of O_2 consumption is measured by recording the amount of O_2 that is added to the air during recirculation. The CO_2 produced by the subject is removed from the recirculated air by an absorber attached to the system and the CO_2 production is measured from the increased weight of the absorber (**Figure 1**).

Open-circuit indirect calorimetry involves a system in which both ends of the breathing system are open to the atmosphere. The inspired and expired air are kept separate by means of a three-way respiratory valve or non-rebreathing mask. The expired gases are collected into an air-tight bag or are frequently sampled or continuously analyzed for O_2 and CO_2 content.

These two terms are also used to describe some of the many other forms of smaller indirect calorimeters that have been developed over the years.

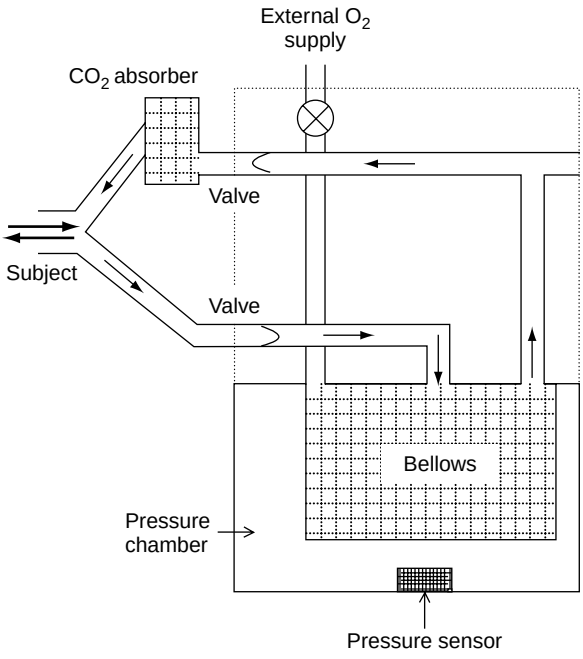


Figure 1 Closed-circuit metabolic chamber in which the subject's oxygen consumption is measured to calculate the corresponding energy expenditure. The change in volume of air in the system is constantly monitored by the sensors and a measured quantity of oxygen is added back to the system. Carbon dioxide is taken out of the recirculated air by a CO_2 absorber.

These can be categorized as laboratory or field techniques based on their portability. The instrumentation used for each varies in complexity and the degree to which they restrict the subject's movement.

Metabolic Carts

Metabolic cart is a common name for a semiportable respiratory gas analyzer that has been made small enough to be placed on a cart with wheels so that it can be rolled to different locations within a building. Two designs are generally available: the ventilated hood and the mouthpiece system.

The ventilatory hood system is an open-circuit indirect calorimeter that usually consists of a pliable plastic or rigid Perspex hood placed over the subject's head with a latex or thin plastic apron providing a rough seal around the neck or chest. These allow air to be drawn across a subject's face while in a reclining or lying position. For longer term measurements, ventilated plastic tents are available that cover all or part of the patient's bed. Since these hoods operate on a suction principle, a tight seal of the hood is not required. For field measurements, whole body transparent plastic ventilated boxes have been used successfully in infants. Many of the ventilatory hoods are constructed by researchers from the components according to the requirements of their study. The components include a pump, a flow meter, and a means of regulating the airflow. Samples of the air drawn from the hood can be directed to gas analyzers, which are usually connected in series to the hood. Respiratory gas exchange is calculated from the difference in O_2 and CO_2 concentration between the air entering and exiting the hood and the controlled rate of airflow (**Figure 2**).

Instruments have been developed to operate in adult and pediatric applications and differ with respect to flow rates and internal volume because of the smaller metabolic rate of children. The expired air enters a mixing chamber within the instrument to eliminate concentration variation resulting from inspiration and expiration before the sample enters O_2 and CO_2 sensor analyzers, which measure the concentration differences between the expired and inspired air. For state-of-the-art instruments, the data are input into a microprocessor providing minute-by-minute calculation of the O_2 consumption, CO_2 production, RER, and energy

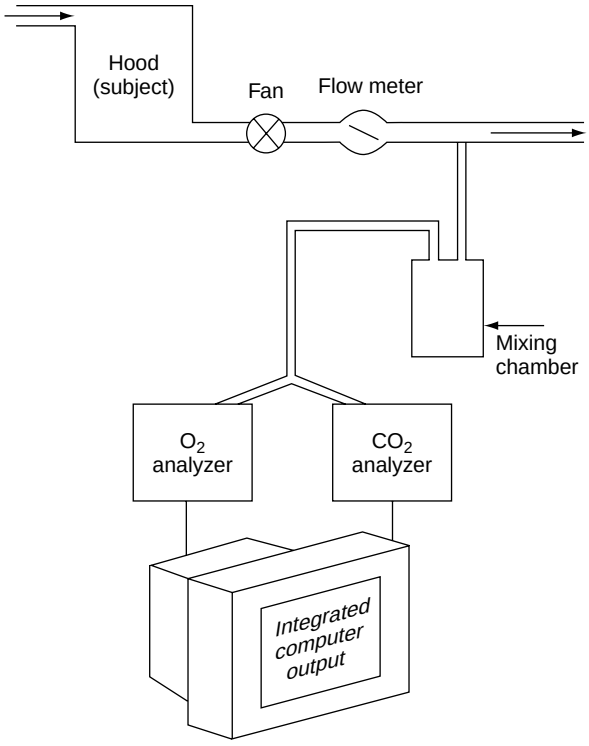


Figure 2 Ventilatory hood system showing a hood that is placed on the subject's head, a mixing chamber, and O_2 and CO_2 analyzers. A fan maintains a slight negative pressure in the hood to pull air into the chamber and also to prevent the escape of the expired air from the system. The air is mixed in the mixing chamber and is analyzed for oxygen and carbon dioxide by the respective analyzers. Results are calculated by the computer.

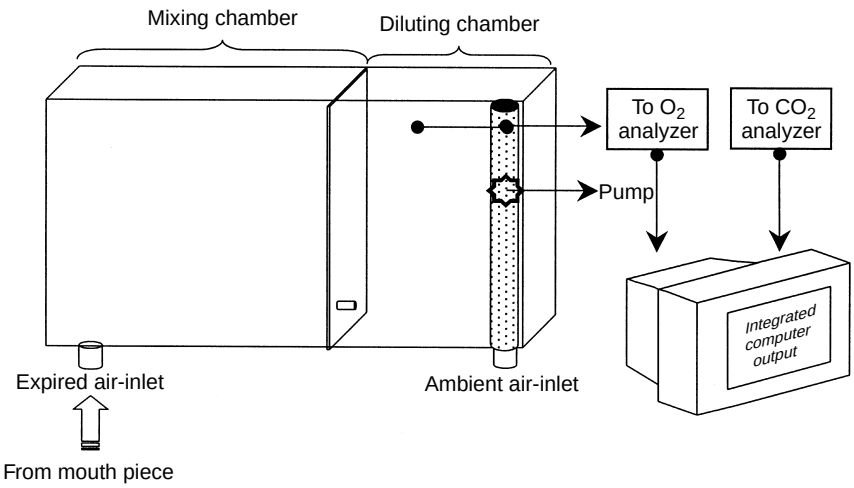


Figure 3 Metabolic unit measuring both O_2 consumption and CO_2 production rates during rest and exercise. In this type of system, expired air is diluted using ambient air before being analyzed by the respective analyzers.

expenditure. These instruments are generally used for measurements of subjects at rest as part of nutritional studies of energy expenditure and macronutrient utilization. These units can also be connected to mechanical ventilators for use in hospitalized patients.

Mouthpiece systems are similar to ventilated hood systems in principle, but instead of placing a hood over the subject's head, the subject wears a mouthpiece connected to the analyzer and nose clips to prevent breathing through the nose. The mouthpiece is connected to a valve system that allows the subject to breath in atmospheric air while directing the exhaled air to the gas analysis system. The expired breath is again subjected to analysis of O_2 and CO_2 concentration, but rather than passing the breath through a mixing chamber to smooth out the changes in concentration gradient of these gases from the start to end of an exhalation, the concentration profile is measured in real time along with the rate of gas flow from the exhalation. Again, the data are fed into a microprocessor for calculation of O_2 consumption and CO_2 production, but in this case the calculation is performed on a breath-by-breath basis. Results are averaged over time, usually provided as minute-by-minute averages of O_2 consumption, CO_2 output, and the rate of energy expenditure. The mouthpiece systems are generally used for studies of gas exchange and energy metabolism during exercise and provide a shorter measurement response time than the ventilated hood systems. The mouthpiece and nose clip used with some of the instruments make long time measurements highly cumbersome. Also, breathing through the mouthpiece often causes untrained subjects to involuntarily hyperventilate leading to inappropriate O_2 and CO_2 rates. It is also often difficult with mask systems to obtain an airtight seal without excessive pressure at the site of contact with the mask and face.

Different types of metabolic carts or monitors are available that are designed for various applications ranging from nutrition to exercise science. Most have built-in gas analyzers and data processing computers, making them highly user-friendly, handy tools for measurement of energy metabolism. They generally provide accurate and reliable data but do require periodic calibration. Ventilated hood systems often use a combination of gases with known concentrations and weighed ethanol or methanol burns for such calibration, whereas breath-by-breath systems use a combination of large volumetric syringes and gases of known O_2 and CO_2 concentration (**Figure 3**).

Field Methods

As for whole body indirect calorimetry, ambulatory and portable systems measure the respiratory gas exchange with the V_{O_2} and V_{CO_2} measurements. Ambulatory methods and less refined laboratory methods often dispense with the measurement of CO_2 to avoid the need for two gas analyzers. The error incurred by assuming a CO_2 production rate is several percentage points, which researchers are prepared to compromise on. When only O_2 consumption is measured, however, it is not possible to compute macronutrient-specific oxidation rates. The accuracy of ambulatory and portable methods is generally between +4 and −2%. Field methods involve the collection of expired air over a fixed period of time as in the Douglas bag or small on-line analysis systems that sample inspired and expired air through a mouthpiece.

Douglas Bag/Tissot Tank

The Douglas bag method is a classical example of collection of expired air to measure energy expenditure in the field during both rest and physical activity. It consists of a gas-impermeable bag with a capacity of ~100 l or a Tissot tank suspended over water, which is used to store the subject's expired air over a fixed, short time interval. A classic Douglas bag is made up of either a rubber sheeting cemented by two layers of canvas or plastic material lined by PVC or aluminum with welded seams. The rubber bags have leakage of CO_2 by diffusion, which is unavoidable, but PVC and metalized bags hold better. If the bags are filled to capacity and analyzed with 20 min of collection, the effects of diffusion can be minimized. The subject wears a nose clip and mouthpiece or a face mask. Outside air or its equivalent is inhaled through the mouthpiece or mask containing a one-way valve and exhaled into a Douglas bag or Tissot tank for a precise period of time. It is important that the mouthpiece and connecting tubing provide minimal resistance to airflow, or the cost of breathing will increase the energy expenditure. Ambient temperature, barometric pressure, and relative humidity are recorded for converting values under conditions of standard temperature and pressure. The volume of air collected in the bag or tank is measured and a sample of exhaled air is obtained to measure the O_2 and CO_2 concentrations using gas analyzers. The volume of oxygen consumed and carbon dioxide exhaled are calculated by analyzing the gas from the Douglas bag for the precise time period during which it was collected. This method is relatively simple and inexpensive yet gives reliable results. It is suitable only for short durations of field measurement, and wearing the mask and nose clip for the whole duration of the study may be cumbersome, interfere with daily activities, and is socially undesirable to the subject.

Spirometers were used in the past for measurement of the volume of the respired air. With the advent of continuous flow electronic analyzers and superior gas flowmeters, spirometers are now rarely used. Ambulatory methods also consist of a mouthpiece incorporating light action-sensitive but robust one-way gas valves, corrugated tubes, and three-way taps. The volume of air respired and the relative concentrations of O_2 and CO_2 in the expired air are measured using O_2 and CO_2 gas analyzers. These small analyzers have replaced the Haldane system or micro-Scholander chemical gas analyzers, which used reagents to absorb the CO_2 and O_2, with the weight of absorbents measured before and after the gases were absorbed.

Max Plank/Kofranyi–Michaels Respirometer

A Max Plank respiration gas meter is a small, compact, and lightweight backpack-mounted respirometer. It combines a gas volume meter and a sampling device for continuous sampling of each breath of expired air. The Max Plank respirometer consists of a dry, bellow-type gas meter for measuring the total volume of expired air during activity. The subject breathes through a low-resistance valve and the expired volume is monitored. A measured quantity of expired air is removed continuously (0.3 or 0.6%) by an aliquoting device to be sent to a small butyl rubber bag. This rubber sampling bag can be connected directly to the oxygen analyzer, eliminating the need for transfer of samples to gas-tight syringes for analysis. The respirometer is suitable for flow rates between 15 and 50 l/min or for periods of 110 min on a slow flow rate and 55 min on a faster rate. It is smaller, more compact, and lighter than the Douglas bag apparatus and can be used in studies involving light to moderate physical activity. Although the system has a low resistance, at higher ventilation rates the resistance increases substantially and hence cannot be used in higher flow rate scenarios. Also, this can be used in studies of shorter duration only. Due to the use of mouthpiece and nose clip, prolonged usage may cause discomfort to the subjects.

Telemetry Systems

The K2 system was the first of a series of portable systems that consists of a soft face mask with a turbine flowmeter attached to it. A transmitter and battery are attached to a chest harness, which transmits signals to a receiver unit. The flowmeter measures the rate of airflow, calculates the volume of expired air per minute, and counts the number of expirations per minute. A small capillary tube passes through to the transmitter unit, which contains an electrochemical gas analyzer used to measure the concentration of oxygen in expired air. The signals from this analyzer are transmitted to the receiver unit by the portable transmitter unit. The receiver unit processes the data and prints it in a desired format. The electrochemical gas analyzer is a polarographic electrode. It has a membrane through which oxygen permeates into an electrolyte solution generating an electrical impulse proportional to the rate of oxygen permeation through the membrane.

Since these systems are portable and easy to use, they have many potential uses in exercise science studies and rehabilitation medicine. They allow a breath-by-breath pulmonary gas exchange

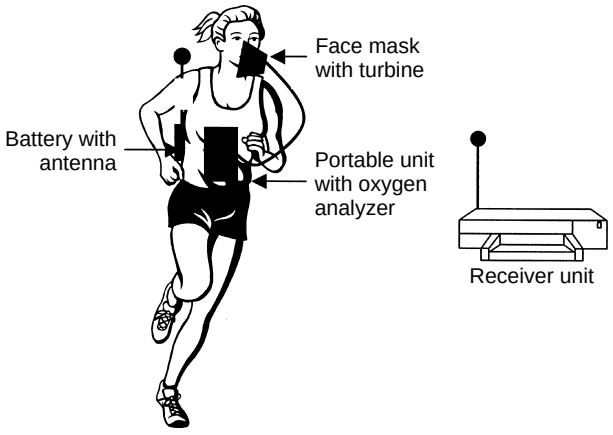


Figure 4 Telemetry system with a face mask attached to a turbine flowmeter, a transmitter, and a receiver unit. The flowmeter measures the rate and volume of airflow and the expiratory cycles per minute. The expired air is analyzed for oxygen concentration by an oxygen analyzer in the transmitter unit. The transmitter then transmits the signals to the receiver unit, which integrates the data and prints the results.

measurement while still being very light and portable, enabling a direct field assessment of human performance and cardiopulmonary limitations. The low-resistance flowmeter allows a wide range of oxygen flow rates to be measured, through these systems face the issue of air leakage from the face masks when subjects are made to exercise at high intensities. The measurement durations usually are limited to 1–5 h. The polarographic electrode membrane is known to have a short life span and hence monitoring of the usage of the instrument is essential. If CO_2 concentrations are essential for a study, this is not a good instrument to use (**Figure 4**).

Tracer Methods of Indirect Calorimetry

These are a third category of techniques that have gained popularity among investigators during the past two decades. These techniques provide a measure of CO_2 production through the use of dilution techniques using isotopic tracers.

Labeled Bicarbonate

A constant infusion-labeled bicarbonate method is useful in estimating the net CO_2 production and hence energy expenditure in animals and humans. This method is based on an isotopic dilution technique whereby the administered label is diluted by the CO_2 produced endogenously by the body. The extent of this isotope dilution is used to measure the rate of CO_2 production and is used to estimate the energy expenditure of the individual. A microinfusion of ^{13}C- or ^{14}C-labeled bicarbonate is given to an individual and the specific activity or enrichment of his or her physiological fluids, especially breath or urine, are measured to estimate the rate of label elimination and hence the rate of endogenous CO_2. Thus, variation in the endogenous CO_2 production rate will be reflected in the dilution of the body pool and consequently in the breath samples.

These measurements are accurate when energy expenditures are measured over a longer duration of time (>1 day) but are subject to effects of label sequestration over shorter periods. Sequestration refers to trapping, or fixation, of the label in tissues that utilize bicarbonate/CO_2 for their metabolic functions. Shorter duration of collection of breath samples requires a correction for the fraction of label that is sequestered. This is based on the assumption that similar amounts of label are sequestered in various individuals. When breath samples are collected over longer durations, the sequestration is often assumed to be negligible.

Some investigators have used a bolus bicarbonate administration rather than the continuous infusion. These investigators measured the rate at which the label concentration decreases with time as a measure of CO_2 turnover and the initial concentration as a measure of the body's bicarbonate pool size. Taken together, these provided a measure of energy expenditure during a short period of constant physical activity.

Doubly Labeled Water

This is an isotope dilution technique wherein deuterium and heavy oxygen-labeled water (doubly labeled water, DLW) are given to individuals and timed urine samples are collected to measure the elimination rates of ^{2}H and ^{18}O in the urine. ^{2}H label from DLW mixes with the body water and is eliminated as water in the urine. Similarly, ^{18}O label from DLW is eliminated as water, but it is also utilized in bicarbonate synthesis and hence is also eliminated in the breath as CO_2. The difference in turnover rates of isotopic ^{2}H-H and ^{18}O-labeled water is proportional to CO_2 production. Energy expenditure, oxygen consumption, water intake, and metabolic water production can be calculated using standard indirect calorimetry equations with an estimated RER (**Figure 5**).

In practice, a measured dose of DLW is given to the subject whose energy expenditure is to be measured. Body water samples, such as blood, urine, saliva, or breath water, are collected before dosing and after equilibrium is attained. The isotopic disappearance rates of ^{18}O and ^{2}H as CO_2 in breath or

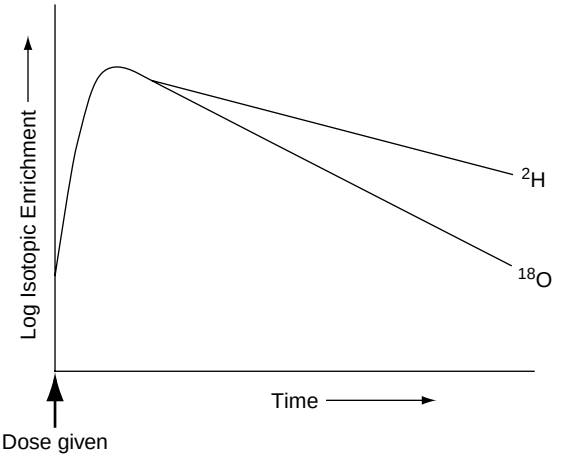


Figure 5 Time course on log scale for the enrichments of the stable isotopes 18-oxygen and deuterium when administered to the subject. Both the tracer enrichments increase rapidly in the body water pool until they reach distribution equilibrium (2–4 h). The enrichments then start to decline as the body water turns over during metabolism. 18-Oxygen is eliminated at a faster rate because it is excreted as water and CO_2 in breath, whereas deuterium is eliminated as water only. The difference in elimination rates of these two tracers is proportional to the rate of CO_2 production by the subject.

H_2O in urine, saliva, or breath water, respectively, are determined from the change in isotopic enrichments of the before dosing and after equilibrium samples.

The doubly labeled water method is both simple and noninvasive. It has been validated in various animals and humans, with the CO_2 production rate showing a mean measurement error of less than 5%. Unlike the majority of the other methods, the doubly labeled water method provides a measure of average energy expended over a period of 3–21 days without restricting the subject's movement and thus provides a better estimate of habitual energy expenditure than the other methods. The doubly labeled water method, however, does not provide any information on the pattern or intensity of any one activity during that time but the overall average energy expenditure. This method is also expensive due to the cost of the ^{18}O and it does require sophisticated mass spectrometric analyses.

Summary

Indirect calorimetry is a noninvasive, reliable, and valuable tool in assessing energy expenditure, evaluating fuel utilization by the body. It has been used extensively for both scientific investigation and medical evaluation and care. Scientists from various fields have used it effectively to measure energy expenditure, establish nutrient requirements, measure physical fitness, and evaluate macronutrient utilization during exercise and rest. Clinicians have used indirect calorimetry to optimize the nutritional support in metabolic disorders as in parenterally fed patients and to quantify the energy expenditure in mechanically ventilated patients. Indirect calorimetry is a reliable, convenient, and accurate diagnostic and prognostic tool in experimental and clinical settings. Indirect calorimetry has such universal appeal because animals and humans derive their energy for sustenance by transforming the chemical energy from the nutrients they consume to heat through respiration, and their existence depends on their ability to balance energy intake and expenditure.

See also: **Energy**: Metabolism; Balance; Requirements.

Further Reading

Elia M, Fuller NJ, and Murgatroyd PR (1992) Measurement of bicarbonate turnover in humans: Applicability to estimation of energy expenditure. *American Journal of Physiology* **263**: E676–E687.

Headley JM (2003) Indirect calorimetry. *AACN Clinical Issues* **14**(2): 155–167.

Jequier E, Acheson K, and Schutz Y (1987) Assessment of energy expenditure and fuel utilization in man. *Annual Review of Nutrition* **7**: 187–208.

Macfarlane DJ (2001) Automated metabolic gas analysis systems. *Sports Medicine* **31**(12): 841–861.

Molnar JA, Cunnigham JJ, Miyatani S *et al.* (1986) Closed-circuit metabolic system with multiple applications. *Journal of Applied Physiology* **61**(4): 1582–1585.

Murgatroyd PR, Shetty PS, and Prentice AM (1993) Techniques for the measurement of human energy expenditure: A practical guide. *International Journal of Obesity* **17**: 549–568.

Peel C and Utsey C (1993) Oxygen consumption using the K2 telemetry system and a metabolic cart. *Medicine and Science in Sports and Exercise* **25**(3): 296–400.

Schoeller DA and Webb P (1984) Five-day comparison of the doubly labeled water method with respiratory gas exchange. *American Journal of Clinical Nutrition* **40**(1): 153–158.

Simonson DC and DeFronzo RA (1990) Indirect calorimetry: Methodological and interpretive problems. *American Journal of Physiology* **258**: E399–E412.

Doubly Labeled Water

W A Coward, MRC Human Nutrition Research, Cambridge, UK

Like methods for the measurement of energy expenditure by respiratory gas analysis, the doubly labeled water (DLW) method is indirect. The disappearance

of stable isotope tracers, given orally, is used to model water and water plus carbon dioxide turnover. Carbon dioxide production rate is then estimated by difference and energy expenditure calculated from it. In practice, this means that subjects merely drink labeled water, samples of body water (e.g., urine, saliva, or blood) are collected over a few days, and these are then passed to the laboratory for tracer analysis and calculation. The method is thus uniquely objective; it is noninvasive and nonrestrictive in that its application does not interfere with normal lifestyles and comparable results can in principle be obtained in any circumstances without subject or observer influence. Complex measurement techniques do not need to be exported to the site where the subjects are located. However, underlying the apparent simplicity are concepts and techniques that are not commonly tools of trade for many potential users of the methodology. In a complete review, these, as well as method practice and results, need to be explained.

Method Fundamentals

Stable Isotopes as Tracers

Although radioactive tracers are familiar tools, the use of tracer elements and compounds to measure metabolic processes was developed first with stable isotopes in the late 1930s by Schoenheimer and Rittenberg soon after ^{2}H and ^{15}N (both stable isotopes) became available. Unlike radioactive isotopes, which are largely man-made, unstable, and decay to other elements, stable isotopes do not decay and are ubiquitous. Virtually all elements exist in nature in at least two stable isotopic forms with the same numbers of electrons and protons but with differing numbers of neutrons in the nucleus. The level of a specific isotopic form in nature is called its natural abundance. For tracer experiments, an element or a simple compound containing it, enriched with one of the isotopes, is prepared by mass-dependent separation on an industrial scale. This is then incorporated into the substrate of interest for biological experiments. In the current context, $^{2}H_2O$ (deuterium oxide, heavy water) is readily available from the electrolysis of water. Water enriched with ^{18}O is prepared directly by fractional distillation or from nitric oxide after its cryogenic distillation.

No radioactivity is involved in the use of stable isotopes in human experiments; thus, the only effects that have to be considered in relation to risk to the subject are related to the physical properties of the isotopic labeled compound. There is inevitably some degree of isotopic discrimination in physical and enzymatic processes, but because stable isotopes are normally present in all biological material at natural abundance levels, the relevant consideration is only by how much and for how long amounts are changed in experimental procedures. Because highly precise measurement techniques are used, it is necessary only to increase isotopic enrichments in body water from natural abundance by very small amounts. In a typical experiment, ^{2}H enrichment might be increased from 150 to 300 parts per million (ppm) and ^{18}O from 2000 to 2400 ppm, and a return to natural abundance levels will occur with a biological half-life of 5–7 days. There is no evidence that amounts many times larger than these have any harmful effects.

Measuring Isotopic Enrichment

Mass spectrometry is a generic name for a family of methodologies in which compounds are ionised and separated on the basis of mass:charge ratio. The method of choice for the measurement of isotopic enrichment with sufficient precision for DLW experiments is isotope ratio mass spectrometry. This technique is applicable only to relatively simple molecules. It separates ions such as $[^{2}H-^{1}H]^{+}$ and $[^{1}H-^{1}H]^{+}$ (mass 3 and 2) or $[^{12}C^{16}O^{18}O]^{+}$ and $[^{12}C^{16}O^{16}O]^{+}$ (mass 46 and 44) and measures isotopic ratios (R) relative to an international standard, such as Vienna Standard Mean Ocean Water (V-SMOW; **Table 1**). For the DLW method, therefore, the isotopic enrichment in water from biological samples has to be measured as hydrogen or carbon dioxide. For hydrogen isotope analysis, a variety of methods have been used for the conversion including reduction by reaction with hot uranium or zinc, but these methods are difficult to automate. Currently favoured methods are the exchange of hydrogen in the water sample with gaseous hydrogen by equilibration in the presence of a platinum catalyst or reduction with hot chromium. Both of these techniques are automated in commercially available equipment. For oxygen isotopes, samples are usually equilibrated

Table 1 Typical isotopic ratios and equivalent enrichments measured in DLW experiments[a]

Sample	*^{2}H isotope ratio (ppm)*	*^{2}H enrichment (‰)*	*^{18}O isotope ratio (ppm)*	*^{18}O enrichment (‰)*
V-SMOW	155.76	0	2005.2	0
Background	152.28	−22.34	1995.74	−4.72
Postdose	342.67	1200	2305.98	150

[a]Enrichment $= 10^{3}\left(\frac{R_{sample}}{R_{V-SMOW}} - 1\right)$.

V-SMOW, Vienna Standard Mean Ocean Water.

with carbon dioxide with exchange of oxygen between the water and carbon dioxide. This procedure is also automated.

Single Pool Kinetics

Considering only hydrogen, **Figure 1** represents a subject, in water balance, with a total body water of N mol with water (tracee) input and output rates of F mol/day containing ^2H at a naturally abundant molar concentration, C_b. A fractional output or rate constant is defined as $K = F/N$.

If a small quantity (D mol) of water labeled with ^2H tracer is added to the pool, it will be removed from it according to the monoexponential relationship

$$q_t - q_b = De^{-Kt}$$

where D is the amount of tracer given, q_t is the total amount (mol) in the body pool at time t (days), and q_b is the amount always present due to inflow at natural abundance. K is a fractional rate constant, sometimes defined in terms of the biological half-life $T_{1/2}$. This can be calculated as $T_{1/2} = \ln 2/K = 0.693/K$.

Since input and output rates are the same and the amount of tracer added is small relative to the pool size, we can write

$$\frac{q_t - q_b}{N} = \left(\frac{D}{N}\right)e^{-Kt} \text{ or } C_t - C_b = (C_0 - C_b)e^{-Kt}$$

where $C_0 - C_b$ is the increment in isotopic concentration resulting from the administration of the dose, and N can be calculated as $N = D/(C_0 - C_b)$.

The foregoing equations have been written in terms of isotopic concentration (e.g., $C = {}^2H/({}^2H + {}^1H)$), but mass spectrometry measurements are in terms of ratio (e.g., $R = {}^2H/{}^1H$) and in practice, for DLW calculations R or enrichment relative to a standard is invariably substituted for C with no effect on results at the low levels of enrichment applied in this methodology.

Principles of the Method

When Lifson first began his physiological experiments with newly available ^{18}O in the mid-1950s, it was already well-known that oral dosing with ^2H_2O and its dilution in body water was a way of measuring body water mass and turnover. Lifson showed that the oxygen in carbon dioxide, the waste product of energy metabolism, was in equilibrium in the body with body water:

$$H_2O + CO_2 \Longleftrightarrow H_2CO_3$$

He realized, therefore, that the greater apparent turnover of body water measured with $H_2{}^{18}O$ in comparison to turnover measured with ^2H_2O (**Figure 2**) was a consequence of carbon dioxide production, as shown in **Figure 3**. Thus, there was potential for a method that would permit the measurement of total CO_2 output and hence energy expenditure over long periods merely by isotopic analysis of samples of body fluids. Initially, the method was applied only to small animals because

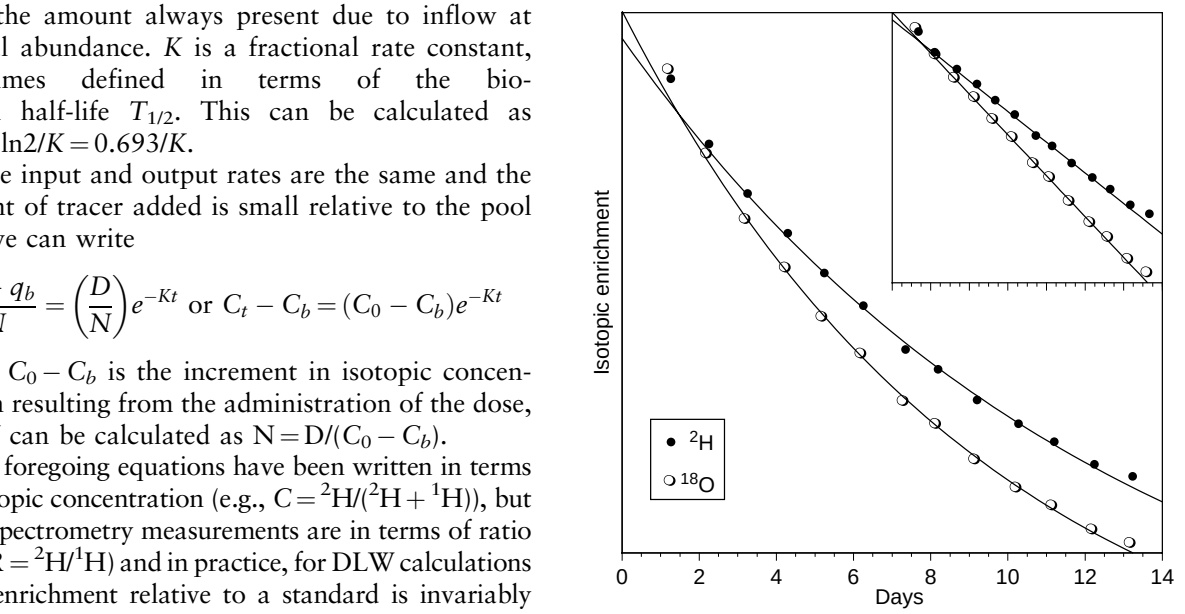


Figure 1 A simple one-compartment model of water turnover.

Figure 2 Exponential loss of ^2H and ^{18}O from body water. The insert shows the data on a log scale.

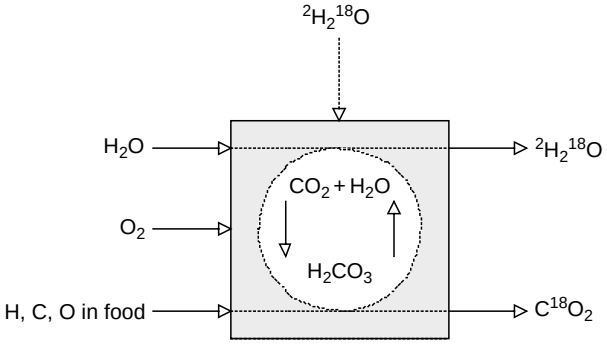


Figure 3 The fate of an oral bolus dose of ^2H and ^{18}O given as water (DLW).

the ^{18}O isotope was (and still is) expensive and instrumental limitations meant that relatively large doses had to be given to achieve adequate measurement precision. However, in the 1980s human studies, which are the focus of this article, became possible and in 1998 a basic unified methodological approach was established as a result of a meeting of the experts in the field (International Dietary Energy Consultancy Group). The publication derived from this meeting remains a valuable tool.

The following are the underlying assumptions of the method:

1. Body water is a single compartment that the isotopes label and from which they are lost.
2. ^{2}H is lost only as water.
3. ^{18}O is lost as water and carbon dioxide.
4. Total body water and output rates of water and carbon dioxide are constant.
5. Water and carbon dioxide loss occurs with the same enrichment as that coexisting in body water.
6. Background isotope intakes are constant.

Taking these in turn, assumption 1 is not correct. Evidence from many studies shows that the single compartments labelled by the isotopes are not the same size; ^{2}H space is approximately 3% larger than ^{18}O space. However, there is no evidence that isotope sequestration is a significant factor in human studies (assumptions 2 and 3). Water and carbon dioxide production rates are unlikely to be constant during a measurement period (assumption 4), but provided variations are random and not unidirectional during the measurement period, justifying the use of mean values for a period in any case, the method will not produce biased results.

Allowing assumptions 1–4, simple equations can be formulated (values of F and N are in mol and K in days^{-1}). F_{H_2O} is measured as

$$F_{H_2O} = K_D N_D$$

and the water plus carbon dioxide output (expressed in mol water equivalents) is

$$F_{H_2O+CO_2} = K_O N_O$$

Carbon dioxide production is then

$$F_{CO_2} = \frac{K_O N_O - K_D N_D}{2}$$

The factor of 2 arises because 2 mol of water is equivalent to 1 mol of carbon dioxide.

These simple relationships are in practice modified to correct for isotopic fractionation that, contrary to assumption 5, does occur. Where evaporative water losses occur, relatively less ^{2}H and ^{18}O leave the body in water vapour compared with liquid water. Fractionation factors are defined as

$$f_1 = \frac{\left(^{2}H/^{1}H\right)_{vapour}}{\left(^{2}H/^{1}H\right)_{liqiud}} = 0.941,$$

$$f_2 = \frac{\left(^{18}O/^{16}O\right)_{vapour}}{\left(^{18}O/^{16}O\right)_{liqiud}} = 0.991,$$

$$f_3 = \frac{\left(^{18}O/^{16}O\right)_{CO_2}}{\left(^{18}O/^{16}O\right)_{H_2O}} = 1.037$$

Thus, water vapour is isotopically depleted in ^{2}H and ^{18}O and carbon dioxide is relatively more enriched in ^{18}O compared to liquid water.

If it is assumed that a constant proportion (x) of water losses is fractionated, carbon dioxide production rate becomes

$$F_{CO_2} = \frac{K_O N_O}{2f_3} - \frac{K_D N_D (xf_2 + 1 - x)}{2f_3 (xf_1 + 1 - x)}$$

This procedure is most frequently used for infants and young children, in whom values of x are assumed to be 0.15–0.20.

For adults, fractionated water losses (F_f) are often defined in terms of F_{CO_2} ($F_f = 2.1F_{CO_2}$), in which case

$$F_{CO_2} = \frac{K_O N_O - K_D N_D}{2f_3 + 2.1(f_2 - f_1)}$$

Assumption 6 relates to the requirement that a predose sample should represent the effect of normal natural abundance isotope input. In most cases, background isotopic enrichment is likely to vary only randomly during a measurement period and so the issues are about the relationship between the background sample measured, the mean background and its random variation during the experimental period, the extent to which background variations in ^{2}H and ^{18}O are covariant, and the size of isotope doses and postdose enrichments in relation to these variations. In most experimental situations investigated with affordable isotopic doses, background variation contributes to the internal errors of the method and limits the extent to which better analytical precision improves results. In some circumstances (e.g., subjects moving from one place to another and use of large amounts of rehydration fluids in hospitalised patients), it is possible that a predose sample taken to represent isotopic background is not at all meaningful and the best advice may be to avoid these circumstances rather than try to correct for them.

Finally, F_{CO_2} values have to be converted into values for energy expenditure based on a fixed relationship between these quantities that depends on metabolic fuels used, expressed as a respiratory quotient (RQ). We can write

$$\text{Energy expenditure (kJ)} = F_{CO_2}\left(\frac{346.7}{RQ} + 124.3\right)$$

where F_{CO_2} is mol. RQ is calculated from dietary information or assumed to have a particular population value, such as 0.85.

Insertion of typical Western adult values ($N_O = 2000$, $N_D = 2066$, $K_O = 0.12$, *and* $K_D = 0.10$) into the relevant equations and 'what if' experimentation will allow the reader to test the effect of making changes to the assumptions and values. **Table 2** provides examples that show that serious errors or bias, for groups or individuals, are unlikely unless the applied population means for assumed values are grossly incorrect or the coefficient of variation (CV) is large.

Experimentation with the data, however, will also show that the magnitude of the difference between $K_O N_O$ and $K_D N_D$ is crucial. The method depends on precisely determining a relatively small difference between these two experimentally measured, larger values. This difference is approximately 20% in the example but can be much less when water turnover is high relative to carbon dioxide production (e.g., very young infants or subjects living in the tropics).

For the slopes (K_O and K_D) a minimum of two time points are required sufficiently far apart in time (two or three biological half-lives) to allow good precision on the slope determination with doses of sufficient magnitude to avoid detrimental effects of natural abundance variations and the limitations of analytical precision, especially at the end of the measurement period. In some protocols, more than two samples are measured, and this permits error calculations based on the goodness of fit of the data. Isotope distribution spaces are calculated from samples taken soon after dose administration (the 'plateau method') or by extrapolation of the disappearance curves to $t = 0$. Distribution spaces may be normalized to population-based estimates (N'_D and N'_O) of their relation to total body water (TBW):

$$TBW = \frac{\frac{N_O}{1.007} + \frac{N_D}{1.041}}{2}$$

$$N'_O = 1.007(TBW)\ N'_D = 1.041(TBW)$$

Figure 4 illustrates some aspects of total imprecision and the origins of the variance for a typical subject defined in **Table 3** when different dosing regimes are applied, with ^{18}O enrichment being varied at a constant initial ^{2}H:^{18}O ratio of 8.

The following are general considerations:

1. Naturally occurring covariance in ^{2}H and ^{18}O enrichment in baseline samples can be used to mitigate errors resulting from physiological variation in these values if dose sizes are suitably tailored to the slope of the variation. Optimum doses in this respect are predicted by

$$\left(\frac{^{2}H}{^{18}O}\right)_{optimal} = S\frac{2^n - 1}{2^{pn} - 1}$$

where $(^{2}H/^{18}O)_{optimal}$ is the ratio of immediate post-dose - background enrichments (rel V-SMOW) for

Table 2 'What if' calculations for a typical subject ($N_O = 2000$, $N_D = 2066$, $K_O = 0.12$, $K_D = 0.10$)

Fractionated water losses defined in terms of F_{CO_2} ($F_f = 2.1\ F_{CO_2}$) for mean and assumed CV = 10%	CO_2 production relative to value for mean
−2 SD = 1.68 F_{CO_2}	1.010
Mean = 2.1 F_{CO_2}	1
+2 SD = 2.58 F_{CO_2}	0.981
Assumed *RQ* (typical mean ±2 SD)	Energy expenditure relative to value for mean
−2 SD = 0.825	1.024
Mean = 0.85	1
+2 SD = 0.875	0.978

CV, coefficient of variation; *RQ*, respiratory quotient.

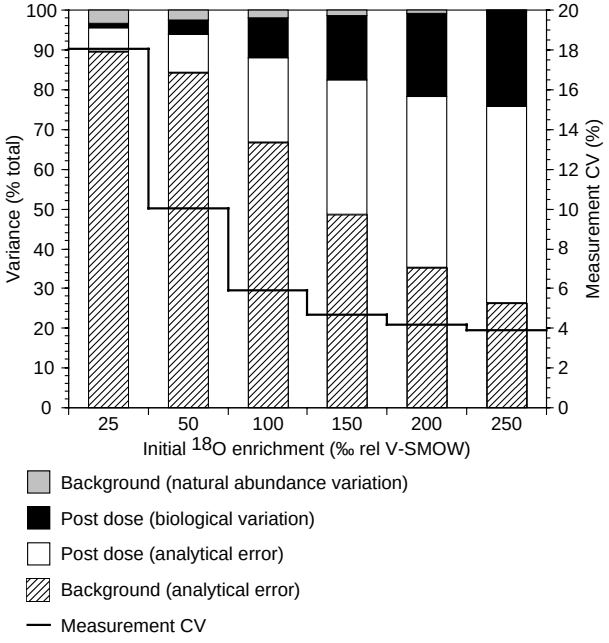


Figure 4 Origin of errors and their size in DLW experiments. The line and right axis show the total CV at different isotope doses in a typical subject defined in **Table 3**. The bars and left axis indicate the proportion of the total variance derived from each source of error.

Table 3 Typical estimates and measurement precision in a DLW experiment lasting 14 d

Parameter	*Value*
N_O	2000 mol
N_D	2066 mol
K_O	0.12 day^{-1}
K_D	0.10 day^{-1}
Proportional error in postdose ^2H samples originating from variations in water turnover (SD)	0.01
Variance in postdose ^{18}O accounted for by variance in ^2H (excluding analytical errors)	90%
^{18}O analytical error at baseline (SD)	0.15‰
^2H analytical error at baseline (SD)	1.5‰
^{18}O analytical error for enriched samples (SD)	0.5% of value + 0.15‰
^2H analytical error for enriched samples (SD)	0.5% of value + 1.5‰
^{18}O background variation (SD)	0.15‰
^2H background variation (SD)	1.2‰
Variance in background ^2H accounted for by variance in ^{18}O (excluding analytical errors)	100%
Slope of background ^2H enrichment on background ^{18}O enrichment	8

^2H and ^{18}O, S is the slope of background ^2H enrichment on background ^{18}O enrichment, n is the experiment duration in terms of the number of biological half-lives for the ^2H isotope, and p is K_O/K_D.

2. Much of the deviation of the ^2H and ^{18}O data from the model for the postdose samples is covariant because it relates to inconstancy of water turnover. Errors thus tend to cancel, and this considerably reduces the potential impact of variance from this source.
3. Although the analytical errors applied in this case are not the lowest reported, they are probably typical and it can be seen that they always account for much of the variance.
4. Errors consequent on background uncertainty become very important when amounts of dose are reduced, but in practice, cost always limits the amount of ^{18}O that can be given. For this example, adequate precision in the total energy expenditure(TEE) measurement is predicted for ^{18}O doses producing initial enrichment in the range of 100–150‰ rel V-SMOW.

Protocols

There are, of course, variations depending on the type of subjects to be investigated, and either exclusively urine or saliva samples can be collected. Typically, for adult subjects, after the collection of a predose sample of urine or saliva, they are asked to drink an accurately weighed mixture of the isotopes to give the required enrichment in body water. A small sample of the dose should be retained for isotope analysis. The dose bottle is then rinsed with a further amount of water (≈50 ml) and this is also drunk. Most investigators fast their subjects for at least 6 h and may restrict food and water intake during the time when the isotopes are equilibrating in body water. If a plateau method is used for the determination of dilution spaces, the requirement is to collect a sample after equilibration is complete but before turnover begins to reduce enrichment. This will usually require a series of three samples collected at successive hourly intervals between 4 and 8 h. If urine samples are used, the first one should be discarded. A further two samples are collected two or three biological half-lives apart. In most adult cases, experiments will last 14 days; however, for both the timing of the plateau samples and the length of time of the study, it is advisable to establish specific times for the population under investigation. If dilution spaces are to be calculated from the intercept of isotope disappearance curves, postdose samples should begin to be collected on day 1 postdose and on subsequent days during the measurement period. Minimally, samples should be collected at the beginning and end of the measurement period (e.g., days 1, 2, 13, and 14). If a plateau method is used, samples are best collected in the presence of the investigation team, but when the intercept method is used subjects can be instructed to collect, label, and store their own samples. A few ml of urine, or saliva are sufficient for analyses, and should be collected and capped immediately to avoid evaporation and possible contamination. For long-term storage, samples should be stored frozen but may be refrigerated in the short term and need not be frozen for shipping.

Experience suggests that often it is the dose administration and sample collection that cause method failures. A good technique and high precision are needed for enrichment measurements but samples can always be reanalysed. Failures consequent on poor technique in subject-related procedures cannot be rectified and can be costly, especially if they are repeated through a whole investigation. New users of the methodology are advised to test all procedures in pilot work before full-scale application in a study.

Enrichment of samples is best calculated in terms of fraction of the dose given; that is,

$$\left(\frac{18.02d}{TD}\right)\left(\frac{E_S - E_P}{E_D - E_T}\right)$$

where E is isotopic enrichment, d is a weight (g) of dose diluted in T (g) tap water, and D is the weight (g) of dose given. Subscripts S, P, D, and T refer to postdose sample, predose sample, diluted dose, and tap water, respectively. The reciprocal of plateau values is the isotope dilution space (N_D or N_O). The reciprocal of the value at the time zero intercept of a plot of its log value vs time provides alternative dilution space estimates. The slope is the rate constant (K_D or K_O).

Validations and Reproducibility

Comparisons between DLW and calorimetry suggest a precision of 4 or 5%, but it should be remembered that studies of this type are highly controlled and may not properly reflect the real-life situation to which the method is intended to apply. The closest useful estimates are therefore perhaps those provided by an analysis of test/retest situations in which the same subjects were measured in more or less the same physiological conditions. **Figure 5** shows a compilation of such data. Apart from the labourers studied in the tropics, where the precision of the estimates may have been limited by known high water turnover rates, the data are quite consistent, with a mean of 8%. Subtraction of a likely contribution of 4% from total measurement error suggests a within-subject variation of 7%.

Applications of DLW in Nutrition

DLW and Energy Intake

Examination of the history of DLW in man suggests that there was an expectation that much would be learned in relation to the development of obesity as an outcome of identified long-term positive energy balance. Certainly in the initial phases of its use in human studies in the late 1980s, experimental protocols were most often designed to measure as accurately as possible the differences between energy intake and energy expenditure, but the findings from these experiments invariably exposed the limitations of energy intake measurements. Probably because the DLW concepts were then somewhat alien to conventional nutrition, the notion that intake measurements were more often than not inaccurate and underestimates was not at first easily accepted, but the most recent of several reviews records a very convincing body of evidence (**Figure 6**). However, although exposing a problem, most of these observations by themselves do nothing

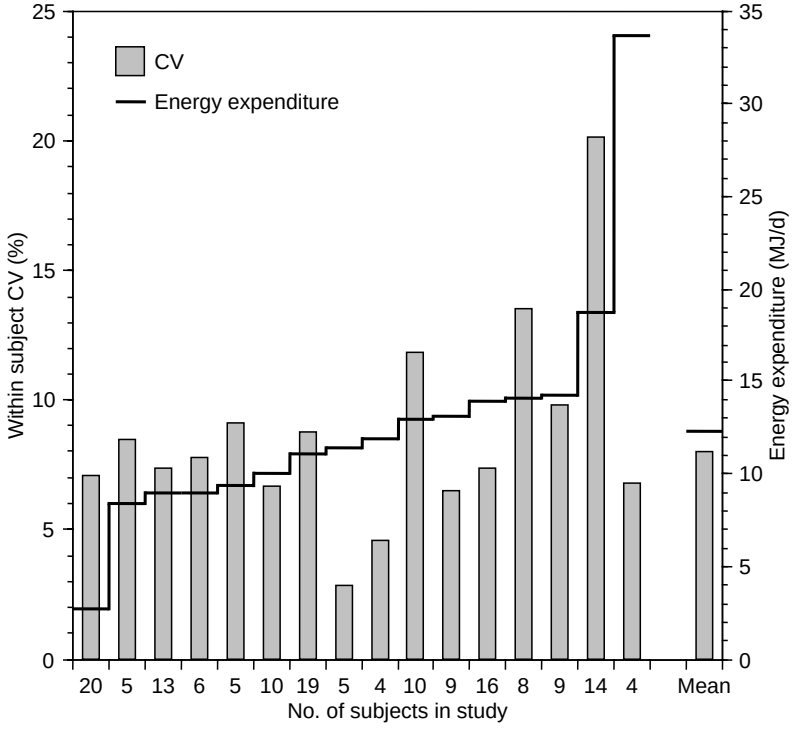
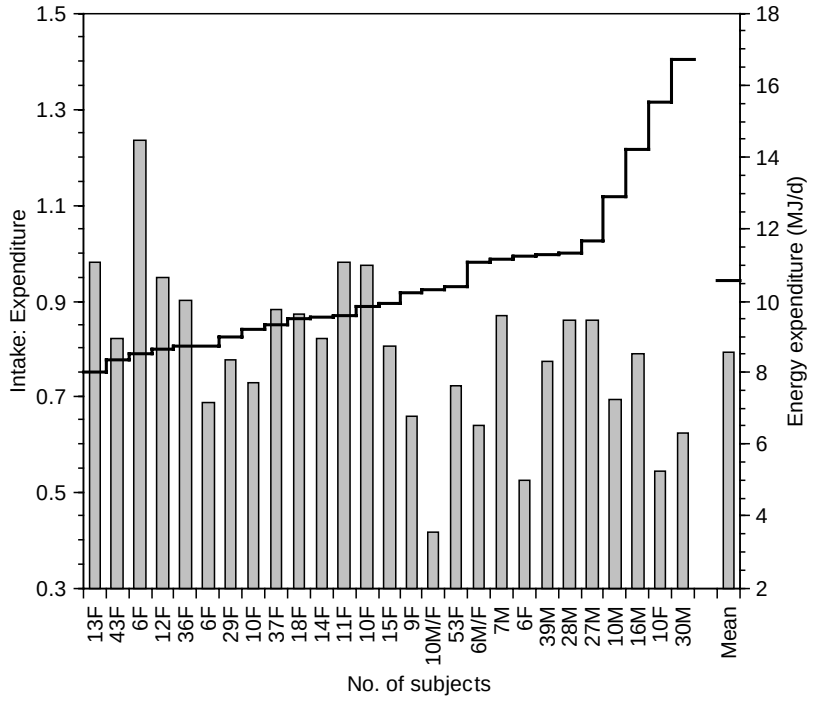
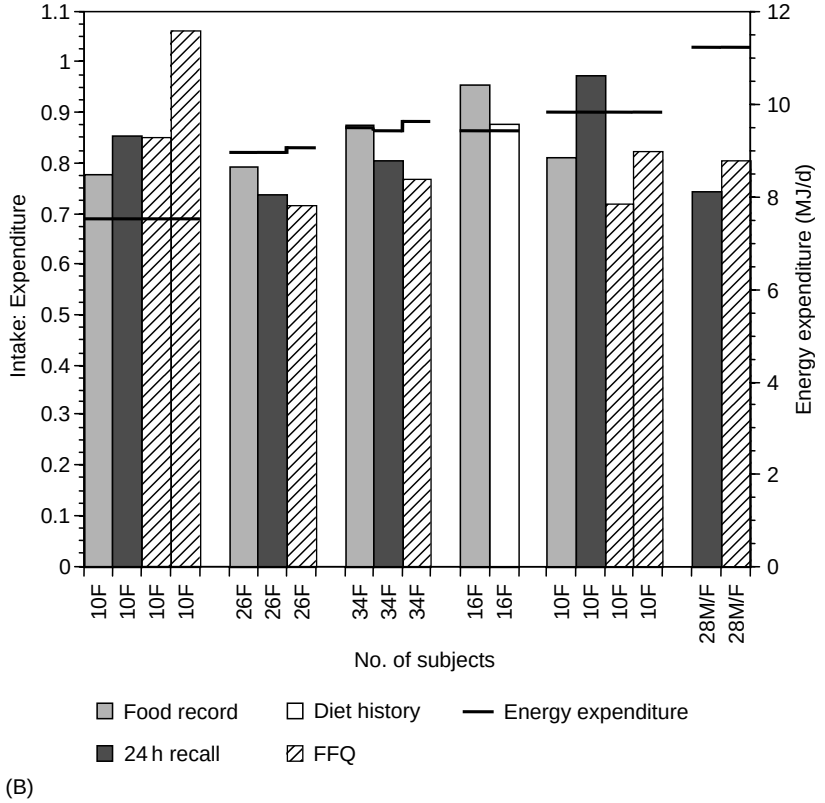


Figure 5 Reproducibility of the DLW method. (Data from Schoeller DA and Hnilicka JM (1996) Reliability of the doubly labeled water method for the measurement of total daily energy expenditure in free-living subjects. *Journal of Nutrition* **126**: 348S–354S.)



(A)



(B)

Figure 6 Accuracy of energy intake measurements assessed by DLW. (A) Dietary record data and (B) simultaneous use of more than one instrument. (Data from Trabulsi J and Schoeller DA (2001) Evaluation of dietary assessment instruments against doubly labeled water, a biomarker of habitual energy intake. *American Journal of Physiology (Endocrinology and Metabolism)* **281**: E891–E899.)

to solve it, not least because the studies are too small and indications of the nature and degree of correlation between DLW energy expenditure measurements and intake have not always been reported. The issue of detecting and correcting for bias in food and specific nutrient intake measurements remains a problem to which DLW is being applied as a biomarker of energy intake in large-scale studies.

DLW and Other Noninvasive Energy Expenditure Measurements

Although DLW can be regarded as the reference noninvasive total energy expenditure measurement, isotope cost and the need for mass spectrometric analyses will always limit it to specialist rather than widescale application. There is thus a need to validate or at least understand the limitations of preexisting methodologies and alternatives under development. A significant consideration is that although DLW measurements in an individual include basal metabolic rate as a component of the total expenditure, in alternatives the focus is most often on activities and their energy cost, and basal metabolic rate is measured separately or derived from prediction equations. This means that comparisons of total energy expenditure derived from DLW and the alternatives include a component representing approximately 70% of the total that is not dependent on the activity measurement method. In these circumstances, it is not surprising that activity-based TEE measurements often show good correlation between DLW and on average tend to be similar, but they should be treated with caution with respect to the validity of the activity measurements. Calculation of the energy cost of activity (TEE - resting metabolic rate) for comparison between methods is a much more useful comparison but is not always available.

DLW and Energy Requirements

The energy requirement of an individual is the intake from food that will balance expenditure when an individual has a body size and composition, and level of physical activity, consistent with long-term good health and that will allow for the maintenance of an economically necessary and socially desirable level of physical activity. In principle, these measurements could be obtained from the measurement of food intake or by factorial methods summing estimates of resting metabolic rate with the energy costs of activity. In practice, neither of these approaches is satisfactory; food intake is generally underestimated and no single instrument for the measurement of activity is sufficiently well validated to justify its general use. However, both in the United States (Standing Committee on the Scientific Evaluation of Dietary Reference Intakes) and internationally (FAO/WHO/UNU) the decision has been made to use DLW estimates of energy expenditure to provide the basis for the estimation of requirements. Given the relatively small number of laboratories involved in this work and its relatively short history, it is quite remarkable that sufficient data are available for this exercise. The normative US databases consist of adults ($n = 407$) and children ($n = 525$), obese adults ($n = 360$) and children ($n = 309$), and subsets for pregnant and lactating women. Regression equations derived from the data sets are used to predict requirements.

Conclusions

This article provided insight into how the DLW method works, showed how it should be used, and highlighted three areas in which it is clear that DLW has made, or at least has begun to make, a significant impact on nutrition research. The method is relatively expensive and uses scarce resources in terms of expertise, instruments, and materials. However, where the research requirement matches method capabilities, in terms of accuracy and precision, it is a uniquely effective tool.

See also: **Energy**: Metabolism; Balance; Requirements. **Energy Expenditure**: Indirect Calorimetry.

Further Reading

Ainslie P, Reilly T, and Westerterp K (2003) Estimating human energy expenditure: A review of techniques with particular reference to doubly labelled water. *Sports Medicine* **33**: 683–698.

Black AE (2000) The sensitivity and specificity of the Goldberg cut-off for EI:BMR for identifying diet reports of poor validity. *European Journal of Clinical Nutrition* **54**: 395–404.

Coward WA and Cole TJ (1991) The doubly labeled water method for the measurement of energy expenditure in humans: Risks and benefits. In: Whitehead RG and Prentice A (eds.) *New Techniques in Nutritional Research*, pp. 139–176. San Diego: Academic Press.

Food and Nutrition Board (2002) Energy. In *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, and Amino Acids (Macronutrients)*, pp. 93–206. Washington, DC: National Academies Press.

Jones PJ and Leatherdale ST (1991) Stable isotopes in clinical research: Safety reaffirmed. *Clinical Science (London)* **80**: 277–280.

Koletzko B, Sauerwald T, and Demmelmair H (1997) Safety of stable isotope use. *European Journal of Pediatrics* **156**(supplement 1): S12–S17.

Lifson N, Gordon GB, and McClintock R (1955) Measurement of total carbon dioxide production by means of $D_2{}^{18}O$. *Journal of Applied Physiology* **7**: 704–710.

Prentice AM (ed.) (1990) *The Doubly-Labelled Water Method for Measuring Energy Expenditure*. Vienna: International Atomic Energy Agency.

Schoenheimer R and Rittenberg D (1939) Studies in protein metabolism. I. General considerations in the application of isotopes to the study of protein metabolism. The normal abundance of nitrogen isotopes in amino acids. *Journal of Biological Chemistry* **127**: 285–290.

Speakman J (1997) *Doubly Labelled Water: Theory and Practice*. Dordrecht, The Netherlands: Kluwer Academic.

Schoeller DA (2002) Validation of habitual energy intake. *Public Health Nutrition* **5**: 883–888.

Schoeller DA and DeLany P (1998) Human energy balance: What have we learned from the doubly labeled water method. *American Journal of Clinical Nutrition* **68**: 930S–979S.

Wong WW (2003) Energy utilization with doubly labelled water. In: Abrams SA and Wong WW (eds.) *Stable Isotopes in Human Nutrition*, pp. 85–106. Cambridge, MA: CABI.

EXERCISE

Contents
Beneficial Effects
Diet and Exercise

Beneficial Effects

C Boreham and M H Murphy, University of Ulster at Jordanstown, Jordanstown, UK

This article examines the roles that physical activity, exercise, and fitness may play in the regulation of energy balance and in the etiology of major diseases such as coronary heart disease, cancer, and osteoporosis. Before proceeding, it is necessary to define the key terms of reference. 'Physical activity' can be defined as "any bodily movement produced by skeletal muscles that results in energy expenditure." 'Exercise' (often used interchangeably with 'physical activity') is defined as "physical activity which is regular, planned, and structured with the aim of improving or maintaining one or more aspects of physical fitness." 'Physical fitness' is "a set of outcomes or traits relating to the ability to perform physical activity."

Exercise and Energy Balance

Energy balance occurs when the total energy expenditure of an individual equals his or her total energy intake from the diet. If intake exceeds expenditure the result is an increase in the storage of energy primarily as body fat. If intake is below expenditure, body energy content or body fat decreases.

In humans, energy is expended in three ways: maintaining the physiological functions of the body at rest, often termed resting metabolic rate (RMR); ingesting food and digesting and assimilating nutrients, or the thermic effect of food (TEF); and skeletal muscular contractions involved in spontaneous physical activity or planned exercise. Of these components, the energy expenditure associated with physical activity and exercise is the factor that accounts for the greatest variability between individuals (**Table 1**). In addition, energy expenditure through physical activity is the only component that may be reasonably

Table 1 Estimated daily energy expenditure (approximate) for individuals of different age, weight, gender, and level of activity[a]

Status	*Estimated daily energy expenditure (kcal)*
Infant, male, age 3 months, body weight 6 kg	760 (3200 kJ)
Child, male, age 4 years, body weight 17 kg	1520 (6400 kJ)
Teenager, male, age 13 years, body weight 46 kg	2200 (9200 kJ)
Sedentary female[b]	1950 (8100 kJ)
Sedentary male[c]	2500 (10 200 kJ)
Female, moderately active[b]	2200 (9200 kJ)
Male, moderately active[c]	3000 (12 500 kJ)
Female, very active[b]	2500 (10 400 kJ)
Male, very active[c]	3200 (13 300 kJ)

[a]Values are based on estimated average requirements from a report by the Committee on Medical Aspects of Food Policy (1991). Dietary reference values are for food energy and nutrients for the *United Kingdom*.
[b]Based on female age 25 years, body weight 60 kg.
[c]Based on male age 25 years, body weight 70 kg.

controlled by an individual, and therefore it may represent an appropriate method for altering energy balance. Physical activity is estimated to make up 5–40% of daily energy expenditure depending on the activity habits of the individual, with RMR and TEF accounting for 60–75 and 10–15%, respectively.

Aside from its direct independent effect on daily energy expenditure, evidence suggests that exercise may also alter RMR, TEF, and the energy expenditure caused by spontaneous physical activity.

Energy Expenditure during Exercise

The magnitude of energy expenditure during exercise is dependent on several factors, including the mode, intensity, and duration of exercise, as well as the body mass of the individual.

When determining the metabolic cost of weight-bearing physical activity, energy expenditure needs to be expressed in relation to body size since a small person will expend less energy performing a given activity (e.g., walking up a flight of stairs) than a larger person performing the same activity. Therefore, to calculate the energy cost of a given activity it is necessary to know the energy cost in kcal (kJ) per kilogram of body weight. The term MET (metabolic equivalent) may also be used to indicate the ratio of the rate of energy expenditure during a given activity to resting metabolic rate (RMR). An example illustrates how METs are used to quantify energy expenditure during exercise. If an individual with a body mass of 70 kg expends 70 kcal (≈300 kJ) per hour at rest (RMR), and walking at a speed of 5.6 km per hour requires 280 kcal (≈1200 kJ) per hour, the energy cost of the activity is 4 METs or four times the RMR of the individual. Since body size is a determinant of both RMR and the energy expenditure during exercise, a heavier individual will have a higher RMR but will still require four times this level of expenditure (or 4 METs) to walk at the same speed. **Table 2** indicates the energy cost in METS of many popular exercise modes.

Table 2 Energy costs of popular physical activities

Activity	*Intensity*	*METs*
Walking	6.4 km h^{-1}	4
Running	10.8 km h^{-1}	11
Cycling	20.9 km h^{-1}	8
Swimming	Front crawl, moderate	8
Tennis	Singles	8
Aerobics	Moderate	6

Adapted from Ainsworth BE, Haskell WL, Leon As *et al.* (1993). Compendium of physical activities: Classification of energy costs of human physical activities. *Medicine and Science in Sports and Exercise* **25**(1): 71–80.

Energy Expenditure after Exercise

In addition to the additional energy consumed during an exercise bout, several researchers have found that energy expenditure remains elevated for a period following exercise. However, conclusions regarding the magnitude and duration of this postexercise elevation in energy expenditure have been equivocal. Studies have found an increase in energy expenditure in the postexercise period varying in magnitude from 5 kcal (21 kJ) to 130 kcal (546 kJ), with some suggesting that this additional energy expenditure lasts a few minutes and others suggesting that the elevated metabolic rate persists for up to 24 h. The divergence in the findings may be accounted for by the various modes, durations, and intensities of exercise employed in the studies as well as the methods used for measuring alterations in energy expenditure and the confounding effects of food ingestion during the recovery period. In addition, alterations in postexercise energy consumption may exhibit intraindividual variations according to the fitness level of subjects. Several mechanisms underlying this increased energy expenditure during the postexercise period have been postulated, including the energy cost of replenishing fuel stores, the cost of dissipating by-products of adenosine triphosphate (ATP) resynthesis, restoration of cellular homeostasis, and the futile cycling of energy substrates. The magnitude of this increase may be related to the intensity and duration of exercise, with longer or more strenuous activity creating a greater perturbation to homeostasis and therefore causing greater energy expenditure in restoring the body to its preexercise condition.

Effects of Exercise Training on Resting Metabolic Rate

Aside from the transient increase in energy expenditure in the period immediately following exercise, several researchers have examined the chronic effect of exercise on RMR. Although findings are far from consistent, some investigators have found that regular exercise causes a persistent augmentation in RMR. The mechanism for effect has yet to be confirmed, but it has been hypothesized that this increase may be due to the high energy turnover associated with the elevated levels of energy intake and expenditure typical of trained individuals. One beneficial effect of exercise training on resting metabolic rate is the maintenance or

increase in lean body mass. As a result of regular resistance exercise, muscle size increases (hypertrophy) or the age-related decline in muscle mass (atrophy) is reduced, contributing to an increase or maintenance of RMR.

Effects of Exercise on the Thermic Effect of Food

The TEF is largely dictated by the composition and energy content of the meal as well as an individual's body composition. However, some studies have indicated that pre- or postprandial exercise may enhance the TEF.

In addition to this acute effect of exercise, regular training may alter the TEF. In males, the thermic effect of a meal is lower in highly trained compared to untrained individuals. In one study, moderate levels of fitness were associated with a greater increase in the TEF than either high or low fitness. The authors suggest that very high or very low levels of fitness may decrease the thermic effect possibly by adaptive mechanisms, such as a lower insulin or lower noradrenaline response to feeding. Interestingly, no equivalent effect has been found in women. Studies on monozygotic twins also suggest a strong genetic factor controlling whether exercise has such an effect.

Effect of Exercise on Energy Expenditure in Spontaneous Physical Activity

In addition to the energy expenditure during planned exercise, other skeletal muscle contraction associated with spontaneous physical activity (including fidgeting) incurs an energy cost. Research indicates that the quantity of energy expended in spontaneous physical activity is highly variable between individuals. Studies show that in addition to its effect on RMR, participation in a planned exercise program increases the energy expenditure of an individual during nonexercising time.

Physiological Adaptations to Exercise Training

Aside from alterations in energy balance, regular exercise brings about many physiological adaptations. The human body is remarkably plastic in response to the increased metabolic demands of exercise training (overload), with many adaptations occurring that enable the body to function more efficiently. The nature and magnitude of these changes are dependent on the volume (duration and frequency), intensity, and type of exercise performed. For this reason, the physiological adaptation to training will be classified according to the nature of the exercise undertaken.

It is important to remember two principles when considering the physiological adaptations to exercise training. First, there is a degree of intraindividual variation in response to exercise training that may be attributed in part to hereditary factors. Second, whereas exercise training will cause adaptation, the removal of this stimulus will result in a reversal of adaptation, or 'detraining.'

Adaptations to Submaximal/Endurance Exercise Training

Submaximal exercise generally refers to an intensity of exercise that requires less than an individual's maximal oxygen uptake. Submaximal exercise challenges the body to deliver and utilise an increased amount of oxygen in the resynthesis of ATP. With training, changes occur that increase the body's ability to utilize oxygen. For simplicity, the adaptations to submaximal exercise training have been grouped according to the site at which they occur.

Central adaptations Central adaptations to regular submaximal exercise include alterations in the morphology and function of the heart and circulatory systems that allow greater delivery of oxygen to the working muscle.

The pulmonary system in healthy individuals does not provide a significant limitation to exercise, and therefore little alteration in the lung volumes, respiratory rate, or pulmonary ventilation and diffusion occurs as a result of training.

Modest cardiac hypertrophy characterized by an increase in left ventricular volume occurs in response to training. This adaptation allows an increase in stroke volume, leading to a reduction in heart rate at rest and during submaximal workloads and an increased cardiac output during maximal workloads.

Finally, an increase in total plasma volume and an increase in the total amount of hemoglobin have been observed in response to submaximal endurance training.

Peripheral Adaptations

Peripheral adaptations refer principally to changes in the structure and function of skeletal muscle that enhance its ability to use oxygen to produce energy aerobically.

As a result of endurance training, there is an increase in blood supply to the working muscle. This is achieved by an increased capillarization in trained muscles, greater vasodilation in existing muscle capillaries, and a more effective redistribution of cardiac output to the working muscle.

An increase in the activity of aerobic enzymes and an increased mitochondrial volume density (approximately 4–8%) within trained muscle have been noted. These are coupled with increased glycogen storage within the muscle and increased fat mobilization allowing a higher rate of aerobic ATP resynthesis from free fatty acids and glucose.

High-Intensity Exercise and Strength Training

High-intensity exercise requires energy utilization rates that exceed the oxidative capabilities of the muscle. Activities such as sprinting require the anaerobic resynthesis of ATP to produce and maintain high levels of muscular force and are therefore limited in duration. Strength training also relies heavily on anaerobic energy sources and requires high force production by specific muscle groups.

Adaptations to High-Intensity Exercise and Strength Training

The main alterations that occur in response to regular high-intensity exercise or strength training are improvements in the structure and function of the neuromuscular system that allow more efficient production of the forces required for these activities and an enhanced ability to produce the energy required through anaerobic processes.

Neuromuscular The initial improvements in performance that occur with high-intensity exercise training are largely a result of improved coordination of the nervous system. Increased nervous system activation, more efficient neuromuscular recruitment patterns, and a decrease in inhibitory reflexes allow the individual to produce greater levels of force.

The maximum force a muscle can exert is largely determined by its cross-sectional area. In addition to the neural adaptations, strength training stimulates an increase in muscle size. This hypertrophy occurs preferentially in fast twitch muscle fibers and is brought about by increased protein synthesis in response to resistance training. The degree to which muscle hypertrophy occurs is dependent on many factors, including gender and body type. Although some researchers have suggested that strength training may increase the number of muscle cells (hyperplasia), the results of these studies are far from conclusive.

Since both high-intensity and strength training rely largely on anaerobic processes for energy production, adaptative alterations in oxygen delivery and utilization, such as increased capillarization or mitochondrial mass of muscle cells, are relatively minor.

Metabolic In addition to the neuromuscular alterations that occur with high-intensity and strength training, several metabolic adaptations improve the ability of the muscle to resynthesize ATP from anaerobic sources. Intramuscular stores of the anaerobic energy intermediates, such as creatine phosphate (CP) and glycogen, increase after a period of supramaximal training. The activity of enzymes involved in anaerobic production of energy, such as creatine kinase and myokinase, is also increased.

Studies on the Role of Exercise/Fitness in the Etiology of Coronary Heart Disease

Coronary heart disease (CHD) has a multifactorial etiology, and major 'biological' risk factors include elevated concentrations of blood total and low-density lipoprotein (LDL) cholesterol, reduced concentration of high-density lipoprotein (HDL) cholesterol, high blood pressure, diabetes mellitus, and obesity. In addition, 'behavioral' risk factors for CHD include cigarette smoking, a poor diet, and low levels of physical activity and physical fitness associated with the modern, predominantly sedentary way of living. Among these risk factors, a sedentary lifestyle is by far the most prevalent according to data from both the United States and England (**Figure 1**).

Scientific verification of a link between an indolent lifestyle and CHD has been forthcoming during the past 40 years, with the publication of more than 100 large-scale epidemiological studies investigating the relationships between physical activity and cardiovascular health. These studies, some of which are summarized in **Figure 2**, have produced consistently compelling evidence that regular physical activity can protect against CHD.

Pooled data and meta-analyses of the 'better' studies indicate that the risk of death from CHD increases about twofold in individuals who are physically inactive compared with their more active counterparts. Relationships between aerobic fitness and CHD appear to be at least as strong. For example, in a cohort of middle-aged men followed up for an average of 6.2 years, the risk of dying was approximately double in those whose exercise capacity at baseline was <5 METS compared with those whose capacity was >8 METS. For both physical activity and fitness, adjustment for a wide range of other risk factors only slightly weakens these associations, suggesting independent relationships.

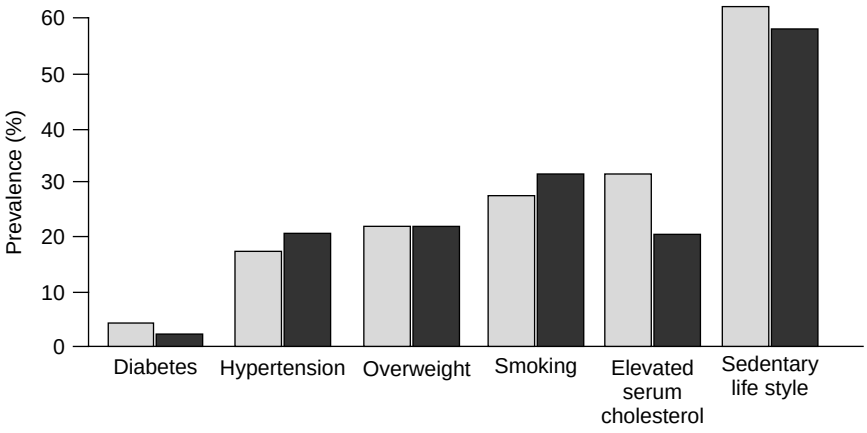


Figure 1 Estimates of the prevalence (%) of the U.S. population with selected risk factors for coronary heart disease and the population from England. In both studies, a sedentary lifestyle was taken as 'no physical activity' or irregular physical activity (i.e., fewer than three times per week and/or less than 20 minutes per session). (From Killoran AJ, Fentem P, and Caspersen C (eds.) (1994) *Moving On. International Perspectives on Promoting Physical Activity*. London: Health Education Authority, with permission.)

A common weakness of such studies is that they often rely on a single measurement of fitness or activity at baseline, with subsequent follow-up for mortality within the cohort. With such a design, it is difficult to discount the possibility that genetic or other confounding factors are influential in the observed relationship between physical activity/fitness and mortality. A further weakness in single baseline studies is that subsequent changes in activity/fitness during the follow-up are not monitored, even though they may affect the observed relationships due to the phenomenon of 'regression to the mean.'

Some prospective studies have overcome these deficiencies by examining the effects of *changes* in physical activity and fitness on mortality. One study reported on the relationship of changes in physical activity and other lifestyle characteristics to CHD mortality in 10 269 alumni of Harvard University. Changes in lifestyle over an 11- to 15-year period were evaluated on the basis of questionnaire

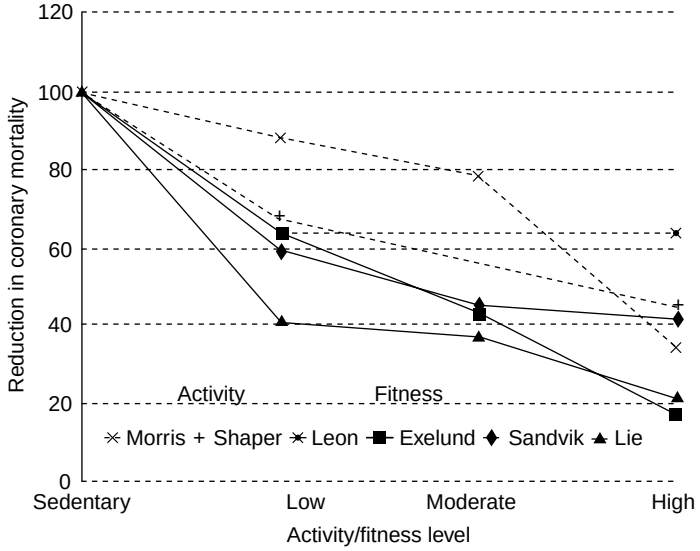


Figure 2 Summary of the results from six studies in which fitness level was determined (three studies) or activity level assessed by questionnaire (three studies) in individual populations. Follow-up was generally between 7 and 9 years except in Sandvik's study, which had a 16-year follow-up. The 'low level' group for each study represented in this figure was the activity/fitness level next to the least active/fit group. The 'high level' represents the group that was the most active/fit for the particular study. If the study participants were grouped by quintile, the 'moderate' group is the average of the third and fourth quintiles. (From Killoran AJ, Fentem P, and Caspersen C (eds.) (1994) *Moving On. International Perspectives on Promoting Physical Activity*. London: Health Education Authority, with permission.)

information, and subsequent mortality was assessed over an 8-year period. In men who were initially sedentary but started participating in moderately vigorous sports (intensity of 4.5 METS or greater), there was a 41% reduced risk of CHD compared to those who remained sedentary. This reduction was comparable to that experienced by men who stopped smoking. The second study examined changes in physical fitness and their effects on mortality. In this study of 9777 men, two clinical examinations (including treadmill tests of aerobic fitness) were administered approximately 5 years apart, with a mean follow-up of 5.1 years after the second examination to assess mortality. Results showed that men who improved their fitness (by moving out of the least fit quintile) reduced their aged-adjusted CHD mortality by 52% compared with their peers who remained unfit. Furthermore, such changes in fitness proved to be the most effective in reducing all-cause mortality when compared with changes in other health risk factors (**Figure 3**).

Mechanisms of Effect

Exercise appears to reduce the risk of CHD through both direct and indirect mechanisms. Regularly performed physical activity may reduce the vulnerability of the myocardium to fatal ventricular arrhythmia and reduce myocardial oxygen requirements. Aerobic training also increases coronary vascular transport capacity via structural adaptations and altered control of vascular resistance. Risk of thrombus formation may also be reduced with regular exercise through its effects on blood clotting and fibrinolytic mechanisms. Regular endurance exercise may also improve the serum lipid profile (particularly in favor of an enhanced HDL: total cholesterol ratio) and have beneficial effects on adipose tissue lipolysis and distribution. Regular exercise may also reduce postprandial lipemia, increase glucose transport into muscle cells, and improve the elasticity of arteries.

Exercise Prescription

For protection against CHD and other diseases associated with inactivty, exercise needs to be habitual, predominantly aerobic in nature, and current. Evidence from work carried out on British civil servants suggests that to be cardioprotective, exercise should be moderately vigorous ($\geq$7.5 kcal min^{-1} ($\geq$31.4 kJ min^{-1}) or 6 METS, equivalent to walking at approximately 3 miles per hour up a gradient of 1 in 20) and performed at least twice weekly. However, other studies have indicated that lower intensity activity is also effective as long as the total accumulated exercise energy expenditure is greater than approximately 2000 kcal week^{-1} ($\geq$8368 kJ week^{-1}).

Thus, recommendations from the U.S. Surgeon General suggest that everyone older than the age of 2 years should accumulate 30 minutes or more of at least moderate-intensity physical activity on most—preferably all—days of the week. Such activity may embrace everyday tasks such as stair climbing and walking, recreational physical activities, and more

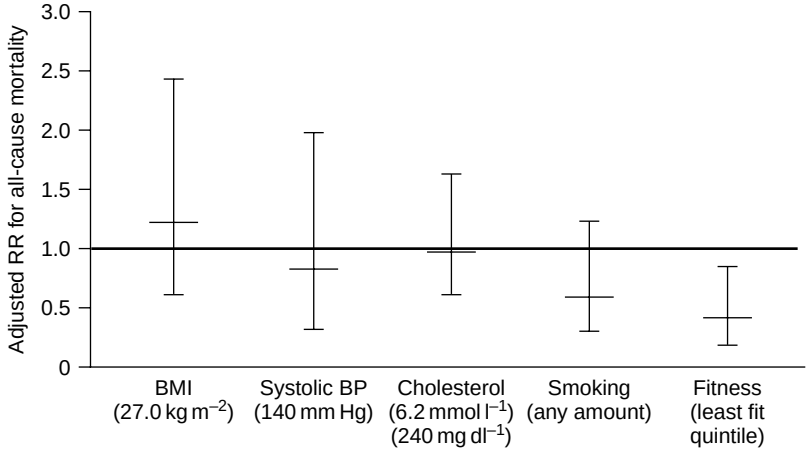


Figure 3 Relative risks (adjusted for age, family history of coronary heart disease, health status, baseline values, and changes for all variables in the figure, and interval in years between examinations) of all-cause mortality by favorable changes in risk factors between first and subsequent examinations. The analyses were for men at risk on each particular variable at the first examination. Cutoff points designating high risk are given parenthetically at the bottom of the figure. The number of men at high risk (and the number of deaths) for each characteristic were as follows: body mass index (BMI), 2691 (66); systolic blood pressure (BP), 1013 (55); cholesterol, 2212 (79); cigarette smoking, 1609 (45); and physical fitness, 1015 (56). (From Blair SN, Kohl HW, Barlow CE, Paffenbarger RS, Gibbons LW, and Macera CA, (1995) Changes in physical fitness and all-cause mortality. A prospective study of healthy and unhealthy men *JAMA*, **273**: 1093–1098, with permission.)

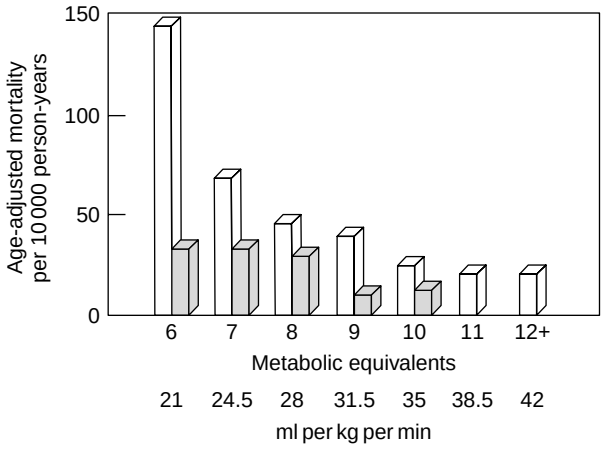


Figure 4 Age-adjusted, all-cause mortality rates per 10 000 person-years of follow-up by physical fitness categories in 3120 women and 10 224 men. Physical fitness categories are expressed as maximal metabolic equivalents (work metabolic rate/resting metabolic rate) achieved during the maximal treadmill exercise test. One metabolic equivalent equals 3.5 ml kg^{-1} min^{-1}. The estimated maximal oxygen uptake for each category is shown also. (From Blair SN *et al.* (1989) Physical fitness and all-cause mortality. A prospective study of healthy men and women. *Journal of the American Medical Association* **262**: 2395–2401, with permission.)

formal aerobic exercise programs and sports. Intermittent or shorter bouts of activity (of at least 10 minutes duration) may be accumulated throughout the day to confer similar benefits to single, continuous 30-minute bouts of exercise. A consistent finding is that previous exercise that has been abandoned confers no benefit.

Desirable aerobic fitness levels have also been described for women (maximal aerobic power of approximately 9 METs [32.5 ml kg^{-1} min^{-1}]) and men (10 METs [35 ml kg^{-1} min^{-1}]) (**Figure 4**).

Studies on the Role of Exercise/Fitness in the Etiology of Other Diseases

Obesity

Obesity is defined as an excess of adipose tissue. This condition plays a central role in the development of diabetes mellitus and confers an increased risk for CHD, high blood pressure, osteoarthritis, dyslipoproteinemia, various cancers, and all-cause mortality. The prevalence of obesity has risen dramatically in recent years, despite a decline in daily energy expenditure during the past two decades in the United Kingdom of approximately 800 kcal day^{-1} (3347 kJ day^{-1}).

Based on the principles of energy balance, such circumstantial evidence indicates that physical inactivity may play a central role in the development of obesity in humans. However, confirmatory data are scarce, particularly from well-designed prospective studies. One large-scale national study in the United States evaluated the relationship of physical activity to weight gain over a 10-year follow-up of 3515 men and 5810 women. Individuals who were sedentary at both baseline and follow-up were much more likely (relative risk, 2.3 (95% confidence interval (CI), 0.9–5.8) in men and 7.1 (95% CI, 2.2–23.3) in women) to experience considerable weight gain (>13 kg) than subjects who were active at both examinations. Evidence suggests that women who gain weight (≥6 kg) over a 1-year period expend on average 212 kcal/day less in light to moderate activities than those who maintain their normal body weight.

Difficulties are also encountered in interpreting results from intervention studies investigating the effects of exercise and/or diet on body weight, body composition, and resting metabolic rate (the latter being the single greatest component of total energy expenditure). Both energy intake and physical activity are notoriously difficult to quantify accurately, as is body fat status and distribution. Methodological differences between studies, a lack of control for possible confounding factors, and the fact that weight loss leads to an enhanced metabolic economy (due to reductions in RMR, energy cost of physical activity, and the TEF) further complicate matters. Nevertheless, exercise, particularly of the moderate-intensity type such as walking or cycling, probably helps to protect fat-free mass while promoting the loss of fat mass, but it does not appear to prevent the decline in RMR during weight loss. Similarly, long-term physical activity has minimal effects on RMR beyond its effect on lean body mass. Although studies have shown that exercise alone can reduce body weight, due to the lower total energy deficit, the rate and amount of weight loss are less than can be achieved through dieting alone.

Although the combination of exercise and dieting might be expected to improve weight loss, most data show only a modest increase (2 or 3 kg). When the total daily deficit is kept constant, diet, exercise, and diet plus exercise result in similar weight loss, but the inclusion of exercise generally results in greater fat loss and an increased lean tissue mass. There is evidence that the long-term maintenance of weight loss may require more regular activity (approximately double the current guidelines of 30 min/day) than that required to prevent weight gain in the first place. The ideal dietary and exercise prescriptions to control body weight in the long-term remain elusive.

Osteoporosis

Osteoporosis-related fractures represent a major public health concern. Once established, osteoporosis may be irreversible, emphasizing the need for primary prevention strategies based on minimizing bone loss and maximizing peak bone mass. Nearly half the variation in bone mineral density (BMD) may be attributable to nonhereditary factors. Behavioural factors of importance include diet (particularly calcium and vitamin D intakes), smoking, and the amount and type of habitual physical activity. These factors may be particularly influential during adolescence when (depending on the site) up to 90% of adult bone mineral content may be deposited, prior to the attainment of peak bone mass in the third decade of life.

Several studies on the relation of physical activity to BMD have been conducted, allowing a few general conclusions to be drawn. Clearly, bone responds positively to the mechanical stresses of exercise. Regular physical activity is likely to boost peak bone mass in young women, probably slows the decline in BMD in middle-aged and older women, and may increase BMD in patients with established osteoporosis. More research is required to clarify the type and amount of exercise that is most effective for enchancing peak bone mass. Evidence favors relatively high-impact, weight-bearing exercises (such as dancing, jumping, and volleyball), particularly during the peripubertal and adolescent years. It is unclear how physical activity and other intervention strategies, such as calcium supplementation and oestrogen replacement therapy, might interact to promote bone health.

In addition to its osteogenic effects, regular exercise may also promote better coordination, balance, and ambulatory muscle strength, thus minimising the risk of falling. The reported reduced risk of fracture (relative risk, 0.41 in men and 0.76 in women) in active individuals compared to sedentary ones is likely due to these combined direct and indirect effects of physical activity.

Cancer

In general, data relating to associations between physical activity and breast, endometrial, ovarian, prostate, and testicular cancers are inconclusive, although the suggestion that activity in adolescence and young adulthood may provide subsequent protection against breast cancer is worthy of further study. To date, the only clear evidence in this field comes from epidemiological studies relating a reduced risk of cancer of the colon to both occupational and leisure time physical activity. One such study investigated 17 148 Harvard alumni, who were assessed for physical activity at two time points, 10–15 years apart. Those who were highly active (exercise energy expenditure $\geq$2500 kcal (10 460 kJ) $\cdot$ week^{-1}) at both assessments displayed half the risk of developing colon cancer than those who were relatively inactive ($\leq$1000 kcal (4184 kJ) $\cdot$ week^{-1}). Interestingly, higher levels of physical activity at one (but not both) assessment were not associated with lower cancer risk, suggesting that consistently higher levels of activity may be necessary to provide a measure of protection. Possible biological mechanisms for this association include exercise-induced alteration of local prostaglandin synthesis (particularly prostaglandin F_2-alpha) and a decreased gastrointestinal transit time—the latter possibly decreasing the duration of contact between the colon mucosa and potential carcinogens.

See also: **Bone**. **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Coronary Heart Disease**: Prevention. **Energy**: Metabolism; Balance. **Energy Expenditure**: Indirect Calorimetry. **Exercise**: Diet and Exercise. **Obesity**: Definition, Etiology and Assessment; Treatment. **Osteoporosis**.

Further Reading

Ainsworth BE, Haskell WL, Leon AS *et al.* (1993) Compendium of physical activities: Classification of energy costs of human physical activities. *Medicine and Science in Sports and Exercise* **25**(1): 71–80.

Booth FW, Gordon SE, Carlson CJ *et al.* (2000) Waging war on modern chronic disease: Primary prevention through exercise biology. *Journal of Applied Physiology* **88**: 774–787.

Bouchard C, Shephard RJ, and Stephens T (eds.) (1994) *Physical Activity, Fitness and Health. International Proceedings and Consensus Statement*. Champaign, III, USA, Human Kinetics.

Goya Wannamethee S and Shaper AG (2001) Physical activity in the prevention of cardiovascular disease. An epidemiologocial perspective. *Sports Medicine* **31**(2): 101–114.

McKenna J and Riddoch C (eds.) (2003) *Perspectives on Health and Exercise*. Basingstoke, UK: Palgrave Macmillan.

Melanson EL, Sharp TA, Seagle HM *et al.* (2002) Effect of exercise intensity on 24-h energy expenditure and nutrient oxidation. *Journal of Applied Physiology* **92**: 1045–1052.

Poehlman ET (1989) A review: Exercise and its influence on resting energy metabolism in man. *Medicine and Science in Sports and Exercise* **21**(s): 510–525.

Poehlman ET, Denino WK, Beckett T *et al.* (2002) Effects of endurance and resistance training on total daily energy expenditure in young women: A controlled randomized trial. *Journal of Clinical Endocrinology and Metabolism* **87**: 1004–1009.

Poehlman ET, Melby CL, and Goran MI (1991) The impact of exercise and diet restriction on daily energy expenditure. *Sports Medicine* **11**(2): 78–101.

U.S. Department of Health and Human Services (1996) *Physical Activity and Health: A Report of the Surgeon General*. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion.

Diet and Exercise

R J Maughan, Loughborough University, Loughborough, UK

Introduction

At an International Consensus Conference held at the offices of the International Olympic Committee in 1991, a small group of experts agreed a consensus statement that began by saying that "Diet significantly influences exercise performance." That is a bold and unambiguous statement, leaving little room for doubt. However, the statement went on to add various qualifications to this opening statement. These largely reflect the uncertainties in our current knowledge, but also reflect the many different issues that arise in considering the interactions between diet and exercise. Exercise may take many forms and may be undertaken for many different reasons: as the emphasis on physically demanding occupations has decreased in most parts of the world, so participation in recreational exercise and sport have increased. Even though physical activity programs have been heavily promoted in most developed countries, however, they rarely involve more than about 30% of the population, leaving a major part of the population who seldom or never engage in any form of strenuous activity.

In considering the interactions between diet and exercise, two main issues must be considered, each of which gives rise to many subordinate questions. The first question is how altered levels of physical activity influence the body's requirement for energy and nutrients: this then has implications for body composition (including the body content of fat, muscle, and bone), for the hormonal environment and the regulation of substrate metabolism, and for various disease states that are affected by body fatness, nutrient intake, and other related factors. The second question is how nutritional status influences the responses to and the performance of exercise. This has implications for those engaged in physically demanding occupations, and also for those who take part in sport on a recreational or competitive basis.

Influence of Physical Activity on Energy Balance

In the simple locomotor activities that involve walking, running, or cycling, the energy cost of activity is readily determined and can be shown to be a function of speed: where body mass is supported, as in running, or where it must be moved against gravity, as in cycling uphill, then body mass is also an important factor in determining the energy cost. For walking, running, and cycling at low speeds, there is a linear relationship between velocity and energy cost, if the energy cost is expressed relative to body mass. Across a range of speeds, the cost of locomotion is approximately $1\,kcal\,kg^{-1}km^{-1}$. Therefore, energy expenditure depends on the distance covered and the body mass and is not influenced by walking speed. In purposeful walking, where the aim is to get from one place to another, the distance is set, but where walking is part of a physical activity program, activity is more often measured by time rather than distance, so walking speed becomes an important factor in determining the energy cost. At higher speeds, the relationship between energy expenditure and speed becomes curvilinear and the energy cost increases disproportionately.

It is often recommended that 20–30 min of moderate intensity exercise three times per week is sufficient exercise to confer some protection against cardiovascular disease: if this exercise is in the form of jogging, aerobics, or similar activities, the energy expenditure will be about 4 MJ (1000 kcal) per week for the average 70-kg individual, or an average of only about $150\,kcal\,day^{-1}$ (**Table 1**). However, even a small daily contribution from exercise to total daily energy expenditure will have a cumulative effect on a long-term basis. For obese individuals, whose exercise capacity is low, the role of physical activity in raising energy expenditure is necessarily limited, but this effect is offset to some degree by the increased energy cost of weight-bearing activity.

Very high levels of daily energy expenditure are now rarely encountered in occupational tasks. The average daily metabolic rate of lumberjacks has

Table 1 Estimated average energy cost of physical activity, expressed as METS (multiples of BMR) and in kJ per kg body mass per h

Activity	*MET*	$kJ\,kg^{-1}\,h^{-1}$
Bicycling, leisure	4.0	17
Bicycling, racing $30\,km\,h^{-1}$, no drafting	16.0	67
Dancing, ballroom	3.0–5.5	13–23
Forestry, fast chopping with axe	17	71
Soccer, casual	7.0	29
Walking, slow	3.5	15
Walking, brisk uphill	5.0–7.0	21–29
Writing, desk work	1.8	7.5

been reported to be about four times the basal metabolic rate, and similar values have been reported for other very demanding occupations, suggesting that this may be close to the upper limit of physical exercise that can be sustained on a long-term basis. In the short term, sporting activities can involve much higher levels of energy output: the world record for distance run in 24 h is 286 km, which requires an energy expenditure of about 80 MJ (20 000 kcal). Such an effort, however, results in considerable depletion of the body's energy reserves, and must be followed by a period of recovery.

For athletes, very high levels of daily energy expenditure are more often a feature of training than of competition, with very high levels of energy intake reported in many sports. Measurements on runners in steady state with regard to training load and body mass show good relationships between energy intake and distance run. There are some competitive events that require high levels of activity to be sustained for many consecutive days, the most obvious examples being the multi-stage cycle tours, of which the most famous is the Tour de France. Measurements on some of the competitors have shown that they manage to maintain body weight in spite of a mean daily energy expenditure of 32 MJ (8000 kcal) sustained over a 3-week period. It was suggested that those cyclists who were unable to meet the daily energy requirement were unable to complete the race.

Measurements of oxygen uptake, heart rate, and other variables made after exercise show that the metabolic rate may remain elevated for at least 12 h and possibly up to 24 h if the exercise is prolonged and close to the maximum intensity that can be sustained. After more moderate exercise, the metabolic rate quickly returns to baseline level. Therefore, it seems likely that the athlete training at near to the maximum sustainable level, who already has a very high energy demand, will find this increased further by the elevation of postexercise metabolic rate: this will increase the difficulties that many of these athletes have in meeting their energy demand. The recreational exerciser, for whom the primary stimulus to exercise is often to control body mass or to reduce body fat content, will not benefit to any appreciable extent from this effect.

The control of food intake in relation to energy expenditure is not well understood, but it is clear that both short-term and long-term regulatory mechanisms exist. These allow the adult body weight to be maintained within fairly narrow limits in spite of wide variations in energy expenditure. It is also clear from the growing prevalence of obesity, that these control mechanisms are not perfect. The acute effects of exercise on appetite and energy intake are also unclear. A period of activity may result in a stimulation of the appetite, leading to an increase in the energy intake: the magnitude of the increased intake may exceed the total energy expenditure of the activity itself. There are, however, reports that exercise may lead to a suppression of appetite, and this is likely to be true especially of high-intensity exercise. A modest training program involving energy expenditure of 200 kcal three times per week has been reported to have no effect on energy intake. In the study of distance runners referred to above, there was a negative association between the training load (expressed as distance run per week) and body fat and a positive association between training load and energy intake: this led to a somewhat paradoxical negative association between energy intake and body fat content (**Figure 1**).

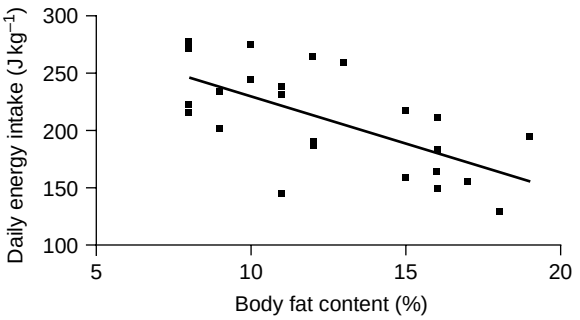


Figure 1 Association between daily energy intake and body fat content. (For further details see Maughan RJ and Piehl Aulin K (1997) Energy needs for physical activity. In: Simopoulos AP and Pavlou KN (eds.) *World Review of Nutrition*, vol. 82, pp. 18–32. Basel: Karger.)

Macronutrients and Physical Activity

Protein

The idea that protein requirements are increased by physical activity is intuitively attractive, and high-protein diets are a common feature of the diets of sportsmen and women. The available evidence does show an increased rate of oxidation of the carbon skeletons of amino acids during exercise, especially when carbohydrate availability is low. Protein contributes only about 5% of total energy demand in endurance exercise, but the absolute rate of protein breakdown is higher than at rest (where protein contributes about the same fraction as the protein content of the diet, i.e., typically about 12–16%) because of the higher energy turnover. Most recommendations suggest that individuals engaged in endurance activities on a daily basis should aim to achieve a protein intake of about

1.2–1.4 g kg^{-1} day^{-1}, whereas athletes engaged in strength and power training may need as much as 1.6–1.7 g kg^{-1} day^{-1}. In strength and power sports such as weightlifting, sprinting, and bodybuilding, the use of high-protein diets and protein supplements is especially prevalent, and daily intakes in excess of 4 g kg^{-1} are not unusual. Scientific support for such high intakes is generally lacking, but those involved in these sports are adamant that such high levels of intake are necessary, not only to increase muscle mass, but also to maintain muscle mass. This apparent inconsistency may be explained by Millward's adaptive metabolic demand model, which proposes that the body adapts to either high or low levels of intake, and that this adjustment to changes in intake occurs only very slowly.

Protein synthesis and degradation are both enhanced for some hours after exercise, and the net effect on muscle mass will depend on the relative magnitude and duration of these effects. Several recent studies have shown that ingestion of small amounts of protein (typically about 35–40 g) or essential amino acids (about 6 g) either before or immediately after exercise will result in net protein synthesis in the hours after exercise, whereas net negative protein balance is observed if no source of amino acids is consumed. These observations have led to recommendations that protein should be consumed immediately after exercise, but the control condition in most of these studies has involved a relatively prolonged (6–12 h) period of fasting, and this does not reflect normal behavior. Individuals who consume foods containing carbohydrate and proteins in the hour or two before exercise may not further increase protein synthesis if additional amino acids or proteins are ingested immediately before, during, or after exercise.

Various low-(40%) carbohydrate, high-(30%) fat, high-(30%) protein diets have been promoted for weight loss and athletic performance. Proposed mechanisms include reduced circulating insulin levels, increased fat catabolism, and altered prostaglandin metabolism. These diets can be effective in promoting short-term weight loss, primarily by restricting energy intake (to 1000–2000 kcal day^{-1}) and by restricting dietary choice. There is no evidence to support improvements in exercise performance, and what evidence there is does not support the concept.

Carbohydrate

Carbohydrate is stored in the body in the form of glycogen, primarily in the liver (about 70–100 g in the fed state) and in the skeletal muscles (about 300–500 g, depending on muscle mass and preceding diet). These stores are small relative to the rate of carbohydrate use during exercise. Fat and carbohydrate are the main fuels used for energy supply in exercise. In low-intensity exercise, most of the energy demand can be met by fat oxidation, but the contribution of carbohydrate, and especially of the muscle glycogen, increases as the energy demand increases. In high-intensity exercise, essentially all of the energy demand is met by carbohydrate metabolism, and carbohydrate oxidation rates of 3–4 g min^{-1} may be sustained for several hours by athletes in training or competition. When the glycogen content of the exercising muscles reaches very low levels, the work rate must be reduced to a level that can be accommodated by fat oxidation. Repeated short sprints will also place high demands on the muscle glycogen store, most of which can be converted to lactate within a few minutes.

Carbohydrate supplies about 45% of the energy in the typical Western diet: this amounts to about 200–300 g day^{-1} for the average sedentary individual, and is the amount that is necessary to get through normal daily activities. In an hour of hard exercise, up to 200 g of carbohydrate can be used, and sufficient carbohydrate must be supplied by the diet to replace the amount used. Replacement of the glycogen stores is an essential part of the recovery process after exercise; if the muscle glycogen content is not replaced, the quality of training must be reduced, and the risks of illness and injury are increased. Low muscle glycogen levels are associated with an increased secretion of cortisol during exercise, with consequent negative implications for immune function.

Replacement of carbohydrate should begin as soon as possible after exercise with carbohydrate foods that are convenient and appealing, and at least 50–100 g of carbohydrate should be consumed within the first 2 h of recovery. Thereafter, the diet should supply about 5–10 g of carbohydrate per kg body mass, including a mixture of different carbohydrate-rich foods. For athletes preparing for competition, a reduction in the training load and the consumption of a high carboydrate diet in the last few days are recommended: this will maximize the body's carbohydrate stores, and should ensure optimum performance, not only in endurance activities, but also in events involving short-duration high-intensity exercise and in field games involving multiple sprints.

The high-carbohydrate diet recommended for the physically active individual coincides with the recommendations of various expert committees that a healthy diet is one that is high in carbohydrate (at least 55% of energy) and low in fat (less than

30% of energy). However, where energy intake is either very high or very low, it may be inappropriate to express the carbohydrate requirement as a fraction of energy intake. With low total energy intakes, the fraction of carbohydrate in the diet must be high, but the endurance athlete with a very high energy intake may be able to tolerate a higher fat intake.

Fat

Fat is an important metabolic fuel in prolonged exercise, especially when the availability of carbohydrate is low. One of the primary adaptations to endurance training is an enhanced capacity to oxidize fat, thus sparing the body's limited carbohydrate stores. Studies where subjects have trained on high-fat diets, however, have shown that a high-carbohydrate diet during a period of training brings about greater improvements in performance, even when a high-carbohydrate diet is fed for a few days to allow normalization of the muscle glycogen stores before exercise performance is measured. It must be recognized, though, that these short-term training studies usually involve relatively untrained individuals and may not reflect the situation of the highly trained elite endurance athlete where the capacity of the muscle for oxidation of fatty acids will be much higher. For the athlete with very high levels of energy expenditure in training, the exercise intensity will inevitably be reduced to a level where fatty acid oxidation will make a significant contribution to energy supply and fat will provide an important energy source in the diet. Once the requirements for protein and carbohydrate are met, the balance of energy intake can be in the form of fat.

Micronutrients and Physical Activity

Many micronutrients play key roles in energy metabolism, and during strenuous physical activity the rate of energy turnover in skeletal muscle may be increased up to 20–100 times the resting rate. Although an adequate vitamin and mineral status is essential for normal health, marginal deficiency states may only be apparent when the metabolic rate is high. Prolonged strenuous exercise performed on a regular basis may also result in increased losses from the body or in an increased rate of turnover, resulting in the need for an increased dietary intake. An increased food intake to meet energy requirements will increase dietary micronutrient intake, but individuals who are very active may need to pay particular attention to their intake of iron and calcium.

Iron deficiency anemia affects some athletes engaged in intensive training and competition, but it seems that the prevalence is the same in athletic and sedentary populations, suggesting that exercise *per se* does not increase the risk. The implications of even mild anemia for exercise performance are, however, significant. A fall in the circulating hemoglobin concentration is associated with a reduction in oxygen-carrying capacity and a decreased exercise performance. Low serum ferritin levels are not associated with impaired performance, however, and iron supplementation in the absence of frank anemia does not influence indices of fitness.

Osteoporosis is now widely recognized as a problem for both men and, more especially, women, and an increased bone mineral content is one of the benefits of participation in an exercise program. Regular exercise results in increased mineralization of those bones subjected to stress and an increased peak bone mass may delay the onset of osteoporotic fractures; exercise may also delay the rate of bone loss. Estrogen plays an important role in the maintenance of bone mass in women, and prolonged strenuous activity may result in low estrogen levels, causing bone loss. Many very active women also have a low body fat content and may also have low energy (and calcium) intakes in spite of their high activity levels. All of these factors are a threat to bone health. The loss of bone in these women may result in an increased predisposition to stress fractures and other skeletal injury and must also raise concerns about bone health in later life. It should be emphasized, however, that this condition appears to affect only relatively few athletes, and that physical activity is generally beneficial for the skeleton.

Water and Electrolyte Balance

Few situations represent such a challenge to the body's homeostatic mechanisms as that posed by prolonged strenuous exercise in a warm environment. Only about 20–25% of the energy available from substrate catabolism is used to perform external work, with the remainder appearing as heat. At rest, the metabolic rate is low: oxygen consumption is about $250\,\mathrm{ml\,min^{-1}}$, corresponding to a rate of heat production of about 60 W. Heat production increases in proportion to metabolic demand, and reaches about 1 kW in strenuous activities such as marathon running (for a 70-kg runner at a speed that takes about $2\frac{1}{2}$ h to complete the race). To prevent a catastrophic rise in core temperature, heat loss must be increased correspondingly and this is achieved primarily by an increased rate of evaporation of sweat from the skin surface. In hard exercise in hot conditions, sweat rates can reach $3\,\mathrm{l\,h^{-1}}$, and trained athletes can sustain sweat rates

in excess of $2\,l\,h^{-1}$ for many hours. This represents a much higher fractional turnover rate of water than that of most other body components. In the sedentary individual living in a temperate climate, about 5–10% of total body water may be lost and replaced on a daily basis. When prolonged exercise is performed in a hot environment, 20–40% of total body water can be turned over in a single day. In spite of this, the body water content is tightly regulated, and regulation by the kidneys is closely related to osmotic balance.

Along with water, a variety of minerals and organic components are lost in variable amounts in sweat. Sweat is often described as an ultrafiltrate of plasma, but it is invariably hypotonic. The main electrolytes lost are sodium and chloride, at concentrations of about 20–70 $mmol\,l^{-1}$, but a range of other minerals, including potassium and magnesium, are also lost, as well as trace elements in small amounts. When sweat losses are high, there can be a substantial electrolyte loss, and intake must increase accordingly.

Failure to maintain hydration status has serious consequences for the active individual. A body water deficit of as little as 1% of total body mass can result in a significant reduction in exercise capacity. Endurance exercise is affected to a greater extent than high-intensity exercise, and muscle strength is not adversely affected until water losses reach 5% or more of body mass. Hypohydration greatly increases the risk of heat illness, and also abolishes the protection conferred by prior heat acclimation.

Many studies have shown that the ingestion of fluid during exercise can significantly improve performance. Adding an energy source in the form of carbohydrate confers an additional benefit by providing an energy source for the working muscles. Addition of small amounts (perhaps about 2–8%) of carbohydrate, in the form of glucose, sucrose, or maltodextrin, will promote water absorption in the small intestine as well as providing exogenous substrate that can spare stored carbohydrate. The addition of too much carbohydrate will slow gastric emptying and, if the solution is strongly hypertonic, may promote secretion of water into the intestinal lumen, thus delaying fluid availability. Voluntary fluid intake is seldom sufficient to match sweat losses, and a conscious effort to drink is normally required if dehydration is to be avoided. Palatability of fluids is therefore an important consideration. If exercise is prolonged and sweat losses high, the addition of sodium to drinks may be necessary to prevent the development of hyponatremia. Ingestion of large volumes of plain water is also likely to limit intake because of a fall in plasma osmolality leading to suppression of thirst.

Replacement of water and electrolyte losses incurred during exercise is an important part of the recovery process in the postexercise period. This requires ingestion of fluid in excess of the volume of sweat lost to allow for ongoing water losses from the body. If food containing electrolytes is not consumed at this time, electrolytes, especially sodium, must be added to drinks to prevent diuresis and loss of the ingested fluid.

Dietary Supplementation for Active Individuals

The use of nutritional supplements in athletes and in the health-conscious recreationally active population is widespread, as it is in the general population. A very large number of surveys have been published. A meta-analysis of 51 published surveys involving 10 274 male and female athletes of varying levels of ability showed an overall prevalence of supplement use of 46%, but the prevalence varies widely in different sports, at different levels of age, performance etc., and in different cultural backgrounds.

A wide variety of supplements are used with the aim of improving or maintaining general health and exercise performance. In particular, supplement use is often aimed at promoting tissue growth and repair, promoting fat loss, enhancing resistance to fatigue, and simulating immune function. Most of these supplements have not been well researched, and anyone seeking to improve health or performance would be better advised to ensure that they consume a sound diet that meets energy needs and contains a variety of foods.

See also: **Anemia**: Iron-Deficiency Anemia. **Appetite**: Physiological and Neurobiological Aspects. **Bone**. **Carbohydrates**: Chemistry and Classification; Regulation of Metabolism; Requirements and Dietary Importance. **Electrolytes**: Water–Electrolyte Balance. **Energy**: Balance. **Exercise**: Beneficial Effects. **Fats and Oils**. **Osteoporosis**. **Protein**: Synthesis and Turnover; Requirements and Role in Diet. **Sports Nutrition**. **Supplementation**: Dietary Supplements; Role of Micronutrient Supplementation; Developing Countries; Developed Countries.

Further Reading

American College of Sports Medicine, American Dietetic Association, and Dietitians of Canada (2000) Joint Position Statement: Nutrition and athletic performance. *Medicine and Science in Sports and Exercise* **32**: 2130–2145.

Devlin JT and Williams C (1992) *Foods, Nutrition and Sports Performance*. London: E and FN Spon.

Henriksson J and Hickner RC (1998) Adaptations in skeletal muscle in response to endurance training. In: Harries M, Williams C, Stanish WD, and Micheli LJ (eds.) *Oxford Textbook of Sports Medicine*, 2nd edn, pp. 45–69. Oxford: Oxford University Press.

Ivy J (2000) Optimization of glycogen stores. In: Maughan RJ (ed.) *Nutrition in Sport*, pp. 97–111. Oxford: Blackwell.

Kiens B and Helge JW (1998) Effect of high-fat diets on exercise performance. *Proceedings of the Nutrition Society* **57**: 73–75.

Maughan RJ (1999) Nutritional ergogenic aids and exercise performance. *Nutritional Research Review* **12**: 255–280.

Maughan RJ and Murray R (eds.) (2000) *Sports Drinks: Basic Science and Practical Aspects*. Boca Raton: CRC Press.

Maughan RJ and Piehl Aulin K (1997) Energy needs for physical activity. In: Simopoulos AP and Pavlou KN (eds.) *World Review of Nutrition*, vol. 82, pp. 18–32. Basel: Karger.

Millward DJ (2001) Protein and amino acid requirements of adults: current controversies. *Canadian Journal of Applied Physiology* **26**: S130–S140.

Nieman DC and Pedersen BK (1999) Exercise and immune function. *Sports Medicine* **27**: 73–80.

Noakes TD and Martin D (2002) IMMDA-AIMS advisory statement on guidelines for fluid replacement during marathon running. *New Studies in Athletics* **17**: 15–24.

Shirreffs SM and Maughan RJ (2000) Rehydration and recovery after exercise. *Exercise and Sports Science Reviews* **28**: 27–32.

Williams C (1998) Diet and sports performance. In: Harries M, Williams C, Stanish WD, and Micheli LJ (eds.) *Oxford Textbook of Sports Medicine*, 2nd edn, pp. 77–97. Oxford: Oxford University Press.

Wolfe RR (2001) Effects of amino acid intake on anabolic processes. *Canadian Journal of Applied Physiology* **26**: S220–S227.

F

FAMINE

K P West Jr, Johns Hopkins University, Baltimore, MD, USA

> There are so many hungry people, that God can not appear to them except in the form of bread.
> Mahatma Gandhi

Famines in History

Famine has afflicted humankind, shaping its demography and history from antiquity. Records of famine in ancient Egypt during the third millennium BC are depicted in bas-relief on the Causeway of the Pyramid of Unas in Saqqura. Biblical accounts of a famine resulting from drought in Egypt during the second millennium BC (Middle Kingdom) that stretched to Mesopotamia describe the devastation wrought on the land and society and the means by which Joseph predicted and managed its consequences. The fall of the Roman Empire followed repeated food shortages and famines from 500 BC to 500 AD. China experienced some 1828 famines, nearly one per year, from 108 BC to 1911 AD. The ranks of the Crusades in the eleventh and twelfth centuries swelled in response to promise of food. The storming of the Bastille and French Revolution followed decades of periodic rises in flour and bread prices that had caused widespread hunger and hardship, and hundreds of 'food riots.'

Recurrent famine motivated the settling of the New World. The Great Irish Famine in the late 1840s caused one and a half million deaths and an equal number of migrations, mostly to America. Decades of Russian famines following crop failures in the late nineteenth century resulted in waves of immigration to the US. Repeated famines led to the overthrow of Czarist Russia that ushered in the Bolshevik Revolution in the early twentieth century. Using food deprivation to wage class warfare and crush the Cossack revolution in the 1930s, Stalinist policies led to the starvation and death of 3.5 million Ukrainians. In China, multiple famines throughout the nineteenth century reportedly led to over 50 million deaths, and these continued throughout the first half of the twentieth century. Maoist communism rose to power in the 1940s understandably amidst promises of land reform and freedom from chronic hunger and periodic famine. However, collectivization of private farms and irrational rural industrialization schemes coupled with monopolistic control of food grain movement, purchase and access, abusive taxation, and repressive policies against the peasantry left China mostly food insecure throughout the 1950s and primed for what has turned out to be the worst single famine in human history (1959–60). During this period an estimated 30 million people perished, in absence of worldview and reaction, following the secretive, cultist policy failures of Mao's 'Great Leap Forward.' Famine was notorious on the Indian subcontinent throughout the mid-twentieth century, with the two final famines both occurring in Bengal in 1943, towards the end of British rule and again in Bangladesh (formerly East Bengal) in 1974–75. An India free from overt famine over the past half-century, despite continuing chronic undernutrition, has been attributed, in part, to the country's economic rise, relative peace, and democratic and popular processes that have included political accountability and a flourishing free press; lessons that still remain to be learnt by some modern states. In North Korea, for example, the effects of repeated floods in the late 1990s that ruined crops, combined with isolation, a collapsed centralized economy, and politicization and diversion of already insufficient international food aid from those most in need led to famine of devastating proportion.

In the late twentieth century famines have inflicted heavy loss of life in Africa, especially in the Greater Horn (i.e., Ethiopia, the Sudan, and Somalia). At least one modern regime's demise, that of Emperor Haile Selassie in 1974, followed famine. Famines of seemingly increased complexity

in Africa have resulted from deteriorating crop production associated with steady rainfall decline, failures in development and commerce, repressive and corrupt governance, and armed conflict leading, at times, to outright anarchy. Tragically, famines over the past 30 years have occurred at a time in human history when general understanding of causes and consequences of famine, and a global ability to monitor antecedents and intervene to avert mass starvation, disease, and death have never been greater. Yet, with conflict, especially internal civil war, rising as the decisive and yet unpredictable trigger of modern famine, stable governance with democratic processes (e.g., free press, people's participation, fair trade, etc.) is increasingly recognized to be one of the most important means for its prevention. History has increased awareness and understanding of the need for a stable, peaceful, and equitable political economy to guide the developing world away from famine in the twenty-first century.

Definition of Famine

Definitions of famine vary but all contain the necessary elements of widespread inaccessibility to food leading to mass numbers of starved individuals. Importantly, lack of access is not equivalent to nonavailability of food within a region, as most famines occur amidst food stocks sufficient to feed the afflicted population. More comprehensive definitions of famine may include elements of time dependency (e.g., steady, continuous erosion of or sudden collapse in food available for consumption), partial causation (e.g., due to natural calamity, armed conflict, or convergence of other complex causal events), class (e.g., affecting certain ethnic, geographic, economic or occupational groups more than others), and health consequence on a population scale (e.g., accompanied by epidemics of disease and high mortality) or other population responses (e.g., mass migration). While poverty-stricken communities tend to view famine as a continuum of increasing loss and oppression that typically begins long before mass casualty, formal 'external' definitions tend to invoke thresholds or shocks involving sudden inflections in trends for events that afflict large numbers of people. These may include spikes in prices of staple grains, levels of violence, destitution, mortality from starvation and infectious disease, and migratory movement. Threshold events tend to distinguish famine, which upon declaration demands a massive relief response, from endemic, chronic food deprivation, which results from extreme poverty, political corruption, developmental neglect and food insecurity and which leads to chronic, high rates of malnutrition, disease, and mortality. Yet, these factors are ones that, often when acting together, predispose underserved populations of the developing world to risk of famine. Such conditioning factors are antecedent causal elements that require more continuous, sensitive, and specific indicators to detect as well as a set of longer term economic, political, and developmental solutions to prevent. Whether continuous and evolving or more sudden, unleashed famine – where thresholds have been transgressed by masses of people – is catastrophic, distinct, and a human tragedy of unparaleled proportion.

Causes of Famine

> Starvation is a matter of some people not having enough food to eat, and not a matter of there being not enough food to eat.
>
> Amartya Sen

Large numbers of people starve during famine, which is usually followed by epidemics of lethal infectious diseases. Typically, a plethora of forces or conditions act within society to deprive people of food to survive. General food decline in a population may be an important factor, but it is neither necessary nor sufficient as a cause, as amply revealed by critical treatises of numerous famines over the past two centuries. This has led analysts to recognize that famines are complex, often with many ('component') causes that vary in their attribution, depending on the classes of society affected, and their timing, severity, duration, and degree of interaction. The constellation of causes and potential solutions of famine can be examined from ecological, economic, social, and public health perspectives, each offering different insights into the ecology of famine. While each view is valid and informative, none are complete or mutually exclusive, making it necessary to integrate these diverse perspectives to understand the complexity of famine and approaches to its prevention. In offering an epidemiologic overview, there appear to be at least three dominant causes of famine that have emerged during the nineteenth and twentieth centuries that appear particularly relevant to understanding modern famine causation (**Figure 1**): (1) market failure; (2) armed conflict; and (3) failure in central planning. Importantly, none are sole-acting causes and, therefore, for each one there are other antecedent factors, sometimes operative for years before, as well as concurrent and late-acting components that together lead to famine.

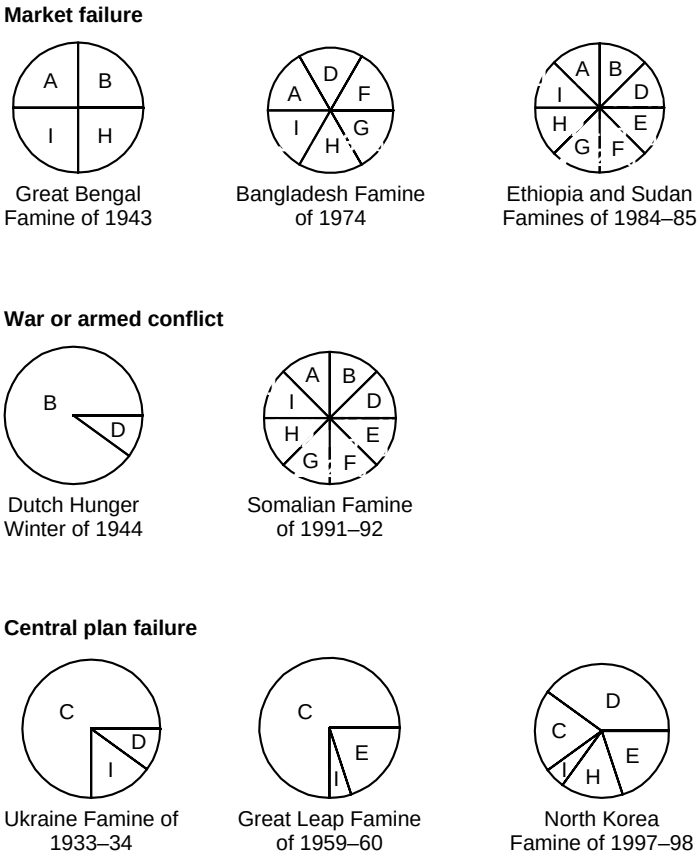


Figure 1 Complex causal networks of selected modern famines, stratified by a dominant cause. Each pie illustrates a complete cause; each wedge illustrates an assumed, essential component cause, without any one of which famine would not occur. Inclusion of causes based on literature reviews; sizes of pie slices are subjective based on descriptions in the literature (causal concepts adapted from Rothman and Greenland, 1998). A: market failure – loss of direct or trade entitlement through a combination of: (1) increased food prices due to food shortage from decreased agricultural production or importation, hoarding and speculation, or other market forces leading to unfavorable terms of exchange; plus (2) loss of means to command food through cash, labor, credit, and other assets (endowment) by vulnerable groups of society. B: war or armed conflict – declared or internal; through siege, blockade, or other expression of force, during a time course leading up to and concurrent with famine. C: central plan failure – occurring within centrally planned states lacking democratic processes, notably in twentieth century communist states; directives that disrupt infrastructure, productivity, and economic well-being, and access to food through heavy taxation, extraction of food grains, livestock and other productive assets and terror, or restrict movement of food stocks outside free-market dynamics, leading to starvation of the masses. D: natural disaster – climatological and environmental catastrophes including floods, or single, repeated or chronic droughts. E: food availability decline – food shortage resulting from poor crop production, lack of trade, poor food transport, storage and marketing sytems. F: weak infrastructure – inadequate systems of finance, credit, roads, communications, agricultural production including irrigation or flood protection systems. G: poor/unstable governance – weak and ineffective forms of governance, including anarchy. H: inadequate aid response/administrative mismanagement – inadequate national or international counter-famine measures, including employment or food procurement policies as well as withheld, slow, ineffectual, or insufficient relief. I: other causes – a catch-all 'causal complement' to those listed above, of interacting prefamine and intrafamine sociological, governmental, environmental, and market forces that render each famine unique.

Market Failure

Market failure famines occur when free, competitive market forces, driven by agriculture, transportation, communication and trade, and enabled by an abiding government fail to assure minimal entitlement to food, either directly (through subsistence) or via trade for a large sector of society. Following Amartya Sen, entitlement failure is an economic phenomenon, broadly defined, in which individuals and households are unable to obtain sufficient amounts of food through all available legal means (cash, labor, skills, credit, and other assets that comprise 'endowment') at the market's existing terms of exchange (costs of securing sufficient amounts of food). Combinations of loss of endowment and adverse shifts in the conditions of exchange (e.g., spikes in grain prices) can lead to certain classes of society being severely deprived of food. Component causes that lead to market failure-driven famine are complex, interacting over an extended time

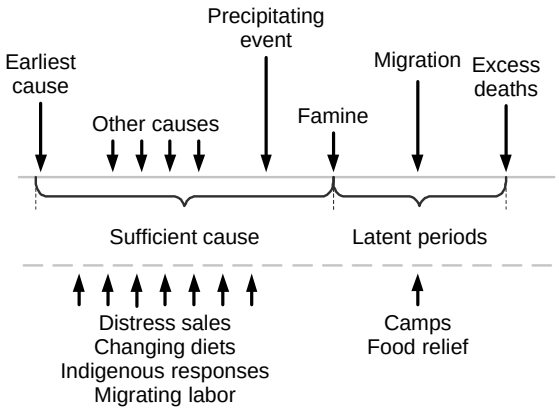


Figure 2 A model depicting actions of individual, or component, causes that can lead to a sufficient cause of famine, and societal, indigenous responses to famine predominantly caused by market failure. Famine may be latent or delayed from external view until migrations or excess deaths occur. Government relief is typically a late response to famine.

(**Figure 2**). Causes acting at various times in the pathway to market failure can be numerous, including long- and short-term adversity in climate leading to drought and excessive floods, pestilence and other causes of lost crop yield, reduced food imports or inefficient transport and marketing infrastructures. These all can lead to a national or, more often, regional declines in food availability, inflationary grain market responses to speculation and hoarding, other aspects of infrastructural neglect, ineffectual trade policies, political instability and corrupt governance, market depressions with year-round or seasonal job losses, and depletion of assets of the poor (endowment). Prior or present conflict can destabilize markets and contribute to such types of famine. Famines that can be classified as those primarily of market failure include the Great Irish Famine from 1844 to 1848, the Great Bengal Famine of 1943, the Bangladesh famine of 1974, and the Sudan famine of 1984–85. The Great Irish Famine was triggered by a potato blight that stripped the country of the only staple that Irish peasantry could afford to grow on their small parcels of land. Peasants who grew other staple grains had to sell them to pay rent to landlords. However, during these same years, there were substantial exports of wheat, barley, oats, and animal products by landowners to English markets. Food did not enter the local Irish markets because the peasants lacked effective demand.

Market or entitlement failures marked the last two great Bengal famines of the twentieth century: The Great Bengal Famine of 1943 and the Bangladesh Famine of 1974–75 (**Figure 1**). The 1943 famine, during which some 3 million people are estimated to have died, was originally judged by a Famine Inquiry Commission to be due to a shortage in rice supply. However, a seminal in-depth analysis years later by Sen showed that the famine occurred in a year during which rice production in Bengal was only 5% lower than the average of the previous 5 years. It was also a year when most economic indicators of Bengal were showing a 'boom' in growth due to World War II. Rural food stocks were being procured by the government to support military needs, subsidize rations for civil servants, and stabilize general prices of rice in Calcutta, which drove up the price of rice in rural areas. This practice, coupled with 'boat blockade' and 'rice denial' policies imposed in regions along the Bay of Bengal for reasons of defense, left certain low wage-earning rural classes (agricultural workers, day laborers, artisans, and fishermen) disentitled, and unable to acquire enough food for their own survival.

In Bangladesh, at least 100 000 people died between 1974 and 1975 in a famine that followed an unusually severe flood. During the several years leading up to the famine there were events that brought the country to a highly vulnerable state, including a devastating cyclone and tidal wave, a civil war that led to the country's independence, and a series of partial crop failures, all superimposed on preexisting high burdens of malnutrition, disease, underdevelopment, and ensuing political chaos. The flood in the middle of 1974 was expected to destroy much of the major 'aman' rice to be harvested a few months later. In anticipation of impending rice shortage, rural traders began to hoard grains in early September of that year causing rice prices to spike across the country's rural markets in a contagious pattern (**Figure 3**). Rice prices remained at about twice their normal level for months thereafter, even after it became evident that the speculated poor rice harvest was, in fact, a normal one. Thus, total and per capita aggregate grain supplies in Bangladesh remained at about average levels throughout the famine. Local area food deficits and hoarding of grains by traders led to the observed points of inflection in the price of rice throughout the country that caused the entitlements of rural wage earners to collapse, initiating a famine that resulted in extremely high mortality and massive migrations to urban centers in search of relief.

The Horn of Africa has been wracked by famine or famine-like conditions, leading to what have become classically defined as 'complex emergencies' for much of the past three decades. Aggregate food shortage has appeared to play a more variable and, at times, prominent role in recent famines in the eastern Horn. In Ethiopia, Sudan, Eritrea, and Somalia large tracts of land are drought-prone, average annual rainfall has been declining since the

Figure 3 Consecutive weekly maps of a contagious spread of spikes in the price of rice in local markets throughout rural Bangladesh from (A) late August 1974 through to (H) the end of October 1974 during a flood-associated period of a famine that reportedly killed from 100 000 to 1 million persons. (Adapted from Seaman J and Holt J (1980) Markets and famines in the third world. *Disasters* **4**(3): 283–297.)

1930s, and robust, indigenous farming and animal husbandry practices have been weakened as agricultural land has increasingly been used for growing export crops. In the Ethiopian famine of 1972–75, in which over 100 000 people died, national crop production dropped to only ~7% below normal levels, a decline that, like in Bengal in 1943 and 1974, would not have been expected to trigger a famine. However, crop production had been severely below normal in Wollo Province, where the famine began. Although the famine subsequently spread to other areas of the country, a reluctance by the government to formally recognize the famine and excessive delays in mobilizing and targeting food aid within country (whether from national or international stocks) were deemed responsible for unleashing a famine that, based on national stocks, should have been averted. Famines during 1982–85 in Ethiopia and the Sudan appeared to be more closely tied to gradual declines in national food security during the preceding decade. These trends were exacerbated by repressive governments enacting targeted, famine-promotive rather than preventive policies, resulting in civil wars and severely deteriorating economic conditions that were compounded by weak international food aid responses.

Armed Conflict

A second major class of famine comprises those precipitated or triggered by declared war or armed insurgency, leading to a siege or food blockade by a foreign power (e.g., Allied blockade of Germany in 1915–18; Nazi blockade of Holland precipitating the Dutch Winter Famine of 1944–45, and the Nazi siege of Lenningrad in 1942–44) or, as occurring more in recent years, severe civil war that disrupts normal markets as well as emergency food delivery systems (e.g., the Somalian civil war and famine of 1991–92). Armed conflict can incapacitate or destroy a country's ability to govern, develop, produce and feed itself domestically or through food aid, as scores of people become displaced, destitute, starve and die from severe malnutrition and epidemic illness. The

famine in Somalia in the early 1990s exemplifies the rapid emergence of military conflict as a precipitating cause of famine. With significant transfers of weaponry to rogue vigilante groups and increased deployments of land mines in other poor, warring countries in recent years, civil violence and lawlessness also pose a major hindrance to the effective provision of short-term relief during the acute phase of famine and to subsequent economic recovery.

Failure in Central Planning

A third class of modern famine, distinct from the other two, has resulted from failure by intent, indifference, ignorance, or incompetence of a centrally planned state to adequately provide food to all sectors of society, often as a result of totalitarian action to advance political goals outside of the rules of free trade or popular processes. Examples of this third type of famine in the twentieth century include those induced by the notorious policies of Stalin in Soviet Russia in the 1920s and 1930s. In an effort to achieve rapid industrial growth, Stalin waged class warfare among rural peasantry, abolished economic incentives, collectivized farms into massive (inefficient) production units and merged villages into socialist agro-towns, seized and exported grain for foreign exchange to fuel industrialization, restricted population movements across municipalities, and brutally suppressed all opposition. Agricultural production plummeted across regions of Russia leading to disastrous shortages (e.g., by 40% in some areas), further intensifying state seizures of food grain, especially in the grain-belt region of the Ukraine where Stalin sought to crush a nationalist revolt by forcibly extracting available food grains from the population. The actions induced the worst famine in Russian history. Between 1930 and 1937 it was estimated that nearly 15 million peasants died, of whom 7–8 million died in the Ukraine in 1933–34.

Under communist rule imposed by Mao Zedong, in 1959–60 China experienced the worst recorded famine in human history that left an estimated 30 million people dead. The Great Leap Famine was provoked through a causal chain of centrally planned policy steps during the preceding decade, modeled after Stalin and motivated by ill-conceived goals to 'Leap forward' MAO's aims were to achieve agricultural sufficiency and superiority through massive agricultural collectivization and the formation of huge peasant communes, and rapid rural industrialization through crash programs to increase steel production. The plight of tens of millions of rural peasants was tightly controlled by the state through brutal force, terror, propaganda, and state control of grain production, procurement and taxation motivated by a blind faith among civil servants in the vision and leadership of Mao. As a result of fabricated inflation of grain production figures, driven by a zeal to demonstrate success, China became a net exporter of more than a million metric tons of grain during the peak of famine mortality in the countryside in 1960, mimicking Stalinist Russia. Thus, in addition to events immediately leading to famine, some component causes contributing to the centrally planned Great Leap Famine can be traced back through the previous one to three decades and to influences beyond the borders of China.

Communist North Korea's inability to avert famine in 1997–98 amounts to the most recent example of a central planning failure, conditioned by chronic food insecurity over the previous decade and precipitated by poorly timed, torrential rains and floods in 1995–96 and drought in 1997. However, some causal elements related to how slowly and secretively the isolationist government responded, actions of governance that date back to the Korean War and Cold War politics, and politicization of food aid.

Coping Strategies

Most is known about household and community coping mechanisms in response to famines due to market failure. In cultures where food shortage or inaccessibility to large sectors of society is chronic, and threat of famine periodic, there exist indigenous responses that enable the local populace to cope, protect their entitlement, and minimize as best it can the risk of starvation as terms of exchange for food deteriorate (illustrated as a concept in **Figure 4**).

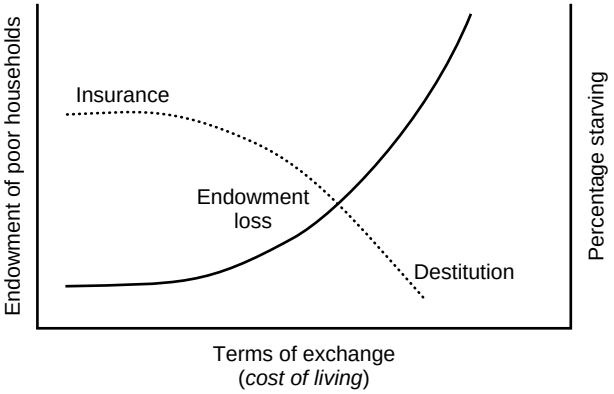


Figure 4 Illustration of collapse in entitlement. As endowment of the poor decreases toward a state of destitution with increasingly severe (costly) terms of exchange for food, the risk of starvation and famine increases.

A first line of responses may be viewed as 'insurance' against uncertainty; these are activities that can stem loss of endowment, such as restructuring the mix of crops grown or pastoral practices in ways that insulate against drought- or flood-induced shortages. Examples include planting more robust crops, dispersing crops across a wider area, staggering plantings, or increasing livestock diversity and mobility. Food preservation practices and dietary changes to include less commonly eaten foods can initially increase the size and diversity of the food base. As terms of exchange become worse, coping mechanisms aimed at survival increasingly cost households their endowment. These responses include working longer and at different jobs for lower wages, migrating far from home to find marginal work, reducing meal frequency, consuming the next planting's seeds, and expanding intake to include 'famine foods' poor in, or lacking, nutritional quality. At first these may include unusual tubers, leaves, flowers, and other plants. Household assets such as pots, utensils, watches, and small animals are increasingly sold as, eventually, are larger assets such as bullock carts, bicycles, and draft animals. Land mortgage or sales transactions become more numerous. With indebtedness and destitution, petty crime and child abandonment increase; famine foods may include tree bark, ground bone, and rodents; suicide and cannablism may occur. An indicator of severe entitlement loss in a community is the livestock-to-grain price ratio in local markets. Normally this ratio is of a figure that reflects the greater asset value of livestock compared to grain. However, it may invert as the cost of grain and feeding animals and the level of animal wasting all continue to rise, such that, at a peak of famine vulnerability, large numbers of animals may be sold at very low prices relative to the costs of grain.

Viewed over time, famine is a continuum. As household and community entitlements erode for increasing numbers due both to deteriorating conditions of exchange and endowment loss, destitution and starvation become more likely. **Figure 5** depicts a hypothetical shift in distribution of starving individuals in a poor population exposed to increasing risk of famine, where under usual conditions a small proportion of individuals routinely face the threat of starvation and wasting malnutrition (top panel). During periods of high or repeated stress, such as those of prolonged drought and internal conflict, while the population faces less food security coping mechanisms continue to protect most vulnerable groups from abject starvation, even as they near such a 'threshold' amidst inevitable losses of human and economic asset (middle panel). During severe distress of famine, entitlement has collapsed for the most vulnerable classes of society, pushing large numbers of persons into a state of starvation, leaving them destitute and migrating or dying (bottom panel). However, not all individuals starve. Some segments of society lose little or no economic ground, or benefit considerably from the plight of others by acquiring property and other assets at low prices, obtain labor at reduced wages or lend money at high interest rates. Still other segments, particularly those trading in famine relief goods and services, stand to gain large profits throughout the famine and recovery periods (depicted by the right skew). Postfamine, the economic landscape is nearly always one of greater polarization of wealth and an increase in size and vulnerability of society's poor and destitute. Peri-urban slums typically remain swollen following famine as a result of permanent migration.

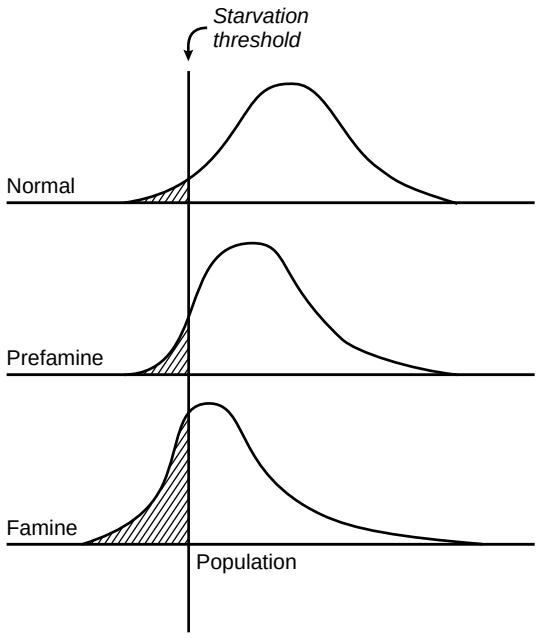


Figure 5 Shifting of a high-risk, undernourished population toward increased starvation during prefamine and famine conditions, particularly those most vulnerable. Truncated left tail area reflects hypothetical effects of coping strategies that prevent starvation. Right skew reflects polarizing of wealth, with some sectors profiting from famine.

Government and International Responses

Famine through the ages has invoked from law abiding governments preventive action, where believed indicated, and relief responses in the face of imminent catastrophe. In Genesis, Pharoah's grain taxes during years of plenty were aimed at relieving dwindling food stores in famine. During China's

Eastern Chou and Ch'in dynasties of the third century BC, as well as in India over 2000 years ago, steps formulated to prevent or relieve famine included disaster reporting procedures, cropping alterations, grain distribution, feeding kitchens, tax remissions, vulnerable group relocation, and public works construction to facilitate irrigation, food shipment or flood control. In sixteenth century England, to counter inflationary effects of speculative grain hoarding, the Tudor First Book of Orders called for enforced extraction and marketing of private grain stocks as a way to control staple prices and thwart famine. Policy response can also amount to inaction. The Great Irish Famine from 1844 to 1848 evoked a different response from the British Government: a flawed 'laissez-faire' policy intending to allow market forces to equilibrate on their own to meet local food needs, a course that never materialized as entitlement collapsed among Irish peasantry. However, learning from a century of repeated famine, Famine Codes emerged in British India in 1880 that called for massive public works coupled with food distribution and feeding centers for vulnerable groups, which served as the core famine relief policy on the subcontinent for more than a half century and have continued to guide famine relief efforts to the present day.

Today, modern preventive response by international agencies and governments can be informed and guided by surveillance systems with regional, national, and local data collection mechanisms. Examples are the Famine Early Warning System (FEWS), which functions across Sub-Saharan Africa and has been supported by the US Agency for International Development over the past two decades and the Global Information Early Warning System (GIEWS) managed by the Food and Agricultural Organization of the United Nations (FAO). The primary aim of surveillance is to detect worsening conditions in high-risk populations in sufficient time to permit effective preventive or pre-emptive action. The task is a 'tall order' given widespread, often complex, component causes that must converge in certain ways to cause famine, against a usual plethora of endemic risk factors. With early, adequate, and effective response serving as the criterion of success, modern surveillance has so far failed to prevent famine. In part, this may reveal a basic epidemiologic dilemma: Against a background of profound, widespread economic and nutritional need throughout the developing world, including numerous prefamine but intact situations arising under surveillance, famine is a rare event. Even with presumed high sensitivity and specificity, low predictive value stemming from infrequent occurrence makes action to prevent a particular famine unlikely given the enormous political and financial resources required to mount preventive responses.

Thus, the most effective preventive action relates to setting and enacting a development agenda that recognizes high risk areas and seeks to strengthen the productivity and well-being of famine-vulnerable population groups in those areas of a country. These can include boosting infrastructural, commercial, education, agricultural, and other inputs into priority areas that improve long-term economic conditions.

Preemptive government policies are directed toward relieving a prefamine condition once it becomes apparent. Setting up famine early warning systems that monitor climatic, agricultural, population mobility, economic, and nutritional indicators is considered preemptive in that such information is intended to identify high-risk trends so that corrective action could be taken long before famine becomes imminent. Normally, early warning surveillance is only possible in high-risk countries with significant international assistance. Another example is a government making large purchases of food on the international market and releasing the commodities through ration shops, food-for-work and other programs that do not disrupt the local food economy but stabilize local grain market prices instead as a means to prevent speculation throughout the period of high risk.

Lagged or relief-oriented responses comprise emergency responses to acute and enormous need that typically are enacted after famine begins and its harsh consequences are already evident in a population. These actions, usually in coordination with major international relief and donor agencies, are typically intended to relieve acute suffering and death and promote the rehabilitation of those masses who have survived to migrate, and reach encampments. By definition, lagged responses represent policy failure for governments intending to minimize the destruction, malnutrition, and mortality of famine.

See also: **Hunger**. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **Nutrition Policies In Developing and Developed Countries**. **Starvation and Fasting**.

Further Reading

Ahmed R, Haggblade S, and Chowdhury TE (2000) *Out of the Shadow of Famine: Evolving Food Markets and Food Policy in Bangladesh* Baltimore: Johns Hopkins University Press.

Aykroyd WR (1974) *The Conquest of Famine*. London: Chatto & Windus.

The Bible. Book of Genesis **47**: 4–26.

Cuny FC (1999) *Famine, Conflict and Response: A Basic Guide* West Harford: Kumarian Press.
Dreze J and Sen A (eds.) (1990) *The Political Economy of Hunger: Famine Prevention*, vol. 2: *WIDER Studies in Developmental Economics*, pp. 1–400. Oxford: Clarendon Press.
Edkins J (1996) Legality with a vengeance: Famines and humanitarian relief in "complex emergencies." *Millenium: Journal of International Studies* **25**: 547–575.
Newman LF (ed.) (1992) *Hunger in History: Food Shortage, Poverty and Deprivation*. Oxford: Blackwell.
Ravallion M (1997) Famines and economics. *Journal of Economic Literature* **35**: 1205–1242.
Rothman K and Greenland S (1998) *Modern Epidemiology*, pp. 7–28. Philadelphia: Lippincott-Raven.
Scrimshaw NS (1987) The phenomenon of famine. Annual Review of Nutrition **7**: 1–21.
Seaman J and Holt J (1980) Markets and famines in the third world. *Disasters* **4**(3): 283–297.
Sen A (1977) Starvation and exchange entitlements: a general approach and its application to the great Bengal famine. *Cambridge Journal of Economics* **1**: 33–59.
Sevoy RE (1986) *Famine in Peasant Societies*. New York: Greenwood Press.
Yang DL (1996) *Calamity and Reform in China: State, Rural Society and Institutional Change since the Great Leap Forward*. Stanford: Stanford University Press.
Yip R (1997) Famine. In: Noji EK (ed.) *Public Health Consequences of Disasters*, pp. 305–335 New York: Oxford University Press.

Fat-Soluble Vitamins *see* **Vitamin A**: Biochemistry and Physiological Role. **Vitamin D**: Physiology, Dietary Sources and Requirements; Rickets and Osteomalacia. **Vitamin E**: Metabolism and Requirements. **Vitamin K**

Fat Stores *see* **Adipose tissue**

Fats *see* **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Lipids**: Chemistry and Classification; Composition and Role of Phospholipids

FATS AND OILS

A H Lichtenstein, Tufts University, Boston MA, USA

Dietary fat is a macronutrient that has historically engendered considerable controversy and continues to do so. Contentious areas include optimal type and amount in the diet, role in body weight regulation, and importance in the etiology of chronic disease(s).

Dietary Fats and Oils: The Good, Bad, and Ugly

Dietary fats and oils are unique in modern times in that they have good, bad, and ugly connotations. The aspects of dietary fat that are classified as good include serving as a carrier of preformed fat-soluble vitamins, enhancing the bioavailability of fat-soluble micronutrients, providing essential substrate for the synthesis of metabolically active compounds, constituting critical structural components of cells membranes and lipoprotein particles, preventing carbohydrate-induced hypertriglyceridemia, and providing a concentrated form of metabolic fuel in times of scarcity. The aspects of dietary fat that can be classified as bad include serving as a reservoir for fat-soluble toxic compounds and contributing dietary saturated and *trans* fatty acids, and cholesterol. Aspects of dietary fat that can be classified as ugly include providing a concentrated form of metabolic fuel in times of excess and comprising the major component of atherosclerotic plaque, the

underlying cause of heart disease, stroke, and phlebitis.

Lipids in Food and in the Body

Fatty Acids

Fatty acids are hydrocarbon chains with a methyl and carboxyl end. The majority of dietary fatty acids have an even number of carbons. The range in chain length of common dietary fatty acids is broad. Fatty acids with 16 and 18 carbons make up the majority of fatty acids present in plants and animals. However, they are by no means the most metabolically active. Long-chain unsaturated fatty acids, such as arachidonic acid (C20:4), are common precursors of regulatory compounds.

Essential nutrients are those that the body cannot synthesize or cannot synthesize in amounts adequate to meet needs. Linoleic acid (18:2) and/or fatty acids that can be derived from linoleic acid are essential fatty acids. These specific fatty acids are essential because humans cannot introduce a double bond above the ninth carbon from the carboxyl end of the acyl chain. To maintain optimal health, they must be supplied by the diet of humans. The metabolism of linoleic acid is represented in **Figure 1**.

A wide range of fatty acids occur in nature. There are a number of features of fatty acids that distinguish one from another. In addition to chain length, they also vary with regard to degree of saturation and location of the double bond(s). Fatty acids with a single double bond are referred to as monounsaturated fatty acids, and those with two or more double bonds are referred to as polyunsaturated fatty acids (**Figure 2**). The double bonds within unsaturated fatty acids can either be in the *cis* (hydrogen atoms on the same side of the acyl chain) or *trans* (hydrogen atoms on opposite sides of the acyl chain) conformation (**Figure 3**). The *cis* conformation is most commonly found in nature. Double bonds can also vary with regard to location within the acyl chain. The presence of double bonds, per se, and their number, position, and conformation, dictates the physical properties of the fatty acids.

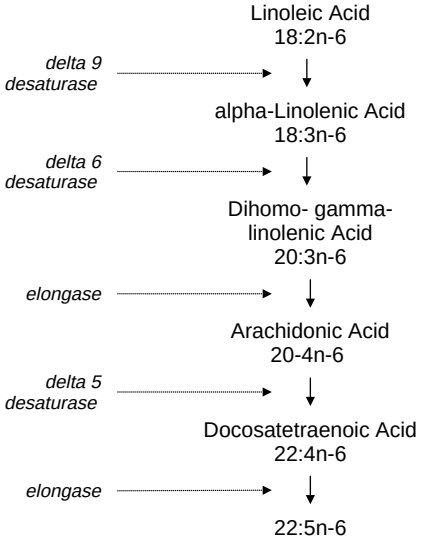


Figure 1 Metabolism of linoleic acid.

Unsaturated fatty acids of the same length with an identical number of double bonds can occur in multiple forms due to variation in the conformation of one or more of the double bonds (*cis* versus *trans*). They are referred to as geometric isomers (**Figure 3**). A common example is oleic acid (18:1c-9) and elaidic acid (18:1t-9). The presence of a *cis* relative to a *trans* double bond results in a greater bend or kink in the hydrocarbon chain. This kink impedes the fatty acids from aligning or packing together, thereby lowering the melting point of the fat. In a cell membrane this will be reflected in increased fluidity. In food this will be reflected in an oil that is liquid or fat that is soft at room temperature.

Unsaturated fatty acids of the same length with an identical number of double bonds and conformation can also occur in multiple forms due to the location of the double bonds within the acyl chain. They are referred to as positional isomers. A common example is alpha-linolenic acid (18:3n-3) and gamma-linolenic acid (18:3n-6). The difference in location of double bonds results in small alterations to the melting point yet large differences in the metabolic properties of the fatty acids. The most common distinction made among positional isomers of fatty acids is the location of the first double bond from the methyl end of the acyl chain. A fatty acid in which the first double bond occurs three carbons from the methyl end is termed an omega-3 fatty acid, frequently denoted n-3 fatty acid. This class of fatty acids is distinguished from the major class of fatty acids in which the first double bond occurs six carbons from the methyl end, termed an omega-6 or n-6 fatty acid. Enzymes that metabolize fatty acids distinguish among both positional and geometric isomers. The metabolic products of the different positional isomers of fatty acids have different and, occasionally, opposite physiological effects.

Most double bonds within fatty acids occur in a nonconjugated sequence, both in the human body and in food. That is, a carbon atom with single carbon–carbon bonds separates the carbons making up the double bonds. Some double bonds occur in

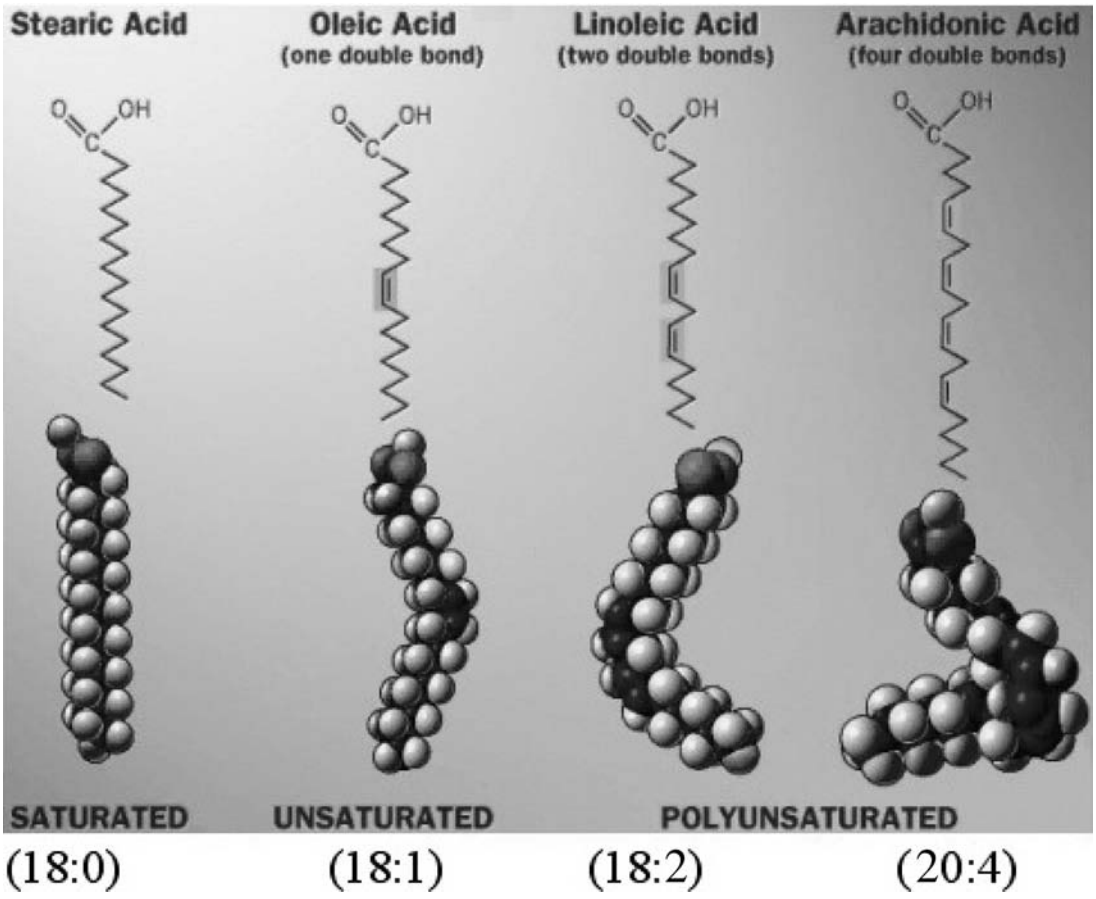


Figure 2 Saturated, monounsaturated, and polyunsaturated (n-3 and n-6) acids.

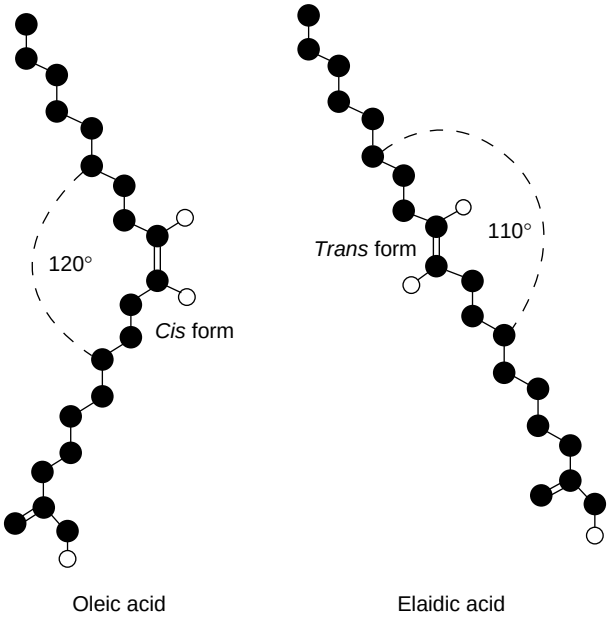


Figure 3 *Cis* and *trans* double-bond-containing fatty acids. (Copyright © The McGraw-Hill Companies, Inc.)

the conjugated form, without an intervening carbon atom separating the double bonds. Conjugated double bonds tend to be more reactive chemically (i.e., more likely to become oxidized). Although there is considerable speculation about the role of conjugated double bond-containing fatty acids and human health, the current state of knowledge is insufficient to draw any firm conclusions.

Triacylglycerol

Triacylglycerol is the major form of dietary lipid in fats and oils, whether derived from plants or animals. Triacylglycerol is composed of three fatty acids esterified to a glycerol molecule (**Figure 4**). The physical properties of the triacylglycerol are determined by the specific fatty acids esterified to the glycerol moiety and the actual position the fatty acid occupies. Each of the three carbons comprising the glycerol molecule allows for a stereochemically distinct fatty acid bond position: sn-1, sn-2, and sn-3. A triacylglycerol with three identical fatty acids is termed a simple triacylglycerol. These are exceedingly rare in nature. A triacylglycerol with two or three different fatty acids is termed a mixed

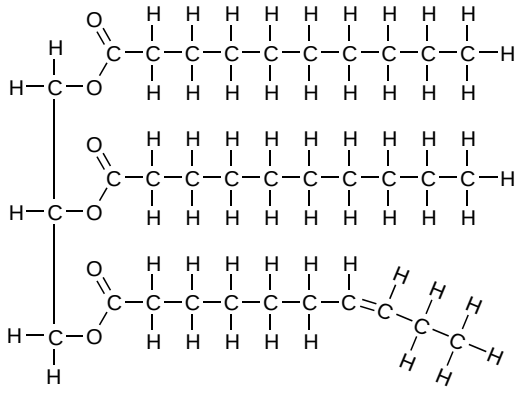


Figure 4 Triacylglycerol.

triacylglycerol, and these make up the bulk of the fat both in the human diet and in the body. The melting point of a triacylglycerol is determined by the position of the fatty acids esterified to glycerol and physical characteristics—their chain length and number, position, and conformation of the double bonds, and the stereochemical position.

Approximately 90% of the molecular weight of triacylglycerol is accounted for by the fatty acids. The fatty acid profile of the diet is reflected, in part, in the fatty acid profile of the adipose tissue triacylglycerol. Such data have been used to approximate long-term food intake patterns of humans. Manipulating the dietary fat provided to domesticated animals is being considered as one approach to modifying the fatty acid profile of meat.

Mono- and diacylglycerols have one and two fatty acids, respectively, esterified to glycerol. They rarely occur in large quantities in nature. Mono- and diacylglycerols are primarily intermediate products of triacylglycerol digestion and absorption, clearance from the bloodstream, or intracellular metabolism. They are frequently added to processed foods because of their ability to act as emulsifiers. Their presence in food products is noted on ingredient labels.

Once consumed, triacylglycerol are hydrolyzed to free fatty acids and monoglycerides in the small intestine prior to absorption. These compounds enter the intestinal cell and are used to resynthesize triacylglycerol. This lipid is then incorporated into a nascent triacylglycerol-rich lipoprotein particle, termed chylomicron, for subsequent release into peripheral circulation. Chylomicrons are secreted directly into the lymph prior to entering the bloodstream. Once in circulation, triacylglycerol are hydrolyzed before crossing the plasma membrane of peripheral cells for subsequent metabolism. The primary enzyme that hydrolyzes triacylglycerol in plasma is lipoprotein lipase. Lipoprotein lipase hydrolyzes triacylglycerol into two free fatty acids and 2-monoacylglycerol. The enzyme is attached to the luminal surface of capillary endothelial cells via a highly charged membrane-bound chain of heparin sulfate–proteoglycans. The ability of lipoprotein lipase to bind both the chylomicron particle and the cell surface ensures the cellular uptake of free fatty acids that are generated from the hydrolysis. Once inside the cell, free fatty acids can be oxidized to provide energy, metabolized to biologically active compounds, incorporated into phospholipid and cholesteryl ester, or resynthesized into triacylglycerol for storage as a potential reservoir of fatty acids for subsequent use.

Phospholipid

There are only trace amounts of phospholipid in dietary fats and oils. However, because the fatty acids in fats and oils provide substrate for the synthesis of phospholipid in the body, it is important to discuss this subtype of fat. Phospholipid is a critical structural component of all cells, both plant and animal. It is composed of two fatty acids on the sn-1 and sn-2 positions and a moiety frequently referred to as a polar head group on the sn-3 position of glycerol, the latter via a phosphate bond (**Figure 5**). Phospholipid molecules are amphipathic—that is, there are both hydrophobic and hydrophilic domains in the molecule. The two fatty acids confer hydrophobic properties and the polar head group hydrophilic properties. The specific fatty acids esterified to the glycerol backbone tend to be unsaturated fatty acids. The different polar head groups, most commonly phosphorylcholine, phosphorylserine, phosphorylinositol, or phosphorylethanolamine, result in phospholipids that vary in size and charge. Due to their amphipathic nature, phospholipids serve as the major structural component of cellular membranes by forming bilayers and in so doing also serve as a reservoir for metabolically active unsaturated fatty acids. Due to their amphipathic properties, in the

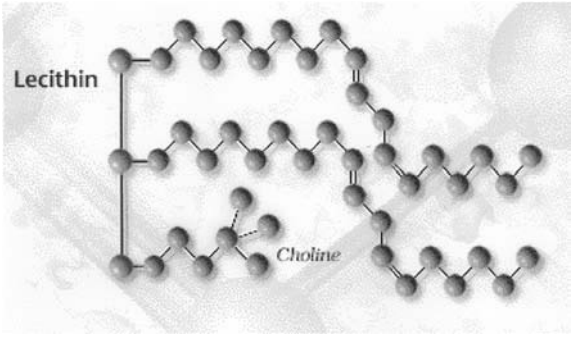


Figure 5 Phospholipid.

small intestine they play an important role in the emulsification and absorption of dietary fat and fat-soluble vitamins. On the surface of lipoprotein particles, they provide a critical component in the packaging and transport of lipid in circulation.

Cholesterol

Dietary sources of cholesterol are limited to foods of animal origin. Cholesterol is an amphipathic molecule that is composed of a steroid nucleus and a branched hydrocarbon tail (**Figure 6**). Cholesterol occurs naturally in two forms, either as free (nonesterified) cholesterol or esterified to a fatty acid (cholesteryl ester). If esterified, the fatty acid is linked to cholesterol at the number 3 carbon of the sterol ring.

Cholesterol serves a number of important functions in the body. Free cholesterol is a component of cell membranes and along with the fatty acid profile of the phospholipid bilayer determines membrane fluidity. The intercalation of free cholesterol into the phospholipid bilayer restricts motility of the fatty acyl chains and hence decreases fluidity. Free cholesterol is critical for normal nerve transmission. It makes up approximately 10% (dry weight) of total brain lipids. Cholesterol is a precursor of steroid hormones (i.e., estrogen and testosterone), vitamin D, adrenal steroids (i.e., hydrocortisone and aldosterone), and bile acids. This latter property is exploited in certain approaches to decrease plasma cholesterol levels by preventing the resorption of bile acids (recycling) and hence forcing the liver to use additional cholesterol for bile acid synthesis and in so doing creating an alternate mechanism for cholesterol excretion.

The receptor-mediated cellular uptake of cholesterol from lipoprotein particles is critical to maintaining intracellular and whole body cholesterol homeostasis. Once internalized, lipoprotein-associated cholesterol that is released from lysosomes has three major effects in the cell. The free cholesterol inhibits the activity of 3-hydroxy 3-methylglutaryl CoA reductase, the rate-limiting enzyme in endogenous cholesterol biosynthesis. This property serves to decrease cholesterol biosynthesis commensurate with the uptake of cholesterol from circulating lipoprotein particles and hence protects the cell from accumulating excess intracellular cholesterol. Free cholesterol inhibits the synthesis of receptors that mediate the uptake of lipoproteins from the bloodstream, thereby limiting the amount of additional cholesterol taken up by the cell. Free cholesterol increases the activity of acyl CoA cholesterol acyltransferase (ACAT), the intracellular enzyme that converts free cholesterol to cholesteryl ester. A high level of intracellular free cholesterol is cytotoxic, whereas cholesteryl ester is a highly nonpolar molecule and coalesces into a lipid droplet within the cell, preventing interaction with intracellular components. Increased ACAT activity is an important mechanism in preventing the accumulation of intracellular free cholesterol.

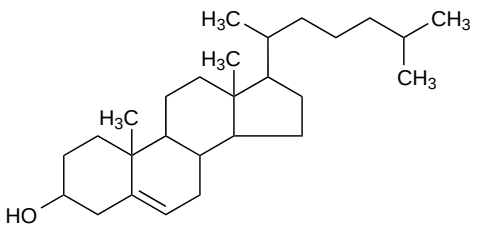


Figure 6 Cholesterol.

Cholesterol can be esterified intracellularly, as previously indicated, by ACAT. ACAT uses primarily oleoyl CoA as substrate and the resulting product is primarily cholesteryl oleate. Cholesterol can also be esterified in plasma by lecithin cholesterol acyltransferase (LCAT). LCAT uses phosphotidylcholine as substrate; the resulting products are primarily cholesteryl linoleate and lysolecithin. Cholesteryl ester is less polar than free cholesterol and this difference dictates how the two forms of cholesterol are handled—intracellularly and within lipoprotein particles.

Approximately one-third of cholesterol in plasma circulates as free cholesterol and approximately two-thirds as cholesteryl ester. Cholesterol in circulation is carried on all the lipoprotein particles (both intestinally derived chylomicrons and hepatically derived very low-density lipoprotein) or those generated during the metabolic cascade (intermediate-density lipoprotein, low-density lipoprotein (LDL), and high-density lipoprotein (HDL)). Free cholesterol is sequestered on the surface of lipoprotein particles within the phospholipid monolayer, whereas cholesteryl ester resides in the core of the lipoprotein particle. The majority of the cholesterol in circulation is carried on LDL particles. Cholesteryl ester is the major component of atherosclerotic plaque. In the arterial wall, cholesteryl ester is derived from the infiltration of lipoprotein-associated cholesteryl ester resulting from LCAT activity or is synthesized in situ as a result of ACAT activity. The fatty acid profile of the cholesteryl ester in arterial plaque can provide some indication of its source.

Other Sterols

Fats and oils derived from plants contain a wide range of phytosterols, compounds structurally similar to cholesterol. The difference between

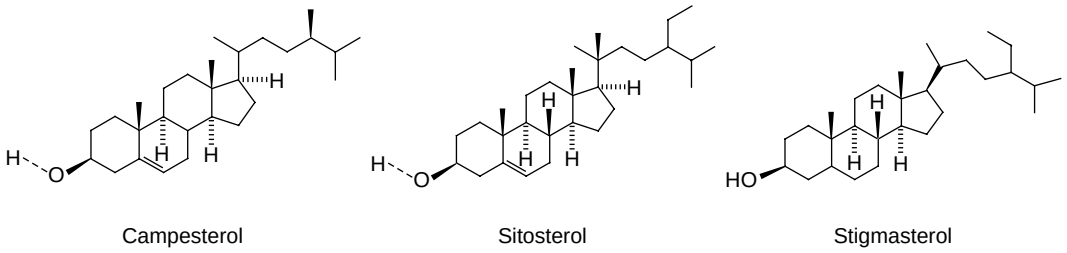


Figure 7 Plant sterols.

phytosterols and cholesterol is related to their side chain configuration and/or steroid ring bond patterns. The most common dietary phytosterols are beta-sitosterol, campesterol, and stigmasterol (**Figure 7**). In contrast to cholesterol, phytosterols are only absorbed in trace amounts. For this reason, plant sterols have been used therapeutically to reduce plasma cholesterol levels. They compete with cholesterol for absorption; hence, they effectively reduce cholesterol absorption efficiency.

The absorption efficiency of cholesterol in humans ranges from approximately 40 to 60%. Because the relative absorption of plant sterols, however low, is correlated with the percentage of cholesterol absorbed in an individual, there is considerable interest in using circulating plant sterol concentrations as a surrogate marker for cholesterol absorption efficiency. Limited data suggest efficiency of cholesterol absorption may have a significant effect on lipoprotein profiles and cardiovascular disease risk. Whether circulating phytosterols have an independent effect on cardiovascular disease risk is under investigation.

Dietary Fats and Oils and Cholesterol

Dietary fat serves critical functions in the human body. It provides a concentrated source of energy, slightly more than twice per gram than protein or carbohydrate. For this reason, the causes of energy imbalances are often attributed to this component of the diet. However, definitive data in this area are lacking.

In addition to providing a source of metabolic energy, dietary fat provides a source of essential fatty acids, linoleic acid (18:2), and/or other fatty acids that are derived from linoleic acid. Dietary fat is the major carrier of preformed fat-soluble vitamins (vitamins A, D, E, and K). The bioavailability of these fat-soluble vitamins is dependent on fat absorption. Dietary fatty acids are incorporated into compounds that serve as structural components of biological membranes and lipoproteins, and as such they serve as a reservoir for fatty acids having subsequent metabolic fates.

Fatty Acid Profile of Common Dietary Fats

Dietary fats and oils derive from both animal and plant sources, primarily in the form of triacylglycerol. The fatty acid profile of dietary fats commonly consumed by humans varies considerably (**Figure 8**). In general, fats of animal origin tend to be relatively high in saturated fatty acids, contain cholesterol, and are solid at room temperature. A strong positive association has been demonstrated in epidemiological, intervention, and animal data between cardiovascular disease risk and intakes of saturated fatty acids. The exception is stearic acid (18:0), a saturated fatty acid of which a large proportion is metabolized to oleic acid (18:1), a monounsaturated fatty acid. Fats and oils of plant origin tend to be relatively high in unsaturated fatty acids (both monounsaturated and polyunsaturated) and are liquid at room temperature. Notable exceptions include plant oils termed tropical oils (palm, palm kernel, and coconut oils) and hydrogenated fat. Tropical oils are high in saturated fatty acids but remain liquid at room temperature because they contain a high proportion of short-chain fatty acids. Hydrogenated plant oils can be relatively high in saturated and/or *trans* fatty acids due to chemical changes induced during processing, including conversion of unsaturated to saturated bonds and *cis* to *trans* double bonds.

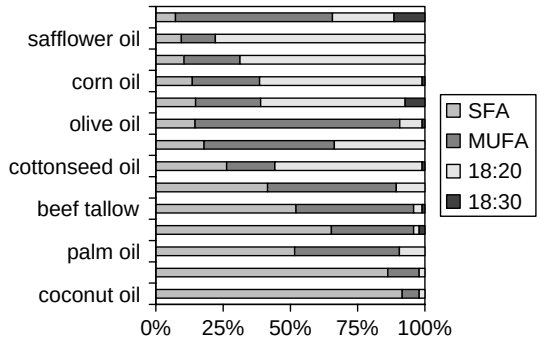


Figure 8 Relative composition of common dietary fats.

Major Contributors of Dietary Saturated, Monounsaturated, and Polyunsaturated Fatty Acids and Cholesterol

The major types of dietary fats and oils are generally broken down on the basis of animal and plant sources. The relative balance of animal and plant foods is an important determinant of the fatty acid profile of the diet. However, with the increasing prominence of processed, reformulated, and genetically modified foods, it is becoming more difficult to predict the fatty acid profile of the diet on the basis of the animal verses plant distinction.

According to the National Health and Nutrition Examination Survey (NHANES) recall data from 1999–2000, the 10 major dietary sources of saturated fatty acids in US diets are regular cheese (6.0% of the total grams of saturated fatty acids consumed), whole milk (4.6%), regular ice cream (3.0%), 2% low-fat milk (2.6%), pizza with meat (2.5%), French fries (2.5%), Mexican dishes with meat (2.3%), regular processed meat (2.2%), chocolate candy (2.1%), and mixed dishes with beef (2.1%). Hence, the majority of saturated fatty acids are contributed by regular dairy products (16%), and the top 10 sources contribute 30% of the total saturated fatty acids consumed. The increased prevalence of fat-free and low-fat dairy products provides a viable option with which to encourage a populationwide decrease in saturated fat intake. To put the value of decreasing populationwide intakes of saturated fat into perspective, it has been estimated that the isocaloric replacement of 5% of energy from saturated fatty acids with complex carbohydrate, on average, would reduce total cholesterol levels by 10 mg/dl (0.26 mmol/l) and LDL cholesterol by 7 mg/dl (0.18 mmol/l). For a person at moderately high risk of developing cardiovascular disease with a total cholesterol level of 220 mg/dl (5.69 mmol/l) and LDL cholesterol level of 140 mg/dl (3.62 mmol/l), such a dietary modification would decrease total and LDL cholesterol levels by 4.5 and 5%, respectively. Each 1% decrease in total cholesterol levels has been associated with a 2% reduction in the incidence of coronary heart disease. Using this example, that would theoretically translate into a 9% decrease in cardiovascular disease risk. However, it is important to note that decreasing the saturated fatty acid content of the diet should not necessarily be done by displacing fat with carbohydrate. As discussed in the next section, the quantity of dietary fat, relative to carbohydrate and protein, also impacts on blood lipid levels and lipoprotein profiles.

The 10 major dietary sources of monounsaturated fatty acids in US diets are French fries (3.3% of the total grams of monounsaturated fatty acids consumed), regular processed meat (2.5%), regular cookies (2.5%), regular miscellaneous snacks (2.4%), pizza with meat (2.4%), regular salad dressing (2.4%), regular cheese (2.3%), Mexican dishes with meat (2.3%), sausage (2.1%), and mixed dishes with beef (2.1%). There is little change in total or LDL cholesterol levels from the isocaloric replacement of monounsaturated fatty acids by complex carbohydrate. However, it is important to note that approximately one-half of the monounsaturated fatty acids consumed in the United States come from animal fats. Therefore, a decrease in saturated fatty acid intake would be predicted to decrease monounsaturated fatty acid intake unless vegetable oils high in monounsaturated fatty acids, such as canola or olive oil, replaced the animal fat.

The 10 major dietary sources of n-6 polyunsaturated fatty acids in US diets are regular salad dressing (8.8% of the total grams of polyunsaturated fatty acids consumed), regular white bread (4.2%), regular mayonnaise (3.0%), French fries (2.6%), regular cake (2.5%), regular cookies (2.1%), mixed dishes with chicken and turkey (2.1%), regular miscellaneous snacks (2.0%), regular potato chips (2.0%), and fried fish (2.0%). The distribution of polyunsaturated fatty acids among commonly consumed foods is wide. It has been estimated that the isocaloric replacement of complex carbohydrate with polyunsaturated fatty acids for 5% of energy, on average, will reduce total cholesterol levels by 5 mg/dl (0.13 mmol/l) and LDL cholesterol by 4 mg/dl (0.11 mmol/l). For a person at moderately high risk of cardiovascular disease with a total cholesterol level of 220 mg/dl (5.69 mmol/l) and LDL cholesterol level of 140 mg/dl (3.62 mmol/l), such a dietary modification would decrease total and LDL cholesterol levels by 2.1 and 3.6%, respectively, and potentially result in a 4% decrease in cardiovascular disease risk.

The 10 major dietary sources of cholesterol in US diets are fried eggs (16.6% of the total milligrams of cholesterol consumed), regular eggs including scrambled eggs (8.4%), mixed dishes with eggs (4.5%), mixed dishes with beef (2.9%), whole milk (2.6%), regular cheese (2.5%), fried fish (2.3%), mixed dishes with chicken and turkey (2.3%), lean cut meat (2.1%), and regular processed meat (2.1%). Eggs or foods high in eggs contribute approximately 30% of the total dietary cholesterol intake. It has estimated that reducing cholesterol intakes by 200 mg/day, on average, will reduce total cholesterol levels by 5 mg/dl (0.13 mmol/l) and LDL cholesterol by 2.6 mg/dl (0.10 mmol/l). Such a change would be predicted to have a similar risk effect as displacing 5% of energy as carbohydrate with polyunsaturated

fatty acids—that is, reducing cardiovascular disease risk by approximately 4%.

Dietary Fat and Cardiovascular Prevention

Amount in Diet

When considering the percentage of energy contributed by dietary fats and oils (amount of fat) and cardiovascular disease prevention and management, there are two major factors—the impact on plasma lipoprotein profiles and body weight. The potential relationship with body weight is important because overweight and obesity are strongly associated with elevated lipid and lipoprotein levels, blood pressure, dyslipidemia, and type 2 diabetes—all potential risk factors for cardiovascular disease. With respect to plasma lipoprotein profiles, the focus is usually on triglyceride and HDL cholesterol levels or total cholesterol:HDL cholesterol ratios.

When body weight is maintained at a constant level, decreasing the total fat content of the diet, expressed as a percentage of total energy, and replacing it with carbohydrate frequently results in an increase in triglyceride levels, decrease in HDL cholesterol levels, and a less favorable (higher) total cholesterol:HDL cholesterol ratio. Low levels of HDL cholesterol are an independent risk factor for cardiovascular disease (<50 mg/dl in females [<1.3 mmol/l] and <40 mg/dl in males, [<1.0 mmol/l]). Very low-fat diets are of particular concern in diabetic and overweight individuals who tend to have low HDL cholesterol and high triglyceride levels or those individuals classified as having metabolic syndrome (having three or more of the following: abdominal obesity, elevated triacylglycerol levels, low HDL levels, hypertension, or elevated fasting glucose levels). Because of these findings, the Adult Treatment Panel of the National Cholesterol Education Program (NCEP) revised its guidelines in 2001 from recommending a diet with less than 30% of energy as fat to a diet with 25–35% of energy as fat. Similarly, the American Heart Association and the USDA/HHS 2000 Dietary Guidelines for Americans changed their recommendations to shift the emphasis from a general recommendation to limit intakes of total and saturated fat to limit saturated and *trans* fat.

With respect to the amount of dietary fats and oils and body weight, two reviews of the long-term data have been published. Both concluded that even a relatively large downward shift in dietary fat intake, approximately 10% of energy, results in only modest weight loss of 1.0 kg during a 12-month period in normal weight subjects and 3 kg in overweight or obese subjects. However, it is important to note that in contrast to what would have been predicted, during the course of the studies included in the reviews, in no case was weight gain reported.

Fatty Acid Profile

Early evidence demonstrated that diets relatively high in saturated fatty acids increased plasma total cholesterol levels. Subsequent work demonstrated that this elevation in total cholesterol levels is contributed to by increases in both LDL and HDL cholesterol levels. It also became clear that not all saturated fatty acids had identical effects on plasma lipoprotein levels. Very short-chain fatty acids (6:0 to 10:0) and stearic acid (18:0) produce little or no change in blood cholesterol levels, whereas saturated fatty acids with short- and intermediate-chain lengths—lauric (12:0), myristic (14:0), and palmitic (16:0) acids—appear to be the most potent in increasing blood cholesterol levels. Because a large proportion of stearic acid (18:0) is rapidly converted to oleic acid (18:1), it appears to have a relatively neutral effect. The underlying mechanism by which saturated fatty acids with 10 or fewer carbon atoms have different effects from those with 12–16 carbons is yet to be determined. The current dietary recommendation as defined in the NCEP Therapeutic Lifestyle Change diet to prevent and treat cardiovascular disease is to limit intakes of saturated fat to less than 7% of total energy. The major contributors of saturated fatty acids were discussed previously.

Compared to saturated fatty acids, unsaturated fatty acids, both monounsaturated and polyunsaturated fatty acids, lower both LDL and HDL cholesterol levels. The absolute magnitude of the change is greater for LDL cholesterol than HDL cholesterol. Most data suggest that monounsaturated fatty acids have a slightly smaller effect than polyunsaturated fatty acids in lowering both LDL and HDL cholesterol levels so that the total cholesterol:HDL cholesterol ratio is similar for both categories of fat.

Quantitatively, the major n-3 polyunsaturated fatty acid in the diet is alpha-linolenic acid (18:3n-3). Major dietary sources include soybean and canola oils (**Figure** 8). Two other n-3 polyunsaturated fatty acids, eicosapentaenoic acid (EPA, 20:5n-3) and docosahexaenoic acid (DHA, 22:6n-3), referred to as very long-chain n-3 fatty acids, are found predominantly in fish, specifically dark flesh fish such as salmon, tuna, and swordfish. Dietary intakes of very long-chain n-3 fatty acids are associated with decreased risk of heart disease and stroke. Intervention studies have substantiated these findings.

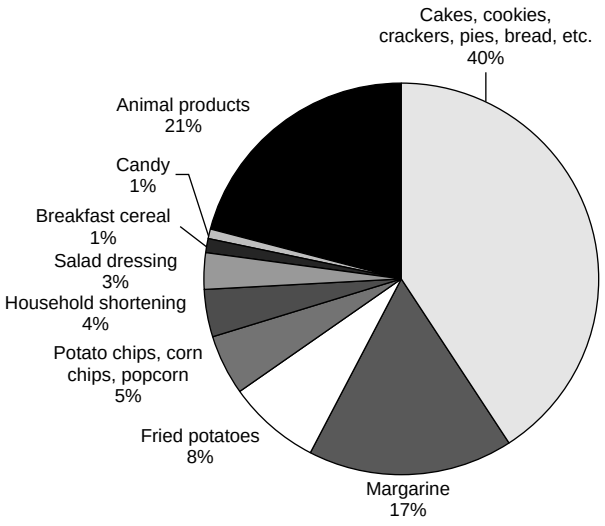


Figure 9 Major food sources of *trans* fat for US adults. (Data from www.cfsan.fda.gov/~dms/qatrans2.)

The beneficial effects of EPA and DHA are attributed to decreased ventricular fibrillation resulting in decreased sudden death, and also decreased triglyceride levels, platelet aggregation, and blood pressure. Evidence suggests that very long-chain n-3 fatty acids may decrease atherosclerotic plaque progression.

Dietary *trans* fatty acids occur naturally in meat and dairy products as a result of anaerobic bacterial fermentation in ruminant animals. *Trans* fatty acids are also introduced into the diet as a result of the consumption of hydrogenated vegetable or fish oils. Hydrogenation results in a number of changes in the fatty acyl chain: conversion of *cis* to *trans* double bonds; saturation of double bonds; and migration of double bonds along the acyl chain, resulting in multiple positional isomers. Oils are primarily hydrogenated to increase viscosity (change a liquid oil into a semiliquid or solid) and extend shelf life (decrease susceptibility to oxidation). Major contributors of *trans* fatty acids are commercially baked products (40%), animal products (21%), margarine (17%), and fried potatoes and chips (13%) (**Figure 9**). In intervention and observational studies, *trans* fatty acid intake has been associated with less favorable total cholesterol:HDL cholesterol ratios and increased risk of cardiovascular disease, respectively.

Composition of Dietary Fats

Types of fat relatively high in saturated fatty acids include butterfat (62%), beef tallow (50%), tropic oils (coconut, 87%; palm kernel, 81%; and palm oil, 49%), and lard (39%) (**Figure 8**). The content of cholesterol in these fats is 33, 14, 0, and 12 mg/tablespoon, respectively. Types of fat relatively high in monounsaturated fatty acids include canola oil (56%), olive oil (73%), and peanut oil (46%). Types of fat relatively high in polyunsaturated fatty acids include soybean oil (51%), corn oil (58%), safflower oil (74%), and sunflower oil (66%). None of the vegetable oils high in monounsaturated or polyunsaturated fatty acids contain cholesterol. The fatty acid profile of diets varies widely among individuals and depends on such factors as availability, cultural and religious dietary patterns, price, and personal preferences.

Summary

Dietary fats and oils have both positive and negative attributes with respect to health outcomes. This makes determining optimal dietary recommendations difficult. Fats and oils are primarily composed of triacylglycerol. The fatty acid composition of the triacylglycerol dictates the physical properties of the fat. During fatty acid biosynthesis, humans are unable to insert a double bond above the ninth carbon of the acyl chain. For this reason, linoleic acid and fatty acids derived from linoleic acid are essential and hence must be consumed preformed. Animal fats are the major contributors of saturated fatty acids to the diet. Vegetable oils such as canola and olive, and also animal fats, are the major contributors of monounsaturated fatty acids to the diet. Vegetable oils such as safflower, sunflower, and corn oils are the major contributors of polyunsaturated fatty acids to the diet. Dietary patterns high in saturated fatty acids have been associated with increased risk of developing cardiovascular disease, whereas dietary patterns high in unsaturated fatty acids (monounsaturated and polyunsaturated) have been associated with decreased risk. Two other subtypes of fatty acids derived from the diet potentially impact on human health: very long-chain n-3 fatty acids and *trans* fatty acids. Very long-chain n-3 fatty acids are derived primarily from fish, and intakes have been associated with decreased risk of developing cardiovascular disease. The majority of *trans* fatty acids are derived from foods made with hydrogenated fat and, to a lesser extent, from animal fats. Higher amounts in the diet have been associated with elevated LDL cholesterol levels and higher total cholesterol to HDL cholesterol ratios. Dietary fatty acid intakes are determined by the sum of individual food choices. Current general dietary recommendations from a number of sources suggest that one should consume a diet high in fruits and vegetables, whole grain products, low-fat and nonfat diary products, legumes, and lean meats. Such a dietary pattern, while accommodating

personal preferences, is consistent with a diet low in saturated and *trans* fatty acids and is predicted to reduce the risk of developing cardiovascular and other chronic diseases.

See also: **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Lipids**: Chemistry and Classification; Composition and Role of Phospholipids. **Lipoproteins**.

Further Reading

Clarke R, Frost C, Collins R *et al.* (1997) Dietary lipids and blood cholesterol: Quantitative meta-analysis of metabolic ward studies. *British Medical Journal* **314**: 112–117.

Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (2001) Executive summary of the third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *Journal of the American Medical Association* **285**: 2486–2497.

Hooper L, Summerbell CD, Higgins JP *et al.* (2001) Dietary fat intake and prevention of cardiovascular disease: Systematic review. *British Medical Journal* **322**: 757–763.

Krauss RM, Eckel RH, Howard B *et al.* (2000) AHA Dietary Guidelines: Revision 2000: A statement for healthcare professionals from the Nutrition Committee of the American Heart Association. *Circulation* **102**: 2284–2299.

Law M (2000) Plant sterol and stanol margarines and health. *British Medical Journal* **320**: 861–864.

Law MR, Wald NJ, and Thompson SG (1994) By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *British Medical Journal* **308**: 367–372.

Lichtenstein AH (2003) Dietary fat and cardiovascular disease risk: Quantity or quality? *Journal of Women's Health* **12**: 109–114.

Sacks FM and Katan M (2002) Randomized clinical trials on the effects of dietary fat and carbohydrate on plasma lipoproteins and cardiovascular disease. *American Journal of Medicine* **113**: 13S–24S.

Schaefer EJ (2002) Lipoproteins, nutrition, and heart disease. *American Journal of Clinical Nutrition* **75**: 191–212.

Tang JL, Armitage JM, Lancaster T *et al.* (1998) Systematic review of dietary intervention trials to lower blood total cholesterol in free-living subjects. *British Medical Journal* **316**: 1213–1220.

Watts GF and Burke V (1996) Lipid-lowering trials in the primary and secondary prevention of coronary heart disease: New evidence, implications and outstanding issues. *Current Opinion in Lipidology* **7**: 341–355.

Willett WC (1998) Is dietary fat a major determinant of body fat? (Comment; erratum appears in Am J Clin Nutr 1999 Aug;70(2):304). *American Journal of Clinical Nutrition* **67**: 556S–562S.

Yao M and Roberts SB (2001) Dietary energy density and weight regulation. *Nutrition Reviews* **59**: 247–258.

FATTY ACIDS

Contents

Metabolism

P A Watkins, Kennedy Krieger Institute and Johns Hopkins University School of Medicine, Baltimore, MD, USA

Introduction

Fatty acids and glucose are the primary metabolic fuels used by higher organisms, including man. As such, fatty acids occupy a central position in human nutrition. Fat, carbohydrate, and protein comprise the macronutrients. When nutritionists speak of fat, they are referring mainly to triacylglycerol (triglyceride), which consists of three fatty-acid molecules covalently linked to a backbone of glycerol. Several properties of fatty acids and triacylglycerol make them highly suited to the storage and provision of energy. When a gram of fatty acid is burned as fuel, about 9 kcal of energy is recovered – more than twice that yielded when a gram of

carbohydrate or protein is utilized. Unlike carbohydrates, fat can be stored in an anhydrous compact state, allowing the organism to amass large quantities of fuel reserves in times of plenty. This property can have unfortunate consequences in prosperous societies, as evidenced by the increasing incidence of obesity. Fatty acids are also fundamental building blocks for the synthesis of most biologically important lipids, including phospholipids, sphingolipids, and cholesterol esters. They are the precursors of bioactive molecules such as prostaglandins and other eicosanoids. In addition, fatty acids and their coenzyme A derivatives have many metabolic regulatory roles.

Fatty-Acid Nomenclature Conventions

In this article, fatty acids will be identified by their chain length, the number of double bonds present, and the position of the first double bond from the methyl end of the molecule. Thus 14:0 denotes a saturated fatty acid with 14 carbon atoms, 16:1*n*-9 denotes a monounsaturated fatty acid with 16 carbon atoms in which one double bond occurs nine carbon atoms from the methyl end, and 20:4*n*-6 denotes a polyunsaturated fatty acid with 20 carbon atoms in which the first of four double bonds is found six carbon atoms from the methyl end. Unless otherwise noted, all double bonds are in the *cis* configuration and double bonds in polyunsaturated fatty acids are separated by a single methylene ($-CH_2-$) group. The carboxyl carbon atom of any fatty acid is carbon-1. The adjacent carbon atom is referred to as either carbon-2 or the α-carbon; the next is carbon-3 or the β-carbon, and so on. Some examples are shown in **Figure 1**.

Physical Properties of Fatty Acids

Fatty acids are aliphatic organic acids with the fundamental structure $CH_3(CH_2)_nCOOH$, where n can range from zero to more than 26. Thus, fatty acids range from the shortest, acetic acid (2:0), to the very long-chain fatty acids containing 26 or more carbon atoms (e.g., 26:0). Although fatty acids with an odd number of carbon atoms exist in nature, most common fatty acids have an even number. The most abundant fatty acids in human lipids and in dietary lipids are the long-chain fatty acids 16:0 (palmitic acid) and 18:1*n*-9 (oleic acid) (**Figure 1**). The hydrophobic nature of the hydrocarbon chain of fatty acids containing more than eight carbon atoms renders them quite insoluble in aqueous media. It has been estimated that for every two additional carbon atoms in the fatty-acid chain its solubility decreases 10-fold.

Owing to the poor solubility of the most abundant fatty acids, free (non-esterified) fatty acids are often found associated with binding and/or transport proteins. Serum albumin has at least six binding sites for fatty acids and is the primary transporter of these molecules through the bloodstream. Several low-molecular-weight fatty-acid binding proteins have been identified and implicated in the intracellular transport of free fatty acids. While free fatty acids can associate with lipophilic cellular and organellar membranes, concentrations of these non-esterified compounds in membranes are typically very low.

Fatty-Acid Activation

Biochemically, fatty acids are rather nonreactive molecules unless they are first activated by thioesterification to coenzyme A (CoA). This reaction is catalyzed by acyl-CoA synthetases (also known as acid : CoA ligases, E.C. 6.2.1.x). The overall acyl-CoA synthetase reaction is

$$\text{RCOOH} + \text{ATP} + \text{CoA-SH} \rightarrow \text{RCO-S-CoA} + \text{AMP} + \text{PPi}$$

where PPi is inorganic pyrophosphate, ATP is adenosine triphosphate, and AMP is adenosine monophosphate. Owing to the wide diversity of fatty-acid chain lengths, many enzymes with varied substrate specificities have been identified.

It is estimated that humans have more than 25 enzymes capable of activating fatty acids and/or fatty-acid-like compounds. Acyl-CoA synthetases that activate fatty acids of similar chain lengths often have different tissue-expression patterns and/or different subcellular locations. Thus, each enzyme may direct its fatty-acid substrates into a particular metabolic pathway.

Mitochondrial Fatty-Acid β-Oxidation

To recover their stored energy, fatty acids must be oxidized. Quantitatively, the most important energy-yielding degradation pathway is mitochondrial β-oxidation (**Figure 2**). Fatty acids must first enter cells or tissues. Serum triacylglycerol, usually associated with lipoproteins, is hydrolyzed by lipoprotein lipase located on the capillary endothelium, releasing fatty acids for cellular uptake. In addition, albumin-bound circulating free fatty acids (e.g., produced by the mobilization of adipocyte fat stores) reach the cell surface. Although

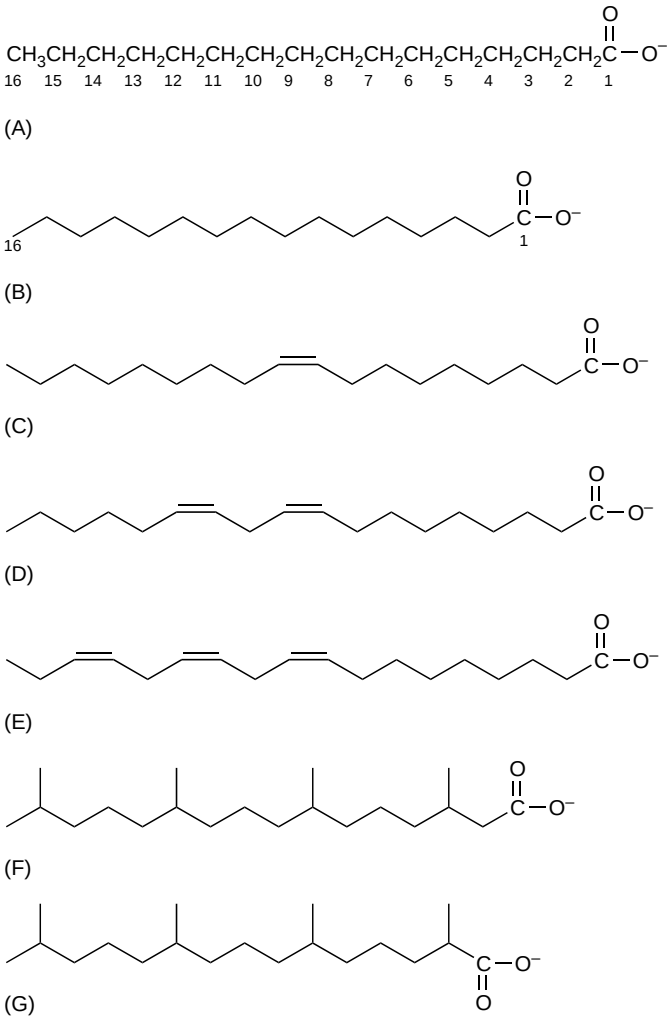


Figure 1 Fatty-acid structure and nomenclature. (A) Chemical formula and carbon atom numbering system for a 16-carbon saturated fatty acid (16:0). (B) Schematic representation of 16:0. (C) A monounsaturated fatty acid, 18:1*n*-9, showing the double bond nine carbon atoms from the methyl end (carbon 18). (D) The essential *n*-6 fatty acid 18:2*n*-6, where the first double bond is found six carbon atoms from the methyl end. The two double bonds are separated by a methylene (–CH_2–) group. (E) The essential *n*-3 fatty acid 18:3*n*-3, where the first double bond is found three carbon atoms from the methyl end. (F) Phytanic acid, a dietary β-methyl-branched-chain fatty acid (3,7,11,15-tetramethyl 16:0). The methyl group on carbon 3 prevents this fatty acid from degradation by β-oxidation. (G) Pristanic acid (2,6,10,14-tetramethyl 15:0) is the product of phytanic acid α-oxidation, in which a single carbon (carbon 1) is lost. The methyl group on carbon 2 does not preclude subsequent degradation by β-oxidation.

hydrophobic fatty acids can traverse the plasma membrane by simple diffusion, a role for membrane transport proteins in this process remains controversial. Once inside the cell, fatty acids are thought to be moved to the mitochondria (or other intracellular sites) by intracellular fatty-acid binding proteins.

Acyl-CoA synthetase activity towards long-chain fatty-acid substrates is present in the outer mitochondrial membrane. However, fatty acyl-CoAs do not readily traverse biological membranes such as the inner mitochondrial membrane. A highly sophisticated transport system has evolved to allow tight regulation of fatty-acid entry into the mitochondrion (**Figure 2**). Carnitine palmitoyl transferase 1 (CPT1), located on the inner aspect of the outer mitochondrial membrane, catalyzes a transesterification reaction:

$$\text{fatty acyl-CoA} + \text{carnitine} \rightarrow \text{fatty acyl-carnitine} + \text{CoA-SH}$$

Carnitine–acylcarnitine translocase (CACT), located in the inner mitochondrial membrane, carries the fatty acyl-carnitine inside the mitochondrion in exchange for a free carnitine molecule. CPT2, located inside the mitochondrion, then catalyzes the reversal of the CPT1 reaction. Thus, the concerted actions of CPT1, CACT, and CPT2

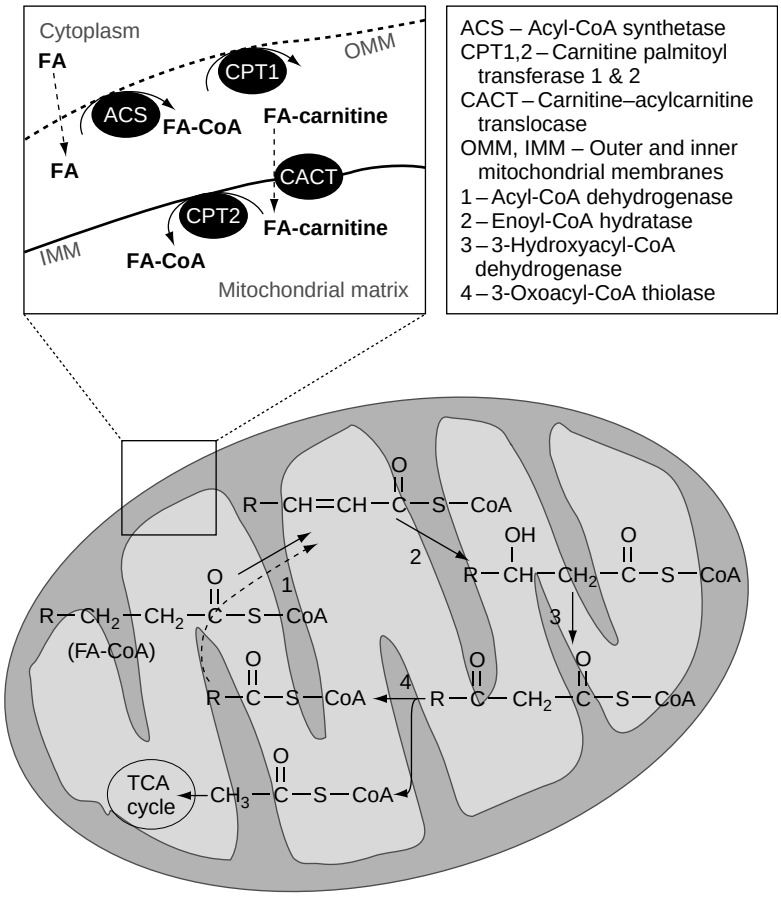


Figure 2 Mitochondrial fatty-acid (FA) β-oxidation pathway. Long-chain fatty acids are activated, converted to carnitine esters, transported across the inner mitochondrial membrane, and re-converted to their CoA thioester once in the mitochondrial matrix. Four sequential mitochondrial enzyme reactions shorten the fatty acyl-CoA (FA-CoA) by two carbon atoms, which are released as acetyl-CoA. The shortened fatty acyl-CoA can undergo additional cycles of degradation until the entire carbon chain has been converted to acetyl-CoA units. $FADH_2$ and NADH, produced in reactions 1 and 3, respectively, can enter the electron transport chain for ATP production. Acetyl-CoA enters the tricarboxylic acid (TCA) cycle, yielding additional NADH and $FADH_2$ for ATP production. Mitochondrial β-oxidation is the primary pathway for recovering the energy stored as triacylglycerol or 'fat'.

effectively translocate fatty acyl-CoA across the inner mitochondrial membrane.

Entry of fatty acids into the mitochondrion is regulated by several mechanisms. Although long-chain fatty acids can readily diffuse across the lipophilic inner mitochondrial membrane, the mitochondrial matrix lacks long-chain acyl-CoA synthetase activity. Thus, long-chain fatty acids cannot be activated intramitochondrially to enter the β-oxidation pathway. Control is also exerted extramitochondrially via malonyl-CoA, a cytoplasmic intermediate in fatty-acid biosynthesis and an indicator of high cellular energy status. Malonyl-CoA is a potent inhibitor of CPT1, prohibiting fatty acids from entering the mitochondria to be degraded.

As depicted in **Figure 2**, the four primary enzymes of mitochondrial β-oxidation act on intramitochondrial fatty acyl-CoA by sequential dehydrogenation, hydration, dehydrogenation, and thiolytic cleavage reactions. The products are (1) fatty acyl-CoA that has been shortened by two carbon atoms, (2) acetyl-CoA, (3) reduced flavin adenine dinucleotide ($FADH_2$), and (4) reduced nicotinamide adenine dinucleotide (NADH). $FADH_2$ and NADH can directly enter the electron transport chain at complex 2 and complex 1, respectively, yielding about five ATP molecules. Acetyl-CoA can be further degraded to carbon dioxide and water by the tricarboxylic acid cycle, yielding additional reducing equivalents that can enter the electron transport chain and produce ATP. Importantly, the entire β-oxidation process can be repeated using the shortened fatty acyl-CoA as a substrate. This process can be repeated until the entire carbon skeleton of the fatty acid has

been degraded to two-carbon acetyl-CoA units. Theoretically, complete oxidation of one molecule of 16:0 (β-oxidation and tricarboxylic acid cycle) will yield more than 160 ATP molecules.

Essentially all cells and tissues can use carbohydrate (glucose) for fuel, and a few (e.g., nerves and erythrocytes) are dependent on this fuel source. An important nutritional consideration is that carbon derived from fatty acids via β-oxidation cannot be converted to glucose in net quantities. In the postprandial state, however, most cell types other than nerves and erythrocytes derive the majority of their energy from fatty-acid oxidation under normal physiologic conditions. Some tissues, e.g., skeletal muscle, completely oxidize fatty acids to carbon dioxide and water. Others, e.g., liver, only partially oxidize fatty acids, using the acetyl-CoA product for biosynthethic needs. In particular, liver uses intramitochondrial acetyl-CoA for the synthesis of ketone bodies, acetoacetate and β-hydroxybutyrate (**Figure 3**). Ketone bodies can be oxidized by all tissues except the liver and provide an alternative fuel source during starvation. In particular, nervous tissue can oxidize ketone bodies. During prolonged starvation, increased ketone-body use spares the brain's requirement for glucose.

Peroxisomal Fatty-Acid β-Oxidation

Like mitochondria, peroxisomes contain pathways for the β-oxidation of fatty acids. The mechanism by which fatty acids enter peroxisomes is unclear but does not appear to involve the CPT1–CACT–CPT2 pathway. Long-chain and very-long-chain acyl-CoA synthetase activities are associated with peroxisomes, but it has not been established whether fatty acids or fatty acyl-CoAs traverse the peroxisomal membrane. The basic reactions of peroxisomal β-oxidation resemble those found in mitochondria, but the peroxisomal and mitochondrial enzymes are distinct proteins (**Figure 4**). In fact, peroxisomes contain two sets of β-oxidation enzymes, which appear to function with distinct substrates.

Unlike mitochondria, peroxisomes do not contain an electron transport chain or tricarboxylic acid cycle, and, thus, peroxisomal fatty-acid degradation is not directly coupled to energy production. Rather, peroxisomes have a more specialized fatty-acid oxidation role, degrading fatty-acid substrates that cannot be catabolized in mitochondria. Peroxisomes are indispensable for the degradation of very-long-chain fatty acids (containing more than 22 carbon atoms),

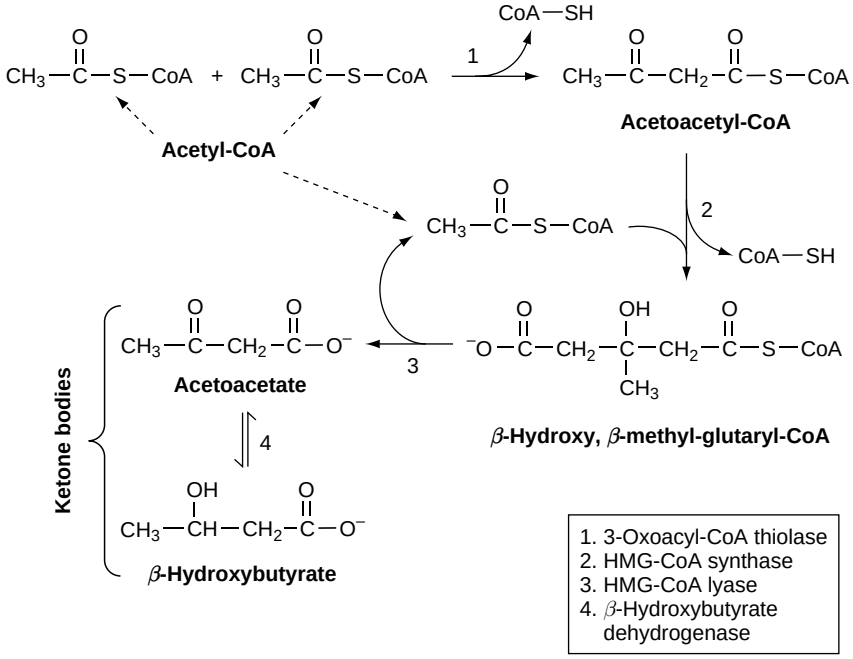


Figure 3 Synthesis of ketone bodies. In the mitochondria of hepatocytes, acetyl-CoA derived from β-oxidation is converted to ketone bodies, primarily acetoacetate and β-hydroxybutyrate, rather than entering the tricarboxylic acid cycle. Two molecules of acetyl-CoA condense in a reversal of the last β-oxidation reaction (3-oxoacyl-CoA thiolase). The product, acetoacetyl-CoA, condenses with another molecule of acetyl-CoA, yielding β-hydroxy, β-methyl-glutaryl-CoA (HMG-CoA), a reaction catalysed by HMG-CoA synthase. Cleavage of HMG-CoA by HMG-CoA lyase yields acetoacetate, regenerating one molecule of acetyl-CoA. Acetoacetate is reversibly reduced to β-hydroxybutyrate via the NAD-dependent enzyme β-hydroxybutyrate dehydrogenase. These ketone bodies can traverse the inner mitochondrial membrane, eventually reaching the bloodstream for ultimate use by the brain and other tissues.

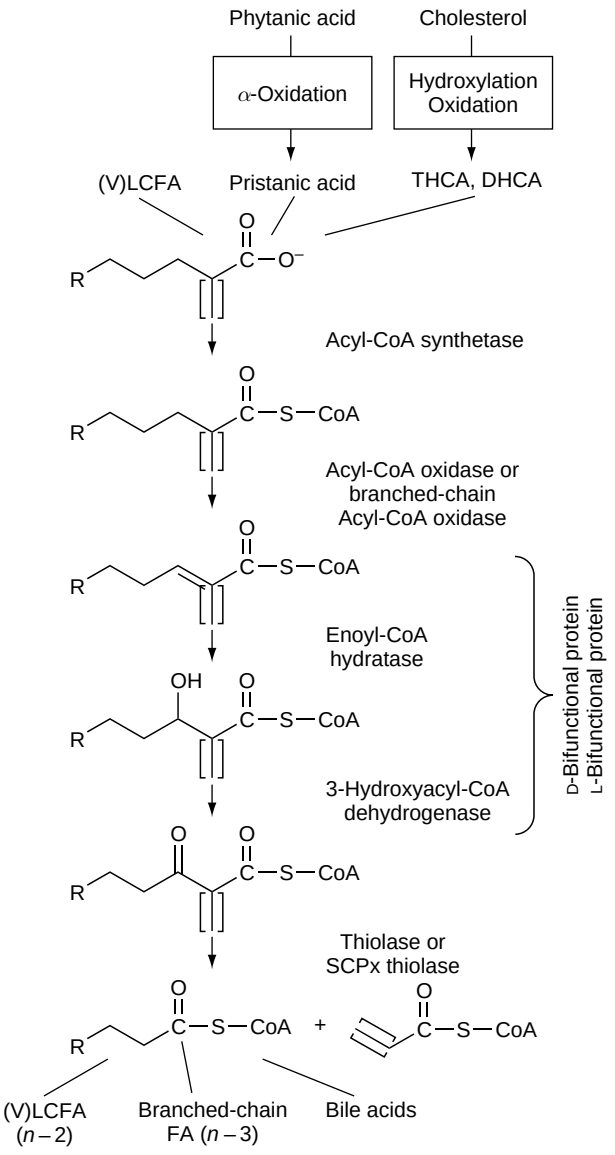


Figure 4 Peroxisomal fatty-acid (FA) β-oxidation pathways. While saturated long-chain fatty acids (LCFA) are preferentially degrade in mitochondria, saturated very-long-chain fatty acids (VLCFA) and some LCFA are shortened by peroxisomal β-oxidation. Degradation of pristanic acid, the product of phytanic acid α-oxidation, and the conversion of the cholesterol-derived 27-carbon bile-acid precursors dihydroxycholestanoic acid (DHCA) and trihydroxycholestanoic acid (THCA) to 24-carbon bile acids also require this pathway. The mechanism by which these substrates enter peroxisomes is unknown. Four enzymatic reactions serve to shorten the substrates by either two (LCFA, VLCFA) or three (pristanic acid, DHCA, THCA) carbon atoms. The 2-methyl group of the latter substrates is shown in brackets. SCPx thiolase refers to the thiolase activity of sterol carrier protein x.

which are neurotoxic if allowed to accumulate. These fatty acids undergo several cycles of peroxisomal β-oxidation until they are between eight and 10 carbon atoms long, after which they go to the mitochondria for further catabolism. Degradation of xenobiotic fatty acyl-like compounds (e.g., sulphur-substituted fatty acids and many nonsteroidal anti-inflammatory drugs) takes place in peroxisomes. Oxidation of dicarboxylic acids (from the diet or from ω-oxidation) and 2-methyl-branched-chain fatty acids (from the diet or from α-oxidation) also occurs in peroxisomes.

The peroxisomal β-oxidation pathway also fulfils an important biosynthetic role. In the hepatic synthesis of bile acids from cholesterol, the aliphatic side chain, which resembles an α-methyl-branched-chain fatty acid, must be shortened. A single cycle of peroxisomal β-oxidation will remove a three-carbon portion of the side chain, converting the 27-carbon bile acid precursors dihydroxycholestanoic and trihydroxycholestanoic acids into the 24-carbon primary bile acids chenodeoxycholate and cholate, respectively.

Fatty-Acid α-Oxidation and ω-Oxidation

Other important fatty-acid catabolic pathways include α-oxidation and ω-oxidation. α-Oxidation is required for degradation of the dietary fatty acid phytanic acid (3,7,11,15-tetramethyl-16:0). This fatty acid cannot be degraded by β-oxidation owing to the methyl group on carbon-3. In the human diet, phytanic acid is obtained from the consumption of ruminant meats, fats, and dairy products. Rumen bacteria hydrolyze chlorophyll, releasing the phytol side chain; phytol is oxidized to phytanic acid and incorporated into triacylglycerol and phospholipids by the animal. Humans typically ingest 50–100 mg of phytanic acid per day. The current view of the α-oxidation pathway, which is found in peroxisomes, is shown in **Figure 5**. After activation to its CoA derivative, phytanoyl-CoA is hydroxylated on the 2-carbon. The next reaction catalyzes the removal of a one-carbon CoA derivative as formyl-CoA. The other product of this reaction is an aldehyde, pristanal, that can be oxidized to form pristanic acid (2,6,10,14-tetramethyl-15:0). This chain-shortening reaction effectively shifts the position of the first methyl group from carbon-3 (in phytanic acid) to carbon-2 (in pristanic acid). The 2-methyl-branched chain fatty acids can then be degraded further via peroxisomal β-oxidation.

Another mechanism for degradation of fatty acids that cannot undergo β-oxidation is known as ω-oxidation. In this process, the terminal methyl group (referred to as the ω-end) of a fatty-acid chain is oxidized to a carboxylic acid via cytochrome P450 isozymes, particularly the CYP52A family, in the endoplasmic reticulum. The resulting

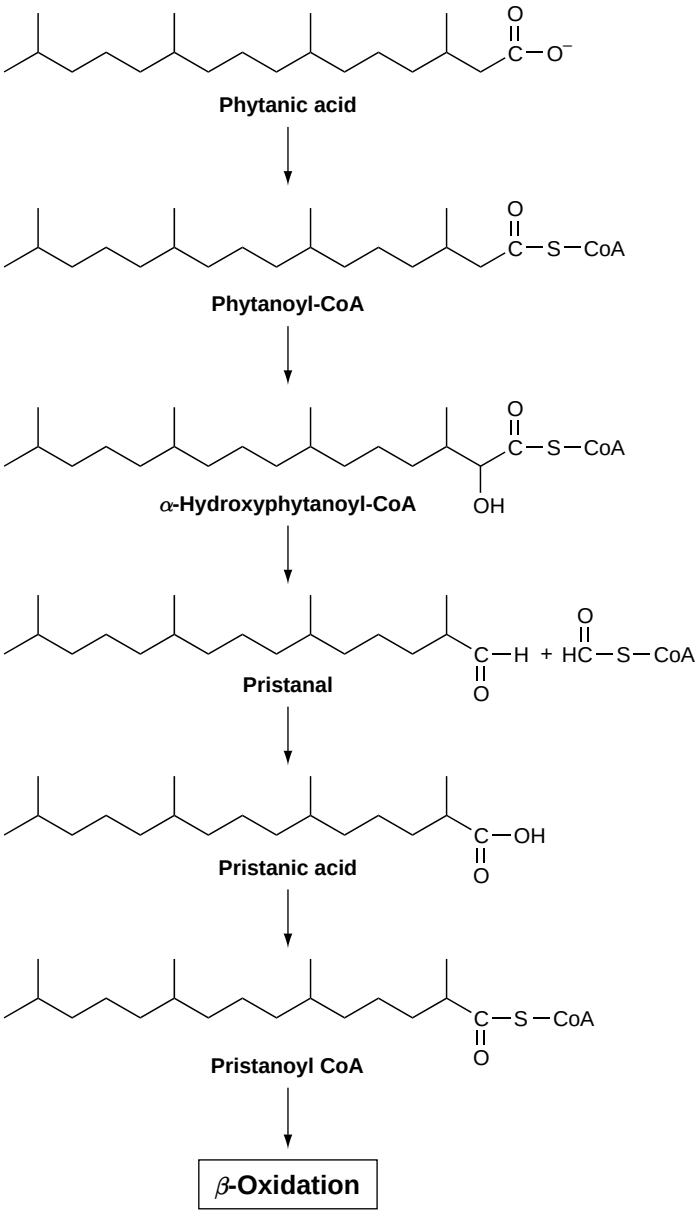


Figure 5 Peroxisomal phytanic acid α-oxidation pathway. The dietary 3-methyl-branched fatty acid phytanic acid is toxic if allowed to accumulate in the tissues. Its 3-methyl group prevents degradation by β-oxidation; therefore, this fatty acid is first shortened by one carbon atom. Like the substrates for peroxisomal β-oxidation, phytanic acid enters peroxisomes by an unknown mechanism. Activated phytanic acid is hydroxylated on carbon 2. Cleavage between carbons 1 and 2 yields a one-carbon CoA compound, formyl-CoA, and an aldehyde, pristanal. After oxidation and reactivation to the CoA derivative, pristanoyl-CoA can be degraded by β-oxidation.

dicarboxylic acids can then be at least partially degraded by β-oxidation from the ω-end, primarily in peroxisomes.

Fatty-Acid *de novo* Synthesis

Much of our need for fatty acids as constituents of phospholipids and other complex lipids is met by the diet. In addition, certain lipogenic tissues are capable of the *de novo* synthesis of fatty acids (**Figure 6**). These tissues include liver (hepatocytes), adipose tissue, and lactating mammary gland. Much of the fatty acids synthesized by all three tissues is incorporated into triacylglycerol. Hepatic synthesis is primarily for export to other tissues (in very low-density lipoproteins), while synthesis in adipocytes and mammary gland is for local storage.

The carbon used for fatty-acid synthesis typically derives from the products of glycolysis. The end product of glycolysis, pyruvate, enters the mitochondria and becomes the substrate for two separate

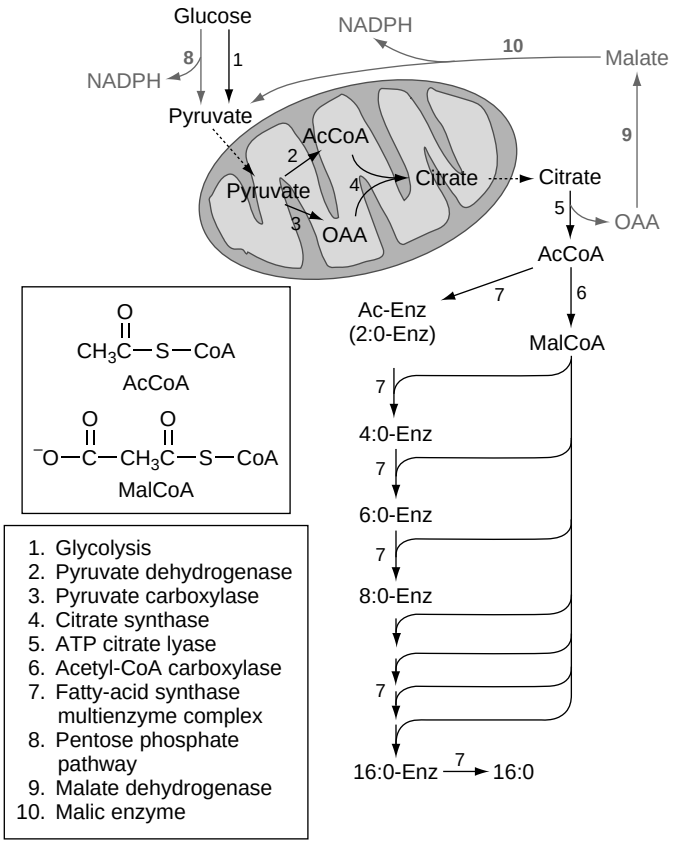


Figure 6 Fatty-acid biosynthesis. Cytoplasmic acetyl-CoA (AcCoA) is the primary substrate for *de novo* fatty-acid synthesis. This two-carbon compound most commonly derives from the glycolytic degradation of glucose, and its formation is dependent upon several reactions in the mitochondria. The mitochondrial enzyme pyruvate carboxylase is found primarily in tissues that can synthesize fatty acids. AcCoA is converted to malonyl-CoA (MalCoA) by acetyl-CoA carboxylase. Using AcCoA as a primer, the fatty-acid synthase multienzyme complex carries out a series of reactions that elongate the growing fatty acid by two carbon atoms. In this process MalCoA condenses with AcCoA, yielding an enzyme-bound four-carbon β-ketoacid that is reduced, dehydrated, and reduced again. The product is enzyme-bound 4:0. This process is repeated six more times, after which 16:0 is released from the complex. The reductive steps require NADPH, which is derived from enzyme reactions and pathways shown in grey. Enz refers to the fatty acid synthase multienzyme complex.

reactions. In one, pyruvate is decarboxylated via the pyruvate dehydrogenase complex, yielding acetyl-CoA. Lipogenic tissues also contain another mitochondrial enzyme, pyruvate carboxylase, which converts pyruvate to the four-carbon acid oxaloacetate (OAA). Acetyl-CoA and oxaloacetate condense to form the six-carbon acid citrate. As citrate accumulates within the mitochondrion, it is exported to the cytoplasm, where it is converted back to oxaloacetate and acetyl-CoA. Cytoplasmic acetyl-CoA is the fundamental building block for *de novo* synthesis of fatty acids.

The first enzyme unique to fatty-acid synthesis is acetyl-CoA carboxylase, which converts the two-carbon substrate acetyl-CoA into the three-carbon product malonyl-CoA. Citrate, in addition to being the precursor of cytoplasmic acetyl-CoA, has a regulatory role. Citrate is an allosteric activator of acetyl-CoA carboxylase and serves as a signal that there is an ample carbon supply for fatty-acid synthesis. As noted above, malonyl-CoA is a potent inhibitor of CPT1. Cytoplasmic malonyl-CoA levels will be high only when there is significant flux through glycolysis, indicative of a high cellular energy state. Under these conditions, entry of fatty acids into the mitochondria (and subsequent β-oxidation) is prevented. Interestingly, there are two isoforms of acetyl-CoA carboxylase. One is found in the above-named lipogenic tissues. The other is found in many tissues that are not capable of synthesizing fatty acids, e.g., the heart. It is thought that the primary role of the second isozyme is to regulate mitochondrial fatty-acid β-oxidation by synthesizing malonyl-CoA when cellular energy needs are being met by carbohydrate metabolism.

The subsequent reactions of fatty-acid synthesis in humans are catalyzed by a multienzyme

Table 1 Distinctions between fatty-acid β-oxidation and fatty-acid synthesis

	Fatty-acid β-oxidation	*Fatty-acid synthesis*
Tissues with active pathway	Nearly all tissues except brain, nerve, and erythrocytes	Liver, adipose, and lactating mammary gland
Subcellular location	Mitochondria	Cytoplasm
Redox cofactors	NAD, FAD	NADPH
Acyl-group carrier	CoA	Enzyme-bound acyl carrier protein
Stereochemistry of 3-hydroxy intermediate	L-	D-

complex, fatty-acid synthase. After binding of one molecule each of acetyl-CoA and malonyl-CoA to unique binding sites within the complex, a condensation reaction occurs in which carbon dioxide is released and an enzyme-bound 4-carbon 3-ketoacid is formed. Subsequent reactions include a reduction step, a dehydration step, and a second reduction step. The intermediates produced in these reactions are similar to those seen in β-oxidation (**Figure 2**), in reverse order. The product (enzyme bound) is the saturated fatty acid 4:0, which can then condense with another molecule of malonyl-CoA to start the process anew. After seven such cycles, the ultimate product is 16:0, which is released from the complex.

The reductive steps in fatty-acid synthesis require reduced nicotinamide adenine dinucleotide phosphate (NADPH). Some NADPH is produced during recycling of the oxaloacetate formed during the cytoplasmic hydrolysis of citrate, described above. Oxaloacetate is first converted to malate (via cytoplasmic malate dehydrogenase). Malate is then decarboxylated to pyruvate in an $NADP^+$-dependent reaction catalyzed by malic enzyme; NADPH is produced in this reaction. NADPH for fatty-acid biosynthesis also comes from reactions in the pentose phosphate pathway (hexose monophosphate shunt).

In several respects, the enzymatic reactions of fatty-acid synthesis are the converse of those in fatty-acid oxidation. However, there are key differences, which are summarized in **Table 1**.

Fatty-Acid Elongation

The primary product synthesized by the *de novo* pathway is 16:0. While 16:0 is an important fatty acid, there is a need to synthesize longer-chain acids. Enzymes for elongation of fatty-acids have been found in membranes of the endoplasmic reticulum and mitochondria. However, these pathways are less well-characterized than that of fatty-acid synthesis. In the endoplasmic reticulum, the reactions involved in fatty-acid elongation are very similar to those of cytoplasmic fatty-acid synthesis. The donor of the added carbon atoms is also malonyl-CoA, indicating that an active acetyl-CoA carboxylase is required for elongation. Whereas the primary reactions of fatty-acid synthesis are found within the fatty-acid synthase multienzyme complex, individual proteins catalyze the four elongation reactions (condensation, reduction, dehydration, and reduction). Like synthesis, elongation in the endoplasmic reticulum requires reducing equivalents in the form of NADPH.

Fatty-acid elongation in mitochondrial membranes is thought to be slightly different from the process in the endoplasmic reticulum. The primary difference is that the donor of elongation units is thought to be acetyl-CoA, not malonyl-CoA. The four elongation reactions are similar, but may require NADH rather than NADPH as source of reducing equivalents. Little is known about how fatty-acid elongation in either the mitochondria or the endoplasmic reticulum is regulated.

Fatty-Acid Unsaturation and the Essential Fatty Acids

Monounsaturated and polyunsaturated fatty acids are extraordinarily important in human health and nutrition. Thus, the insertion of double bonds into the carbon skeleton of a fatty acid is a vital metabolic function. However, humans are in general not capable of inserting double bonds closer than nine carbon atoms from the methyl end of a fatty acid. Thus, we are incapable of the *de novo* synthesis of two important classes of fatty acids, the *n*-3 fatty acids such as docosahexaenoic acid (22:6*n*-3) and the *n*-6 fatty acids such as arachidonic acid (20:4*n*-6). The *n*-3 fatty acids have proven to be beneficial in the prevention of coronary artery disease. The fatty acid 22:6*n*-3 has been shown to be important for the normal development of the brain and retina, leading some manufacturers to include this fatty acid in their infant formula preparations. The *n*-6 fatty acids are important constituents of membrane lipids. The fatty acid 20:4 is also the well-known precursor of prostaglandins and other bioactive eicosanoids. Since we cannot synthesize these fatty acids *de novo*, we are dependent on the presence of

at least some *n*-3 and some *n*-6 fatty acids in the diet. Linoleic acid (18:2*n*-6) and α-linolenic acid (18:3*n*-3) are the precursors of most biologically important *n*-3 and *n*-6 fatty acids; thus, they are referred to as essential fatty acids.

As noted earlier, the most abundant fatty acids in humans include a saturated fatty acid (16:0) and a monounsaturated fatty acid (18:1*n*-9). Humans can readily insert a *cis*-double bond nine carbons from the carboxyl carbon atom of a fatty acid (Δ9) in a reaction catalyzed by stearoyl-CoA desaturase (SCD1; so-named because the preferred substrate is the CoA derivative of 18:0, stearic acid). Because SCD1 is involved in the synthesis of such an abundant fatty acid, 18:1, the importance of this enzyme in metabolism was initially overlooked. However, 18:1 produced by SCD1 appears to be directed specifically towards triacylglycerol synthesis. Mice in which the SCD1 gene is disrupted have decreased adiposity. Furthermore, genetically obese leptin-deficient (ob–/ob–) mice in which the SCD1 gene is also disrupted have significantly reduced body weight compared with ob–/ob– mice, leading to the hypothesis that leptin regulates the synthesis of SCD1. Interestingly, dietary 18:1 seems to be more readily incorporated into lipids other than triacylglycerols, implying that the dietary and the SCD1-produced pools of this fatty acid are metabolically distinct. As with the *n*-3 fatty acids, dietary ingestion of monounsaturated fatty acids such as 18:1 has been associated with benefits to cardiovascular health.

Humans are also capable of inserting *cis*-double bonds either five or six carbon atoms from the carboxyl carbon atom of a fatty acid (Δ5 desaturase and Δ6 desaturase activity, respectively). These activities, when combined with the elongation pathways described above, form a powerful mechanism for synthesis of highly polyunsaturated fatty acids such as 20:4*n*-6 and 22:6*n*-3 from the dietary essential fatty acids. Previously, it was thought that humans also had the ability to insert a double bond four carbon atoms from the carboxyl carbon (Δ4 desaturase activity), as this activity was thought to be necessary for the conversion of 18:3*n*-3 to 22:6*n*-3. However, attempts to measure Δ4 desaturase activity experimentally were not successful. It is now thought that, through a series of elongation and desaturation reactions, 18:3*n*-3 is converted to the penultimate intermediate, 22:5*n*-3. Rather than using a Δ4 desaturase to complete the synthesis, 22:5*n*-3 is elongated to 24:5*n*-3, converted to 24:6*n*-3 by Δ6 desaturase, and finally chain-shortened to 22:6*n*-3 by one cycle of peroxisomal β-oxidation.

Fatty Acids as Components of Complex Lipids

Fatty acids are important building blocks for various cellular complex lipids (**Figure 7**). For simplicity, the pathways for incorporation of fatty acids into these lipids are outlined only briefly. More details can be found in any good biochemistry text. In most cases, fatty acyl-CoA and not free fatty acid participates in these biosynthetic reactions. Nearly all cells synthesize phospholipids, which are essential membrane constituents. Phospholipid synthesis takes place in the endoplasmic reticulum. It begins by fatty acylating the two free hydroxyl groups in α-glycerophosphate, a triose derived from glycolytic intermediates, yielding phosphatidic acid. Various head groups (e.g., choline, ethanolamine, inositol, or serine) can then be linked to the phosphate group. For synthesis of triacylglycerol, this phosphate moiety is removed, yielding diacylglycerol, and a third fatty acyl group is esterified to the free hydroxyl group.

Another type of lipid, the ether-linked phospholipids (e.g., plasmalogens), comprises about 20% of membrane phospholipids (**Figure 7**). Plasmalogen synthesis requires enzymes present in both peroxisomes and the endoplasmic reticulum. These lipids are thought to be part of the cellular defense mechanism against oxidative injury.

Fatty acids are also found esterified to the 3-hydroxyl group of cholesterol (cholesterol esters; ChE). ChE, which are more hydrophobic than free cholesterol, are a transport and storage form of cholesterol. ChE are found in high concentrations in low-density lipoproteins. Intracellular lipid droplets containing ChE are found in steroidogenic tissues and are thought to be a reservoir of cholesterol for steroid-hormone synthesis. The fatty acid most commonly found in ChE is 18:1. It must be activated to its CoA derivative before transfer to cholesterol in a reaction catalyzed by acyl-CoA cholesterol acyltransferase. ChE are also formed within lipoproteins by the transfer of one fatty acyl chain from phosphatidyl choline to cholesterol, a reaction catalyzed by circulating lecithin: cholesterol acyltransferase.

Synthesis of sphingolipids, which include sphingomyelin, ceramides, cerebrosides, and gangliosides, begins by the condensation of palmitoyl-CoA (16:0-CoA) with serine. The amino group of serine is then acylated by a second fatty acyl-CoA to form ceramide; the chain length of the second fatty acid can be variable. Transfer of phosphorylcholine (from the phospholipid phosphatidyl choline) to the hydroxyl group of ceramide yields

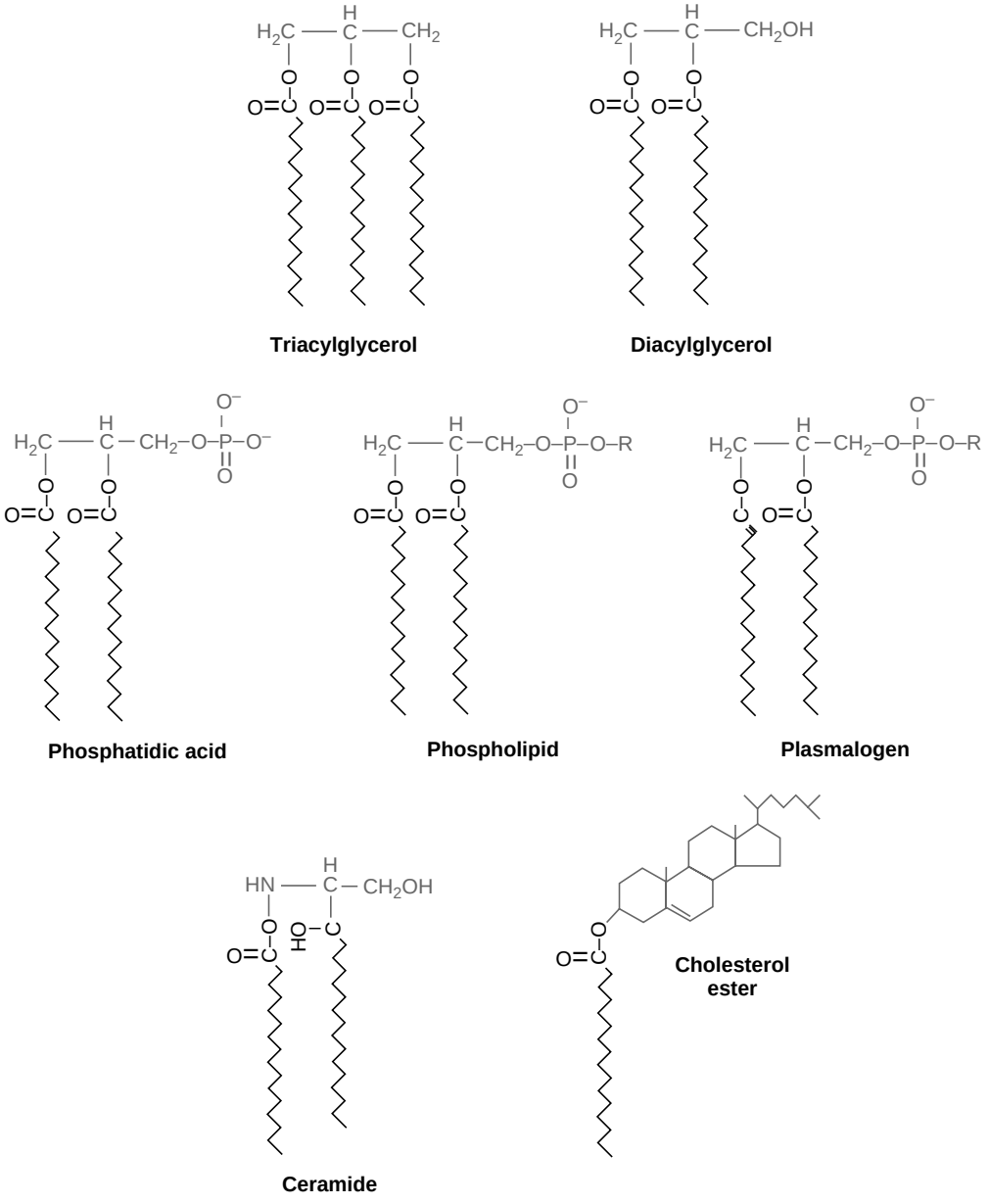


Figure 7 Fatty acids form the basis of most complex lipids. The part of the molecule derived from fatty acids is shown in black, and the part derived from other sources is shown in grey. For phospholipids and plasmalogens, R = choline, ethanolamine, inositol, serine, or a similar head group.

sphingomyelin. Alternatively, sugars (from sugar nucleotide donors) are added to produce the cerebrosides, gangliosides, and related lipids.

Eicosanoid Synthesis

The fatty acid 20:4*n*-6 (arachidonic acid) is the precursor of most eicosanoids, which include the prostaglandins, leukotrienes, and thromboxanes. Because it is an *n*-6 fatty acid, 20:4 must be derived from dietary lipids or synthesized by elongation and unsaturation of the essential fatty acid 18:2*n*-6. As with other fatty acids, cellular concentrations of unesterified 20:4 are low. Conversion of 20:4 to eicosanoids begins with an agonist-induced release of the fatty acid from the sn-2 position of membrane phospholipids via the action of phospholipase A2. Unlike most reactions of fatty acids, eicosanoid synthesis appears to use free 20:4 rather than its CoA derivative as the substrate. Cyclooxygenases (COX1 and COX2) catalyze a complex molecular oxygen-requiring reaction that converts 20:4 to

prostaglandin G2. This reaction involves carbon atoms in the middle of the acyl chain, rather than the methyl carbon (such as occurs in ω-oxidation) or the carboxyl carbon (such as occurs in nearly all other reactions of fatty acids). Prostaglandin G2 can subsequently be converted to other prostaglandins or to thromboxanes. As these compounds have potent biological effects, including mediation of inflammation, COX inhibitors are an important class of anti-inflammatory drugs. Free 20:4 is also the primary substrate for the enzyme 5-lipoxygenase, which is the first step in the synthesis of leukotrienes.

Fatty Acylation of Proteins

Covalent modification of proteins is a more recently discovered role of fatty acids. Fatty acylation of proteins frequently serves as a means of targeting or anchoring a protein to a membrane. Myristoylation, the addition of 14:0 to a protein, occurs at N-terminal glycine residues after removal of the initiator methionine. This process is generally co-translational and irreversible. N-myristoyl proteins include many signal-transduction-associated proteins, e.g., *src* and ADP-ribosylation factors. The enzyme N-myristoyltransferase catalyzes the reaction and uses 14:0-CoA as substrate.

Palmitoylation, the addition of 16:0 to a protein, is also commonly observed. This modification to the sulfydryl side chain of cysteine residues occurs post-translationally and is reversible. Both membrane-associated proteins and integral membrane proteins can be palmitoylated; examples are ion channels, neurotransmitter receptors, and sonic hedgehog. Protein palmitoyl transferases also use the CoA derivative of the fatty acid as a substrate. Several proteins are modified with both an N-terminal 14:0 and an S-linked 16:0 elsewhere in the protein chain. α-subunits of heterotrimeric G-proteins and endothelial nitric oxide synthase are examples of dually acylated proteins.

There are instances of acylation by fatty acids with chain lengths other than 14 or 16 carbon atoms. One nutritionally important example is the recently identified orexigenic peptide ghrelin. The active form of this 28-amino-acid peptide hormone has the medium-chain fatty acid 8:0 covalently esterified to the hydroxyl group of serine-3. Octanoylated ghrelin is believed to act at the level of the hypothalamus to stimulate appetite, perhaps via neuropeptide Y.

Vitamins and Fatty-Acid Metabolism

Several of the B vitamins are essential for normal fatty-acid metabolism (**Table 2**). Pantothenic acid is a constituent of CoA and is thus required for numerous reactions of fatty acids. Niacin and riboflavin are necessary for the synthesis of oxidized and reduced NAD(P) and FAD, respectively. These compounds play essential roles in fatty-acid oxidation, synthesis, and elongation. Biotin is a constituent of acetyl-CoA carboxylase and pyruvate carboxylase, both of which are involved in the synthesis of fatty acids from glucose. Thiamine is required for activity of the pyruvate dehydrogenase complex, which also participates in fatty-acid synthesis from glucose.

Regulation of Fatty-Acid Metabolism

A few specific aspects of the regulation of fatty-acid metabolism have been described above. More global regulatory mechanisms that deserve mention include those mediated by insulin and glucagon, sterol regulatory element-binding protein (SREBP) 1c, and peroxisome proliferator-activated receptor (PPAR) α. In the fed and fasted states, control of fuel metabolism is mediated to a large extent by insulin and glucagon, respectively. Effects of glucagon are mediated via cyclic adenosine monophosphate (cAMP)-dependent kinases and serve to decrease flux through glycolysis,

Table 2 Vitamins associated with fatty-acid metabolism

Vitamin	*Active form*	*Enzymes*	*Pathways*
Pantothenic acid	CoA	Many enzymes	Most reactions involving fatty acids
Niacin	NAD, NADH, NADP, NADPH	Dehydrogenases; reductases	Many pathways, particularly β-oxidation and fatty-acid synthesis and elongation
Riboflavin	FAD, $FADH_2$	Oxidases	β-Oxidation
Thiamine	Thiamine pyrophosphate	Pyruvate dehydrogenase complex; β-hydroxyphytanoyl-CoA lyase	Fatty-acid synthesis from glucose; phytanic acid β-oxidation
Biotin	Biocytin	Acetyl-CoA carboxylase; pyruvate carboxylase	Fatty-acid synthesis from glucose

thus decreasing the rate of *de novo* fatty-acid biosynthesis and increasing rates of mitochondrial β-oxidation and ketogenesis. Insulin effects are mediated through activation of its receptor tyrosine kinase and are in general opposite to those of glucagon, stimulating glycolysis and fatty-acid synthesis while inhibiting fatty-acid degradation. Insulin and glucagon have both acute and long-term effects on fatty-acid metabolism. The transcription factor SREBP1c is thought to mediate the action of insulin in upregulating genes involved in fatty-acid synthesis. Activation of PPARα on the other hand increases rates of fatty-acid oxidation and ketogenesis. Endogenous ligands for this nuclear receptor are thought to include polyunsaturated fatty acids and branched-chain fatty acids. The PPARs heterodimerize with the retinoid X receptor, and both receptors must be ligand-bound for transcriptional activation. Several mitochondrial, microsomal, and peroxisomal genes associated with fatty-acid catabolism are upregulated via PPARα stimulation.

See also: **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Fatty Acids**: Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Lipids**: Chemistry and Classification; Composition and Role of Phospholipids. **Obesity**: Definition, Etiology and Assessment.

Further Reading

Frohnert BI and Bernlohr DA (2000) Regulation of fatty acid transporters in mammalian cells. *Progress in Lipid Research* **39**: 83–107.

Gibbons GF (2003) Regulation of fatty acid and cholesterol synthesis: co-operation or competition? *Progress in Lipid Research* **42**: 479–497.

Gunstone FD, Harwood JL, and Padley FB (eds.) (1994) *The Lipid Handbook*, 2nd edn. London: Chapman & Hall.

Kunau WH, Dommes V, and Schulz H (1995) Beta-oxidation of fatty acids in mitochondria, peroxisomes, and bacteria: a century of continued progress. *Progress in Lipid Research* **34**: 267–342.

McGarry JD and Foster DW (1980) Regulation of hepatic fatty acid oxidation and ketone body production. *Annual Review of Biochemistry* **49**: 395–420.

Numa S (ed.) (1984) *Fatty Acid Metabolism and its Regulation*. New York: Elsevier.

Vance DE and Vance JE (eds.) (2002) *Biochemistry of Lipids, Lipoproteins and Membranes*, 4th edn. New Comprehensive Biochemistry 36. New York: Elsevier.

Wanders RJ, Vreken P, Ferdinandusse S *et al.* (2001) Peroxisomal fatty acid alpha- and beta-oxidation in humans: enzymology, peroxisomal metabolite transporters and peroxisomal diseases. *Biochemical Society Transactions* **29**: 250–267.

Watkins PA (1997) Fatty acid activation. *Progress in Lipid Research* **36**: 55–83.

Monounsaturated

P Kirk, University of Ulster, Coleraine, UK

Introduction

Fatty acids are described according to two characteristics: chain length and degree of saturation with hydrogen. Monounsaturated fatty acids (MUFA) have, as the name suggests, only one unsaturated bond attached to the carbon chain. This double bond is fixed in nature and is positioned on the ninth carbon atom counting from the methyl (omega) end of the fatty-acid chain. Four of these MUFA are found in significant quantities in food, the most common being oleic acid ($C_{18:1}$) (**Figure 1**). This n-9 fatty acid is capable of being synthesized by animals, including humans, but is predominantly incorporated via the diet. While butter and animal fats contain only small amounts of oleic acid, olive oil is a rich source. Olive oil, which comprises up to 70% of the fat intake in Mediterranean diets, is postulated to be effective in decreasing the risk of certain chronic diseases. These include such diseases as coronary heart disease, cancers, and inflammatory disorders, particularly rheumatoid arthritis.

Cholesterol Metabolism

Cholesterol metabolism is of fundamental biological importance. All vertebrates require cholesterol as a precursor for bile acids and hormones, including corticosteroids, sex steroids, and vitamin D. The amount of cholesterol found in tissues greatly exceeds the requirement for production of these hormones and bile acids, and the bulk of this excess is associated with the cell-membrane structure, where it is believed to modulate the physical state of phospholipid bilayers.

Cholesterol circulates in plasma as a component of lipoproteins. There are several distinct classes of plasma lipoprotein, which differ in several respects, including type of apolipoprotein and relative content of triacylglycerol and cholesterol.

Cholesterol transport

Chylomicron remnants deliver dietary cholesterol to the liver. It is then incorporated into very low-density lipoproteins (VLDL), which are secreted in

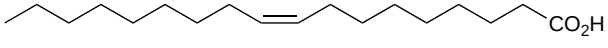


Figure 1 Structure of oleic acid, $C_{18:1}$.

plasma. The VLDL acquire cholesteryl esters and apolipoprotein E (apo E) from high-density lipoproteins (HDL) to produce intermediate-density lipoproteins (IDL), which are rapidly taken up by the liver or are further catabolized into low-density lipoproteins (LDL). These cholesterol-rich LDL particles are catabolized only slowly in human plasma and are therefore present at relatively high concentrations. Elimination of cholesterol from these extrahepatic cells is achieved by the delivery of cholesterol from cell membranes to plasma HDL in the first step of a pathway known as reverse cholesterol transport. This process allows for esterification of cholesterol and its delivery back to the liver.

LDL, HDL, and atherosclerosis

Membrane function is compromised if it contains either too much or too little cholesterol. Epidemiological studies have classified raised plasma cholesterol levels as a risk factor for atherosclerosis, and it is one of the more important predictors of coronary heart disease (CHD). Elevated plasma cholesterol concentration (hypercholesterolemia) is associated with an increased concentration of LDL, owing to either an increased rate of LDL formation or a decrease in the rate at which they are cleared from plasma, and usually a decreased concentration of HDL. Numerous dietary-intervention studies have aimed both to prevent CHD and to reduce total mortality, but almost all have been ineffective.

MUFA and CHD

Many of the trials conducted concentrated on the substitution of polyunsaturated vegetable oils for saturated fat from animal sources and on decreasing the amount of dietary cholesterol. These studies followed the reasoning that fats rich in saturated fatty acids (SFA) raised plasma cholesterol mainly by increasing plasma LDL cholesterol levels, and oils rich in polyunsaturated fatty acids (PUFA) lowered plasma cholesterol mainly by decreasing LDL cholesterol. The MUFA were first considered neutral in regard to their influence on plasma cholesterol, but more recent findings suggest a decrease in total LDL cholesterol concentration following substitution of SFA by MUFA. Moreover, clinical trials have also shown that a MUFA-rich diet does not decrease concentrations & HDL, the lipoprotein inversely correlated with CHD.

Although important links exist between cholesterol metabolism and aspects of cell function, other complicating factors must be considered. Cholesterol metabolism is sensitive to the inflammatory response that accompanies most pathological events. Tumor necrosis factor (TNF) reduces LDL and HDL cholesterol levels and inhibits lipoprotein lipase, resulting in a fall in cholesterol and an increase in triacylglycerol levels. These changes may be perpetuated beyond the acute phase if an inflammatory process is present. Cholesterol metabolism is also sensitive to genetic and environmental factors, which may have independent effects on noncardiovascular disease. As a consequence, the relationship between cholesterol levels and the presence or absence of a disease state must be interpreted with caution.

Atherogenesis and Endothelial Dysfunction

Atherosclerosis can be considered as a chronic inflammatory disease, which slowly progresses over a period of decades before clinical symptoms become manifest. The atherogenic process comprises interactions between multiple cell types, which initiate a cascade of events involving alterations in vascular production of autocoids, cytokines, and growth factors. The endothelium, because of its location between blood and the vascular wall, has been implicated in the atherogenic process from the initial stages.

Function of endothelial cells

Owing to the strategic location of the endothelium, it is able to perform many different functions. In addition to acting as a protective barrier, endothelial cells have been shown to play important roles in control of homeostasis, capillary transport, and, more importantly, regulation of the tone of underlying vascular smooth muscle. The endothelium evokes relaxation of these muscle cells, allowing vasodilation via the chemical factor endothelium-derived relaxing factor (EDRF), which has been identified as nitric oxide (NO). The EDRF or NO is vital for maintaining the vasodilatory capacity of vascular muscle and also controls levels of platelet function and monocyte adhesion. Any endothelial injury or dysfunction could therefore be an important factor in atherosclerosis.

Endothelial dysfunction

Decrease in the production, release, or action of NO may lead to enhanced expression of adhesion molecules and chemotactic factors at the endothelial surface. The exact nature of endothelial dysfunction is unknown, although possibilities include a decreased expression of NO synthase, imbalance between the production of endothelium-derived constricting and

relaxing factors, production of an endogenous NO synthase inhibitor, and overproduction of oxygen-derived free radicals including O_2^-. The release of the free radical O_2^- from smooth muscle cells is believed to be responsible for the oxidation of LDL cholesterol. Raised cholesterol levels and – more importantly – increased levels of oxidatively modified LDL cholesterol (OxLDL) are considered to be among the most powerful inhibitors of normal endothelial function and hence contribute to the process of atherogenesis.

Lipid peroxidation and atherosclerosis

Lipid peroxidation apparently plays a major role in the pathology of atherosclerosis. Atherosclerosis, which is usually a precondition for CHD, is a degenerative process leading to the accumulation of a variable mixture of substances including lipid in the endothelium of the arteries. This disease is characterized by the formation of a fatty streak and the accumulation of cells loaded with lipid: the foam cells. These cells are believed to arise from white blood cell-derived macrophages or arterial smooth muscle cells. Most of the lipid in the foam cells is in the form of LDL particles. Although research has determined that LDL receptors are responsible for the uptake of LDL by cells, the arterial uptake of LDL, which leads to development of foam cells, occurs by a different pathway. It is only when the LDL particles have undergone oxidative modification that they are available for uptake by macrophages via the scavenger receptor. During the course of oxidative modification, LDL cholesterol acquires various biological properties not present in native LDL that make it a potentially important mediator, promoting atherogenesis. The LDL, once oxidized, becomes cytotoxic and causes local cellular damage to the endothelium. This process, which enhances LDL uptake to generate foam cells, is considered one of the earliest events in atherogenesis (**Figure 2**).

MUFA and atherogenesis

Studies have looked at the oxidizability *in vitro* of LDL using nonphysiological oxidizing conditions to evaluate the susceptibility of LDL to oxidation and hence its atherogenic potential. It is well known that modification of LDL is inhibited by various antioxidants commonly present within plasma LDL particles. More recent studies, however, indicate that raising the ratio of $C_{18:1}$ to $C_{18:2}$ (linoleic acid) may also reduce the susceptibility of LDL to oxidation. The LDL is particularly vulnerable to peroxidation once PUFA form part of the lipoprotein fraction of cell membranes, as these fatty acids

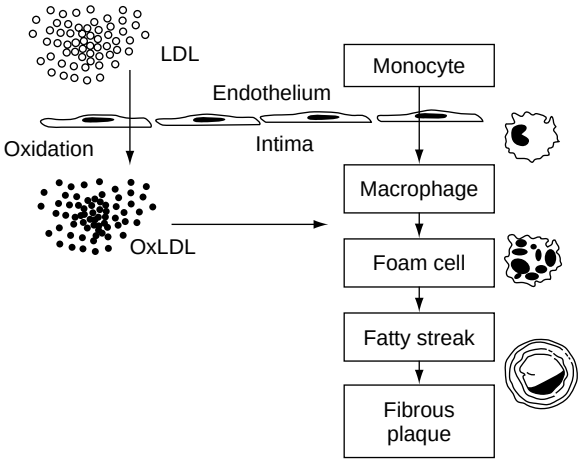


Figure 2 The role of oxidized LDL (OxLDL) in formation of foam cells. Reproduced with permission from Ashwell M (1993) *Diet and Heart Disease – A Round Table of Factors*. London: British Nutrition Foundation.

have reactive double bonds in their structures. The MUFA are much less easily oxidized as they have only one double bond. This property may confer a protective effect against CHD by generating LDL particles more resistant to oxidation. Further protection may be afforded from MUFA as they do not lower HDL. It is postulated that oxidized HDL, in contrast to oxidized LDL, is not avidly taken up by macrophages but instead inhibits the modification of LDL, thereby substantially decreasing oxidized LDL cellular uptake.

Oxidation of LDL cholesterol is, therefore, clearly linked to damage to the endothelium and hence to the process of atherogenesis. It has, however, more far-reaching effects, as it has also been linked to activation and aggregation of platelets. This process is involved in the production of occlusive thrombosis, which contributes significantly to the fibrous atherosclerotic plaque.

Thrombosis and Fibrinolysis

The importance of thrombosis in causing heart disease is receiving increasing attention. Thrombosis, in contrast to atherosclerosis, is an acute event resulting in the formation of a thrombus or blood clot, which is an aggregate of fibrin, platelets, and red cells. Blood clotting or coagulation is an important process as it is responsible for repairing tissues after injury. Under normal physiological conditions, a blood clot forms at the site of injury. Platelets are attracted to the damaged tissue and adhere to the surface. They are then activated to release substances that attract more platelets, allowing platelet aggregation and triggering coagulation mechanisms.

The coagulation cascade

The process of blood coagulation involves two pathways: the extrinsic and intrinsic pathways (**Figure 3**). The cascade is dependent on a series of separate clotting factors, each of which acts as a catalyst for the next step in the system. The process results in the formation of insoluble fibrin from the soluble protein fibrinogen. This then interacts with a number of blood components, including red blood cells, to form the thrombus. Any damage to the endothelium, therefore, causes platelet aggregation and adherence to the lining of the blood-vessel walls, thereby triggering the coagulation cascade. An imbalance of this process, by increasing the rate of thrombus formation, could increase the risk of CHD, and data have shown that levels of factor VII and fibrinogen are particularly important in balancing the coagulation cascade.

Factor VII and fibrinogen

There is accumulating evidence the factor VII is involved in arterial thrombosis and atherogenesis. The physiology of the factor VII system is intricate, not least since it can potentially exist in several forms. Activation of factor VIIc is generally achieved by tissue factor and initiates blood coagulation by subsequent activation of factors IX and X. It has been further suggested that tissue factor associated with the lipoproteins LDL and VLDL, but not HDL, may possibly generate factor VIIc activity, and a direct relationship is believed to exist between the level of factor VII complex in plasma and the dietary influence on plasma triacylglycerol concentration.

Several mechanisms have been suggested whereby an increase in plasma fibrinogen concentration may be linked to CHD. These include the involvement of fibrinogen and fibrin in the evolution of the atheromatous plaque through fibrin deposition and in

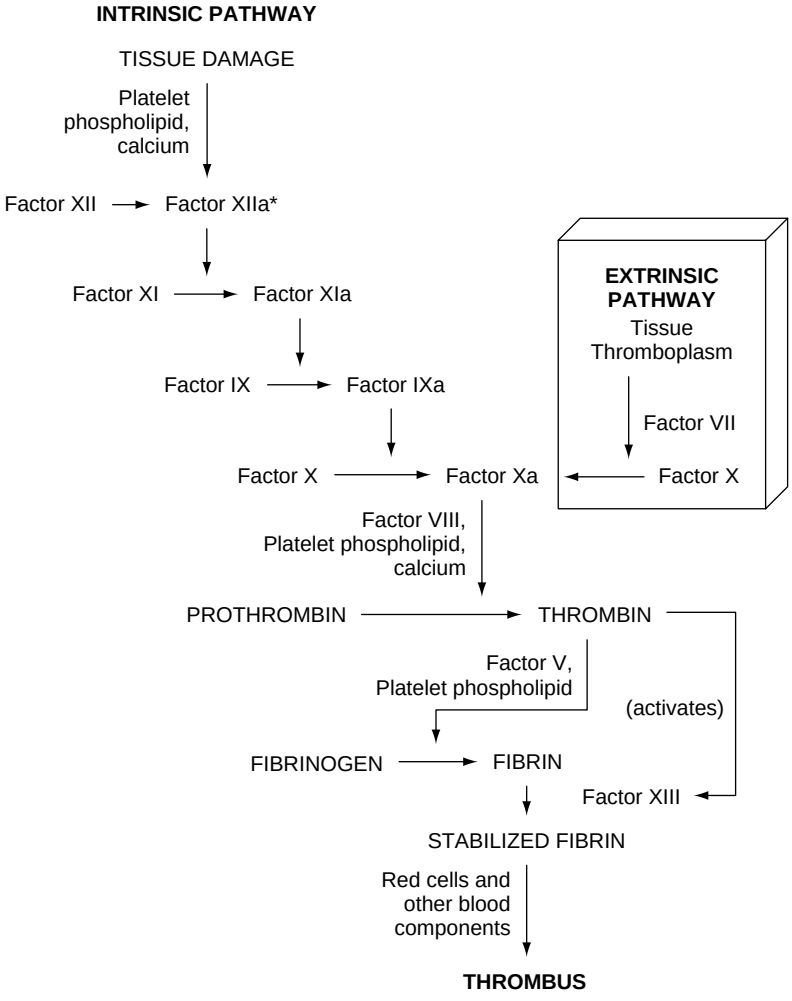


Figure 3 The coagulation cascade; a, active. Reproduced with permission from Buttriss JL and Gray J (1992) *Coronary Heart Disease II, Fact File 8*. London: National Dairy Council. Copyright National Dairy Council.

platelet aggregation through its impact on blood viscosity, which in turn is related to the risk of thrombosis. A mechanism exists to dissolve the thrombus by the breakdown or lysis of the fibrin meshwork (fibrinolysis). Plasminogen, which is generated by plasmin, is the zymosan that ultimately effects fibrinolysis. Failure of the mechanism to activate will cause obstruction of blood vessels and prevent normal blood supply.

Fibrinolysis

Investigators have shown that a decrease in the release of tissue plasminogen activator (tPA) and an elevation of plasminogen activator inhibitor 1 (PAI-1) will reduce fibrinolytic function. It has emerged that triacylglycerol-rich lipoproteins stimulate PAI-1 secretion from endothelial cells, and furthermore it has been shown that OxLDL induces secretion, whereas native LDL has no detectable effect. Lipoprotein (a) (Lp(a)) has also been linked with a decrease in fibrinolysis. Lp(a) is an LDL-like particle consisting of the protein apo(a). It is believed that apo(a) competes with plasminogen and plasmin for binding to fibrin, thus interfering with fibrinolysis; LDL and Lp(a) may represent, therefore, an important link between thrombotic and lipid mechanisms in atherogenesis.

MUFA and thrombosis

Results from animal studies have shown an elevation in platelet activation and hence greater risk of thrombosis as a result of feeding saturated fat. Platelet aggregation thresholds, however, decrease when total fat intake is decreased or when dairy and animal fats are partially replaced with vegetable oils rich in PUFA. These studies failed, however, to keep the intakes of SFA and total fat constant. More recent work has shown that, in fact, diets high in PUFA significantly increase platelet aggregation in animals, compared with MUFA-rich diets. The changes in fatty-acid composition may affect blood clotting because the increase in PUFA allows for oxidation of LDL. As previously mentioned, OxLDL is cyto-toxic, and this can cause endothelial damage leading to the activation of platelets, generation of factor VII, and hence thrombus formation. Increased dietary intakes of MUFA may also increase the rate of fibrinolysis by lowering levels of LDL cholesterol and reducing the susceptibility of LDL to oxidation, thereby affecting both PAI-1 secretion and apo(a) activity.

It must be noted that both atherosclerosis and thrombosis are triggered by inflammation, and evidence suggests that several hemostatic factors other than the glycoprotein fibrinogen not only have an important role in thrombotic events but are also recognized as potentially important CHD risk factors.

Inflammation and Oxidative Damage

Many diseases that have an inflammatory basis such as cancer, sepsis, and chronic inflammatory diseases such as rheumatoid arthritis (RA) have symptoms mediated by pro-inflammatory mediators named cytokines. These mediators, which include interleukins (IL) 1–8, tumor necrosis factors (TNF), and interferons, are essential for protection from invading bodies. They act by producing a situation in which immune cells are attracted to the inflammatory site and are activated. An inflammatory stimulus, such as tissue damage incurred by trauma or invasion of tissue by bacteria or viruses, induces production of IL-1, IL-6, and TNF from a range of immune cells, including phagocytic leucocytes and T and B lymphocytes. Once induced, IL-6, IL-1, and TNF further induce each other's production, leading to a cascade of cytokines, which are capable of producing metabolic and immune effects. Inflammatory stimuli also bring about the activation of neutrophils to release free radicals, which enhance the production of TNF and other cytokines. Overproduction of these pro-inflammatory mediators may, therefore, allow excessive release of reactive oxygen species (ROS) into extracellular fluid to damage its macromolecular components.

Oxidative damage

Free radicals are any species capable of an independent existence that contain one or more unpaired electrons. ROS is a collective term, referring not only to oxygen-centered radicals such as superoxide ($O_2^{\cdot}$) and the hydroxy radical ($^{\cdot}OH$), but also to hydrogen peroxide (H_2O_2), ozone (O_3), and singlet oxygen ($^1O_2$). These are produced as the by-products of normal metabolism and, as such, are highly reactive in chemical terms. In order to become more stable chemically, the free radical reacts with other molecules by either donating or taking an electron, in either case leaving behind another unstable molecule, and hence this becomes a chain reaction. So, although oxygen is essential for life, in certain circumstances it may also be toxic. Damage caused by ROS to cellular target sites includes oxidative damage to proteins, membranes (lipid and proteins), and DNA. PUFA are particularly vulnerable to ROS attack because they have unstable double bonds in their structure. This process is termed 'lipid

peroxidation'; because PUFA are an essential part of the phospholipid fraction of cell membranes, uncontrolled lipid peroxidation can lead to considerable cellular damage. The balance of MUFA in cell membranes is also critical to cell function, but, as already noted, MUFA are far less vulnerable to lipid peroxidation.

MUFA and inflammation

Oxidative damage by ROS to DNA and lipids contributes significantly to the etiology of cancer and atherosclerosis. A decrease in production of pro-inflammatory mediators would, therefore, be beneficial by decreasing the release of ROS. Diminishing the production of cytokines is also believed to improve the symptoms of RA. It has been suggested that olive oil may have anti-inflammatory properties as it can reduce the production of these proinflammatory mediators. Although few studies have been carried out on the benefits of olive oil on symptoms of inflammation, it is possible that olive oil produces a similar effect to fish oil. Fish oils and butter have both been shown to reverse the proinflammatory effects of one cytokine, TNF. Further research, where $C_{18:1}$ was added to a diet containing coconut oil, resulted in responses to TNF that were similar to those seen in animals fed butter. It was assumed that, as the anti-inflammatory effects of butter appeared to be due to its oleic acid content, olive oil should be more anti-inflammatory. This was put to the test, and, while both butter and olive oil reduced the extent of a number of symptoms of inflammation, olive oil showed a greater potency than butter. From this, it can be concluded that dietary factors such as olive oil may play a significant protective role in the development severity of RA.

Carcinogenesis

Cancer is second only to CHD as a cause of death in Western countries. Cancer in humans is a multistep disease process in which a single cell can develop from an otherwise normal tissue into a malignancy that can eventually destroy the organism. Carcinogenesis is believed to proceed through three distinct stages. Initiation is brought about when carcinogens mutate a single cell. This mutation provides a growth advantage, and cells rapidly proliferate during the second stage, promotion. Tumor promotion produces relatively benign growths, which can be converted into cancer in the third stage, malignant conversion. While the causes of cancer are not known with certainty, both initiation and conversion require some form of genetic alteration, and ROS and other free radicals have long been known to be mutagenic (**Figure 4**).

Oxidation and cancer

Although PUFA are the most reactive of substrates for ROS attack leading to lipid peroxidation, interest is centering on the detection of oxidized nucleic acids as an indicator of pro-oxidant conditions. It

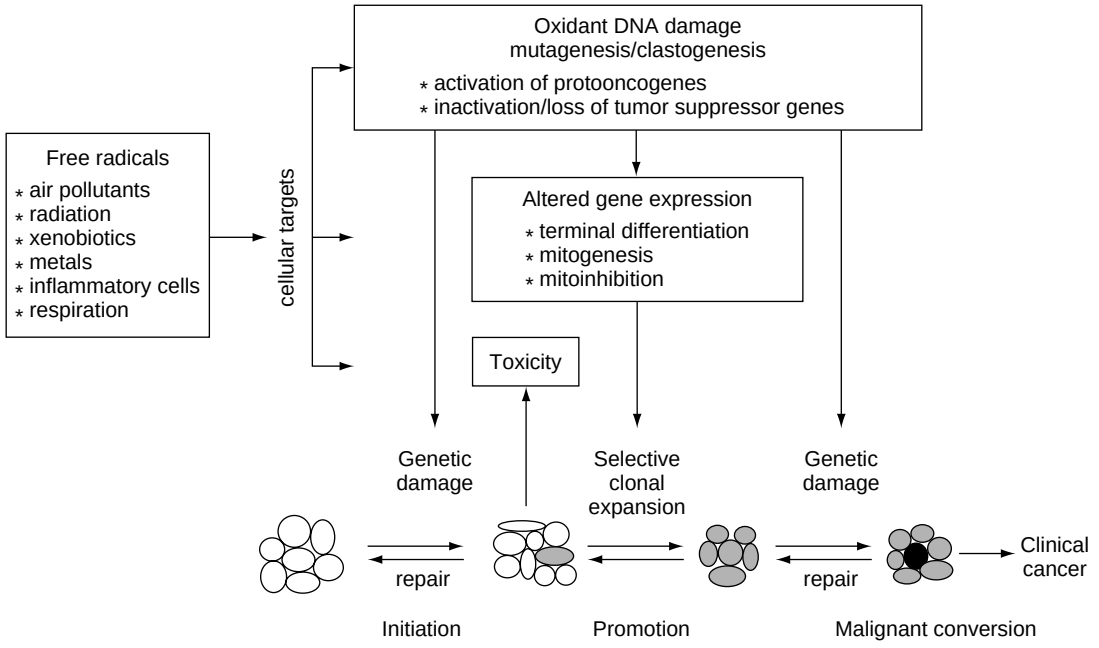


Figure 4 The role of oxidants in multistage carcinogenesis. Reproduced with permission from Guyton KZ and Kensler TW (1993) Oxidative mechanisms in carcinogenesis. *British Medical Bulletin* **49**: 523–544.

has been indicated that significant oxidative damage occurs *in vitro* and contributes to the etiology of cancer. It has become apparent that many genotoxic agents act through the common mechanism of oxidative damage to DNA. Oxidative processes may be responsible for initiating carcinogenic changes via DNA oxidative damage and may also act as tumor promoters, modulating the expression of genes that regulate cell differentiation and growth and act synergistically with the initiators. Animal studies have indicated diets containing high levels of $C_{18:2}$ as strong promoters of tumors, and this may be as a result of increased oxidative stress. The fact that MUFA are much less readily oxidized may therefore confer a protective effect against carcinogenesis.

Immune function and cancer

The diet is believed to play an important role in the onset of carcinogenesis, and there are a number of carcinogens present in food, including mycotoxins, polycyclic hydrocarbons, and pesticides. Associations have been made between dietary fat intake and morbidity and mortality from breast and colon cancer. Another possible mechanism for the proposed protective effects against cancer of olive oil compared with sunflower oil involves diet-induced alterations in host immune responses. Both the type and concentration of dietary fats have been reported to influence immune status in several animal models. The PUFA $C_{18:2}$ is necessary for T-cell-mediated immunity, but high intakes will suppress immune function and may therefore increase the risk of cancer. Furthermore, comparisons between the effects of diets rich in $C_{18:2}$ and those rich in $C_{18:1}$ on varying indicators of immune function in mice have shown that, while dietary $C_{18:2}$ predisposed animals to suppression of certain T-cell-mediated reactions, diets rich in $C_{18:1}$ did not. MUFA may therefore have a significant effect in humans against cancer, by lowering the risk of suppression of T-cell activity.

Other Physiological Effects

Because many, sometimes competing, mechanisms appear to mediate the relation between intake of MUFA and CHD incidence, no single surrogate biochemical or physiological response can predict with confidence the effect of a particular dietary pattern. For this reason examinations of the relation between specific dietary factors and CHD incidence itself are particularly valuable because such studies integrate the effects of all known and unknown mechanisms. The extremely low rate of CHD in countries with high consumption of olive oil, for instance, suggests the benefits of substituting this fat for other fats. This kind of analysis has been expanded further by noting that MUFA intake is inversely associated with total mortality as well as with CHD. Some effects may well be because of the amount of antioxidant vitamins olive oil contains. Vegetable oils are the most important source of α-tocopherol in most diets, and olive oil contains about 12 mg per 100 g. Evidence indicates that α-tocopherol functions as a free-radical scavenger to protect cellular membranes from oxidative destruction. Oxidative stress has been linked to an increased risk of many chronic diseases, including atherosclerosis, cancer, and inflammatory disorders. Other injuries such as cataract and reperfusion injury are also associated with an increase in oxidative stress and a decrease in antioxidant activity.

A large body of evidence suggests a beneficial effect of MUFA in the diet. Although much remains to be learned about the mechanisms by which $C_{18:1}$ acts, it is believed to lower risks of CHD, several common cancers, cataracts, and other inflammatory disorders. It is suggested, therefore, that consuming MUFA, for instance in the form of olive oil as used widely in the Mediterranean diet, is likely to enhance long-term health.

See also: **Antioxidants**: Diet and Antioxidant Defense. **Arthritis**. **Cancer**: Epidemiology and Associations Between Diet and Cancer; Effects on Nutritional Status. **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Coronary Heart Disease**: Hemostatic Factors; Lipid Theory; Prevention. **Cytokines**. **Dairy Products**. **Fats and Oils**. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Immunity**: Physiological Aspects. **Lipids**: Chemistry and Classification. **Lipoproteins**.

Further Reading

Ashwell M (1993) *Diet and Heart Disease – A Round Table of Factors*. London: British Nutrition Foundation.

Barter P (1994) Cholesterol and cardiovascular disease: basic science. *Australia and New Zealand Journal of Medicine* **24**: 83–88.

Besler HT and Grimble RF (1993) Modulation of the response of rats to endotoxin by butter and olive and corn oil. *Proceedings of the Nutrition Society* **52**: 68A.

Cerutti PA (1985) Prooxidant states and tumor promotion. *Science* **227**: 375–381.

Daae LW, Kierulf P, Landass S, and Urdal P (1993) Cardiovascular risk factors: interactive effects of lipid, coagulation, and fibrinolysis. *Scandinavian Journal of Clinical Laboratory Investigation* **532**: 19–27.

Dunnigan MG (1993) The problem with cholesterol. No light at the end of the tunnel. *British Medical Journal* **306**: 1355–1356.
Ernst E (1993) The role of fibrinogen as a cardiovascular risk factor. *Atherosclerosis* **100**: 1–12.
Guyton KZ and Kensler TW (1993) Oxidative mechanisms in carcinogenesis. *British Medical Bulletin* **49**: 523–544.
Halliwell B (1989) Tell me about free radicals, doctor: a review. *Journal of the Royal Society of Medicine* **82**: 747–752.
Hannigan BM (1994) Diet and immune function. *British Journal of Biomedical Science* **51**: 252–259.
Hoff HF and O'Neil J (1991) Lesion-derived low density lipoprotein and oxidized low density lipoprotein share a lability for aggregation, leading to enhanced macrophage degradation. *Arteriosclerosis and Thrombosis* **11**: 1209–1222.
Linos A, Kaklamanis E, and Kontomerkos A (1991) The effect of olive oil and fish consumption on rheumatoid arthritis – a case control study. *Scandinavian Journal of Rheumatology* **20**: 419–426.
Mensink RP and Katan MB (1989) An epidemiological and an experimental study on the effect of olive oil on total serum and HDL cholesterol in healthy volunteers. *European Journal of Clinical Nutrition* **43**(supplement 2): 43–48.
Morel DW, Dicorleto PE, and Chisolm GM (1984) Endothelial and smooth muscle cells alter low density lipoprotein *in vitro* by free radical oxidation. *Arteriosclerosis* **4**: 357–364.
National Dairy Council (1992) *Coronary Heart Disease*. Fact File No. 8. London: NDC.
Visioli F and Galli C (1995) Natural antioxidants and prevention of coronary heart disease: the potential role of olive oil and its minor constituents. *Nutrition Metabolism and Cardiovascular Disease* **5**: 306–314.

Omega-3 Polyunsaturated

A P Simopoulos, The Center for Genetics, Nutrition and Health, Washington, DC, USA

Introduction

Over the past 20 years many studies and clinical investigations have been carried out on the metabolism of polyunsaturated fatty acids (PUFAs) in general and on n-3 fatty acids in particular. Today we know that n-3 fatty acids are essential for normal growth and development. Research has been carried out in animal models, tissue cultures, and human beings. The original observational studies have given way to controlled clinical intervention trials. Great progress has taken place in our knowledge of the physiologic and molecular mechanisms of the n-3 fatty acids in health and disease. Specifically, their beneficial effects have been shown in the prevention and management of coronary heart disease, hypertension, type 2 diabetes, renal disease, rheumatoid arthritis, ulcerative colitis, Crohn's disease, and chronic obstructive pulmonary disease. This chapter focuses on the sources, desaturation and elongation of n-6 and n-3 fatty acids; evolutionary aspects of diet relative to n-3 fatty acids and the n-6:n-3 balance; eicosanoid metabolism and biological effects of n-6 and n-3 fatty acids; nutrigenetics – interaction between the n-6:n-3 fatty acids and the genome; effects of dietary α-linolenic acid compared with long-chain n-3 fatty acid derivatives on physiologic indexes; human studies in growth and development; coronary heart disease; inflammation – a common base for the development of coronary heart disease, diabetes, arthritis, mental health and cancer; the need to return the n-3 fatty acids into the food supply for normal homeostasis; and future considerations.

n-6 and n-3 Fatty Acids: Sources, Desaturation and Elongation

Unsaturated fatty acids consist of monounsaturates and polyunsaturates. There are two classes of PUFA: n-6 and n-3. The distinction between n-6 and n-3 fatty acids is based on the location of the first double bond, counting from the methyl end of the fatty acid molecule. In the n-6 fatty acids, the first double bond is between the 6th and 7th carbon atoms and in the n-3 fatty acids the first double bond is between the 3rd and 4th carbon atoms. Monounsaturates are represented by oleic acid an n-9 fatty acid, which can be synthesized by all mammals including humans. Its double bond is between the 9th and 10th carbon atoms (**Figure 1**).

n-6 and n-3 fatty acids are also known as essential fatty acids (EFAs) because humans, like all mammals, cannot make them and must obtain them in their diet. n-6 fatty acids are represented by linoleic acid (LA; 18:2n-6) and n-3 fatty acids by α-linolenic acid (ALA; 18:3n-3). LA is plentiful in nature and is found in the seeds of most plants except for coconut, cocoa, and palm. ALA, on the other hand, is found in the chloroplasts of green leafy vegetables and in the seeds of flax, rape, chia, perilla, and in walnuts (**Tables 1**, **2**, and **3**). Both EFAs are metabolized to longer chain fatty acids of 20 and 22 carbon atoms. LA is metabolized to arachidonic acid (AA; 20:4n-6) and LNA to eicosapentaenoic acid (EPA; 20:5n-3) and docosahexaenoic acid (DHA; 22:6n-3), increasing the chain length and degree of unsaturation by adding extra double bonds to the carboxyl end of the fatty acid molecule (**Figure 2**).

Humans and other mammals, except for carnivores such as lions, can convert LA to AA

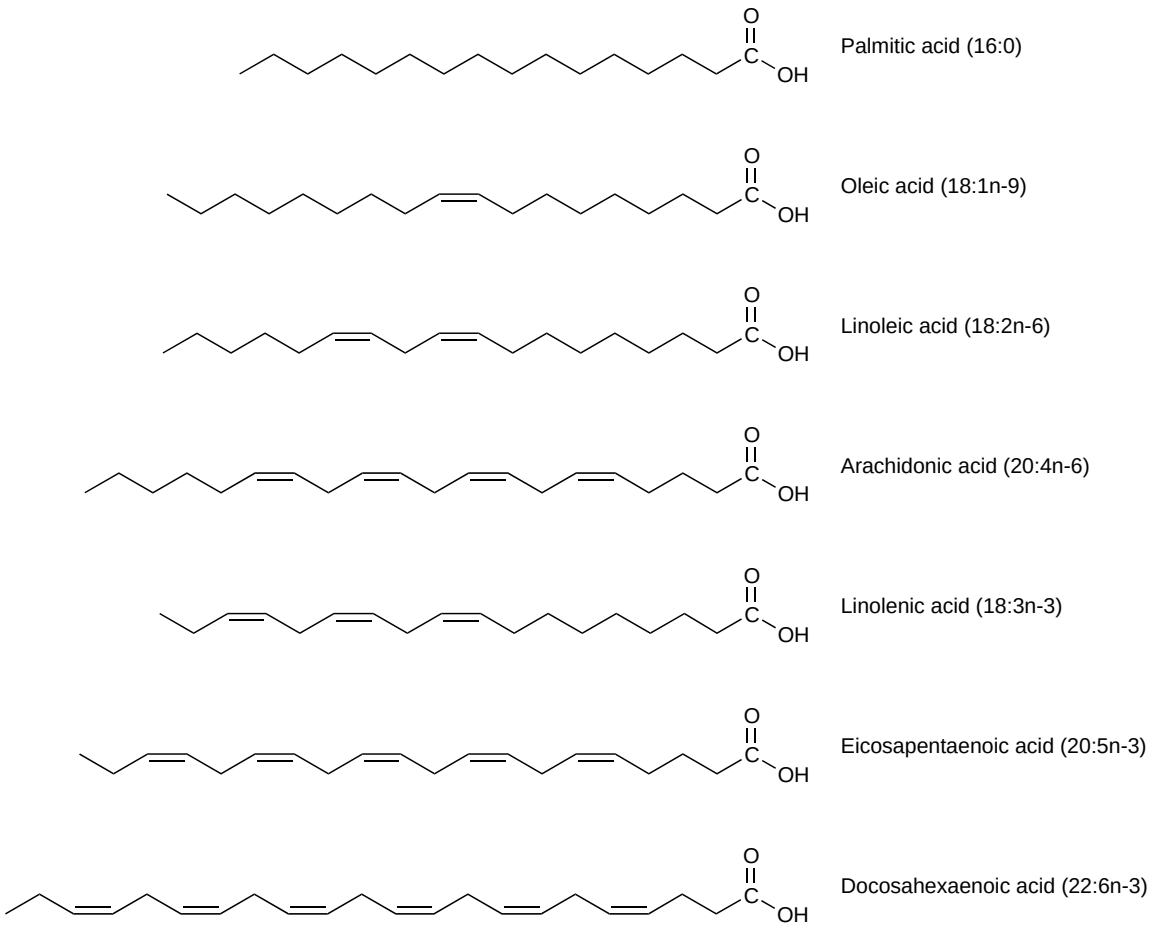


Figure 1 Structural formulas for selected fatty acids.

and ALA to EPA and DHA. This conversion was shown by using deuterated ALA. There is competition between n-6 and n-3 fatty acids for the desaturation enzymes. However, both Δ-4 and Δ-6 desaturases prefer n-3 to n-6 fatty acids. But a high LA intake interferes with the desaturation and elongation of ALA. *Trans*-fatty acids interfere with the desaturation and elongation of both LA and ALA.

Table 1 Polyunsaturated oils high in n-6 and n-3 fatty acids

n-6 oils	*n-3 oils*
Corn oil	Fish oil
Safflower oil	Chia oil
Sunflower seed oil	Perilla oil
Cottonseed oil	Flaxseed oil
Soybean oil	Canola oil
Peanut oil	Walnut oil
Sesame oil	Soybean oil[a]
Grapeseed oil	
Borage oil	
Primrose oil	

[a]note: soybean oil is higher in n-6 fatty acids than most n-3 oils, so it belongs in both categories.

Δ-6 desaturase is the limiting enzyme and there is some evidence that it decreases with age. Premature infants, hypertensive individuals, and some diabetics are limited in their ability to make EPA and DHA from ALA. These findings are important and need to be considered when making dietary recommendations. EPA and DHA are found in the oils of fish, particularly fatty fish (**Table 4**). AA is found predominantly in the phospholipids of grain-fed animals and eggs.

LA, ALA, and their long-chain derivatives are important components of animal and plant cell membranes. In mammals and birds, the n-3 fatty acids are distributed selectively among lipid classes. ALA is found in triglycerides, in cholesteryl esters, and in very small amounts in phospholipids. EPA is found in cholesteryl esters, triglycerides, and phospholipids. DHA is found mostly in phospholipids. In mammals, including humans, the cerebral cortex, retina, and testis and sperm are particularly rich in DHA. DHA is one of the most abundant components of the brain's structural lipids. DHA, like EPA, can be

Table 2 Comparison of dietary fats (fatty acid content normalized to 100%)

Dietary fat	*Saturated fat*	*Polyunsaturated fat*			*Monounsaturated fat*	*Cholesterol*
		LA	*ALA*	*LA:ALA*		
Flaxseed oil	10	16	53	(0.3)	20	0
Canola (rapeseed) oil	6	22	10	(2.2)	62	0
Walnut oil	12	58	12	(4.8)	18	0
Safflower oil	10	77	Trace	(77)	13	0
Sunflower oil	11	69	–	(69)	20	0
Corn oil	13	61	1	(61)	25	0
Olive oil	14	8	1	(8.0)	77	0
Soybean oil	15	54	7	(7.7)	24	0
Margarine	17	32	2	(16)	49	0
Peanut oil	18	33	–	(33)	49	0
Palm oil[a]	51	9	0.3	(30)	39	0
Coconut oil[a]	92	2	0	(2.0)	7	0
Chicken fat	31	21	1	(21)	47	11
Lard	41	11	1	(11)	47	12
Beef fat	52	3	1	(3.0)	44	14
Butter fat	66	2	2	(1.0)	30	33

[a]palm oil has arachidic of 0.2 and coconut oil has arachidic of 0.1.
Data on canola oil from data on file, Procter & Gamble. All other data from Reeves JB and Weihrauch JL (1979) *Composition of Foods, Agriculture Handbook No. 8-4*. Washington, DC: US Department of Agriculture.

derived only from direct ingestion or by synthesis from dietary EPA or ALA.

Evolutionary Aspects of Diet Relative to n-3 Fatty Acids and the n-6:n-3 Balance

On the basis of estimates from studies in Paleolithic nutrition and modern-day hunter-gatherer populations, it appears that human beings evolved consuming a diet that was much lower in saturated fatty acids than today's diet. Furthermore, the diet contained small and roughly equal amounts of n-6 and n-3 PUFAs (ratio of 1–2:1) and much lower amounts of *trans*-fatty acids than today's diet (**Figure 3**). The current Western diet is very high in n-6 fatty acids (the ratio of n-6 to n-3 fatty acids ranges between 10:1 and 30:1) because of the recommendation to substitute vegetable oils high in n-6 fatty acids for saturated fats to lower serum cholesterol concentrations. Furthermore, intake of n-3 fatty acids is much lower today because of the decrease in fish consumption and the industrial production of animal feeds rich in grains containing n-6 fatty acids, leading to production of meat rich in n-6 and poor in n-3 fatty acids. The same is true for cultured fish and eggs. Even cultivated vegetables contain fewer n-3 fatty acids than do plants in the wild. In summary, modern agriculture, with its emphasis on production, has decreased the n-3 fatty acid content in many foods: green leafy vegetables, animal meats, eggs, and even fish, while it has increased the amount of n-6 fatty acids in foods, leading to high n-6 intake for the first time in the history of human beings in many countries around the world (**Table 5**). The traditional diet of Crete (Greece) is consistent with the Paleolithic diet relative to the n-6:n-3 ratio. The Lyon Heart Study, which was based on a modified diet of Crete, had an n-6:n-3 ratio of 4:1 resulting in a 70% decrease in risk for cardiac death. As shown in **Table 6**, the higher the ratio of n-6 to n-3 fatty acids in platelet phospholipids, the higher the death rate from cardiovascular disease. As the ratio of n-6 PUFAs to n-3 PUFAs increases, the prevalence of type 2 diabetes also increases (**Figure 4**). As will be discussed below, a balance between the n-6 and n-3 fatty acids is a more physiologic state in terms of gene expression, eicosanoid metabolism, and cytokine production.

Further support for the need to balance the n-6:n-3 PUFAs comes from studies that clearly show the ability of both normal rat cardiomyocytes and human breast cancer cells in culture to form all the n-3 fatty acids from n-6 fatty acids when fed the cDNA encoding n-3 fatty acid desaturase obtained from the roundworm *Caenorhabditis elegans*. The n-3 desaturase efficiently and quickly converted the n-6 fatty acids that were fed to the cardiomyocytes in culture to the corresponding n-3 fatty acids. Thus, n-6 LA was converted to n-3 ALA and AA was converted to EPA, so that at equilibrium, the ratio of n-6 to n-3 PUFAs was close to 1:1. Further studies

Table 3 Terrestrial sources of n-3 (18:3n-3) fatty acids (grams per 100 g edible portion, raw)

Nuts and seeds	
Butternuts, dried	8.7
Walnuts, English/Persian	6.8
Chia seeds, dried	3.9
Walnuts, black	3.3
Beechnuts, dried	1.7
Soya bean kernels, roasted and toasted	1.5
Hickory nuts, dried	1.0
Oils	
Linseed oil	53.3
Rapeseed oil (canola)	11.1
Walnut oil	10.4
Wheat germ oil	6.9
Soya bean oil	6.8
Tomato seed oil	2.3
Rice bran oil	1.6
Vegetables	
Soya beans, green, raw	3.2
Soya beans, mature seeds, sprouted, cooked	2.1
Seaweed, Spirulina, dried	0.8
Radish seeds, sprouted, raw	0.7
Beans, navy, sprouted, cooked	0.3
Beans, pinto, sprouted, cooked	0.3
Kale, raw	0.2
Leeks, freeze-dried	0.2
Broccoli, raw	0.1
Cauliflower, raw	0.1
Lettuce, butterhead	0.1
Spinach, raw	0.1
Fruits	
Avocados, raw, California	0.1
Raspberries, raw	0.1
Strawberries	0.1
Legumes	
Soya beans, dry	1.6
Beans, common, dry	0.6
Cowpeas, dry	0.3
Lima beans, dry	0.2
Peas, garden, dry	0.2
Chickpeas, dry	0.1
Lentils, dry	0.1
Grains	
Oats, germ	1.4
Wheat, germ	0.7
Barley, bran	0.3
Corn, germ	0.3
Rice, bran	0.2
Wheat, bran	0.2
Wheat, hard red winter	0.1

Data from United States Department of Agriculture. Provisional table on the content of n-3 fatty acids and other fat components in selected foods from Simopoulos AP, Kifer RR, and Martin RE (eds.) (1986) *Health Effects of Polyunsaturated Fatty Acids in Seafoods*. Orlando, FL: Academic Press.

demonstrated that the cancer cells expressing the n-3 desaturase underwent apoptotic death whereas the control cancer cells with a high n-6:n-3 ratio continued to proliferate.

Eicosanoid Metabolism and Biological Effects of n-6 and n-3 Fatty Acids

When humans ingest fish or fish oil, the ingested EPA and DHA partially replace the n-6 fatty acids (especially AA) in cell membranes, particularly those of platelets, erythrocytes, neutrophils, monocytes, and liver cells.

Because of the increased amounts of n-6 fatty acids in the Western diet, the eicosanoid metabolic products from AA, specifically prostaglandins, thromboxanes, leukotrienes, hydroxy fatty acids, and lipoxins, are formed in larger quantities than those formed from n-3 fatty acids, specifically EPA. As a result (**Figure 5**), ingestion of EPA and DHA from fish or fish oil leads to: (1) decreased production of prostaglandin E2 metabolites; (2) decreased concentrations of thromboxane A2, a potent platelet aggregator and vasoconstrictor; (3) decreased formation of leukotriene B4, an inducer of inflammation and a powerful inducer of leukocyte chemotaxis and adherence; (4) increased concentrations of thromboxane A3, a weak platelet aggregator and vasoconstrictor; (5) increased concentrations of prostacyclin prostaglandin I3 (PGI3), leading to an overall increase in total prostacyclin by increasing PGI3 without decreasing PGI2 (both PGI2 and PGI3 are active vasodilators and inhibitors of platelet aggregation); and (6) increased concentrations of leukotriene B5, a weak inducer of inflammation and a chemotactic agent. The eicosanoids from AA are biologically active in small quantities and if they are formed in large amounts, they contribute to the formation of thrombi and atheromas; the development of allergic and inflammatory disorders, particularly in susceptible people; and cell proliferation. Thus, a diet rich in n-6 fatty acids shifts the physiologic state to one that is prothrombotic and proaggregatory, with increases in blood viscosity, vasospasm, and vasoconstriction and decreases in bleeding time. Bleeding time is shorter in groups of patients with hypercholesterolemia, hyperlipoproteinemia, myocardial infarction, other forms of atherosclerotic disease, type 2 diabetes, obesity, and hypertriglyceridemia. Atherosclerosis is a major complication in type 2 diabetes patients. Bleeding time is longer in women than in men and in younger than in older persons. There are ethnic differences in bleeding time that appear to be related to diet. The hypolipidemic, antithrombotic, anti-inflammatory, and anti-arrhythmic effects of n-3 fatty acids have been studied extensively in animal models, tissue cultures, and cells (**Table 7**).

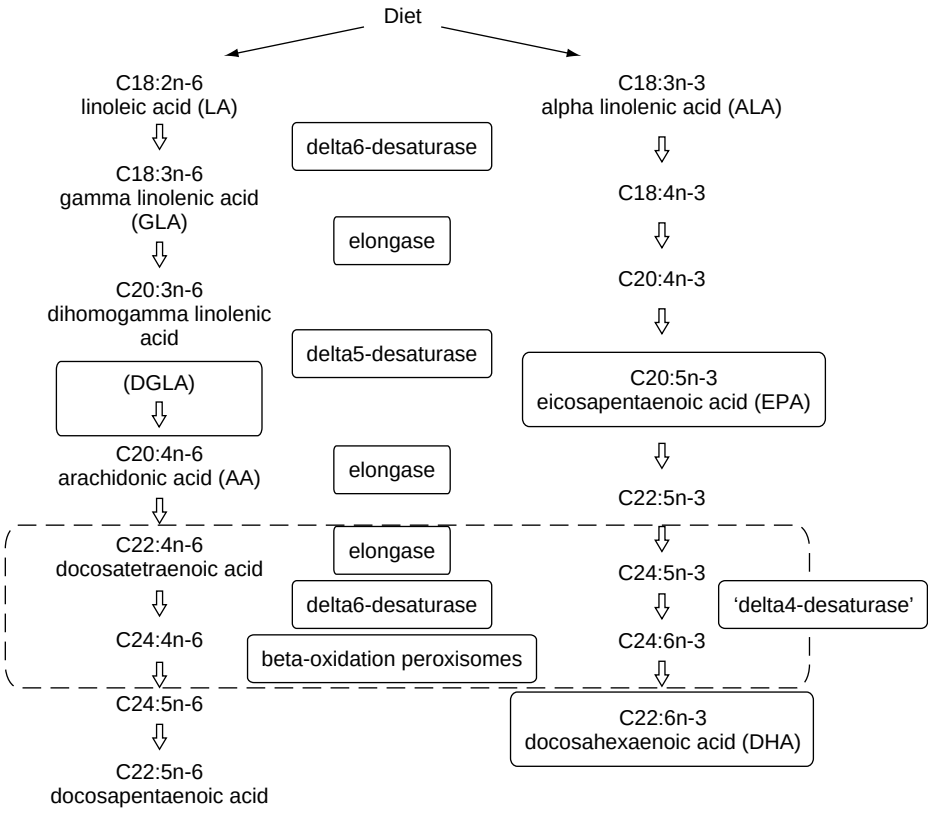


Figure 2 Essential fatty acid metabolism: desaturation and elongation of n-6 and n-3.

Nutrigenetics: Interaction between the n-6:n-3 Fatty Acids and the Genome

As expected, earlier studies focused on mechanisms that involve eicosanoid metabolites. More recently, however, the effects of fatty acids on gene expression have been investigated and this focus of interest has led to studies at the molecular level (**Tables 8**, **9**). Previous studies have shown that fatty acids, whether released from membrane phospholipids by cellular phospholipases or made available to the cell from the diet or other aspects of the extracellular environment, are important cell signaling molecules. They can act as second messengers or substitute for the classic second messengers of the inositide phospholipid and cyclic AMP signal transduction pathways. They can also act as modulator molecules mediating responses of the cell to extracellular signals. It has been shown that fatty acids rapidly and directly alter the transcription of specific genes.

5-Lipoxygenase and Atherosclerosis: An Example of Nutrigenetics/Nutrigenomics

Leukotrienes are eicosanoids derived through the action of 5-lipoxygenase (5-LO). It has been recently shown that genetic variants of the 5-LO promoter, already known to be associated with variable sensitivity to anti-asthmatic medications, also influence atherosclerosis. Variant genotypes of the 5-LO gene were found in 6% of a cohort of 470 healthy middle-aged men and women. Carotid intima-media thickness (IMT), taken as a marker of the atherosclerotic burden, was significantly increased, by 80% in the variant group compared to carriers of the common allele, suggesting increased 5-LO promoter activity associated with the mutant (variant) allele. Furthermore, dietary AA intake significantly enhanced the proatherogenic effect of 5-LO gene variants, while intake of EPA and DHA decreased (blunted) the effect of 5-LO and was associated with less IMT. EPA and DHA decrease the formation of leukotrienes of the 4-series by competing with AA (**Figure 5**) as substrates for 5-LO and generate weaker leukotrienes of the 5-series. The results of this study suggest that person with genetic variants are at higher risk for atherosclerosis at higher AA intake. It also suggests that the effects of EPA and DHA may be stronger in individuals with genetic variants associated with increased 5-LO activity. Therefore, clinical trials in the future should be controlled for genetic variation.

Table 4 Content of n-3 fatty acids and other fat components in selected fish (grams per 100 g edible portion, raw)

Fish	*Total fat*	*Fatty acids (g/100 g)*			*18:3*	*20:5*	*22:6*	*Cholesterol (mg/100 g)*
		Total saturated	*Total monounsaturated*	*Total polyunsaturated*				
Anchovy, European	4.8	1.3	1.2	1.6	–	0.5	0.9	–
Bass, striped	2.3	0.5	0.7	0.8	Tr	0.2	0.6	80
Bluefish	6.5	1.4	2.9	1.6	–	0.4	0.8	59
Carp	5.6	1.1	2.3	1.4	0.3	0.2	0.1	67
Catfish, brown Bullhead	2.7	0.6	1.0	0.8	0.1	0.2	0.2	75
Catfish, channel	4.3	1.0	1.6	1.0	Tr	0.1	0.2	58
Cod, Atlantic	0.7	0.1	0.1	0.3	Tr	0.1	0.2	43
Croaker, Atlantic	3.2	1.1	1.2	0.5	Tr	0.1	0.1	61
Flounder, unspecified	1.0	0.2	0.3	0.3	Tr	0.1	0.1	46
Grouper, red	0.8	0.2	0.1	0.2	–	Tr	0.2	–
Haddock	0.7	0.1	0.1	0.2	Tr	0.1	0.1	63
Halibut, Greenland	13.8	2.4	8.4	1.4	Tr	0.5	0.4	46
Halibut, Pacific	2.3	0.3	0.8	0.7	0.1	0.1	0.3	32
Herring, Pacific	13.9	3.3	6.9	2.4	0.1	1.0	0.7	77
Herring, round	4.4	1.3	0.8	1.5	0.1	0.4	0.8	28
Mackerel, king	13.0	2.5	5.9	3.2	–	1.0	1.2	53
Mullet, striped	3.7	1.2	1.1	1.1	0.1	0.3	0.2	49
Ocean perch	1.6	0.3	0.6	0.5	Tr	0.1	0.1	42
Plaice, European	1.5	0.3	0.5	0.4	Tr	0.1	0.1	70
Pollock	1.0	0.1	0.1	0.5	–	0.1	0.4	71
Pompano, Florida	9.5	3.5	2.6	1.1	–	0.2	0.4	50
Salmon, Chinook	10.4	2.5	4.5	2.1	0.1	0.8	0.6	–
Salmon, pink	3.4	0.6	0.9	1.4	Tr	0.4	0.6	–
Snapper, red	1.2	0.2	0.2	0.4	Tr	Tr	0.2	–
Sole, European	1.2	0.3	0.4	0.2	Tr	Tr	0.1	50
Swordfish	2.1	0.6	0.8	0.2	–	0.1	0.1	39
Trout, rainbow	3.4	0.6	1.0	1.2	0.1	0.1	0.4	57
Tuna, albacore	4.9	1.2	1.2	1.8	0.2	0.3	1.0	54
Tuna, unspecified	2.5	0.9	0.6	0.5	–	0.1	0.4	–

Dashes denote lack of reliable data for nutrient known to be present; Tr, trace (<0.05 g/100 g food). Adapted from the United States Department of Agriculture Provisional Table on the Content of Omega-3 Fatty Acids and Other Fat Components in Seafoods as presented by Simopoulos AP, Kifer RR, and Martin RE (eds.) (1986) *Health Effects of Polyunsaturated Fatty Acids in Seafoods*. Orlando, FL: Academic Press.

Effects of Dietary ALA Compared with Long-Chain n-3 Fatty Acid Derivatives on Physiologic Indexes

Several clinical and epidemiologic studies have been conducted to determine the effects of long-chain n-3 PUFAs on various physiologic indexes. Whereas the earlier studies were conducted with large doses of fish or fish oil concentrates, more recent studies have used lower doses. ALA, the precursor of n-3 fatty acids, can be converted to long-chain n-3 PUFAs and can therefore be substituted for fish oils. The minimum intake of long-chain n-3 PUFAs needed for beneficial effects depends on the intake of other fatty acids. Dietary amounts of LA as well as the ratio of LA to ALA appear to be important for the metabolism of ALA to long-chain n-3 PUFAs. While keeping the amount of dietary LA constant (3.7 g) ALA appears to have biological effects similar to those of 0.3 g long-chain n-3 PUFAs with conversion of 11 g ALA to 1 g long-chain n-3 PUFAs. Thus, a ratio of 4 (15 g LA:3.7 g ALA) is appropriate for conversion. In human studies, the conversion of deuterated ALA to longer chain metabolites was reduced by ≅50% when dietary intake of LA was increased from 4.7% to 9.3% of energy as a result of the known competition between n-6 and n-3 fatty acids for desaturation. After ALA supplementation there is an increase in long-chain n-3 PUFAs in plasma and platelet phospholipids and a decrease in platelet aggregation. ALA supplementation does not alter triacylglycerol concentrations. Only long-chain n-3 PUFA have triacylglycerol-lowering effects. Supplementation with ALA to lower the n-6:n-3 ratio from 13:1 to 1:1 led to a 50% reduction in C-reactive protein (CRP), a risk factor for coronary heart disease.

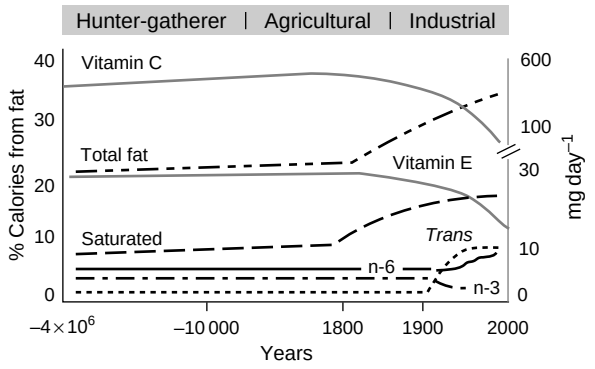


Figure 3 Hypothetical scheme of fat, fatty acid (n-6, n-3, *trans* and total) intake (as per cent of calories from fat) and intake of vitamins E and C ($mg\,day^{-1}$). Data were extrapolated from cross-sectional analyses of contemporary hunter-gatherer populations and from longitudinal observations and their putative changes during the preceding 100 years. *Trans*-fatty acids, the result of the hydrogenation process, have increased dramatically in the food supply during this century. (Reproduced with permission from Simopoulos AP (1999) Genetic variation and evolutionary aspects of diet. In: Papas A (ed.) *Antioxidants in Nutrition and Health*, pp. 65–88. Boca Raton: CRC Press.)

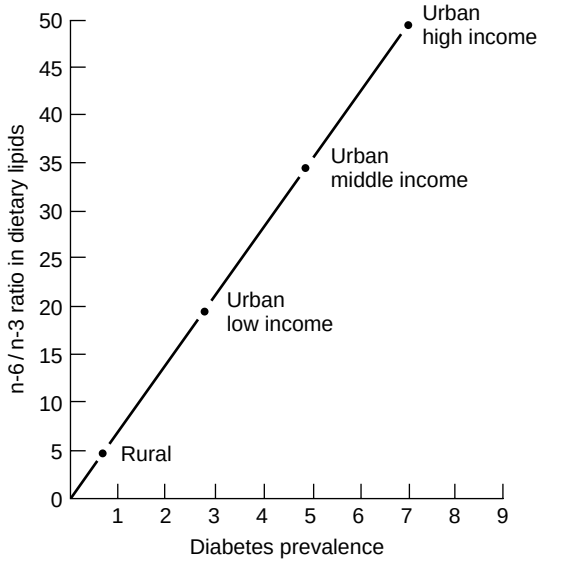


Figure 4 Relation between the ratio of n-6 to n-3 fatty acids in dietary lipids in the Indian diet and the prevalence of type 2 diabetes. (Reproduced with permission from Raheja BS, Sadikot SM, Phatak RB, and Rao MB (1993) Significance of the n-6/n-3 ratio for insulin action in diabetes. *Annals of the NewYork Academy of Science* **683**: 258–271.)

Table 5 n-6:n-3 ratios in various populations

Population	*n-6:n-3*
Paleolithic	0.79
Greece prior to 1960	1.00–2.00
Current Japan	4.00
Current India, rural	5–6.1
Current UK and northern Europe	15.00
Current US	16.74
Current India, urban	38–50

Reproduced with permission from Simopoulos AP (2003) Importance of the ratio of omega-6/omega-3 essential fatty acids: Evolutionary aspects. *World Review of Nutrition and Diet* **92**: 1–22.

Table 6 Ethnic differences in fatty acid concentrations in thrombocyte phospholipids and percentage of all deaths from cardiovascular disease

	Europe and US	*Japan*	*Greenland Eskimos*
Arachidonic acid (20:4n-6)	26%	21%	8.3%
Eicosapentaenoic acid (20:5n-3)	0.5%	1.6%	8.0%
Ratio of n-6:n-3	50%	12%	1%
Mortality from cardiovascular disease	45%	12%	7%

Modified from Weber PC (1989) Are we what we eat? Fatty acids in nutrition and in cell membranes: cell functions and disorders induced by dietary conditions. In: *Fish, Fats and your Health*, Report no. 4, pp. 9–18. Norway: Svanoybukt Foundation.

In Australian studies, ventricular fibrillation in rats was reduced with canola oil as much or even more efficiently than with fish oil, an effect attributable to ALA. Further studies should be able to show whether this result is a direct effect of ALA per se or whether it occurs as a result of its desaturation and elongation to EPA and possibly DHA.

The diets of Western countries have contained increasingly larger amounts of LA, which has been promoted for its cholesterol-lowering effect. It is now recognized that dietary LA favors oxidative modification of low-density lipoprotein (LDL) cholesterol, increases platelet response to aggregation, and suppresses the immune system. In contrast, ALA intake is associated with inhibitory effects on the clotting activity of platelets, on their response to thrombin, and on the regulation of AA metabolism. In clinical studies, ALA contributed to lowering of blood pressure. In a prospective study, ALA was inversely related to the risk of coronary heart disease in men.

ALA is not equivalent in its biological effects to the long-chain n-3 fatty acids found in fish oils. EPA and DHA are more rapidly incorporated into plasma and membrane lipids and produce more rapid effects than does ALA. Relatively large reserves of LA in body fat, as are found in vegans or in the diet of omnivores in Western societies, would tend to slow down the formation of long-chain n-3 fatty acids from ALA. Therefore, the role of ALA in human nutrition becomes important in terms of long-term

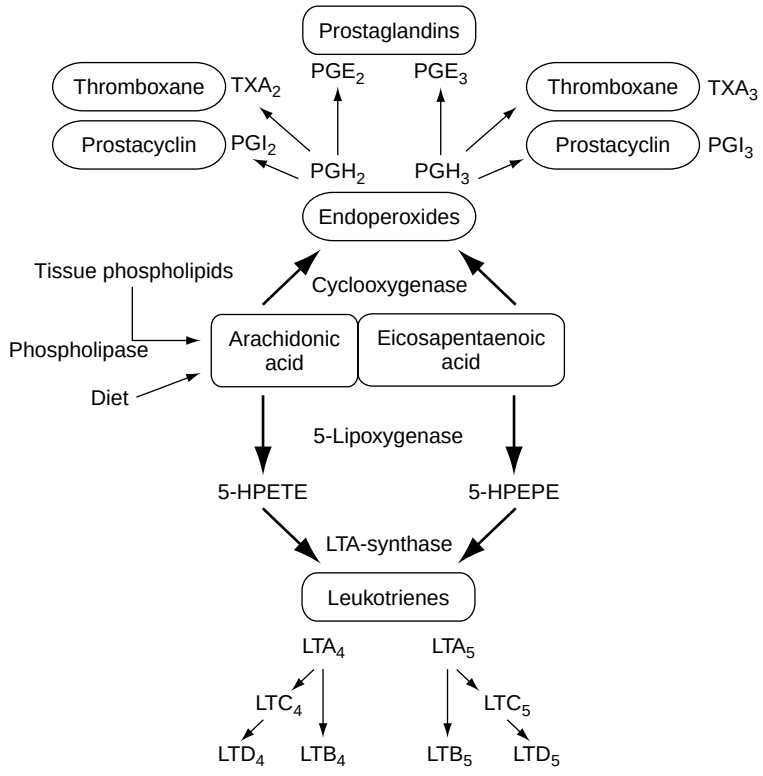


Figure 5 Oxidative metabolism of arachidonic acid and eicosapentaenoic acid by the cyclooxygenase and 5-lipoxygenase pathways. 5-HPETE denotes 5-hydroperoxyeicosatetranoic acid and 5-HPEPE denotes 5-hydroxyeicosapentaenoic acid.

dietary intake. One advantage of the consumption of ALA over n-3 fatty acids from fish is that the problem of insufficient vitamin E intake does not exist with a high intake of ALA from plant sources.

Human Studies in Growth and Development

Pregnancy and Fetal Growth

Since World War II, the role of maternal nutrition in fetal growth and development has been extensively studied in the context of protein-calorie malnutrition. The role of n-3 fatty acids has only recently come into focus, despite the evidence of its importance having been demonstrated in a series of studies between 1928 and 1930 involving rats and primates. Lipid nutrition during pregnancy and lactation is of special relevance to human development, because brain development in the human takes place during fetal life and in the first 2 years after birth. DHA is found in larger amounts in the gray matter of the brain and in the retinal membranes, where it accounts for 30% or more of the fatty acids in the ethanolamine and serine phospholipid. DHA accumulates in the neurons of the brain between weeks 26 and 40 of gestation in humans.

During the third trimester of human development, rapid synthesis of brain tissue occurs in association with increasing neuromotor activity. The increase in cell size, number, and type requires de novo synthesis of structural lipids, leading to accumulation of DHA in the brain of the human infant during the last trimester. The levels of ALA and LA are low in the brain, whereas marked accretion of long-chain desaturation products, specifically DHA and AA, occurs. More recent data indicate that the main developmental changes in the brain seem to be an increase in DHA at the end of gestation and a decrease in oleic acid (18:1n-9) and AA in phosphatidlyethanolamine (PE). Similar changes occur in the liver. Therefore, a premature infant (prior to 37 weeks' gestation) has much lower amounts of DHA in the brain and liver and is at risk of becoming deficient in DHA unless it is supplied in the diet. In the full-term newborn, about half of the DHA accumulates in the brain before birth and the other half after birth.

There is epidemiologic evidence that the birth weights of newborns in the Faroe Islands (where fish intake is high) are higher than those in Denmark, as is the length of gestation: 40.3 ± 1.7 weeks for the Faroese versus 39.7 ± 1.8 weeks for the Danish pregnant women. The average birth weight of primiparas is 194 g higher for the Faroe Islands. The higher dietary n-3 fatty acid intake quite possibly influences endogenous prostaglandin metabolism. It is

Table 7 Effects of n-3 fatty acids on factors involved in the pathophysiology of atherosclerosis and inflammation

Factor	*Function*	*Effect of n-3 fatty acid*
Arachidonic acid	Eicosanoid precursor; aggregates platelets; stimulates white blood cells	↓
Thromboxane A_2	Platelet aggregation; vasoconstriction; increase of intracellular Ca^{++}	↓
Prostacyclin ($PGI_{2/3}$)	Prevent platelet aggregation; vasodilation; increase cAMP	↑
Leukotriene (LTB_4)	Neutrophil chemoattractant; increase of intracellular Ca^{++}	↓
Fibrinogen	A member of the acute phase response; and a blood clotting factor	↓
Tissue plasminogen activator	Increase endogenous fibrinolysis	↑
Platelet activating factor (PAF)	Activates platelets and white blood cells	↓
Platelet-derived growth factor (PDGF)	Chemoattractant and mitogen for smooth muscles and macrophages	↓
Oxygen free radicals	Cellular damage; enhance LDL uptake via scavenger pathway; stimulate arachidonic acid metabolism	↓
Lipid hydroperoxides	Stimulate eicosanoid formation	↓
Interleukin 1 and tumor necrosis factor	Stimulate neutrophil O_2 free radical formation; stimulate lymphocyte proliferation; stimulate PAF; express intercellular adhesion molecule-1 on endothelial cells; inhibit plasminogen activator, thus, procoagulants	↓
Interleukin-6	Stimulates the synthesis of all phase proteins involved in the inflammatory response: C-reative protein; serum amyloid A; fibrinogen; α_1-chymotrypsin; and haptoglobin	↓
C-reactive protein (CRP)	An acute phase reactant and an independent risk factor for cardiovascular disease	↓
Endothelial-derived relaxation factor	Reduces arterial vasoconstrictor response	↑
Insulin sensitivity		↑
VLDL		↓
HDL	Decreases the risk for coronary heart disease	↑
Lp(a)	Lipoprotein(a) is a genetically determined protein that has atherogenic and thrombogenic properties	↓
Triglycerides and chylomicrons	Contribute to postprandial lipemia	↓

Source: Updated and modified from Weber PC, Leaf A. Cardiovascular effects of omega-3 fatty acids. Atherosclerosis risk factor modification by omega-3 fatty acids. World Rev Nutr Diet 1991, **66**: 218–32. With permission.

hypothesized that the dietary n-3 fatty acids inhibit the production of the dienoic prostaglandins, especially PGF_{2a} and PGE_2, because they are involved in the mediation of uterine contractions and the ripening of the cervix that lead to labor and delivery. These important observations need to be further investigated, as the prevention of prematurity is one of the most critical issues to be overcome in perinatal medicine.

Human Milk and Infant Feeding

A number of studies from around the world indicate that human milk contains both LNA and LA and their long-chain n-6 and n-3 fatty acids, whereas cow's milk does not. The long-chain fatty acid composition of red blood cell membrane phospholipids may reflect the composition of phospholipids in the brain. Therefore, determination of red blood cell membrane phospholipids has been carried out by many investigators to determine the long-chain PUFA content in breast-fed and bottle-fed infants. As expected, the fatty acids 22:5n-3 and 22:6n-3 were higher in the erythrocytes from breast-fed infants than those from bottle-fed babies and the 20:3n-9 was lower in the erythrocytes of the breast-fed infants.

Following birth, the amount of red blood cell DHA in premature infants decreases; therefore the amount of DHA available to the premature infant assumes critical importance. Preterm infants have a limited ability to convert LNA to DHA (**Figure 2**); therefore, a number of studies have been carried out on the DHA status of the premature infant. Premature babies have decreased amounts of DHA, but human milk contains enough DHA to support normal growth of the premature baby. The amount of n-3 fatty acids in human milk varies with the mother's diet; in particular, DHA is lower in vegetarians than in omnivores. One can increase the amount of DHA in human milk by giving fish oil rich in DHA to the mother.

The need to supplement infant formula with n-3 fatty acids and, particularly, DHA for the premature is now recognized and many countries have licensed infant formula enriched with n-3 fatty acids. DHA is essential for normal visual function and visual

Table 8 Effects of polyunsaturated fatty acids on several genes encoding enzyme proteins involved in lipogenesis, glycolysis, and glucose transport

Function and gene	*Linoleic acid*	*α-Linolenic acid*	*Arachidonic acid*	*Eicosapentaenoic acid*	*Docosahexaenoic acid*
Hepatic cells					
Lipogenesis					
FAS	↓	↓	↓	↓	↓
S14	↓	↓	↓	↓	↓
SCD1	↓	↓	↓	↓	↓
SCD2	↓	↓	↓	↓	↓
ACC	↓	↓	↓	↓	↓
ME	↓	↓	↓	↓	↓
Glycolysis					
G6PD	↓				
GK	↓	↓	↓	↓	↓
PK	—	↓	↓	↓	↓
Mature adiposites					
Glucose transport					
GLUT4	—	—	↓	↓	—
GLUT1	—	—	↑	↑	—

↓ = Suppress or decrease; ↑ = induce or increase

Source: Modified from Simopoulos AP. The role of fatty acids in gene expression: Health implications. Ann Nutr Metab 1996, **40**: 303–311. With permission.

maturation, particularly of the premature infant. Studies are currently in progress comparing the growth and development of both the premature and full-term infant who are fed mother's milk with those who are receiving formula supplemented with n-3 fatty acids and those whose formula is not supplemented, to define precisely the effects of DHA on intelligence quotient (IQ) and overall neuromotor development. **Tables 10** and **11** show the EFA dietary recommendations for adults, pregnant women, and infants made by a scientific group at a workshop held at the National Institutes of Health in Bethesda, Maryland in 1999.

Aging

ALA deficiency has been found in patients on long-term gastric tube-feeding that included large amounts of skim milk without ALA supplementation. These patients, who were in nursing homes,

Table 9 Effects of polyunsaturated fatty acids on several genes encoding enzyme proteins involved in cell growth, early gene expression, adhesion molecules, inflammation, β-oxidation, and growth factors[a]

Function and gene	*Linoleic acid*	*α-Linolenic acid*	*Arachidonic acid*	*Eicosapentaenoic acid*	*Docosahexaenoic acid*
Cell growth and early gene expression					
c-fos	—	—	↑	↓	↓
Egr-1	—	—	↑	↓	↓
Adhesion molecules					
VCAM-1 mRNA[b]	—	—	↓	[c]	↓
Inflammation					
IL-1β	—	—	↑	↓	↓
β-oxidation					
Acyl-CoA oxidase[d]	↑	↑	↑	↑	↑↑
Growth factors					
PDGF	—	—	↑	↓	↓

[a]VCAM, vascular cell adhesion molecule; IL, interleukin: PDGF, platelet-derived growth factor. ↓ suppresses or decreases, ↑ induces or increases.

[b]Monounsaturated fatty acids (MONOs) also suppress VCAM1 mRNA, but to a lesser degree than does DHA. AA also suppresses to a lesser extent than DHA.

[c]Eicosapentachoic acid has no effect by itself but enhances the effect of docosahexachoic acid (DHA)

[d]MONOs also induce acyl-CoA oxidase mRNA

Source: Modified from Simopoulos AP. The role of fatty acids in gene expression: Health implications. Ann Nutr Metab 1996, **40**: 303–311. With permission.

Table 10 Adequate intake (AI) for adults

Fatty acid	*Grams/day (2000 kcal diet)*	*% Energy*
LA	4.44	2.0
(upper limit)[a]	6.67	3.0
ALA	2.22	1.0
DHA + EPA	0.65	0.3
DHA to be at least[b]	0.22	0.1
EPA to be at least	0.22	0.1
TRANS-FA		
(upper limit)[c]	2.00	1.0
SAT		
(upper limit)[d]	–	<8.0
MONOs[e]	–	–

[a]Although the recommendation is for AI, the Working Group felt that there is enough scientific evidence to also state an upper limit (UL) for LA of 6.67 g day^{-1} based on a 2000 kcal diet or of 3.0% of energy.
[b]For pregnant and lactating women, ensure 300 mg day^{-1} of DHA.
[c]Except for dairy products, other foods under natural conditions do not contain *trans*-FA. Therefore, the Working Group does not recommend *trans*-FA to be in the food supply as a result of hydrogenation of unsaturated fatty acids or high-temperature cooking (reused frying oils).
[d]Saturated fats should not comprise more than 8% of energy.
[e]The Working Group recommended that the majority of fatty acids are obtained from monounsaturates. The total amount of fat in the diet is determined by the culture and dietary habits of people around the world (total fat ranges from 15% to 40% of energy) but with special attention to the importance of weight control and reduction of obesity.
If sufficient scientific evidence is not available to calculate an estimated average requirement, a reference intake called an adequate intake is used instead of a recommended dietary allowance. The AI is a value based on experimentally derived intake levels or approximations of observed mean nutrient intakes by a group (or groups) of healthy people. The AI for children and adults is expected to meet or exceed the amount needed to maintain a defined nutritional state or criterion of adequacy in essentially all members of a specific healthy population.
LA, linoleic acid; ALA, α-linolenic acid; DHA, docosahexaenoic acid; EPA, eicosapentaenoic acid; *TRANS*-FA, *trans*-fatty acids; SAT, saturated fatty acids; MONOs, monounsaturated fatty acids.
Reproduced with permission from Simopoulos AP, Leaf A, and Salem N Jr (1999) Essentiality of and recommended dietary intakes for omega-6 and omega-3 fatty acids. *Annals of Nutrition and Metabolism* **43**: 127–130.

Table 11 Adequate intake (AI) for infant formula/diet

Fatty acid	*Per cent of fatty acids*
LA[a]	10.00
ALA	1.50
AA[b]	0.50
DHA	0.35
EPA[c]	
(upper limit)	<0.10

[a]The Working Group recognizes that in countries like Japan the breast milk content of LA is 6–10% of fatty acids and the DHA is higher, about 0.6%. The formula/diet composition described here is patterned on infant formula studies in Western countries.
[b]The Working Group endorsed the addition of the principal long-chain polyunsaturates, AA and DHA, to all infant formulas.
[c]EPA is a natural constituent of breast milk, but in amounts more than 0.1% in infant formula may antagonize AA and interfere with infant growth.
If sufficient scientific evidence is not available to calculate an estimated average requirement, a reference intake called an adequate intake is used instead of a recommended dietary allowance. The AI is a value based on experimentally derived intake levels or approximations of observed mean nutrient intakes by a group (or groups) of healthy people. The AI for children and adults is expected to meet or exceed the amount needed to maintain a defined nutritional state or criterion of adequacy in essentially all members of a specific healthy population.
LA, linoleic acid; ALA, α-linolenic acid; AA, arachidonic acid; DHA, docosahexaenoic acid; EPA, eicosapentaenoic acid; *TRANS*-FA, *trans*-fatty acids; SAT, saturated fatty acids; MONOs, monounsaturated fatty acids.
Reproduced with permission from Simopoulos AP, Leaf A, and Salem N Jr (1999) Essentiality of and recommended dietary intakes for omega-6 and omega-3 fatty acids. *Annals of Nutrition and Metabolism* **43**: 127–130.

developed skin lesions diagnosed as scaly dermatitis, which disappeared with ALA supplementation. A number of other patients were reported to have n-3 fatty acid deficiency, again patients on long-term gastric tube-feeding or prolonged total parenteral nutrition because of chronic illnesses. If a deficiency of total n-3 fatty acid intake is suspected, its concentration in plasma should be measured. A decrease in the concentration of 20:5n-3, 22:5n-3, and particularly 22:6n-3 in plasma or erythrocyte phospholipids indicates that the dietary intake of n-3 fatty acids has been low. The presence of clinical symptoms, along with the biochemical determinations, provides additional support for the diagnosis. To verify the diagnosis, it is essential that the clinical symptoms disappear upon supplementation of the deficient diet with n-3 fatty acids.

With the increase in the number of elderly persons in the population, and the proliferation of nursing homes, particular attention must be given to the nutritional requirements of the elderly, especially those who are fed enterally or parenterally.

Coronary Heart Disease

Most epidemiologic studies and clinical trials using n-3 fatty acids in the form of fish or fish oil have been carried out in patients with coronary heart disease. However, studies have also been carried out on the effects of ALA in normal subjects and in patients with myocardial infarction.

The hypolipidemic effects of n-3 fatty acids are similar to those of n-6 fatty acids, provided that they replace saturated fats in the diet. n-3 fatty acids have the added benefit of not lowering high-density

lipoprotein (HDL) and consistently lowering serum triacylglycerol concentrations, whereas the n-6 fatty acids do not and may even increase triglyceride levels.

Another important consideration is the finding that during chronic fish oil feeding postprandial triacylglycerol concentrations decrease. Furthermore, consumption of high amounts of fish oil blunted the expected rise in plasma cholesterol concentrations in humans. These findings are consistent with the low rate of coronary heart disease found in fish-eating populations. Studies in humans have shown that fish oils reduce the rate of hepatic secretion of very low-density lipoprotein (VLDL) triacylglycerol. In normolipidemic subjects, n-3 fatty acids prevent and rapidly reverse carbohydrate-induced hypertriglyceridemia. There is also evidence from kinetic studies that fish oil increases the fractional catabolic rate of VLDL (**Table 7**).

The effects of different doses of fish oil on thrombosis and bleeding time have been investigated. A dose of 1.8 g EPA day^{-1} did not result in any prolongation in bleeding time, but 4 g day^{-1} increased bleeding time and decreased platelet count with no adverse effects. In human studies, there has never been a case of clinical bleeding, even in patients undergoing angioplasty, while the patients were taking fish oil supplements. Clinical investigations indicate that n-3 fatty acids prevent sudden death. A series of intervention trials have clearly shown that the addition of n-3 fatty acids in the form of fish oil (EPA and DHA) decrease the death rate in the secondary prevention of coronary heart disease by preventing ventricular arrhythmias that lead to sudden death.

Antiarrhythmic Effects of n-3 Fatty Acids (ALA, EPA, and DHA)

Studies have shown that n-3 fatty acids, more so than n-6 PUFA, can prevent ischemia-induced fatal ventricular arrhythmias in experimental animals. n-3 fatty acids make the heart cells less excitable by modulating the conductance of the sodium and other ion channels. Clinical studies further support the role of n-3 fatty acids in the prevention of sudden death due to ventricular arrhythmias which, in the US, account for 50–60% of the mortality from acute myocardial infarction and cause 250 000 deaths a year. In the intervention trials, there was no change in lipid concentration, suggesting that the beneficial effects of n-3 fatty acids were due to their antithrombotic and antiarrhythmic effects.

The antiarrhythmic effects of n-3 fatty acids are supported by clinical intervention trials (Diet and Reinfarction Trial (DART), Lyon Heart Study, Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto miocardico (GISSI)-Prevenzione Trial, Indo-Mediterranean Diet Heart Study). Their results strongly support the role of fish or fish oil in decreasing total mortality and sudden death in patients with one episode of myocardial infarction. Therefore, the addition of 1 g/d of n-3 fatty acids is highly recommended for the primary and secondary prevention of coronary heart disease.

Inflammation: a Common Base for the Development of Coronary Heart Disease, Diabetes, Arthritis, Mental Health, Neurodegenerative Diseases and Cancer

Anti-inflammatory Aspects of n-3 Fatty Acids

Many experimental studies have provided evidence that incorporation of alternative fatty acids into tissues may modify inflammatory and immune reactions and that n-3 fatty acids in particular are potent therapeutic agents for inflammatory diseases. Supplementing the diet with n-3 fatty acids (3.2 g EPA and 2.2 g DHA) in normal subjects increased the EPA content in neutrophils and monocytes more than sevenfold without changing the quantities of AA and DHA. The anti-inflammatory effects of fish oils are partly mediated by inhibiting the 5-lipoxygenase pathway in neutrophils and monocytes and inhibiting the leukotriene B_4 (LTB_4)-mediated function of LTB_5 (**Figure 5**). Studies show that n-3 fatty acids influence interleukin metabolism by decreasing IL-1β and IL-6. Inflammation plays an important role in both the initiation of atherosclerosis and the development of atherothrombotic events. An early step in the atherosclerotic process is the adhesion of monocytes to endothelial cells. Adhesion is mediated by leukocyte and vascular cell adhesion molecules (CAMs) such as selectins, integrins, vascular cell adhesion molecule 1 (VCAM-1), and intercellular adhesion molecule 1 (ICAM-1). The expression of E-selectin, ICAM-1, and VCAM-1, which is relatively low in normal vascular cells, is upregulated in the presence of various stimuli, including cytokines and oxidants. This increased expression promotes the adhesion of monocytes to the vessel wall. The monocytes subsequently migrate across the endothelium into the vascular intima, where they accumulate to form the initial lesions of atherosclerosis. Atherosclerosis plaques have been shown to have increased CAM expression in animal models and human studies.

Diabetes is a major risk factor for coronary heart disease. EPA and DHA increase sensitivity to insulin and decrease the risk of coronary heart disease. Rheumatoid arthritis has a strong inflammatory component characterized by an increase in

interleukin (IL)-1β. n-3 fatty acids decrease IL-1β as well as the number of swollen and painful joints. Supplementation with EPA and DHA, changing the ratio of n-6:n-3 of the background diet by increasing the n-3 and decreasing the n-6 intake, is now standard treatment for patients with rheumatoid arthritis along with medication in a number of centers around the world. Similarly, changing the background diet in patients with asthma has led to decreases in the dose of nonsteroidal anti-inflammatory drugs.

These studies suggest the potential for complementarity between drug therapy and dietary choices and that increased intake of n-3 fatty acids and decreased intake of n-6 fatty acids may lead to drug sparing effects. Therefore, future studies need to address the fatty acid composition and the ratio of n-6:n-3 of the background diet, and the issue of concurrent drug use. A diet rich in n-3 fatty acids and low in n-6 fatty acids provides the appropriate background biochemical environment in which drugs function.

Psychologic stress in humans induces the production of proinflammatory cytokines such as interferon gamma (IFNγ), TNFα, IL-6, and IL-10. An imbalance of n-6 and n-3 PUFA in the peripheral blood causes an overproduction of proinflammatory cytokines. There is evidence that changes in fatty acid composition are involved in the pathophysiology of major depression. Changes in serotonin (5-HT) receptor number and function caused by changes in PUFAs provide the theoretical rationale connecting fatty acids with the current receptor and neurotransmitter theories of depression. The increased 20:4n-6/20:5n-3 ratio and the imbalance in the n-6:n-3 PUFA ratio in major depression may be related to the increased production of proinflammatory cytokines and eicosanoids in that illness. Studies have shown that EPA and DHA prolong remission, that is, reduce the risk of relapse in patients with bipolar disorder. There are a number of studies evaluating the therapeutic effect of EPA and DHA in major depression.

Earlier studies in rodents showed that ALA intake improved learning, memory and cognition. In Zellweger's syndrome (a genetic neurodegenerative disease) high amounts of DHA early in life decreased somewhat the rate of progression of the disease. A number of studies have suggested that people who eat a diet rich in fish are less likely to develop Alzheimer's disease. Learning and memory depend on dendritic spine action assembly and DHA. High DHA consumption is associated with reduced risk for Alzheimer's disease, yet mechanisms and therapeutic potential remain elusive. In an Alzheimer's disease mouse model, reduction of dietaty n-3 fatty acid resulted in 80%-90% losses of the p85 alpha subunit of phosphoinositol 3-kinase and the postsynaptic action-regulating protein drebrin as in the brain of patients with Alzheimer's disease. The loss of postsynaptic proteins was associated with increased oxidation without concomitant neuron or presynaptic protein loss. Treatment of the n-3 fatty acid restricted mice with DHA protected against these effects and behavioral deficits. Since n-3 fatty acids are essential for p85-mediated central nervous system insulin signaling and selective protection of postsynaptic proteins, these findings have implications for neurodegenerative diseases, where synaptic loss is critical, especially in Alzheimer's disease. A few case control studies suggest that higher EPA and DHA intake is associated with lower risk of Alzheimer's disease and severity of the disease. Inflammation is a risk factor for Alzheimer's disease. It remains to be determined whether low n-3 fatty acids, especially low DHA status, in patients with Alzheimer's disease is a causal factor in the pathogenesis and progression of Alzheimer's disease and other neurodegenerative diseases.

Cancer is characterized by inflammation, cell proliferation, and elevated IL-6 levels. Since EPA and DHA suppress IL-6, fish oil supplementation suppresses rectal epithelial cell proliferation and PGE_2 biosynthesis. This was achieved with a dietary n-6:n-3 ratio of 2.5:1, but not with the same absolute level of fish oil intake and an n-6:n-3 ratio of 4:1. Case control studies in women with breast cancer support the hypothesis that the balance between n-6 and n-3 in breast adipose tissue plays an important role in breast cancer and in breast cancer metastasis.

Future Work, Conclusions, and Recommendations

n-3 fatty acids should be added to foods rather than be used solely as dietary supplements, which is a quasi-pharmaceutical approach. Furthermore, the development of a variety of n-3-rich foodstuffs would allow increased n-3 dietary intakes with little change of dietary habits. n-3 fatty acids maintain their preventative and therapeutic properties when packaged in foods other than fish. Efficient use of dietary n-3 fatty acids will require the simultaneous reduction in the food content of n-6 fatty acids and their substitution with monounsaturated oils. Dietary n-3 fats give rise to higher tissue levels of EPA when the 'background' diet is low in n-6 fats. Compared to n-6 fatty acids, olive oil increases the incorporation of n-3 fatty acids into tissues.

In the past, industry focused on improvements in food production and processing to increase shelf life of the products, whereas now and in the future the focus will be on nutritional quality in product

development. This will necessitate the development of research for the nutritional evaluation of the various food products and educational programs for professionals and the public. The definition of food safety will have to expand in order to include nutrient structural changes and food composition. The dawn of the twenty-first century will enhance the scientific base for product development and expand collaboration among agricultural, nutritional, and medical scientists in government, academia, and industry. This should bring about a greater involvement of nutritionists and dieticians in industrial research and development to respond to an ever-increasing consumer interest in the health attributes of foods.

Today, more is known about the mechanisms and functions of n-3 fatty acids than other fatty acids. It is evident that Western diets are relatively deficient in n-3 fatty acids and that they contain much higher amounts of n-6 fatty acids than ever before in the evolution of human beings. Research has shown that DHA is essential for the development of the premature infant relative to visual acuity, visual function, and maturation. In the full-term infant, DHA may influence visual acuity and neural pathways associated with the developmental progression of language acquisition. These findings have led to the inclusion of DHA and AA in infant formulas in most countries around the world.

Most of the research on the role of n-3 fatty acids in chronic diseases has been carried out in patients with coronary heart disease. Intervention trials have clearly shown that n-3 fatty acids decrease sudden death and all cause mortality in the secondary prevention of coronary heart disease and in one study also in the primary prevention of coronary heart disease. The decrease in sudden death is most likely due to the anti-arrhythmic effects of n-3 fatty acids.

Most recent research suggests that the response to n-3 fatty acids may be genotype dependent, since certain individuals respond more than others. The time has come to take genetic variation into consideration when setting up clinical intervention trials. We need to move away from the long-term prospective studies and proceed with genotype-specific clinical intervention trials.

Inflammation and cell proliferation are at the base of many chronic diseases and conditions, especially atherosclerosis and cancer, but also diabetes, hypertension, arthritis, mental health, and various autoimmune diseases. Individuals carrying genetic variants for these conditions are much more prone to develop them because the high n-6:n-3 ratio leads to proinflammatory and prothrombotic states.

The time has come to return to high n-3 fatty acid levels in the diet and to decrease the n-6 intake. There is good scientific evidence from studies on the Paleolithic diet, the diet of Crete, other traditional diets (Okinawa), intervention studies, and finally studies at the molecular level using transgenic rodents that the physiologic n-6:n-3 ratio should be 1:1 or 2:1. Japan has already recommended a ratio of 2:1. Industry has moved in the direction of including n-3 fatty acids in various products starting with n-3 enriched eggs, which are based on the *Ampelistra* (Greek) egg as a model obtained under completely natural conditions and which has a ratio of n-6:n-3 of 1:1.

It is essential that Nutrition Science drives Food Science and the production of foods rather than Food Technology. This is of the utmost importance in the development of novel foods. The scientific evidence is strong for decreasing the n-6 and increasing the n-3 fatty acid intake to improve health throughout the life cycle. The scientific basis for the development of a public policy to develop dietary recommendations for EFA, including a balanced n-6:n-3 ratio, is robust. What is needed is a scientific consensus, education of professionals and the public, the establishment of an agency on nutrition and food policy at the national level, and willingness of governments to institute changes. Education of the public is essential to demand changes in the food supply.

Abbreviations

ALA	α-linolenic acid
CAM	cell adhesion molecule
CRP	C-reactive protein
DHA	docosahexaenoic acid
EFA	essential fatty acid
EPA	eicosapentaenoic acid
FAS	fatty acid synthase
GK	glucokinase
GLUT	glucose transporter
ICAM	intercellular adhesion molecule
IFN	interferon
IL	interleukin
IMT	intima-media thickness
LA	linoleic acid
LO	lipoxygenase
ME	malic enzyme
PDGF	platelet-derived growth factor
PE	phosphatidylethanolamine
PG	prostaglandin
PK	pyruvate kinase
PUFA	polyunsaturated fatty acid
TNF	tumor necrosis factor
VCAM	vascular cell adhesion molecule

See also: **Aging**. **Arthritis**. **Breast Feeding**. **Cancer**: Effects on Nutritional Status. **Coronary Heart Disease**:

Hemostatic Factors; Lipid Theory; Prevention. **Diabetes Mellitus**: Dietary Management. **Fatty Acids**: Omega-6 Polyunsaturated. **Lactation**: Dietary Requirements. **Pregnancy**: Nutrient Requirements; Safe Diet for Pregnancy.

Further Reading

Burr ML, Fehily AM, Gilbert JF, Rogers S, Holliday RM, Sweetnam PM, Elwood PC, and Deadman NM (1989) Effect of changes in fat fish and fibre intakes on death and myocardial reinfarction: diet and reinfarction trial (DART). *Lancet* **2**: 757–761.

Calon F, Lim GP, Yang F, Morihara T, Teter B, Ubeda O, Rostaing P, Triller A, Salem N Jr, Ashe KH, Frautschy SA, and Cole GM (2004) Docosahexaenoic acid protects from dendritic pathology in an Alzheimer's disease mouse model. *Neuron* **43**: 633–645.

de Lorgeril M, Renaud S, Mamelle N, Salen P, Martin JL, Monjaud I, Guidollet J, Touboul P, and Delaye J (1994) Mediterranean alpha-linolenic acid rich-diet in the secondary prevention of coronary heart disease. *Lancet* **343**: 1454–1459.

Dwyer JH, Allayee H, Dwyer KM, Fan J, Wu H, Mar R, Lusis AJ, and Mehrabian M (2004) Arachidonate 5-lipoxygenase promoter genotype, dietary arachidonic acid, and atherosclerosis. *New England Journal of Medicine* **350**: 29–37.

GISSI-Prevenzione Investigators (1999) Dietary supplementation with n-3 polyunsaturated fatty acids and vitamin E after myocardial infarction: results of the GISSI-Prevenzione trial. *Lancet* **354**: 447–455.

Kang JX, Wang J, Wu L, and Kang ZB (2004) *Fat-1* mice convert n-6 to n-3 fatty acids. *Nature* **427**: 504.

Maes M, Smith R, Christophe A, Cosyns P, Desynder R, and Meltzer H (1996) Fatty acid composition in major depression: decreased omega 3 fractions in cholesteryl esters and increased C20:4 omega 6/C20:5 omega 3 ratio in cholesteryl esters and phospholipids. *Journal of Affective Disorders* **38**(1): 35–46.

Mechanisms of Action of LCPUFA (2003) Effects on infant growth and neurodevelopment. Proceedings of a conference held in Arlington, Virginia, May 14–15, 2002. *Journal of Pediatrics* **143**(supplement 4): S1–S109.

Simopoulos AP (2001) N-3 fatty acids and human health: defining strategies for public policy. *Lipids* **36**: S83–S89.

Simopoulos AP (2002) Omega-3 fatty acids in inflammation and autoimmune diseases. *Journal of American College of Nutrition* **21**: 494–505.

Simopoulos AP and Cleland LG (eds.) (2003) *Omega-6/Omega-3 Essential Fatty Acid Ratio: The Scientific Evidence. World Review of Nutrition and Dietetics*, vol. 92 Basel: Karger.

Simopoulos AP, Leaf A, and Salem N Jr (1999) Essentiality of and recommended dietary intakes for omega-6 and omega-3 fatty acids. Annals of Nutrition and Metabolism **43**: 127–130.

Simopoulos AP and Nestel PJ (eds.) (1997) *Genetic Variation and Dietary Response. World Review of Nutrition and Dietetics*, vol. 80, Basel: Karger.

Simopoulos AP and Robinson J (1999) In *The Omega Diet. The Lifesaving Nutritional Program Based on the Diet of the Island of Crete*. New York: Harper Collins.

Simopoulos AP and Visioli F (eds.) (2000) *Mediterranean Diets. World Review of Nutrition and Dietetics*, vol. 87. Basel: Karger.

Singh RB, Dubnov G, Niaz MA, Ghosh S, Singh R, Rastogi SS, Manor O, Pella D, and Berry EM (2002) Effect of an Indo-Mediterranean diet on progression of coronary artery disease in high risk patients (Indo-Mediterranean Diet Heart Study): a randomised single-blind trial. *Lancet* **360**(9344): 1455–1461.

Yehuda S (2003) Omega-6/omega-3 ratio and brain-related functions. *World Review of Nutrition and Dietetics* **92**: 37–56.

Omega-6 Polyunsaturated

J M Hodgson and T A Mori, University of Western Australia, Perth, WA, Australia
M L Wahlqvist, Monash University, Victoria, VIC, Australia

Structure, Function, and Nutritional Requirements

Omega-6 (n-6) fatty acids are a class of polyunsaturated fatty acids (PUFA). They have two or more *cis* double bonds, with the position of the first double bond six carbon atoms from the methyl end of the molecule. The general formula of n-6 fatty acids is $CH_3(CH_2)_4(CH{=}CHCH_2)_x(CH_2)_yCOOH$ [where $x = 2–5$]. Linoleic acid (*cis*-9, *cis*-12-octadecadienoic acid, 18:2n-6, LA) and α-linolenic acid (*cis*-9, *cis*-12, *cis*-15-octadecatrienoic acid, 18:3n-3, ALA) are the precursor fatty acids of the n-6 and omega-3 (n-3) fatty acids, respectively. These two fatty acids cannot be made by mammals and are therefore termed essential fatty acids (EFA). In addition, mammals are unable to interconvert LA and ALA, or any of the n-6 and n-3 fatty acids, because mammalian tissues do not contain the necessary desaturase enzyme. Plant tissues and plant oils tend to be rich sources of LA. ALA is also present in plant sources such as green vegetables, flaxseed, canola, and some nuts. Once consumed in the diet, LA can be converted via chain elongation and desaturation to γ-linolenic acid (GLA, 18:3n-6), dihomo-γ-linolenic acid (DGLA, 20:3n-6), and arachidonic acid (AA, 20:4n-6) (**Figure 1**). The same enzymes involved in elongation and desaturation of the n-6 fatty acids are common to the n-3 series of fatty acids (**Figure 1**). Thus, ALA can be converted to eicosapentaenoic acid (EPA, 20:5n-3) and docosahexaenoic acid (DHA, 22:6n-3). EPA and DHA are found in relatively high proportions in marine oils.

The n-6 and n-3 fatty acids are metabolically and functionally distinct and often have important opposing physiological functions. Indeed, the balance of EFA is important for good health and normal development. Historically, human beings evolved on a diet in which the ratio of n-6 to n-3 fatty acids was about 1:1. In contrast, Western diets have a ratio of

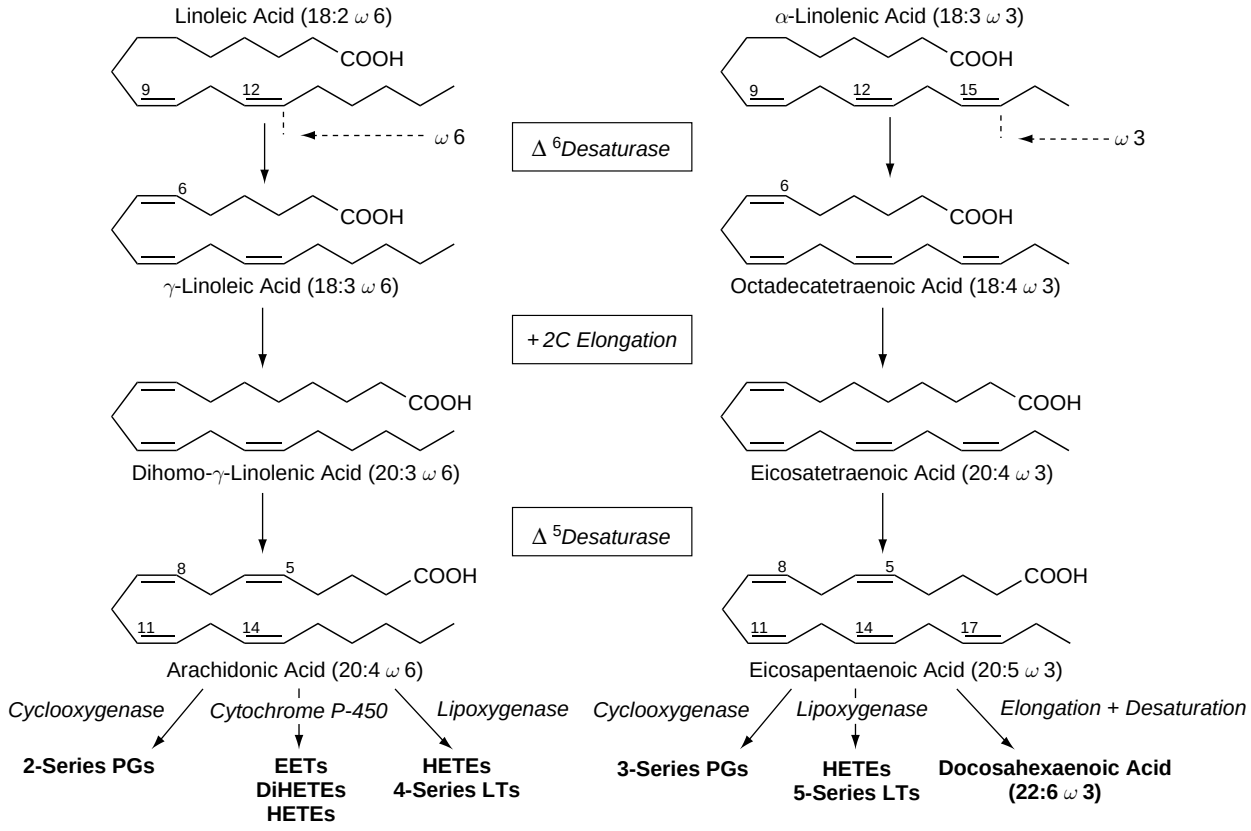


Figure 1 Essential fatty acid metabolism.

approximately 15:1. Evidence for this change in diet through history comes from studies on the evolutionary aspects of diet, modern-day hunter–gatherers, and traditional diets. Modern agriculture has led to a substantial increase in n-6 fatty acids at the expense of n-3 fatty acids, which has resulted in excessive consumption of n-6 fatty acids by humans.

The n-6 EFAs have two main functions. First, they act as structural components of membranes forming the basis of the phospholipid component of the lipid bilayer of plasma membranes in every cell in the body, thus providing a membrane impermeable to most water-soluble molecules. The length and degree of saturation of the fatty acids determine how the phospholipid molecules pack together and consequently affect membrane fluidity, signal transduction, and the expression of cellular receptors. The second role of n-6 fatty acids is as precursors to the eicosanoids (**Figure 1**). The eicosanoids are a family of 'hormone-like' compounds including prostaglandins (PGs), leukotrienes (LTs), and hydroxy- (HETEs), dihydroxy- (DiHETEs), and epoxy- (EETs) fatty acids. Eicosanoids, however, are distinct from most hormones in that they act locally, near their sites of synthesis, and they are catabolized extremely rapidly. Thus, they are considered to be locally acting hormones. The eicosanoids modulate renal and pulmonary function, vascular tone, and inflammatory responses. The enzymes involved in AA metabolism include the cyclooxygenases and lipoxygenases, which yield the 2-series PGs and 4-series LTs, respectively. Lipoxygenase also utilizes AA for the formation of the HETEs. A third pathway for the utilization of AA involves the cytochrome P-450 enzymes found in the liver, kidney, lung, intestines, heart, small blood vessels, and white blood cells. AA metabolized via cytochrome P-450 yields EETs, DiHETEs, as well as HETEs. The cytochrome P-450 metabolites play an important role as paracrine factors and second messengers in the regulation of pulmonary, cardiac, renal, and vascular function and modulate inflammatory and growth responses.

Endothelial Function, Atherosclerosis, and Cardiovascular Disease

Differences in n-6 fatty acid intake have the potential to influence several chronic diseases and disorders. This article will focus on the effects of n-6 fatty acids on cardiovascular disease and atherosclerosis.

The vascular endothelium is the most important organ controlling vascular function and consists of a single layer of epithelial cells lining blood vessels. Its primary function is to regulate vascular tone, but it plays a critical role in modulating coagulation and fibrinolysis, inflammation, smooth muscle cell proliferation, and macrophage function. Many of these functions are regulated through the release of various mediators including eicosanoids. There is multiple and close interaction of the endothelial cells with circulating cells, smooth muscle cells, and macrophages. There is also evidence that endothelial dysfunction precedes clinically apparent atherosclerosis.

Atherosclerosis is an inflammatory disease involving multiple cellular and molecular responses that lead to an alteration in vascular function and structure, and the development and progression of cardiovascular disease. Atherosclerosis is characterized by degenerative changes, deposition of cholesterol, proliferation of smooth muscle cells, involvement of a range of circulating proinflammatory cell types, and fibrosis. Resulting atheromatous plaques cause narrowing of arteries and increase the likelihood of thrombosis and occlusion. When this process occurs in the coronary arteries, the outcome is myocardial infarction and with possible death.

Eicosanoids: Relevance to Endothelial Function, Thrombosis, Inflammation, and Atherosclerosis

In general, the eicosanoids derived from AA have potent prothrombotic and proinflammatory activity. In contrast, the eicosanoids derived from EPA have reduced biological activity and are less prothrombotic and proinflammatory. Eicosanoid production is generally tightly controlled through homeostatic mechanisms. However, eicosanoid production can be significantly altered in situations in which endothelial dysfunction, atherosclerosis and plaque rupture, or various thrombotic or inflammatory conditions are present.

Prostaglandins and Leukotrienes

Prostaglandins have a central role in the regulation of platelet aggregation and vascular tone. In this regard, two of the major prostaglandins derived from AA are thromboxane A_2, produced in platelets, and prostacyclin I_2, produced in endothelial cells. Thromboxane A_2 promotes platelet aggregation and blood vessel constriction, while prostacyclin I_2 has the opposite effects. An increase in availability of EPA can decrease platelet thromboxane A_2 and increase thromboxane A_3, the latter having considerably less physiological activity. EPA supplementation also stimulates formation of prostacyclin I_3, while prostacyclin I_2 is unaffected. Prostacyclin I_3 and prostacyclin I_2 are equipotent in their biological activity. The net result following intake of n-3 fatty acids is a shift in the thromboxane/prostacyclin balance toward a reduced prothrombotic state.

Leukotriene B_4 is a potent inflammatory mediator produced by neutrophils from 20:4n-6 at the site of injury. Leukotriene B_4 is also a powerful chemotactic factor responsible for attracting neutrophils to the site of injury. Leukotriene B_5, which is produced from EPA, has significantly lower biological activity. Therefore an increased availability of EPA has the potential to reduce inflammation.

Fatty Acid Intake and Eicosanoids

The proportional concentration of the eicosanoid precursor fatty acids both circulating and in tissues depends on dietary intake. DGLA and AA can be obtained from animal meat and fat, and by desaturation and chain elongation of LA. The major dietary source of EPA is fish. EPA can also be obtained indirectly from ALA, although desaturation and chain elongation of ALA appears to be a less important pathway in humans.

Only the free form of the fatty acid precursors of eicosanoids can be utilized by the enzymes for conversion to the biologically active metabolites. However, the amount of precursor free fatty acid in the cytoplasm and circulating is usually low and so too is basal eicosanoid formation. Furthermore, basal eicosanoid formation may depend on dietary and adipose tissue fatty acid composition. The amount of eicosanoid precursor free fatty acids is controlled to a large extent by incorporation and release from cellular phospholipids. Which eicosanoids are produced during stimulated synthesis may depend on membrane fatty acid composition as well as the cell type involved. Dietary fatty acid composition, therefore, has the potential to effect basal and stimulated synthesis of eicosanoids and influence endothelial function and thrombotic and inflammatory responses.

n-6 Fatty Acids and Risk of Cardiovascular Disease

Evidence that differences in n-6 fatty acid intake can influence cardiovascular disease risk derives from several sources. Population studies may provide useful data for establishing optimal intakes of n-6 fatty acids. However, valuable information on the potential mechanisms and effects of these fatty acids is

derived from studies focusing on their impact on thrombosis, inflammation, endothelial function, and other cardiovascular risk factors.

Cardiovascular Disease: Population Studies

The incidence of cardiovascular disease within populations with either very high or very low intakes of n-6 fatty acids may provide some indication for optimal intakes of n-6 fatty acids. Within populations with low n-6 fatty acid intakes ($\leq$3%) there would appear to be a benefit of having a higher n-6 fatty acid intake on cardiovascular disease risk reduction. These observations suggest that very low n-6 fatty acid intakes increase the risk for cardiovascular disease. The presence of EFA deficiency in a significant proportion of such populations may explain the increased risk. Several populations, including the Israelis, Taiwanese, and !Kung bushmen in the African Kalahari desert, have high to very high intakes of n-6 fatty acids. The contribution of n-6 fatty acids to total energy intake is about 10% in the Israelis and Taiwanese and about 30% in the !Kung bushmen. Rates of cardiovascular disease are low in the Taiwanese, where dietary n-6 fatty acids are obtained mainly from soybean oil, and estimated to be very low in the !Kung bushmen, where dietary n-6 fatty acids were obtained mainly from the monongo fruit and nut. In the Taiwanese, the soybean oil is refined but is accompanied by a diet rich in antioxidant polyphenols, notably from tea, fruits, and vegetables. In the !Kung bushmen the oil is unrefined and is therefore likely to contain a range of phytochemicals. There is, however, a high prevalence of cardiovascular disease in the Israeli population, where n-6 PUFAs are obtained largely from refined sources. These observations suggest that a high n-6 fatty acid intake can be compatible with low risk of cardiovascular disease, but the dietary context may be very important. Given that n-6 fatty acids are susceptible to lipid peroxidation, high n-6 fatty acid intake may increase risk for cardiovascular disease when consumed against a background diet low in antioxidants. The potential impact on eicosanoid metabolism remains uncertain.

Several factors may need to be considered in the interpretation of the results of population studies. First, the effect of LA on atherosclerosis and cardiovascular disease may depend on the background intake in the population being studied. Second, any relationships observed may be confounded by intake of other foods from which LA derives. Third, LA may have differential effects on aspects of the aetiology of cardiovascular disease, including endothelial function, thrombosis, arrhythmia, and atherosclerosis.

Thrombosis

Dietary fatty acids influence thrombosis by altering the activity and function of endothelial cells, platelets, and other circulating cells—effects that can be mediated, in part, by alterations in eicosanoid metabolism. Replacement of dietary saturated fatty acids with unsaturated fatty acids, including n-6 fatty acids, generally lowers the risk of thrombosis and cardiovascular disease. Furthermore, studies have shown that an increase in n-3 fatty acid intake can increase vasodilation, attenuate platelet aggregation, and alter circulating concentrations of factors involved in coagulation and fibrinolysis. The net effect of increasing n-3 fatty acid intake is a tendency toward reduced risk for thrombosis. These findings are supported by population studies demonstrating that n-3 fatty acids may reduce the risk of thrombosis. It remains uncertain whether the major factor influencing these functions is the absolute increase in n-3 fatty acids or the relative proportions of n-6 and n-3 fatty acids in the diet and cell membranes. There is evidence, however, that increased n-3 fatty acid intake may be more beneficial in populations consuming relatively small quantities of fish, which includes many Western populations.

Much of the evidence for a potential impact of n-6 fatty acids on thrombosis derives from research on platelet function. The role of platelets in thrombosis is established and the influence of fatty acid intake on platelet function has been assessed in many studies. Platelets play a part in thrombosis by adhering to, and aggregating at, the site of injury. Platelet reactivity and increased platelet activation may increase the risk of thrombosis. *In vitro* and *in vivo* studies assessing effects of n-6 fatty acids on platelet aggregation are inconsistent. To date there is little evidence that a high n-6 fatty acid diet in humans decreases platelet aggregation and some studies are suggestive of increased aggregation with high n-6 fatty acid diets, primarily in the form of LA. The effects of AA on platelet aggregation are also not clear. One of the main difficulties in interpreting these studies is the unresolved issue as to how the *in vitro* aggregation test reflects platelet function *in vivo*.

Inflammation

Conditions of increased inflammation, such as inflammatory arthritis, dermatological conditions such as psoriasis and atopic dermatitis, chronic

inflammatory bowel disease, autoimmune diseases, and bronchial asthma, appear to be beneficially influenced by n-3 fatty acids but not by n-6 fatty acids.

Whether or not increased intake of n-6 fatty acids can exacerbate inflammation *via* increased production of proinflammatory eicosanoids remains uncertain. Results of *in vitro* studies and intervention studies in humans are generally consistent with this theoretical potential of n-6 fatty acids to enhance inflammation, at least in comparison to n-3 fatty acids and probably n-9 monounsaturated fatty acids. The importance of absolute and relative intakes of n-6 fatty acids to inflammatory processes also remains unclear. The effects of changes in n-6 fatty acid intake on inflammatory processes may depend on the background dietary fatty acid intake, as well as proportional and absolute intake of n-3 fatty acids.

Cholesterol and Lipoproteins

The major classes of circulating lipoproteins in human plasma are chylomicrons, very low-density lipoproteins (VLDL), low-density lipoproteins (LDL), and high-density lipoproteins (HDL). High fasting plasma concentrations of LDL cholesterol and triglycerides—predominantly circulating as part of VLDL—and low plasma concentrations of HDL cholesterol are associated with increased risk of cardiovascular disease. Dietary fatty acids can influence lipoprotein metabolism and therefore have the potential to influence atherosclerosis and cardiovascular disease risk. Most studies examining the effects of n-6 PUFAs on cholesterol metabolism have focused on LA, the major dietary n-6 fatty acid.

It is now established that LDL cholesterol lowering reduces the risk of cardiovascular disease. In the fasting state LDL is the major cholesterol carrying lipoprotein in human plasma. The mechanisms through which raised plasma LDL cholesterol concentrations increase cardiovascular disease risk are not entirely understood but oxidative modification of LDL is thought to be involved. An increase in LA intake results in a lowering of plasma LDL cholesterol concentrations and therefore has the potential to reduce cardiovascular disease risk. These effects may not be linear over the entire range of LA intake and most of the benefits appear to be gained by moving from lower (<2% of energy) to moderate (~4–5% of energy) intakes. In addition, it is worthy of note that the effects of dietary n-6 PUFAs are less than half that of lowering dietary saturated fatty acids. Therefore, if total fat intake is maintained, the LDL cholesterol lowering effects of increasing n-6 PUFA intake are greatly enhanced if saturated fatty acid intake is decreased.

HDL cholesterol is inversely associated with cardiovascular disease risk. The mechanism by which HDL reduces cardiovascular disease risk may involve reverse cholesterol transport and reductions in cholesterol accumulation in the arterial wall. Intakes of LA within the normal ranges of intakes in most populations do not appear to alter HDL cholesterol concentrations. However, very high intakes—above 12% of energy—can lower HDL cholesterol concentrations.

Oxidative Stress

Several lines of evidence suggest that oxidatively modified LDL plays an important role in the development of atherosclerosis. Oxidative modification of LDL involves peroxidation of PUFAs. LDL particles enriched in PUFAs have been shown to be more susceptive to oxidative modification compared to LDL particles rich in monounsaturated fatty acids. Others have also suggested that a diet high in PUFAs may overwhelm the antioxidant defenses of cells. In particular, studies have shown that LA-enriched LDL is more prone to *in vitro* oxidation than oleic acid-enriched LDL. Concern also remains with respect to the potential for increased lipid peroxidation following n-3 fatty acids. To date, however, the data *in vivo* are inconclusive, with observations of increased, unchanged, and decreased lipid peroxidation. The most plausible explanation relates to differences in the methodologies employed to assess lipid peroxidation. Much of the literature relating to PUFAs and lipid peroxidation is based on indirect and nonspecific assays, including measurement of LDL oxidative susceptibility, which relies on the isolation of LDL from plasma. In this regard, the recent discovery of F_2-isoprostanes, which are nonenzymatic prostaglandin-like products of free radical peroxidation of arachidonic acid, has allowed for the direct assessment of *in vivo* lipid peroxidation. There is now good evidence that quantitation of F_2-isoprostanes provides a reliable measure of *in vivo* oxidative stress. Using measurement of F_2-isoprostanes, recent data have demonstrated that n-3 fatty acids decrease oxidative stress. It has also been suggested that the concentration of PUFAs may be a more important factor affecting lipid peroxidation than the degree of unsaturation. Further research using better markers of lipid peroxidation is required before definitive statements can be made relating to the effect of n-6 fatty acids, and indeed PUFAs in general, on oxidative stress.

Blood Pressure

The possible effects of dietary fatty acids on blood pressure have been explored in population studies and dietary intervention trials. With the exception of studies comparing vegetarian and nonvegetarian populations, from which there is a suggestion of a blood pressure lowering effect of diets high in PUFAs, including LA, and lower in saturated fatty acids, the results of most within- and between-population studies have generally not found significant associations. The results of intervention studies suggest that n-6 fatty acids, LA in particular, may be responsible for a small blood pressure lowering effect. However, these studies are also inconsistent, with several failing to find a significant blood pressure lowering effect.

Conclusions

Diets low in n-6 fatty acids, principally LA, appear to be associated with an increased risk of cardiovascular disease. The results of studies examining the effects of LA on risk factors for atherosclerosis and cardiovascular disease are consistent with this observation. An increase in n-6 PUFA intake from a low to a moderate intake level, in conjunction with decreases in total and saturated fat intake, may beneficially influence lipoprotein metabolism, lower blood pressure, and reduce cardiovascular disease risk. Observations in populations with high n-6 PUFA intake indicate that high intakes of n-6 fatty acids (>10%) can occur together with low rates of cardiovascular disease and possibly also cancer. However, where antioxidant composition of the diet is low, there is the potential for increased risk of cardiovascular disease. An increased susceptibility of PUFAs to oxidative damage, particularly in the presence of low concentrations of protective antioxidants, may be an important factor involved. The source of n-6 PUFAs in the diet, refined versus unrefined, and the composition of the background diet may therefore be important determinants of whether high n-6 fatty acid intake increases or decreases risk of cardiovascular disease. In addition, the proportion of n-6 to n-3 fatty acids in the diet may also play an important role in determining cardiovascular risk.

The available evidence suggests that n-6 fatty acid-derived eicosanoids are generally proinflammatory and prothrombotic. In contrast, eicosanoids derived from n-3 fatty acids have attenuated biological activity on cardiovascular risk factors. The effects of altering n-6 PUFA intake, in conjunction with changes in other polyunsaturated fatty acids, as well as other classes of fatty acids, on endothelial function, thrombosis, and inflammation are not understood. The relative proportion of all the classes of fatty acids in the diet may well be more important and relevant to cardiovascular risk reduction than any single class of fatty acids. Clearly such research warrants further investigation.

See also: **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Coronary Heart Disease**: Lipid Theory. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Fish**. **Lipoproteins**. **Prostaglandins and Leukotrienes**.

Further Reading

Grundy SM (1996) Dietary fat. In: Ziegler EE and Filer LJ Jr. (eds.) *Present knowledge in Nutrition*, 7th edn., pp. 44–57. Washington, DC: ILSI Press.

Hodgson JM, Wahlqvist ML, Boxall JA, and Balazs NDH (1993) Can linoleic acid contribute to coronary artery disease? *American Journal of Clinical Nutrition* **58**: 228–234.

Hodgson JM, Wahlqvist ML, and Hsu-Hage B (1995) Diet, hyperlipidaemia and cardiovascular disease. *Asia Pacific Journal of Clinical Nutrition* **4**: 304–313.

Hornsrtra G, Barth CA, Galli C *et al.* (1998) Functional food science and the cardiovascular system. *British Journal of Nutrition* **80**(supplement 1): S113–S146.

Horrobin DF (ed.) (1990) *Omega-6 Essential Fatty Acids: Pathophysiology and Roles in Clinical Medicine*. New York: Wiley-Liss.

Jones GP (1997) Fats. In: Wahlqvist ML (ed.) *Food and Nutrition, Australia, Asia and the Pacific*, pp. 205–214. Sydney: Allen & Unwin.

Jones PJH and Kubow S (1999) Lipids, sterols and their metabolism. In: Shils ME, Olson JA, Shike M, and Ross AC (eds.) *Modern Nutrition in Health and Disease*, 9th edn., pp. 67–94. Baltimore: Williams & Wilkins.

Knapp HR (1997) Dietary fatty acids in human thrombosis and hemostasis. *American Journal of Clinical Nutrition* **65**(supplement 5): 1687S–1698S.

Lyu LC, Shieh MJ, Posner BM *et al.* (1994) Relationship between dietary intake, lipoproteins and apolipoproteins in Taipei and Framingham. *American Journal of Clinical Nutrition* **60**: 765–774.

Mensink R and Connor W (eds.) (1996) *Nutrition. Current Opinion in Lipidology* **7**: 1–53.

National Health and Medical Research Council (1992) *The role of polyunsaturated fats in the Australian diet: Report of the NHMRC working party*. Canberra: Australian Government Publishing Service.

Salem N, Simopoulos AP, Galli, Lagarde M, and Knapp HR (eds.) (1996) Fatty acids and lipids from cell biology to human disease: Proceedings of the 2nd International Congress of the International Society for the Study of Fatty Acids and Lipids. *Lipids* **31** (supplement).

Truswell AS (1977) Diet and nutrition of hunter gatherers. *Ciba Foundation Symposium*, 213–221.

Yam D, Eliraz A, and Berry EM (1996) Diet and disease—The Israeli paradox: Possible dangers of a high omega-6 polyunsaturated fatty acid diet. *Israeli Journal of Medical Sciences* **32**: 1134–1143.

Saturated

R P Mensink, Maastricht University, Maastricht, The Netherlands
E H M Temme, University of Leuven, Leuven, Belgium

Fats and oils always consist of a mixture of fatty acids, although one or two fatty acids are usually predominant. **Table 1** shows the fatty acid composition of some edible fats rich in saturated fatty acids. In the Western diet, palmitic acid ($C_{16:0}$) is the major saturated fatty acid. A smaller proportion comes from stearic acid ($C_{18:0}$), followed by myristic acid ($C_{14:0}$), lauric acid ($C_{12:0}$), and short-chain and medium-chain fatty acids (MCFA) ($C_{10:0}$ or less).

When discussing the health effects of the total saturated fat content of diets, this class of fatty acids has to be compared with some other component of the diet that provides a similar amount of energy (isoenergetic). Otherwise, two variables are being introduced: changes in total dietary energy intake and, as a consequence, changes in body weight. Normally, an isoenergetic amount from carbohydrates is used for comparisons.

Cholesterol Metabolism

Lipoproteins and their associated apoproteins are strong predictors of the risk of coronary heart disease (CHD). Concentrations of total cholesterol, low-density lipoproteins (LDL), and apoprotein B are positively correlated with CHD risk; high-density lipoprotein (HDL) and apoprotein Al concentrations are negatively correlated. Controlled dietary trials have now demonstrated that the total saturated fat content and the type of saturated fatty acid in the diet affect serum lipid and lipoprotein levels.

Total Saturated Fat Content of Diets

Using statistical techniques, results from independent experiments have been combined to develop equations that estimate the mean change in serum lipoprotein levels for a group of subjects when carbohydrates are replaced by an isoenergetic amount of a mixture of saturated fatty acids. The predicted changes for total LDL and HDL cholesterol and triacylglycerols are shown in **Figure 1**. Each bar represents the predicted change in the concentration of that particular lipid or lipoprotein when a particular fatty acid class replaces 10% of the daily energy intake from carbohydrates. For a group of adults with an energy intake of 10 MJ daily, 10% of energy is provided by about 60 g of carbohydrates or 27 g of fatty acids.

A mixture of saturated fatty acids strongly elevates serum total cholesterol levels. It was predicted that when 10% of dietary energy provided by carbohydrates was exchanged for a mixture of saturated fatty acids, serum total cholesterol concentrations would increase by $0.36\,\mathrm{mmol\,l^{-1}}$. This increase in total cholesterol will result from a rise in both LDL and HDL cholesterol concentrations. Saturated fatty acids will also lower fasting triacylglycerol concentrations compared with carbohydrates. Besides affecting LDL and HDL cholesterol concentrations, a mixture of saturated fatty acids also changes the concentrations of their associated apoproteins. In general, strong associations are observed between changes in LDL cholesterol and changes in apo-B and between changes in HDL cholesterol and apo-Al.

Figure 1 also shows that total and LDL cholesterol concentrations decrease when saturated fatty acids are replaced by unsaturated fatty acids. In addition, slight decreases of HDL cholesterol concentrations are then predicted.

Effects of Specific Saturated Fatty Acids

Cocoa butter raises total cholesterol concentrations to a lesser extent than palm oil. This difference in

Table 1 Composition of fats rich in satured fatty acids

	Weight per 100 g of total fatty acids (g)								
	$\leq C_{10:0}$	$C_{12:0}$	$C_{14:0}$	$C_{16:0}$	$C_{18:0}$	$C_{18:1}$	$C_{18:2}$	$C_{18:3}$	*Other*
Butterfat	9	3	17	25	13	27	3	1	2
Palm kernel fat	8	50	16	8	2	14	2		
Coconut fat	15	48	17	8	3	7	2		
Palm oil			1	45	5	39	9		1
Beef fat			3	26	22	38	2	1	8
Pork fat (lard)			2	25	12	44	10	1	6
Cocoa butter				26	35	35	3		1

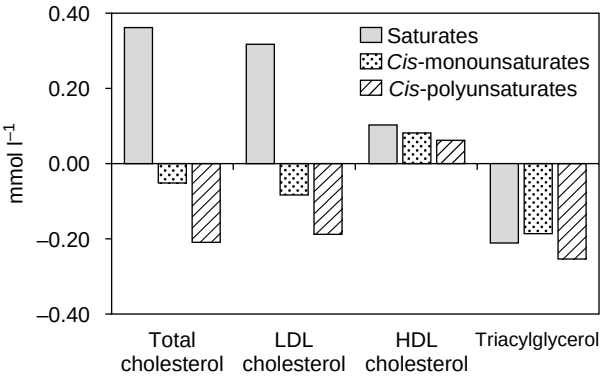


Figure 1 Predicted changes in serum lipids and lipoproteins when 10% of energy from dietary carbohydrates is replaced by an isoenergetic amount of saturated fatty acids. From Mensink *et al.* (2003) *American Journal of Clinical Nutrition* **77**: 1146–1155. Reproduced with permission by the *American Journal of Clinical Nutrition*. © Am J Clin Nutr. American Society for Clinical Nutrition.

the serum cholesterol-raising potency of two fats high in saturated fatty acids (see **Table 1**) showed that not all saturated fatty acids have equal effects on cholesterol concentrations. **Figure 2** illustrates the effects of lauric, myristic, palmitic, and stearic acids on LDL and HDL cholesterol concentrations. Compared with other saturated fatty acids, lauric and myristic acids have the strongest potency to increase serum total and LDL cholesterol concentrations and also HDL cholesterol concentrations. Effects of lauric acid on HDL are stronger than those of myristic acid.

Scientists are not unanimous about the cholesterol-raising properties of palmitic acid, the major dietary saturated fatty acid. Many studies have indicated that, compared with carbohydrates, palmitic acid raises serum total and LDL cholesterol levels but has less effect on HDL cholesterol (**Figure 2**). However, a few studies indicated that palmitic acid might not raise total and LDL cholesterol concentrations compared with carbohydrates. It has been proposed that this negative finding is only present when the linoleic acid content of the diet is adequate (6–7% of energy). It is hypothesized that the increased hepatic apo-B100 production caused by palmitic acid, and the consequent elevation of concentrations of serum very low density lipoproteins (VLDL) and LDL particles, is counteracted by an increased uptake of LDL particles by the LDL receptor which is upregulated by linoleic acid. To explain the discrepancy with other studies, it has been suggested that in some situations, such as hypercholesterolaemia or obesity, linoleic acid is unable to increase LDL receptor activity sufficiently to neutralize the cholesterol-raising effects of palmitic acid. This theory, however, awaits confirmation, and for now it seems justified to classify palmitic acid as a cholesterol-raising saturated fatty acid.

Stearic acid, a major fatty acid in cocoa butter, does not raise total, LDL, and HDL cholesterol levels compared with carbohydrates. Also, MCFA have been reported not to raise LDL and HDL cholesterol concentrations compared with carbohydrates, but data are limited. Like carbohydrates, diets containing large amounts of MCFA increase fasting triacylglycerol concentrations compared with the other saturated fatty acids. However, such diets are the sole energy source only in parenteral or enteral nutrition or in sports drinks. Other saturated fatty acids have not been reported to raise

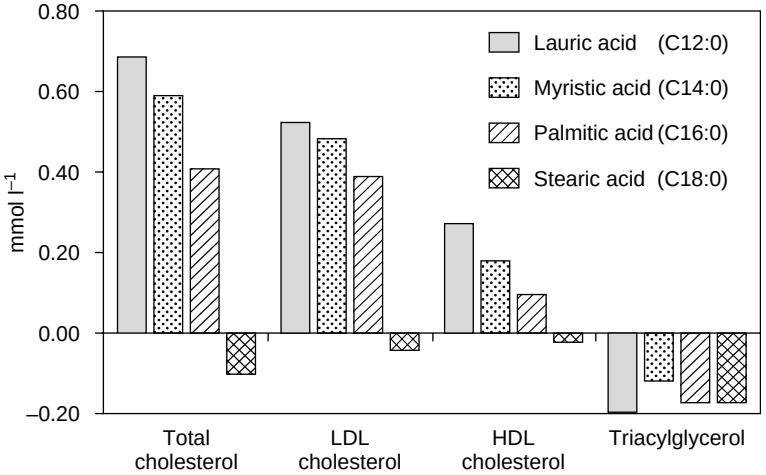


Figure 2 Overview of the effects of particular fatty acids on serum total, LDL, and HDL cholesterol concentration when 10% of energy from dietary carbohydrates is replaced by an isoenergetic amount of a particular saturated fatty acid. From Mensink *et al.* (2003) *American Journal of Clinical Nutrition* **77**: 1146–1155. Reproduced with permission by the *American Journal of Clinical Nutrition*. © Am J Clin Nutr. American Society for Clinical Nutrition.

triacylglycerol concentrations compared with each other, but lower triacylglycerol concentrations compared with carbohydrates.

Platelet Aggregation

Increased platelet aggregation may be an important risk marker for the occurrence of cardiovascular disease, and different types of fatty acids can modify platelet aggregation *in vitro*. However, reports of research on this topic are confusing. All measurements have their limitations, and it is not known whether measurement *in vitro* of platelet aggregation reflects the reality of platelet reactivity *in vivo*.

Many methods are available to measure platelet aggregation *in vitro*. First, the blood sample is treated with an anticoagulant to avoid clotting of the blood in the test tube or in the aggregometer; many different anticoagulants are used, which all differ in their mechanism of action. Second, platelet aggregation can be measured in whole blood, in platelet-rich plasma, or (to remove the influence of the plasma constituents) in a washed platelet sample. Finally, the platelet aggregation reaction in the aggregometer can be initiated with many different compounds, such as collagen, ADP, arachidonic acid, and thrombin. Platelet aggregation can also be studied by measuring the stable metabolites of the proaggregatory thromboxane A_2 (TxA_2), thromboxane B_2 (TxB_2), the stable metabolite of the antiaggregatory prostaglandin (prostacyclin: PGI_2), or 6-keto-PGF1α.

Total Saturated Fat Content of Diets

Platelet aggregation and clotting activity of plasma were studied in British and French farmers, who were classified according to their intake of saturated fatty acids. A positive correlation was observed between thrombin-induced aggregation of platelet-rich plasma and the intake of saturated fatty acids. Aggregation induced by ADP or collagen, however, did not correlate with dietary saturated fat intake. In a follow-up study, a group of farmers consuming high-fat diets were asked to replace dairy fat in their diets with a special margarine rich in polyunsaturated fatty acids. Besides lowering the intake of saturated fatty acids, this intervention also resulted in a lower intake of total fat. A control group of farmers did not change their diets. After this intervention the thrombin-induced aggregation of platelet-rich plasma decreased when saturated fat intake decreased. Aggregation induced by ADP, however, increased in the intervention group. From these studies, it is not clear whether the fatty acid composition of the diets or the total fatty acid content is responsible for the changes in platelet aggregation. Furthermore, it is not clear if one should favor increased or decreased platelet aggregation after decreasing the saturated fat content of diets as effects did depend on the agonist used to induce platelet aggregation. Saturated fatty acids from milk fat have also been compared with unsaturated fatty acids from sunflower and rapeseed oils. Aggregation induced by ADP or collagen in platelet-rich plasma was lower with the milk fat diet than with either oil.

One of the mechanisms affecting platelet aggregation is alteration of the proportion of arachidonic acid in the platelet phospholipids. Arachidonic acid is a substrate for the production of the proaggregatory TxA_2 and the antiaggregatory PGI_2, and the balance between these two eicosanoids affects the degree of platelet activation. The proportion of arachidonic acid in membranes can be modified through changes in dietary fatty acid composition. Diets rich in saturated fatty acids increase the arachidonic acid content of the platelet phospholipids, but this is also dependent on the particular saturated fatty acid consumed (see below).

Diets rich in saturated fatty acids have also been associated with a lower ratio of cholesterol to phospholipids in platelet membranes, which may affect receptor activity and platelet aggregation. However, these mechanisms have been described from studies *in vitro* and on animals and have not adequately been confirmed in human studies.

Effects of Specific Saturated Fatty Acids

Diets rich in coconut fat have been reported to raise TxB_2 and lower 6-keto-$PGF_{1\alpha}$ concentrations in collagen-activated plasma compared with diets rich in palm or olive oils, indicating a less favourable eicosanoid profile. The main saturated fatty acids of coconut fat — lauric and myristic acids — did not, however, change collagen-induced aggregation in whole-blood samples compared with a diet rich in oleic acid. Also, diets rich in MCFA or palmitic acid did not change collagen-induced aggregation in whole-blood samples. Compared with a diet rich in a mixture of saturated fatty acids, a stearic acid diet increased collagen-induced aggregation in platelet-rich plasma. In addition, a decreased proportion of arachidonic acid in platelet phospholipids was demonstrated after a cocoa butter diet compared with a diet rich in butterfat. Changes in eicosanoid metabolite concentrations in urine, however, were not observed after either diet. These results are

conflicting and it is debatable whether measurement *in vitro* of platelet aggregation truly reflects the situation *in vivo*.

Coagulation and Fibrinolysis

Processes involved in thrombus formation include not only those required for the formation of a stable thrombus (platelet aggregation and blood clotting) but also a mechanism to dissolve the thrombus (fibrinolysis). Long-term prospective epidemiological studies have reported that in healthy men factor VII coagulant activity (factor VIIc) and fibrinogen concentrations were higher in subjects who developed cardiovascular diseases at a later stage of the study. Factor VIIc in particular was associated with an increased risk of dying from cardiovascular disease. A high concentration of plasminogen activator inhibitor type 1 (PAI-1) indicates impaired fibrinolytic capacity of the plasma and is associated with increased risk of occurrence of coronary events.

Saturated fatty acids can affect the plasma activity of some of these coagulation and fibrinolytic factors and thus the prethrombotic state of the blood. However, the effects of saturated fatty acids on coagulation and fibrinolytic factors in humans, unlike effects on cholesterol concentrations, have received little attention, and few well-controlled human studies have been reported. Also, regression equations derived from a meta-analysis, which predict the effects on coagulation and fibrinolytic factors of different fatty acid classes compared with those of carbohydrates, do not exist. Therefore, the reference fatty acid is dependent on the experiment discussed. In the epidemiological studies that have found associations between CHD risk and factors involved in thrombogenesis or atherogenesis, subjects were mostly fasted. Also, the effects of saturated fatty acids on cholesterol metabolism, platelet aggregation, and coagulation and fibrinolysis have been studied mainly in fasted subjects. It should be noted, however, that concentrations of some coagulation factors (e.g., factor VIIc) and fibrinolytic factors change after a meal.

Total Saturated Fat Content of Diets

Coagulation Results of studies on the effects of low-fat diets compared with high-fat diets provide some insight into the effects of decreasing the saturated fat content of diets. However, in these studies multiple changes are introduced which makes interpretation of results difficult.

Figure 3 demonstrates that decreased factor VIIc levels were observed in subjects on low-fat diets compared with those on high-saturated fat diets. In many of these studies, the low-fat diet provided smaller quantities of both saturated and unsaturated fatty acids and more fiber than the high-saturated fat diets. The combined results, however, suggest that, apart from a possible effect of dietary fiber, saturates increase factor VII levels compared with carbohydrates. Effects on other clotting factors are less clear. Measurements of markers of *in vivo* coagulation (e.g., prothrombin fragment $1+2$) might have provided more information on the effect of saturates on blood coagulation but were unfortunately not measured in most experiments.

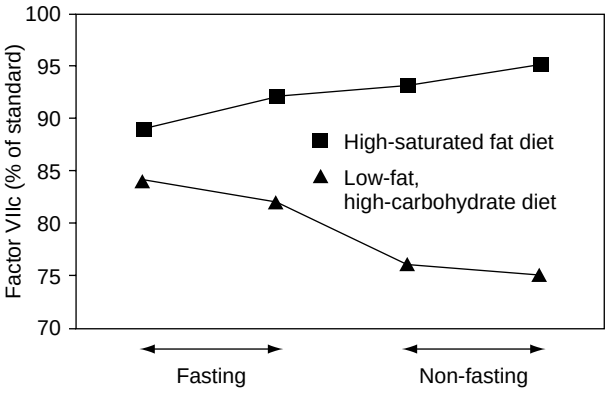


Figure 3 Effects of a high-saturated fat diet on fasting and postprandial factor VIIc activity From Miller (1998) *American Journal of Clinical Nutrition* **67**(supplement): 542S–545S. Reproduced with permission by the *American Journal of Clinical Nutrition*. © Am J Clin Nutr. American Society for Clinical Nutrition.

Fibrinolysis Effects of low-fat and high-fat diets on the fibrinolytic capacity of the blood have also been studied. A similar problem, as stated before, is that multiple changes were introduced within a single experiment. Results of longer term and shorter term studies with dietary changes of total fat (decrease of saturated and unsaturated fatty acids contents) and increased fiber content indicate beneficially increased euglobulin fibrinolytic capacity of the blood. However, when the saturated fatty acid and fiber content of two diets were almost identical and only the unsaturated fatty acid content was changed, no significant differences in fibrinolytic capacity were observed.

Little is known about the relative effects on fibrinolytic capacity of saturated fatty acids compared with unsaturated fatty acids. It has been reported, however, that diets rich in butterfat decreased PAI-1 activity compared with a diet rich in partially hydrogenated soybean oil, but whether this is because of changes in the saturated acid or the *trans* fatty acid content is not clear from this study.

As for coagulation factors, the findings on the fibrinolytic effects of saturates are still inconclusive and need to be examined by more specific assays, measuring the activities of the separate fibrinolytic factors such as tPA and PAI-1.

Effects of Specific Saturated Fatty Acids

Coagulation The interest in the effects of particular fatty acids on coagulation and fibrinolytic factors has increased since the observation that different saturated fatty acids raise serum lipids and lipoproteins in different ways (see section on cholesterol metabolism). Although results are conflicting, some studies indicate that the most potent cholesterol-raising saturated fatty acids also increase factor VII activity.

Diets rich in lauric plus myristic acids compared with a diet rich in stearic acid also increase concentrations of other vitamin K-dependent coagulation proteins. In addition, this mixture of saturated fatty acids raised $F1+2$ concentrations, indicating increased *in vivo* turnover of prothrombin to thrombin. This agreed with a study in rabbits where increased $F1+2$ concentrations were associated with increased hepatic synthesis of vitamin K-dependent clotting factors.

Diets rich in certain saturated fatty acids (lauric acid and palmitic acid) and also diets rich in butterfat have been reported to raise fibrinogen concentrations, but increases were small.

Postprandially, increased factor VIIc concentrations have been demonstrated after consumption of diets rich in fat compared with fat-free meals (**Figure 3**). The response is stronger when more fat is consumed, but this occurs regardless of whether the fat is high in saturated or unsaturated fatty acids. Only meals with unrealistically high amounts of MCFA have been reported not to change factor VIIc levels in comparison with a meal providing a similar amount of olive oil.

Fibrinolysis Increased PAI-1 activity of a palmitic acid-rich diet has been observed compared with diets enriched with oleic acid, indicating impaired fibrinolytic capacity of the plasma. However, this was not confirmed by other experiments on the effects of particular saturated fatty acids (including palmitic acid), which did not indicate changes in fibrinolytic capacity of the blood, measured as tPA, PAI-1 activity, or antigen concentrations of tPA and PAI-1.

Conclusion

Saturated fatty acids as a group affect factors involved in cholesterol metabolism. Relative to the carbohydrate content of the diet, a decrease in saturated fat content induces a favorable decrease in serum total and LDL cholesterol concentrations but unfavorably reduces HDL cholesterol concentrations. Both increasing and decreasing effects of saturates on platelet aggregation have been observed, as well as the absence of effect, so results are inconsistent and difficult to interpret. Whether the beneficial effect of a diet low in saturated fat on the prethrombotic state of blood depends on the dietary fiber content is still unclear.

Of the saturated fatty acids, lauric and myristic acids have the strongest potency to raise total and LDL cholesterol concentrations. In addition, both of these saturated fatty acids raise HDL cholesterol levels. Palmitic acid raises total and LDL cholesterol levels compared with carbohydrates but is less potent than lauric and myristic acids. Stearic acid does not raise LDL and HDL cholesterol concentrations compared with carbohydrates. Lauric, myristic, and palmitic acids increase factor VII activity in a similar way, whereas the effects of MCFA and stearic acid seem limited.

See also: **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Coronary Heart Disease**: Hemostatic Factors; Lipid Theory; Prevention. **Fats and Oils**. **Fatty Acids**: Metabolism. **Lipids**: Chemistry and Classification. **Lipoproteins**.

Further Reading

Hornstra G and Kester ADM (1997) Effect of the dietary fat type on arterial thrombosis tendency: Systematic studies with a rat model. *Atherosclerosis* **131**: 25–33.

Khosla P and Sundram K (1996) Effects of dietary fatty acid composition on plasma cholesterol. *Progress in Lipid Research* **35**: 93–132.

Kris-Etherton PM, Kris-Etherton PM, Binkoski AE *et al.* (2002) Dietary fat: Assessing the evidence in support of a moderate-fat diet; The benchmark based on lipoprotein metabolism. *Proceedings of the Nutrition Society* **61**: 287–298.

Masson LF, McNeill G, and Avenell A (2003) Genetic variation and the lipid response to dietary intervention: A systematic review. *American Journal of Clinical Nutrition* **77**: 1098–1111.

Mensink RP, Zock PL, Kester AD, and Katan MB (2003) Effects of dietary fatty acids and carbohydrates on the ratio of serum total to HDL cholesterol and on serum lipids and apolipoproteins: A meta-analysis of 60 controlled trials. *American Journal of Clinical Nutrition* **77**: 1146–1155.

Miller GJ (1998) Effects of diet composition on coagulation pathways. *American Journal of Clinical Nutrition* **67**(supplement): 542S–545S.

Mutanen M and Freese R (1996) Polyunsaturated fatty acids and platelet aggregation. *Current Opinion in Lipidology* **7**: 14–19.

Mutanen M and Freese R (2001) Fats, lipids and blood coagulation. *Current Opinion in Lipidology* **12**: 25–29.

Sacks FM and Katan M (2002) Randomized clinical trials on the effects of dietary fat and carbohydrate on plasma lipoproteins and cardiovascular disease. *American Journal of Medicine* **113**(Supplement 9B): 13S–24S.

Temme EHM, Mensink RP, and Hornstra G (1998) Saturated fatty acids and effects on whole blood aggregation in vitro. *European Journal of Clinical Nutrition* **52**: 697–702.

Temme EHM, Mensink RP, and Hornstra G (1999) Effects of diets enriched in lauric, palmitic or oleic acids on blood coagulation and fibrinolysis. *Thrombosis and Haemostasis* **81**: 259–263.

Tholstrup T, Miller GJ, Bysted A, and Sandström B (2003) Effect of individual dietary fatty acids on postprandial activation of blood coagulation factor VII and fibrinolysis in healthy young men. *American Journal of Clinical Nutrition* 77: 1125–1132.

Trans Fatty Acids

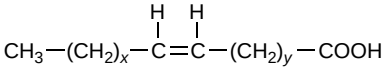
M J Sadler, MJSR Associates, Ashford, UK

Chemistry

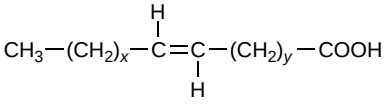
The *trans* fatty acids are unsaturated fatty acids that contain one or more ethylenic double bonds in the *trans* geometrical configuration, i.e., on opposite sides of the carbon chain (**Figure 1**). The *trans* bond is more thermodynamically stable than the *cis* bond and is therefore less chemically reactive.

Trans bonds have minimal effect on the conformation of the carbon chain such that their physical properties more closely resemble those of saturated fatty acids than of *cis* unsaturated fatty acids. The conformation remains linear, compared with *cis* fatty acids, which are kinked (**Figure 2**). Hence, *trans* isomers can pack together more closely than their *cis* counterparts.

Trans fatty acids have higher melting points than their *cis* counterparts, while saturated fatty acids have higher melting points than both *trans* and *cis* fatty acids. For example, the melting points of C_{18}

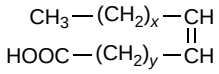


cis configuration

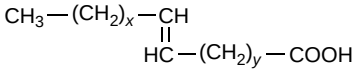


trans configuration

Figure 1 The *trans* and *cis* configurations of unsaturated bonds. Reproduced with kind permission of the British Nutrition Foundation.



cis conformation



trans conformation

Figure 2 Conformation of the carbon chain with *trans* and *cis* bonds. Reproduced with kind permission of the British Nutrition Foundation.

fatty acids are 69.6 °C for stearic acid (18:0), 44.8 °C for elaidic acid (*trans*-18:1), and 13.2 °C for oleic acid (*cis*-18:1). The relative proportion of these different types of fatty acids influences the physical properties of cooking fats and their suitability for different uses in the food processing industry.

In addition to geometrical isomerism (*cis* and *trans*), unsaturated fatty acids also exhibit positional isomerism, where the double bonds can occur in different positions along the chain in fatty acids which have identical chemical formulae. As with *cis* fatty acids, *trans* fatty acids also occur as mixtures of positional isomers.

Occurrence

Trans fatty acids present in the diet arise from two origins. The first is from bacterial biohydrogenation in the forestomach of ruminants, which is the source of *trans* fatty acids present in mutton and beef fats. These are present at a concentration of 2–9% of bovine fat. *Trans*-11-octadecenoic acid is the main isomer produced although *trans*-9- and *trans*-10-octadecenoic acid are also produced. Thus, *trans* fatty acids occur in nature and cannot be considered to be foreign substances.

The second origin is from the industrial catalytic hydrogenation of liquid oils (mainly of vegetable origin, but also of fish oils). This produces solid fats and partially hydrogenated oils and is undertaken to increase the thermal stability of liquid oils and to alter their physical properties. The margarines, spreads, shortenings, and frying oils produced are thus more useful in the food processing industry than liquid oils. Chemically, a range of *trans* isomers is produced that, for vegetable oils containing predominantly C_{18} unsaturated fatty acids, is qualitatively similar to those produced by biohydrogenation, although the relative proportions of the

isomers may differ. Use of fish oils containing a high proportion of very long-chain (C_{20} and C_{22}) fatty acids with up to six double bonds produces more complex mixtures of *trans*, *cis*, and positional isomers. However, the use of hydrogenated fish oils in food processing is declining, owing to a general fall in edible oil prices and to consumer preference for products based on vegetable oils.

Analysis

Methods available for the estimation of total *trans* unsaturation and to determine individual *trans* fatty acids are outlined in **Table 1**. At present there is no one simple and accurate method suitable for both research applications and for use in the food industry. In dietary studies data for *trans* fatty acid intake are generally expressed as the sum of the fatty acids containing *trans* double bonds, and there is generally no differentiation between the different isomers.

A report from the British Nutrition Foundation (BNF) in 1995 highlighted concerns over the variations in estimations of *trans* fatty acid concentrations in some food products provided by different analytical techniques. A thorough review of the available analytical techniques was called for.

Sources and Intakes

The main sources of *trans* fatty acids in the UK diet are cereal-based products (providing 27% of total *trans* fatty acid intake), margarines, spreads, and frying oils (22%), meat and meat products (18%), and milk, butter, and cheese (16%). In the USA, the main sources of intake are baked goods (28%), fried foods (25%), margarine, spreads, and shortenings (25%), savory snacks (10%), and milk and butter (9%).

Typical ranges of *trans* fatty acids in foods are shown in **Table 2**. *Trans* isomers of $C_{18:1}$ (elaidic acid) are the most common *trans* fatty acids, accounting for 65% of the total *trans* fatty acids in the UK diet.

Intakes of *trans* fatty acids are difficult to assess because of:

- analytical inaccuracies;
- difficulties of obtaining reliable information about food intake.

A number of countries have attempted to assess intakes of *trans* fatty acids (**Table 3**). Reliable intake data are available for the UK, based on a 7-day weighed intake of foods eaten both inside and outside the home, for 2000 adults aged 16–64 years (**Table 3**). Data from the UK National Food Survey,

Table 1 Analytical methods for *trans* fatty acids

General method	*Determines*	*Advantages*	*Disadvantages*
Infrared (IR) absorption spectrometry	Total *trans* unsaturation	Inexpensive; reliable results provided concentrations of *trans* isomers exceed 5%; can analyze intact lipids	Unreliable results if concentrations of *trans* isomers less than 5%; interpretive difficulties—need to apply correction factors
Fourier transform IR spectroscopy	Total *trans* unsaturation	Reliable results if concentrations of *trans* isomers less than 2%	Does not distinguish between two esters each with one *trans* bond or between one ester with two *trans* bonds and one with none
Gas–liquid chromatography (GLC)	Individual *trans* fatty acids		Presence of unidentified compounds can give false estimates of *trans* fatty content
Argentation—GLC	Individual *trans* fatty acids	Saturated, monounsaturated, and diunsaturated fatty acids can be resolved	Method is time-consuming
Capillary column GLC	Individual *trans* fatty acids which can be summated to give total *trans* unsaturation	Accurate resolution of fatty acid esters including *cis* and *trans* isomers	Great skill required for preparing columns and interpretation of chromatograms
High-performance liquid chromatography	Individual *trans* fatty acids	*cis,cis*- and *trans,trans*-dienoic fatty acids can be separated	
Nuclear magnetic resonance (NMR)	Individual *trans* fatty acids	Intact lipids can be analyzed; can identify *trans*-diene isomers by use of proton (^{1}H) NMR	Equipment is costly; more use as a research tool than for general analysis

Table 2 Typical content of *trans* fatty acids in a range of foods

Food	*Content of* trans *fatty acids per 100 g product (g)*
Butter	3.6
Soft margarine, not high in PUFA	9.1
Soft margarine, high in PUFA	5.2
Hard margarine	12.4
Low-fat spread, not high in PUFA	4.5
Low-fat spread, high in PUFA	2.5
Blended vegetable oil	1.1
Vegetable oil (sunflower, safflower, soya, sesame)	0
Commercial blended oil	6.7
Potato crisps	0.2
Whole wheat crisps	0.2
Low-fat crisps	0.3
Beefburger, 100% beef frozen, fried, or grilled	0.8
Sausage, pork, fried	0.1
Sausage roll, flaky pastry	6.3
Hamburger in bun with cheese, take-away	0.5
Biscuits, cheese-flavored	0.2
Biscuits, chocolate, full coated	3.4
Chocolate cake and butter icing	7.1
Chips, old potatoes, fresh, fried in commercial blended oil	0.7
Chips, frozen, fine cut, fried in commercial blended oil	0.7

PUFA, polyunsaturated fatty acids. *Trans* fatty acid methyl esters were determined by capillary gas chromatography.
Reproduced with kind permission of the British Nutrition Foundation.

which does not include food eaten outside the home, show a steady decline in intake of *trans* fatty acids from 5.6 g per person per day in 1980 to 4.8 g in 1992. In the UK *trans* fatty acids account for approximately 6% of dietary fat, and in the USA for approximately 7–8% of dietary fat. Estimates of *trans* fatty acid intake are likely to show a downward trend because of:

- improved analytical techniques which give lower but more accurate values for the *trans* fatty acid content of foods;
- the availability of values for *trans* fatty acids in a wider range of foods which allows more accurate estimation of intakes;
- the reformulation of some products which has led to a reduction in the concentration of *trans* fatty acids in recent years.

Advances in food technology that are enabling a gradual reduction in the *trans* fatty acid content include:

1. refinements in hydrogenation processing conditions which will enable the reduction and in the future, the elimination of *trans* fatty acids;
2. the interesterification (rearrangement of fatty acids within and between triacylglycerols) of liquid oils with solid fats;
3. the future genetic modification of oils.

Table 3 Estimated intakes of *trans* fatty acids in various countries

Country	*Estimated daily intake of total* trans *fatty acids (g)*	*Year published and basis for estimation*
UK	5.6 (men) 4.0 (women)	1990: 7-day weighed intake undertaken in 1986–87 including food eaten outside the home
USA	8.1 3.8	1991: availability data 1994: food frequency questionnaire
Denmark	5.0	1995: availability data
Finland	1.9	1992: duplicate diets
Spain	2.0–3.0	1993: calculated from food consumption data
Norway	8.0	1993: food frequency questionnaire

Physiology of *trans* Fatty Acids

Extensive reviews of the health effects of *trans* fatty acids conducted in the 1980s found no evidence for any adverse effects of *trans* fatty acids on growth, longevity, reproduction, or the occurrence of disease, including cancer, from studies conducted in experimental animals.

Digestion, Absorption, and Metabolism

Trans fatty acids are present in the diet in esterified form, mainly in triacylglycerols but those from ruminant sources may also be present in phospholipids. Before absorption into the body, triacylglycerols must be digested by pancreatic lipase in the upper small intestine. There is no evidence of differences in the hydrolysis and absorption of *trans* fatty acids, in comparison with that of *cis* fatty acids. *Trans* fatty acids are transported from the intestine mainly in chylomicrons, but some are also incorporated into cholesteryl esters and phospholipids.

Trans fatty acids are incorporated into the lipids of most tissues of the body and are present in all the major classes of complex lipids. The positional distribution of *trans* fatty acids tends to show more similarity to that of saturated fatty acids than to that of the corresponding *cis* fatty acids. Some selectivity between tissues results in an uneven distribution of *trans* fatty acids throughout the body.

Trans fatty acids occur mainly in positions 1 and 3 of triacylglycerols, the predominant lipids in adipose tissue. The concentration of *trans* fatty acids in adipose tissue is approximately proportional to long-term dietary intake, and determination of the concentrations in storage fat is one method used to estimate *trans* fatty acid intake. However, this is not entirely straightforward as variation has been reported in the composition of adipose tissue obtained from different sites and depths, and factors that influence adipose tissue turnover rates such as dieting and exercise are also complicating factors. *Trans*-18:1 isomers account for approximately 70% of the *trans* fatty acids found in adipose tissue, and *trans*-18:2 isomers (*trans,trans, trans,cis*, and *cis,trans*) account for about 20%.

In heart, liver, and brain, *trans* fatty acids occur mainly in membrane phospholipids. The position of the double bond as well as the conformation of the carbon chain may determine the pattern of *trans* fatty acid esterification in phospholipids, but there is evidence that *trans*-18:1 fatty acids are preferentially incorporated into position 1 of the phosphoacylgly-cerols, as are saturated fatty acids; in contrast, oleic acid is randomly distributed.

The turnover of *trans* fatty acids parallels that of other types of fatty acids in the body, and *trans* fatty acids are readily removed from the tissues for oxidation. Studies in which human subjects were fed labelled carbon-13 isotope have demonstrated that the whole-body oxidation rate for *trans*-18:1 is similar to that for *cis*-18:1. *Trans* fatty acids are a minor component of tissue lipids, and their concentrations in tissues are much lower than their concentrations in the diet. However, research has focused on C_{18} *trans* fatty acids, and more studies are needed to investigate the effects of very long-chain *trans* fatty acids derived from the hydrogenation of fish oils.

Interactions with Metabolism of Essential Fatty Acids

From experiments mainly with laboratory animals, it has been demonstrated that relatively high intakes of *trans* fatty acids in the diet in conjunction with marginal intakes of essential fatty acids (less than 2% dietary energy from linoleic acid) can lead to the presence of Mead acid (*cis*-5,8,11–20:3) in tissue lipids and an increase in the ratio of 20:3 n-9 to 20:4 n-6. This has been interpreted to suggest early signs of essential fatty acid deficiency, with potentially increased requirements for essential fatty acids. Mead acid can accumulate in the presence of linoleic acid, if large amounts of nonessential fatty acids are also present. Two mechanisms have been suggested to explain these observations in relation to intake of *trans* fatty acids:

- that *trans* fatty acids may compete with linoleic acid in metabolic pathways;
- that *trans* fatty acids may inhibit enzymes involved in elongation and further desaturation of linoleic acid.

The consensus is that the significance of Mead acid production in humans has not been established, and further research is needed in this area. It is unlikely that a competitive effect between polyunsaturated fatty acids (PUFA) and *trans* fatty acids would arise, because of the relatively high intakes of linoleic acid in people freely selecting their own diets. Also, as there is a large body pool of linoleic acid available for conversion to long-chain PUFA, it is unlikely that the *trans* fatty acids in the body would interfere even at relatively low ratios of dietary linoleic acid to *trans* fatty acids. The appearance of Mead acid is not specifically induced by *trans* fatty acids, and experiments in animals have not demonstrated any adverse health effects of its production.

Effect of *trans* Fatty Acids on Plasma Lipoproteins

Raised plasma concentrations of low-density lipoprotein (LDL) are considered to be a risk factor for coronary heart disease (CHD); in contrast, reduced concentrations of high-density lipoprotein (HDL) are considered to increase risk. It therefore follows that to help protect against CHD, diets should ideally help to maintain plasma concentrations of HDL cholesterol and to lower those of LDL cholesterol. Dietary factors that raise LDL and lower HDL concentrations would be considered to be undesirable in this context.

Several trials have evaluated the effects of C_{18} *trans* monounsaturated fatty acids on plasma lipoproteins (**Figure 3**). The results have been relatively consistent, and the following general conclusions have been drawn from these studies:

- C_{18} monounsaturated *trans* fatty acids raise LDL cholesterol concentration; the cholesterol-raising effect is similar in magnitude to that of the cholesterol-raising saturated fatty acids, i.e., myristic (14:0) and palmitic (16:0) acids.
- C_{18} monounsaturated *trans* fatty acids decrease HDL cholesterol concentration; this is in contrast to saturated fatty acids which produce a small rise in HDL levels.
- In comparison with the effects of oleic and linoleic fatty acids, C_{18} monounsaturated *trans* fatty acids raise LDL cholesterol and lower HDL cholesterol levels.

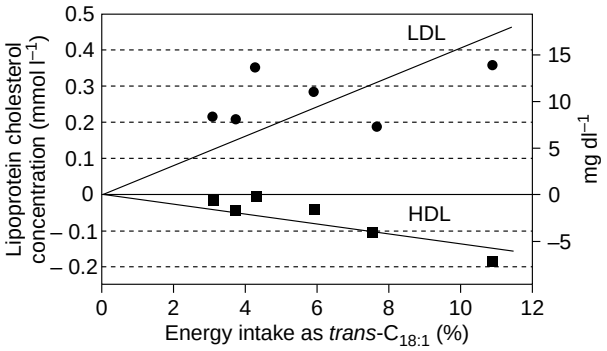


Figure 3 Effects of monounsaturated C_{18} *trans* fatty acids on lipoprotein cholesterol concentrations relative to oleic acid (*cis*-$C_{18:1}$). Data are derived from six dietary comparisons between *trans* and *cis* monounsaturated fatty acids; differences between diets in fatty acids other than *trans* and *cis* monounsaturated fatty acids were adjusted for by using regression coefficients from a meta-analysis of 27 controlled trials. The regression lines were forced through the origin because a zero change in intake will produce a zero change in lipoprotein concentrations. From Zock *et al.* (1995), reproduced with kind permission of the *American Journal of Clinical Nutrition*.

It has been calculated that 'theoretically', each 1% increase in energy from *trans* fatty acids (18:1) in place of oleic acid (*cis*-18:1) would raise plasma LDL concentration by 0.040 mmol l^{-1} (an approximately 1% increase based on average UK plasma cholesterol concentration); HDL would be decreased by 0.013 mmol l^{-1} (a 1% decrease).

The 1995 BNF Task Force calculated that, in the UK, replacing 2% energy from *trans* fatty acids with 2% energy from oleic acid would reduce mean plasma LDL cholesterol concentration by 0.08 mmol l^{-1}; plasma HDL concentration would rise by 0.026 mmol l^{-1}, and the HDL ratio would fall from 3.92 to 3.77. From estimates of the effect of changes in LDL and HDL concentrations on CHD risk, this was predicted to reduce the risk of CHD by 5–15%. In comparison, replacing *trans* fatty acids with either saturated fatty acids or carbohydrate would decrease risk by up to 8%.

The influence of *trans* fatty acids on plasma lipoproteins in relation to CHD risk would thus appear to be more unfavorable than that of saturated fatty acids, as determined by the effect on the ratio of LDL to HDL cholesterol. However, the overall magnitude of the effect would be dependent on the relative intakes of *trans* fatty acids and saturated fatty acids. In the UK *trans* fatty acids contribute about 2% of dietary energy, in contrast to saturated fatty acids which contribute about 15% dietary energy, and this needs to be considered when formulating dietary advice. The Task Force also estimated, on the same basis, that a reduction of 6% in energy from saturated fatty acids would decrease risk by 37%.

However, these conclusions of the adverse effect of *trans* fatty acid on plasma lipoprotein concentrations are not universally accepted. It has been commented that some trials used an inappropriate basis for comparison of the different diets and did not always control for other fatty acids that are known to influence blood cholesterol levels.

Several studies have suggested that *trans* fatty acids raise the plasma concentration of lipoprotein(a), particularly in individuals with already raised levels. Lipoprotein(a) has been suggested to be an independent risk marker for CHD, although this is not universally accepted.

Atherosclerosis and Hemostasis

Despite the reported effects of *trans* fatty acids on blood lipoproteins, experiments with laboratory animals have not provided evidence that dietary *trans* fatty acids are associated with the development of experimental atherosclerosis, provided that the diet contains adequate levels of linoleic acid. Similarly, there is no evidence that *trans* fatty acids raise blood pressure or affect the blood coagulation system. However, there has been no thorough evaluation of the effect of *trans* fatty acids on the coagulation system, and this is an area worthy of investigation.

The Role of *trans* Fatty Acids in Coronary Heart Disease

A number of epidemiological studies have suggested an association between *trans* fatty acids and CHD.

Case–Control Studies

A study by Ascherio in 1994 demonstrated that in subjects who had suffered acute myocardial infarction (AMI), past intake of *trans* fatty acids, assessed from a food frequency questionnaire, was associated with increased risk. *Trans* fatty acid intake per day in the top quintile was 6.5 g compared with 1.7 g in the lowest quintile. After adjusting for age, energy intake, and sex, relative risk of a first AMI for the highest compared with the lowest quintile was 2.44 (95% confidence interval, 1.42–4.10). However, there was not a clear dose–response relationship.

A case–control study of sudden cardiac death found that higher concentrations of *trans* isomers of linoleic acid in adipose tissue, compared with lower concentrations, were associated with increased risk of sudden death. After controlling

for smoking and making an allowance for social class, this relationship became insignificant.

A multicenter study in eight European countries plus Israel found that the risk of AMI was not significantly different across quartiles of the concentration of *trans*-18:1 fatty acids in adipose tissue, the multivariate odds ratio being 0.97 (95% confidence interval, 0.56–1.67) for the highest compared with the lowest quartiles. However, there were significant differences within countries. In Norway and Finland, relative risk was significantly increased in the highest compared with the lowest quartiles, but in Russia and Spain relative risk was significantly decreased in these groups. Exclusion from the multicenter analysis of the Spanish centers, which had particularly low intakes of *trans* fatty acids, resulted in a tendency to increased risk of AMI in the highest quartiles of *trans*-18:1 concentration. However, the trend was not statistically significant, and adjustment for confounding factors had no effect on the results.

A Prospective Study

The relationship between *trans* fatty acid intake and subsequent CHD events was investigated in approximately 85 000 US nurses (the Nurses Health Study). *Trans* fatty acid intake was calculated from food frequency questionnaires for women who had been diagnosed free from CHD, stroke, diabetes, and hypercholesterolemia. The subjects were followed up for 8 years and CHD events were recorded. The relative risk of CHD in the highest compared with the lowest quintile was 1.5 (95% confidence interval, 1.12–2.0), after adjustment for age, energy intake, social class, and smoking. However, there was no clear dose–response relationship between the highest and the lowest intake groups. The intake of *trans* fatty acids in the top quintile was 3.2% dietary energy compared with 1.3% in the lowest quintile.

It has been commented on that the benefit predicted by the authors, that individuals in the top quintile of intake could halve their risk of myocardial infarction by reducing their intake of *trans* fatty acids to that of the lowest quintile, seems a large effect in view of the small difference in intakes between these groups (3.3 g). The changes in plasma lipoprotein cholesterol concentrations that would be predicted to occur as a result of lowering *trans* fatty acid intake would not explain all of the observed increase in risk. Also, the study was carried out in a selected population of women and it is unclear that the findings are applicable to the whole population or to other population groups.

Cancer

Although there is much evidence concerning the effect of different intakes of different types of fats on experimental carcinogenesis, data for *trans* fatty acids are limited and are hampered by confounding due to the lack of a suitable control diet.

Studies using different tumor models in mice and rats have shown no effect of *trans* fatty acids on tumor development. Increasing the intake of *trans* fatty acids, in place of *cis* fatty acids, has not demonstrated an adverse outcome with regard to cancer risk. In humans, there is little to suggest that *trans* fatty acids are adversely related to cancer risk at any of the major cancer sites. Early studies did not generally find that *trans* fatty acids were an important risk factor for malignant or benign breast disease. One study did report an association between the incidence of cancer of the colon, breast, and prostate and the use of industrially hydrogenated vegetable fats in the USA; however, other known risk factors were not allowed for.

Cancer of the Breast

Some epidemiological evidence suggests that total fat intake may be related to increased risk of cancer of the breast, although this is by no means conclusive. There is no strong evidence that intake of *trans* fatty acids *per se* is related to increased risk of breast cancer and many studies have not reported examining this relationship. A study in which adipose tissue concentrations of *trans*-18:1 fatty acids were assessed in 380 women with breast cancer at various stages and in controls revealed no consistent pattern of association. A similar, smaller study suggested an increased risk with higher body stores of *trans* fatty acids, but it was concluded that any such association may be modified by adipose tissue concentrations of polyunsaturated fatty acids.

Cancer of the colon

Epidemiological data from the Nurses Health Study suggested a link between intake of meat and meat products and colon cancer. The data indicated that high intakes of total, animal, saturated, and monounsaturated fat were associated with increased risk. Consuming beef, pork, or lamb as a main dish was positively associated with risk; though beef and lamb contain *trans* fatty acids, there was no evidence that high intakes of *trans* fatty acids increased risk. The Health Professionals Follow-up Study (a prospective study in male health professionals of parallel design to the Nurses Health Study) found similar dietary associations for colon cancer risk,

with no suggestion of any link with intake of *trans* fatty acids.

Prostate Cancer

Dietary associations with risk of prostate cancer were also assessed from the Health Professionals Follow-up Study. High intakes of total, saturated, and monounsaturated fatty acids and of α-linolenic acid were associated with increased risk, whereas high intakes of saturated fatty acids and linoleic acid were found to be protective. Intake of *trans* fatty acids was not found to be associated with risk of prostate cancer.

Dietary Guidelines

The details of population dietary guidelines for the quality and quantity of fat intake differ between countries. However, in consideration of prevention of CHD, dietary guidelines generally reflect advice to reduce average total fat intakes to 30–35% dietary energy and to lower saturated fat intakes to approximately 10% of dietary energy. Though the effect of *trans* fatty acids on the plasma LDL/HDL ratio is less favorable than that of saturated fatty acids, dietary advice needs to reflect the relative intakes of these two types of fatty acids. Since the contribution of saturated fat intake to dietary energy is approximately 5–7 times higher than that of *trans* fatty acids, advice on *trans* fatty acids should not assume more importance than advice to lower saturated fatty acids. However, because of the unfavorable effect of *trans* fatty acids on plasma lipoprotein concentrations, the 1995 BNF Task Force report concluded that the average intake of *trans* fatty acids in the UK diet (2% of energy) should not rise, and that dietary advice should continue to focus on reducing intake of saturated fatty acids as a priority.

Extreme consumers of *trans* fatty acids may be at greater risk, and individuals with high intakes may benefit from advice to lower their intake. It has been calculated that lowering total fat and increasing carbohydrate intake (which reduces plasma HDL cholesterol) will have minimal effect on risk of CHD. Substituting *cis* unsaturated fatty acids for saturated and *trans* fatty acids would be predicted to have a greater impact. For individuals at risk of CHD, high intakes of *trans* fatty acids would appear to be undesirable. Most dietary guidelines call for a reduction in *trans* fatty acids "as much as possible" recognizing that small amounts of *trans* fats are naturally present in the food chain. The food industry in general has reduced or eliminated *trans* in many products by improving manufacturing techniques and reducing *trans* fats generated during the hydrogenation process. Additionally, several countries now have regulations requiring that *trans* fatty acid be listed on products' labels. In some cases, like in the US, *trans* must be added to the saturated fat content reported on the Nutrition Facts label, based on their similar adverse effects on health. Although listing both fats together may not be correct in chemical terms, it is a practical way to allow consumers to quickly assess the content of unhealthy fat in a product.

Conclusions

Several lines of evidence indicate that *trans* fats have an adverse effect on the lipoprotein profile and likely on risk of cardiovascular disease. Although reducing intake of *trans* fats is a desirable goal, public health policy should keep as a central recommendation the reduction of saturated fats, which constitute four to six times more percent calories in the diet than *trans* fats. Since the largest proportion of *trans* fats is generated during food processing, industry bears the main responsibility for reducing the *trans* content of its products, thus helping the general public to lower their intake of this type of fat.

See also: **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Cholesterol**: Factors Determining Blood Levels. **Coronary Heart Disease**: Lipid Theory. **Dairy Products**. **Dietary Guidelines, International Perspectives**. **Dietary Intake Measurement**: Methodology; Validation. **Dietary Surveys**. **Food Composition Data**. **Hyperlipidemia**: Nutritional Management. **Lipids**: Chemistry and Classification; Composition and Role of Phospholipids. **Lipoproteins**. **Meat, Poultry and Meat Products**. **Socio-economic Status**.

Further Reading

AIN/ASCN (1996) Position paper on *trans* fatty acids. *American Journal of Clinical Nutrition* **63**: 663–670.

Aro AV, Kardinal AFM, Salminen I *et al.* (1995) Adipose tissue isomeric *trans* fatty acids and the risk of myocardial infarction in different countries: the EURAMIC study. *Lancet* **345**: 273–278.

Ascherio A, Hennekens CH, Buring JE *et al.* (1994) *Trans* fatty acids intake and risk of myocardial infarction. *Circulation* **89**: 94–101.

Berger KG (1996) In *Lipids and Nutrition: Current Hot Topics.* Bridgwater: PJ Barnes.

British Nutrition Foundation (1995) *Trans* Fatty Acids. Report of the British Nutrition Foundation Task Force. London: BNF.

Giovannucci E, Rimm E, Colditz GA *et al.* (1993) A prospective study of dietary fat and risk of prostate cancer. *Journal of the National Cancer Institute* **85**: 1571–1579.

Giovannucci E, Rimm EB, Stampfer MJ *et al.* (1994) Intake of fat, meat and fiber in relation to risk of colon cancer in men. *Cancer Research* **54**: 2390–2397.

Gurr MI (1996) Dietary fatty acids with *trans* unsaturation. *Nutrition Research Reviews* **9**: 259–279.

International Life Sciences Institute (1995) *Trans* fatty acids and coronary heart disease risk. Report of the expert panel on *trans* fatty acids and coronary heart disease. *American Journal of Clinical Nutrition* **62**: 655S–707S.

Ip C and Marshall JR (1996) *Trans* fatty acids and cancer. *Nutrition Reviews* **54**: 138–145.

Mensink RP and Katan MB (1990) Effect of dietary fatty acids on high density and low density lipoprotein levels in healthy subjects. *New England Journal of Medicine* **323**: 439–444.

Willett WC, Stampfer MJ, Manson JE *et al.* (1993) Intake of *trans* fatty acids and risk of coronary heart disease among women. *Lancet* **341**: 581–585.

Zock PL, Katan MB, and Mensink RP (1995) Dietary *trans* fatty acids and lipoprotein cholesterol. *American Journal of Clinical Nutrition* **61**: 617.

FERTILITY

R E Frisch, Harvard Center for Population and Development Studies, Cambridge, MA, USA

Linking Body Fat and Reproduction

It is well documented that women who are underweight, or too lean, because of injudicious dieting, excessive athletic activity, or both, experience disruption of their reproductive ability. It is also well documented that moderate weight loss, approximately 10–15% of normal weight for height, unassociated with anorexia nervosa (where weight loss is approximately 30% below ideal weight), results in amenorrhea due to hypothalamic dysfunction. Weight loss in this moderate range is equivalent to a loss of one-third of body fat. If the excessive leanness occurs before menarche, menarche may be delayed until as late as the age of 19 or 20 years.

In addition to the disruptive effects of weight loss and athletic activity on the menstrual cycle, women who exercise moderately or who are regaining weight into the normal range may have a menstrual cycle that appears to be normal but that actually has a shortened luteal phase or is anovulatory. All of these partial or total disruptions of reproductive ability are usually reversible, after varying periods of time, following weight gain, decreased athletic training, or both.

Excessive fatness is also associated with infertility in women; fertility is restored by loss of weight. Too little or too much fat are thus both associated with infertility. It is hypothesized that these associations are causal, and that the high percentage of body fat—26–28% in women after completion of growth—is necessary for and may directly influence reproduction.

The basic question is how does the hypothalamus, the part of the brain controlling reproduction, 'know' how much fat is stored in the body? The discovery of leptin, a protein hormone made by body fat cells of women and men, provides the biochemical link between body fat and reproductive ability. Friedman and colleagues cloned a gene that encoded leptin. Leptin controls food intake, energy metabolism, and reproduction. Receptors for leptin are located in the hypothalamus, the part of the brain that controls all three of these functions in addition to temperature control and stress.

Even before the leptin discovery, the 'critical fatness' hypothesis of Frisch and McArthur led to the prediction of minimum or threshold weights for height for the onset and maintenance of regular ovulatory menstrual cycles. These weights have been found to be useful clinically as target weights for the restoration of ovulatory cycles in cases of amenorrhea due to weight loss. Both the absolute and the relative amounts of fat are important since the lean mass and the fat must be in a particular absolute range as well as a relative range (i.e., the woman must be large enough to reproduce successfully).

Why Fat? The Energy Cost of Reproduction

A human pregnancy and lactation each have a high energy cost: A pregnancy requires approximately 336 MJ (74 000 kcal) in addition to normal metabolic requirements. Lactation requires approximately 2.5 MJ (600 kcal) per day. In premodern times, lactation was an essential part of reproduction.

While the reproductive system is slowly maturing during growth, the body changes in composition as well as in size and proportions. Direct measurements of body water of girls from birth to completion of growth at ages 16–18 years show a continuous

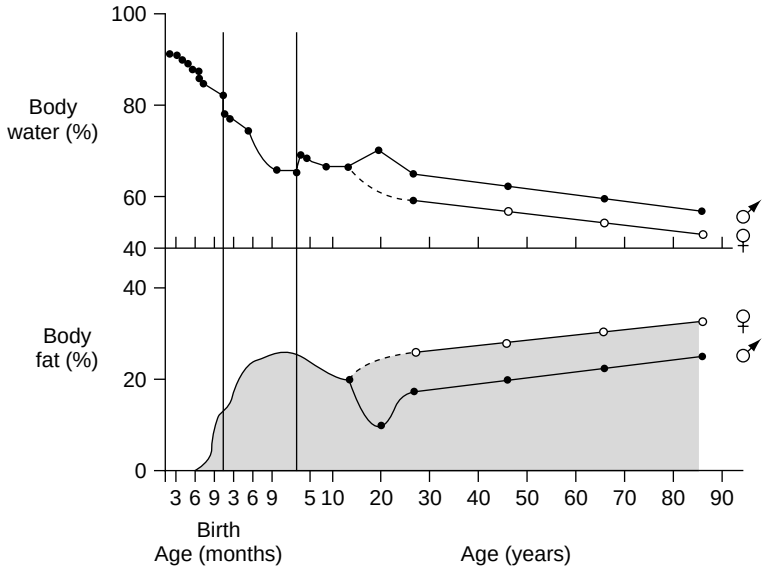


Figure 1 Changes in body water as percentage of body weight throughout the life span, and corresponding changes in the percentage of body fat. (Adapted from Friis-Hansen B (1965) Hydrometry of growth and aging. In: Brozek J (ed.) *Human Body Composition*, vol. 7, Symposia of the Society for the Study of Human Biology, pp. 191–209. Oxford: Pergamon Press.)

decline in the proportion of body water because girls have a large relative increase in body fat (**Figure 1**). This decrease is particularly rapid during the adolescent growth spurt in height and weight, which precedes menarche.

At the completion of growth, between ages 16 and 18 years, the body of a well-nourished woman contains approximately 26–28% fat and approximately 52% water, whereas the body of a man at completion of growth contains approximately 12% fat and 61% water. A young girl and boy of the same height and weight (**Table 1**) differ markedly in the percentages of body water and fat. The main function of the 16 kg of stored female fat, which is equivalent to more than 600 MJ (144 000 kcal), may be to provide energy for a pregnancy and for approximately 3 months of lactation. In prehistoric times when the food supply was scarce or fluctuated seasonally, stored fat would have been necessary for successful reproduction. Fat is the most labile component of body weight. Body fat therefore would reflect environmental changes in food supplies more rapidly than other tissues.

Table 1 Total body water as percentage of body weight: an index of fatness[a]

Variable	*Girl*	*Boy*
Height (cm)	165.0	165.0
Weight (kg)	57.0	57.0
Total body water (l)	29.5	36.0
Lean body weigt (kg)	41.0	50.0
Fat (kg)	16.0	7.0
Fat/body weight (%)	28.0	12.0
Total body water/body weight (%)	51.8	63.0

[a]Comparison of an 18-year-old girl and a 15-year-old boy of the same height and weight.
Lean body weight = total body water/0.72.
Fatness/body weight % = 100 – [(total body water/body wt %)/0.72].

Body Weight and Infant Survival

Infant survival is correlated with birth weight, and birth weight is correlated with the prepregnancy weight of the mother and, independently, her weight gain during pregnancy. From a teleologic and evolutionary view, it is economical to hypothesize that the physical ability to deliver a viable infant and the hypothalamic control of reproduction are synchronized. Adipose tissue may be the synchronizer.

Other Inputs of Adipose Tissue on Female Reproductive Ability

In addition to the message of leptin to the hypothalamus on the amount of body fat, adipose tissue makes other hormones that may directly affect ovulation and the menstrual cycle and, hence, fertility:

1. Adipose tissue is a significant extragonadal source of estrogen. Conversion of androgen to estrogen takes place in the adipose tissue of the breast and abdomen, the omentum, and the fatty marrow of the long bones. This conversion accounts for approximately one-third of the circulating estrogen of premenopausal women and is the main source of

estrogen in postmenopausal women. Men also convert androgen into estrogen in body fat.
2. Body weight, and hence fatness, influences the direction of estrogen metabolism to more potent or less potent forms. Very thin women have an increase in the 2-hydroxylated form of estrogen, which is relatively inactive and has little affinity for the estrogen receptor. Lean female athletes also have an increase in the 2-hydroxylated form of estrogen. In contrast, obese women metabolize less of the 2-hydroxylated form and have a relative increase in the 16-hydroxylated form, which has potent estrogenic activity.
3. Obese women and young girls who are relatively fatter have a diminished capacity for estrogen to bind to serum sex hormone-binding globulin (SHBG); this results in an elevated percentage of free serum estradiol. Since SHBG regulates the availability of estradiol to the brain and other target tissues, the changes in the proportion of body fat to lean mass may influence reproductive performance through the intermediate effects of SHBG.
4. The adipose tissue of obese women stores steroid hormones.

Changes in relative fatness may also affect reproductive ability indirectly through disturbance of the regulation of body temperature and energy balance by the hypothalamus. Very lean women, both anorexic and nonanorexic, display abnormalities of temperature regulation in addition to delayed response, or lack of response, to exogenous luteinizing hormone-releasing hormone.

Hypothalamic Dysfunction, Gonadotropin Secretion, and Weight Loss

It is now recognized that the amenorrhea of underweight and excessively lean women is due to hypothalamic dysfunction. Hypothalamic dysfunction has also been implicated in the amenorrhea of athletes. Consistent with the view that this type of amenorrhea is adaptive, the pituitary–ovarian axis is apparently intact and functions when exogenous gonadotropin-releasing hormone (GnRH) is given in pulsatile form or in a bolus.

Women with this type of hypothalamic amenorrhea have both quantitative and qualitative changes in the secretion of the gonadotropins, luteinizing hormone (LH) and follicle-stimulating hormone (FSH), and of estrogen:

1. Levels of LH, FSH, and estradiol levels are low.
2. The secretion of LH and the response to GnRH are reduced in direct correlation with the amount of weight loss.
3. Underweight patients respond to exogenous GnRH with a pattern of secretion similar to that of prepubertal children; the FSH response is greater than the LH response. The return of LH responsiveness is correlated with weight gain.
4. The maturity of the 24-h LH secretory pattern and body weight are related. Weight loss results in an age-inappropriate secretory pattern resembling that of prepubertal or early pubertal children. Weight gain restores the postmenarcheal secretory pattern.
5. A reduced response or absence of response to clomiphene, a pituitary hormone which stimulates ovulation, is correlated with the degree of the loss of body weight and hence of fat. A normal response occurs after weight gain to the normal range.

Supportive of the view that this type of hypothalamic amenorrhea is adaptive is the finding of one study that women in whom ovulation had been induced had a higher risk of having babies who were small for their age, and this risk was greatest (54%) in those who were underweight. The authors of this study concluded that the most suitable treatment for infertility secondary to weight-related amenorrhea is dietary rather than induction of ovulation.

Physiological Basis of Reproductive Ability

Weight at Menarche

The idea that relative fatness is important for female reproductive ability followed findings that the events of the adolescent growth spurt, particularly menarche in girls, were closely related to an average critical body weight. This result was unexpected for human beings, although it was well-known for rats and monkeys that puberty (defined by vaginal opening or, more precisely, by first estrus) was more closely related to body weight than to chronological age.

In the United States, the mean weight at menarche for girls was 47.8 ± 0.5 kg at a mean height of 158.5 ± 0.5 cm and mean age of 12.9 ± 0.1 years. This mean age included girls from Denver, who had a slightly later age of menarche than that of sea-level populations due to the slowing effect of altitude on prenatal and postnatal weight growth.

Secular Trend Toward an Earlier Age of Menarche

Even before the meaning of the critical weight for an individual girl was analyzed, the idea that menarche is associated with a critical weight for a population

explained simply many observations associated with early or late menarche. Observations of earlier menarche are associated with attaining the critical weight more quickly. The most important example is the secular (long-term) trend to an earlier menarche of approximately 3 or 4 months per decade in Europe in the past 100 years (**Figure 2**). The explanation is that children become larger sooner; therefore, girls on average reach 46 or 47 kg, the mean weight at menarche of US and many European populations, more quickly. Theoretically, the secular trend should end when the weight of children of successive cohorts remains the same because of the attainment of maximum nutrition and child care; this has happened in the United States.

Conversely, a late menarche is associated with body weight growth that is slower prenatally, postnatally, or both so that the average critical weight is reached at a later age: Malnutrition delays menarche, twins have later menarche than do singletons of the same population, and high altitude delays menarche.

Components of Weight at Menarche

Individual girls have menarche at varied weights and heights. To make the notion of a critical weight meaningful for an individual girl, the components of body weight at menarche were analyzed. We investigated body composition at menarche because total body water (TW) and lean body weight (LBW; TW/0.72) are more closely correlated with metabolic rate than is body weight since they represent the metabolic mass as a first approximation. Metabolic rate was considered to be an important clue since Kennedy hypothesized a food intake–lipostat–metabolic signal to explain his elegant findings on weight and puberty in the rat.

The greatest change in estimated body composition of both early and late-maturing girls during the adolescent growth spurt was a large increase in body fat from approximately 5 to 11 kg (a 120% increase) compared to a 44% increase in lean body weight. There was thus a change in the ratio of lean body weight to fat from 5:1 at initiation of the spurt to 3:1 at menarche. The shortest, lightest girls at menarche had a smaller absolute amount of fat (8.9 ± 0.4 kg) compared to the tallest, heaviest girls (12.3 ± 0.6 kg) (the mean of all subjects was 11.5 ± 0.3 kg). However, both extreme groups had approximately 22% of their body weight as fat at menarche, as did all subjects, and the ratio of lean body weight to fat of both groups was in the range of 3:1, as it was for all subjects.

Since adipose tissue can convert androgens to estrogens, the relative degree of fatness can be

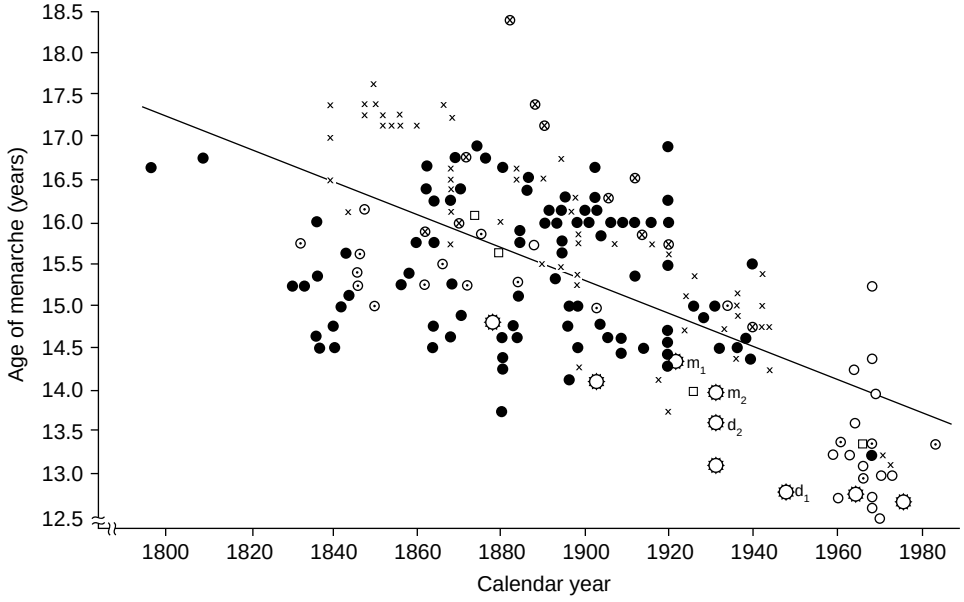


Figure 2 Mean or median age of menarche as a function of calendar year from 1790 to 1980. The symbols refer to England ⊙; France •; Germany ⊗; Holland □; Scandinavia (Denmark, Finland, Norway, and Sweden) X; Belgium, Czechoslovakia, Hungary, Italy, Poland (rural), Romania (urban and rural), Russia (15.2 years at an altitude of 2500 m and 14.4 years at 700 m), Spain, and Switzerland, all labeled ○; and the United States ☼ (data not included in the regression line). Twenty-seven points were identical and do not appear on the graph. The regression line cannot be extended indefinitely. The age of menarche has already leveled off in some European countries, as it has in the United States. (From Wyshak and Frisch (1982) Evidence for a secular trend in age of menarche. *New England Journal of Medicine* **306**: 1033–1035, with permission from the *New England Journal of Medicine*.)

directly related to the quantity of circulating estrogen. The biological effectiveness of the estrogen is also related to body weight. Rate of fat gain, therefore, is a neat mechanism for relating rate of growth, nutrition, and physical work to the energy requirements for reproduction.

Fatness as a Determinant of Minimal Weights for Menstrual Cycles

As shown in **Table 1** and **Figure 1**, total body water as a percentage of body weight is an index of fatness. In a study in the United States, this index in each of the 181 girls followed from menarche to the completion of growth at ages 16–18 years provided a method of determining a minimal weight for height necessary for menarche in primary amenorrhea and for the resumption of normal ovulatory cycles in cases of secondary amenorrhea, when the amenorrhea was due to undernutrition or intensive exercise. These weights have been found useful in the evaluation and treatment of patients with primary or secondary amenorrhea due to weight loss.

Percentiles of total body water/body weight, which are percentiles of fatness, were made at menarche and for the 181 girls at age 18 years, the age at which body composition stabilized. Patients with amenorrhea due to weight loss, other possible causes having been excluded, were studied in relation to the weights indicated by the diagonal percentile lines of total water/body weight percent (**Figure 3**). It was found that 56.1% of total water/body weight, the 10th percentile at age 18 years (equivalent to approximately 22% fat of body weight), indicated a minimal weight for height necessary for the restoration and maintenance of menstrual cycles. For example, a 20-year-old woman whose height is 165 cm (65 in.) should weigh at least 49 kg (108 lb) before menstrual cycles would be expected to resume (**Figure 3**).

The weights at which menstrual cycles ceased or resumed in postmenarcheal patients age 16 years and older were approximately 10% greater than the minimal weights for the same height observed at menarche (**Figure 4**). The explanation was that both early and late-maturing girls gain an average of 4.5 kg of fat from menarche to age 18 years. Almost all of this gain is achieved by age 16 years, when mean fat is 15.7 ± 0.3 kg, 27% of body weight. At age 18 years, mean fat is 16.0 ± 0.3 kg, 28% of the mean body weight of 57.1 ± 0.6 kg. Reflecting this increase in fatness, the total water/body weight percent decreases from $55.1 \pm 0.2\%$ at menarche (12.9 ± 0.1 years in this sample) to $52.1 \pm 0.2\%$ (standard deviation 3.0) at age 18 years.

Because girls are less fat at menarche than when they achieve stable reproductive ability, the minimal weight for onset of menstrual cycles in cases of primary amenorrhea due to undernutrition or exercise is indicated by the 10th percentile of fractional body water at menarche, 59.8%, which is equivalent to approximately 17% of body weight as fat. For example, a 15-year-old girl whose completed height is 165 cm (65 in.) should weigh at least 43.6 kg (96 lb) before menstrual cycles can be expected to begin (**Figure 4**).

The minimum weights indicated in **Figure 4** would be used also for girls who become amenorrheic as a result of weight loss soon after menarche, as often occurs in cases of anorexia nervosa in adolescent girls.

The absolute and relative increase in fatness from menarche to ages 16–18 years coincides with the period of adolescent subfecundity. During this time, there is still rapid growth of the uterus, ovaries, and oviducts.

Other factors such as emotional stress affect the maintenance or onset of menstrual cycles. Therefore, menstrual cycles may cease without weight loss and may not resume in some subjects even though the minimum weight for height has been achieved. Also, these standards apply only to Caucasian US females and European females since different races have different critical weights at menarche and it is not known whether the different critical weights represent the same critical body composition of fatness.

Since the prediction of the minimum weights for height is based on total water/body weight percent (not fat to body weight percent), successful prediction may be related to the ratio of lean mass to fat, which is normally approximately 3:1 at menarche and 2.5:1 at the completion of growth at age 18 years. No prediction can be made above the threshold weight for a particular height.

Physical Exercise, Delayed Menarche, and Amenorrhea

Does intense exercise cause delayed menarche and amenorrhea of athletes, or do late maturers choose to be athletes and dancers? We found that the mean age of menarche of 38 college swimmers and runners was 13.9 ± 0.3 years, significantly later ($p < 0.001$) than that of the general population (12.8 ± 0.05 years), in accord with other reports. However, the mean menarcheal age of the 18 athletes whose training began before menarche was 15.1 ± 0.5 years, whereas the mean menarcheal age

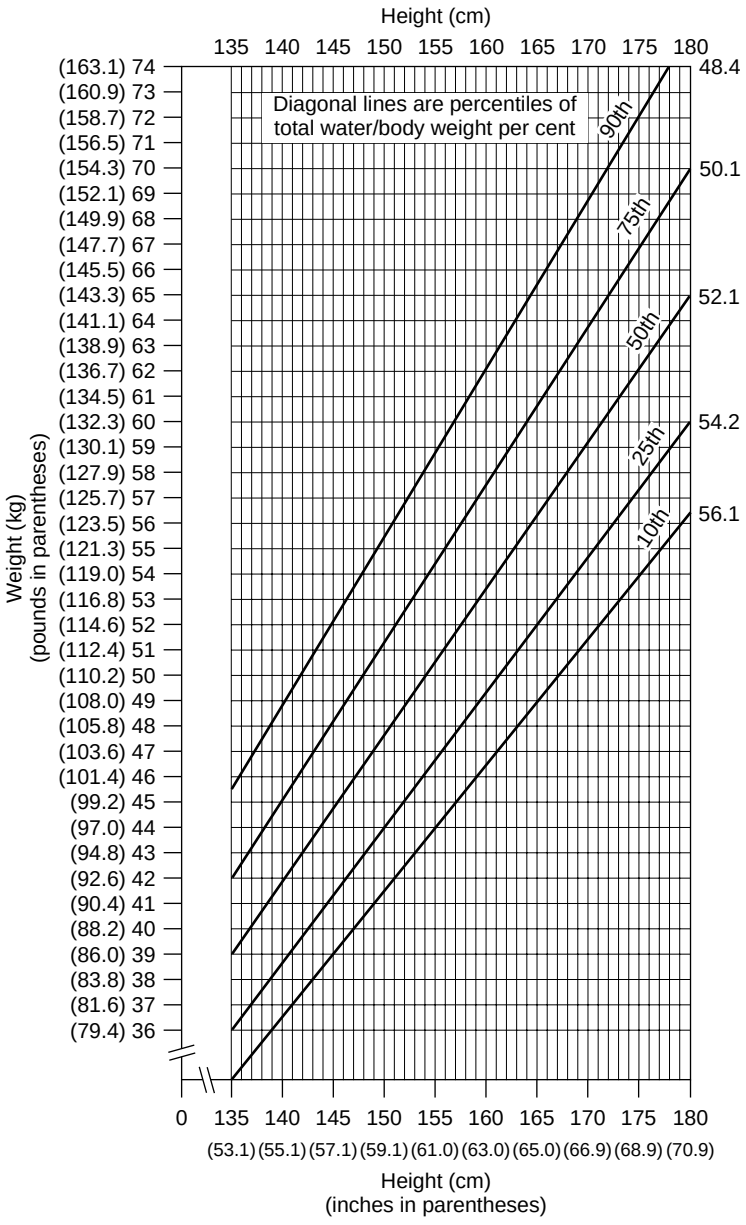


Figure 3 The minimal weight necessary for a particular height for restoration of menstrual cycles is indicated on the weight scale by the 10th percentile diagonal line of total water/body weight percent, 56.1%, as it crosses the vertical height line. For example, a 20-year-old woman whose height is 165 cm (65 in.) should weigh at least 49 kg (108 lb) before menstrual cycles would be expected to resume. (Adapted from Frisch and McArthur (1974) Menstrual cycles: Fatness as a determinant of minimum weight for height necessary for their maintenance or onset. *Science* **185**: 949–951, with permission from *Science*.)

of the 20 athletes whose training began after menarche was 12.8 ± 0.2 years ($p < 0.001$). The latter mean age was similar to that of the college controls (12.7 ± 0.4 years) and the general population. Therefore, training, not preselection, is the delaying factor. Each year of premenarcheal training delayed menarche by 5 months (0.4 years). This suggests that one constructive way to reduce the incidence of teenage pregnancy would be to have girls join teams at age 8 or 9 years and maintain regular moderate exercise. Such a program might reduce the risk of serious diseases of women in later life, as discussed later.

Training also directly affected the regularity of the menstrual cycles during the training year. Of the premenarche trained athletes, only 17% had regular cycles; 61% were irregular and 22% were amenorrheic. In contrast, 60% of the postmenarche trained athletes were regular, 40% were irregular, and none were amenorrheic. However, during intense

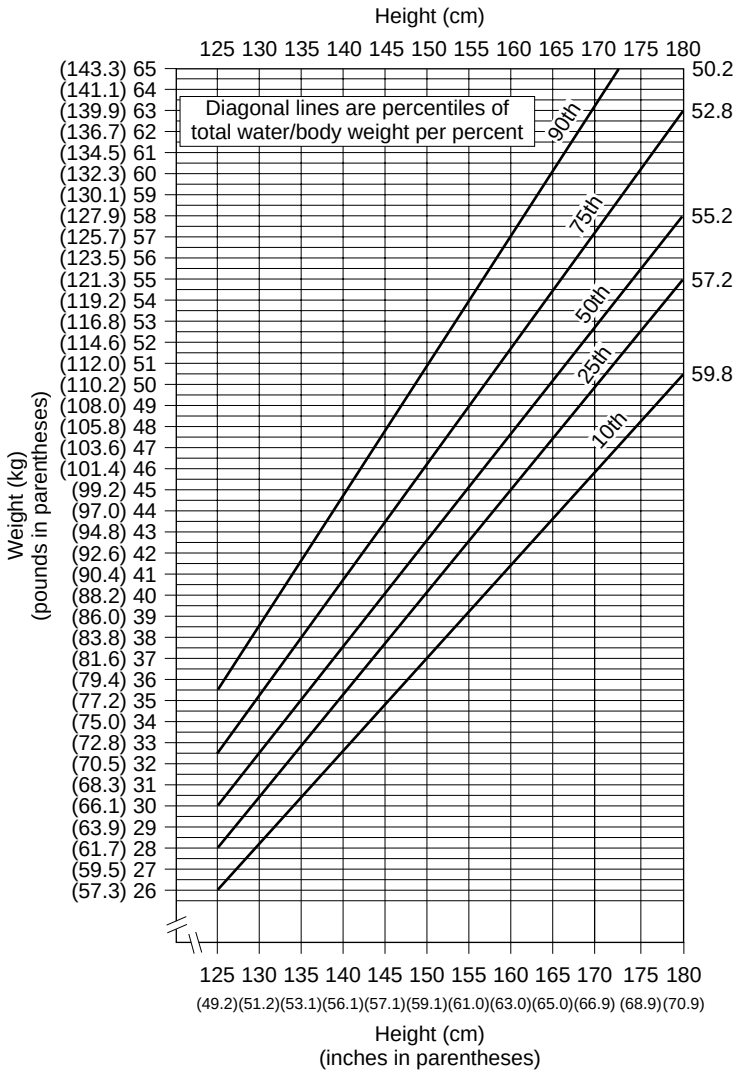


Figure 4 The minimal weight necessary for a particular height for onset of menstrual cycles is indicated on the weight scale by the 10th percentile diagonal line of total water/body weight percent, 59.8%, as it crosses the vertical height lines. The height growth of girls must be completed or approaching completion. For example, a 15-year-old girl whose completed height is 165 cm (65 in.) should weigh at least 43.6 kg (96 lb) before menstrual cycles can be expected to start. (Adapted from Frisch and McArthur (1974) Menstrual cycles: Fatness as a determinant of minimum weight for height necessary for their maintenance or onset. *Science* **185**: 949–951, with permission from *Science*.)

training, the incidence of oligomenorrhea and amenorrhea increased in both groups.

As other workers have found, plasma gonadotropins and estrogen levels were in the low-normal range in this study for the athletes with irregular cycles or amenorrhea. Progesterone was at the follicular phase level. Thyroid hormones, however, were in the normal range. These athletes had increased muscularity and decreased adiposity compared to nonathletes. The explanation for their menstrual disturbances may therefore be the same as that for dieting, nonathletic women—too little fat in relation to lean mass. Some of the swimmers and track and field athletes were above average weight for height. A raised lean mass:fat ratio may nevertheless have caused their menstrual problems because their body weight represented a greater amount of muscle and less adipose tissue than the same weight of nonathletic women.

Psychologic Stress and Changes in Weight

The psychologic stress of competition, which may increase the secretion of adrenal corticosteroids and catecholamines, thus affecting the hypothalamic control of gonadotropins, may be involved, but stress does not seem to be the main factor in many individuals.

Nutrition and Male Reproduction

Undernutrition delays the onset of sexual maturation in boys in a similar way to the delaying effect of undernutrition on menarche. Undernutrition and weight loss in men also affect their reproductive ability. The sequence of effects, however, is different from that in females. In men, loss of libido is the first effect of a decrease in energy intake and subsequent weight loss. Continued energy reduction and weight loss result in a loss of prostate fluid and then decreases in sperm motility and sperm longevity. Sperm production ceases when weight loss is approximately 25% of normal body weight. Refeeding results in a restoration of function in the reverse order of loss.

Effects of Exercise on Men

Men marathon runners have been shown to have decreased hypothalamic GnRH secretion. Also reported are changes in serum testosterone levels with weight loss in wrestlers, a reduction in serum testosterone and prolactin levels in male distance runners, and changes in reproductive function and development in relation to physical activity.

Body Mass Index

Recognition of the importance of relative fatness levels for general health and successful reproductive outcome has led to the use of the body mass index (BMI) to estimate relative fatness levels.

BMI is calculated as weight (kg)/height $(m)^2$.

A BMI of 20–25 kg/m^2 is recommended for good health and is associated with normal fertility.

A weight for height equivalent to a BMI of 18 kg/m^2 and lower is considered too low for successful reproductive ability. Research indicates that the hormonal environment for a successful pregnancy outcome will be improved with a BMI higher than 19 kg/m^2.

A BMI of 25–27 kg/m^2 is associated with a slight reduction in fertility, and that higher than 27 kg/m^2 is associated with a significantly reduced fertility.

BMI standards may not apply to athletes who train regularly because of their increased muscle mass.

Magnetic Resonance Imaging of Body Fat of Athletes and Controls

Using magnetic resonance imaging for direct quantification of body fat showed that athletes who did not differ in body weight from nonathletes had 30–40% less fat than the nonathletes. Muscles are heavy (80% water), so the body weight of an athlete does not necessarily indicate body composition. Athletes had a more sensitive insulin response to a glucose tolerance test compared to controls. The insulin area under the curve of athletes and controls was significantly related to their total fat as a percentage of total volume, determined by magnetic resonance imaging.

Athletes with menstrual disorders had significantly decreased subcutaneous and internal fat, overall and at all regional sites, compared to controls. The extent of estradiol 2-hydroxylation to 2-hydroxyoestrone, determined by radiometric analysis, was significantly ($p = 0.005$) inversely related to total fat as a percentage of total volume and to subcutaneous fat as a percentage of total volume ($p = 0.004$) overall and at each of the regional fat depots. This inverse relationship may be a determinant of the anovulatory cycles and amenorrhea of excessively lean women by a feedback to the hypothalamus since 2-hydroxyoestrone is antiestrogenic.

Long-Term Regular Exercise Lowers the Risk of Sex Hormone-Sensitive Cancers

The amenorrhea and delayed menarche of athletes raised the question: Are there differences in the long-term reproductive health of athletes with moderate training compared to nonathletes?

A study of 5398 college graduates ages 20–80 years, of whom 2622 were former athletes and 2776 were nonathletes, showed that the former athletes had a significantly lower lifetime occurrence of breast cancer and cancers of the reproductive system compared to the nonathletes. More than 82.4% of the former college athletes began their training in high school or earlier, compared to 24.9% of the nonathletes. The analysis controlled for potential confounding factors, including age, age of menarche, age of first birth, smoking, and cancer family history. The relative risk (RR) for nonathletes compared to athletes for cancers of the reproductive system was 2.53 (95% confidence limit (CL), 1.17–5.47) (**Figure 5**). The RR for breast cancer was 1.86 (95% CL, 1.00–3.47). The former college athletes were leaner in every age group compared to the nonathletes.

Although one can only speculate as to the reasons for the lower risk, the most likely explanation is that long term, the former athletes had lower levels of estrogen because they were leaner, and more estrogen was metabolized to the nonpotent catechol estrogens. Also, the former athletes may have consumed diets lower in fat and saturated fat. Such diets shift the pattern of estrogen metabolism toward the less active catechol estrogens.

Compared to the nonathletes, the former college athletes also had a lower lifetime occurrence

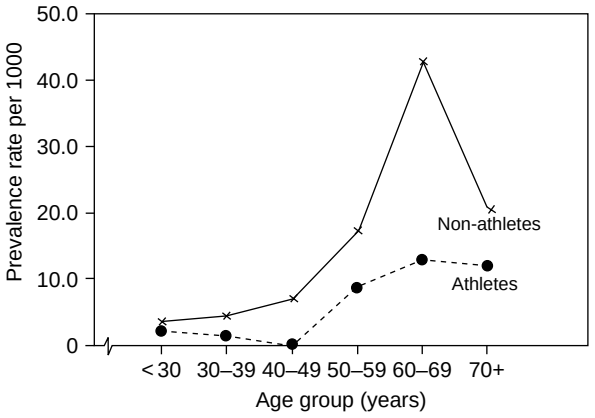


Figure 5 Prevalence rate of cancers of the reproductive system for athletes (circles) and nonathletes (crosses) by age group. (From Frisch RE, Wyshak G, Albright NL *et al.* (1985) Lower prevalence of breast cancer and cancers of the reproductive system among former college athletes compared to non-athletes. *British Journal of Cancer* **52**: 885–891, with permission from the *British Journal of Cancer*.)

(prevalence) of benign tumors of the breast and reproductive system; a lower prevalence of diabetes, particularly after age 40 years; and no greater risk of bone fractures, including risk of wrist and hip fractures, in the menopausal period.

These data indicate that long-term exercise, which was not at Olympic or marathon level but moderate and regular, reduces the risk of sex hormone-sensitive cancers and the risk of diabetes for women in later life. Data showing that moderate exercise also reduces the risk of nonreproductive system cancers suggest that other factors, such as changes in immunosurveillance, may also be involved.

Nutrition, Physical Work, and Natural Fertility

The effects of hard physical work and nutrition on reproductive ability suggest that differences in the fertility of populations, historically and today, may be explained by a direct pathway from food intake to fertility (**Figure 6**), in addition to the classic Malthusian pathway through mortality. Charles Darwin described this commonsense direct relationship between food supplies and fertility, observing the following:

1. Domestic animals that have regular, plentiful food without working to get it are more fertile than the corresponding wild animals.
2. "Hard living retards the period at which animals conceive."
3. The amount of food affects the fertility of the same individual.
4. It is difficult to fatten a cow that is lactating.

All of Darwin's dicta apply to human beings.

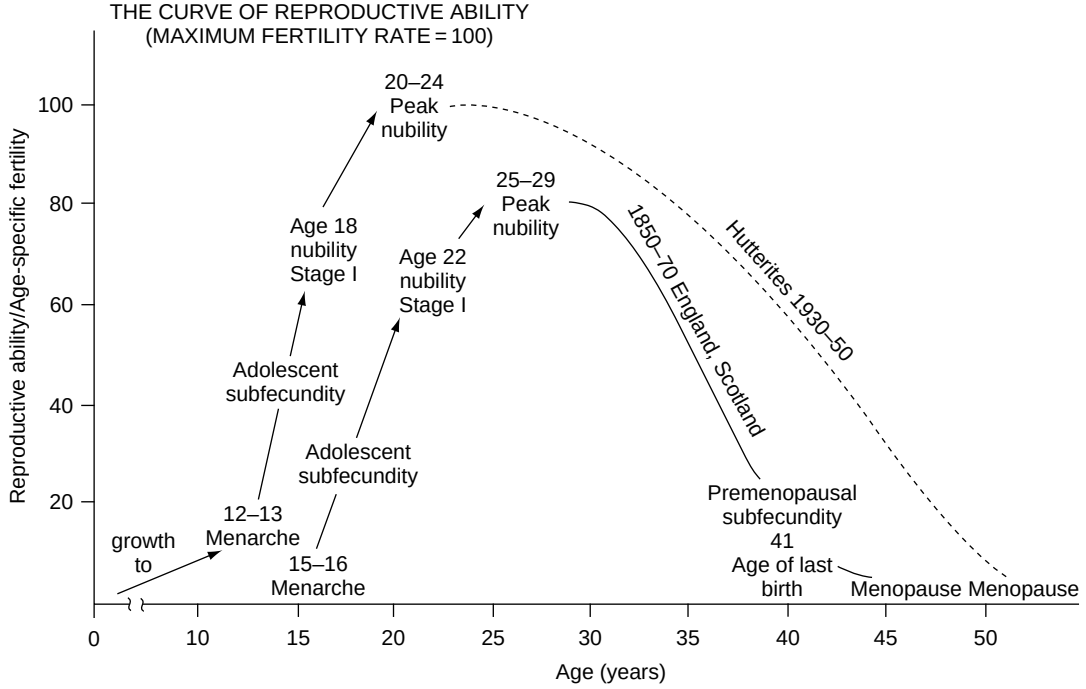


Figure 6 The mid-nineteenth century curve of female reproductive ability (variation of the rate of childbearing with age) compared with that of the well-nourished, modern Hutterites who do not use contraception. The Hutterite fertility curve (broken line) results in an average of 10–12 children; the 1850–1870 fertility curve (solid line) in approximately 6–8 children. (From Frisch (1978) Population, food intake and fertility. *Science* **199**: 22–30, with permission from *Science*.)

The Paradox of Rapid Population Growth in Undernourished Populations

In many historical populations with slow population growth, poor couples living together to the end of their reproductive lives had only 6 or 7 living births. Most poor couples in many developing countries today also have 6 or 7 living births during their reproductive life span. This total fertility rate is far below the human maximum of 11 or 12 children observed among well-nourished couples not using contraception, such as the Hutterites. However, 6 children per couple today in developing countries results in a very rapid rate of population growth because of decreased mortality rates due to the introduction of modern public health procedures. The difference between the birth rate per 1000 and the death rate per 1000, which gives the percentage growth rate, is currently as high as 2–4%. Populations growing at 2, 3, and 4% double in 35, 23, and 18 years, respectively.

British data from the mid-nineteenth century on growth rates, food intake, age-specific fertility, sterility, and ages of menarche and menopause show that females who grew relatively slowly to maturity, completing height growth at ages 20 or 21 years (instead of 16–18 years, as in well-nourished contemporary populations), also differed from well-nourished females in each event of the reproductive span: Menarche was later, for example, 15.0–16.0 years compared with 12.8 years; adolescent sterility was longer, and the age of peak nubility was later; the levels of specific fertility were lower; pregnancy wastage was higher; the duration of lactational amenorrhea was longer; the birth interval was longer; and the age of menopause was earlier, preceded by a more rapid period of perimenopausal decline (**Figure 6**). Thus, the slower, submaximal growth of women to maturity is subsequently associated with a shortened and less efficient reproductive span. The differences in the rate of physical growth of women and men result not only in a displacement of the age-specific fertility curve in time but also in a difference in the ultimate level: The faster the growth of females and males, the earlier and more efficient the reproductive ability.

Endocrinological data show that undernourished women have a longer lactational amenorrhea than do well-nourished women. The amount of suckling is not the only factor, as has been suggested in explaining reduced natural fertility. In addition, age of menarche and the other events of the reproductive span, which are known to be affected by the nutritional state, are pertinent to overall fertility.

See also: **Exercise**: Beneficial Effects; Diet and Exercise. **Growth and Development, Physiological Aspects**. **Low Birthweight and Preterm Infants**: Causes, Prevalence and Prevention. **Obesity**: Definition, Etiology and Assessment; Complications. **Pregnancy**: Role of Placenta in Nutrient Transfer; Nutrient Requirements; Energy Requirements and Metabolic Adaptations; Weight Gain; Safe Diet for Pregnancy; Prevention of Neural Tube Defects; Pre-eclampsia and Diet. **Weight Management**: Weight Maintenance.

Further Reading

Chehab FF, Mounzih K, Ronghva L *et al.* (1997) Early onset of reproductive function in normal female mice treated with leptin. *Science* **275**: 88–90.

Clement K, Vaisse C, Lahlou N *et al.* (1998) A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature* **392**: 398–401.

Friis-Hansen B (1965) Hydrometry of growth and aging. In: Brozek J (ed.) *Human Body Composition*, vol. 7, Symposia of the Society for the Study of Human Biology, pp. 191–209. Oxford: Pergamon Press.

Frisch RE (1978) Population, food intake and fertility. *Science* **199**: 22–30.

Frisch RE (1981) What's below the surface? *New England Journal of Medicine* **305**: 1019–1020.

Frisch RE (ed.) (1990) *Adipose Tissue and Reproduction.* Basel: Karger.

Frisch RE (2002) *Female Fertility and the Body Fat Connection* Chicago: University of Chicago Press.

Frisch RE and McArthur JW (1974) Menstrual cycles: Fatness as a determinant of minimum weight for height necessary for their maintenance or onset. *Science* **185**: 949–951.

Frisch RE, Snow RC, Johnson L *et al.* (1993) Magnetic resonance imaging of overall and regional body fat, estrogen metabolism and ovulation of athletes compared to controls. *Journal of Clinical Endocrinology and Metabolism* **77**: 441–477.

Frisch RE, Wyshak G, Albright NL *et al.* (1985) Lower prevalence of breast cancer and cancers of the reproductive system among former college athletes compared to non-athletes. *British Journal of Cancer* **52**: 885–891.

Frisch RE, Wyshak G, and Vincent L (1980) Delayed menarche and amenorrhea of ballet dancers. *New England Journal of Medicine* **303**: 17–19.

Halaas JL, Gajiwala KS, Maffei M *et al.* (1995) Weight-reducing effects of the plasma protein encoded by the obese gene. *Science* **269**: 543–546.

Hileman SM, Pierroz DD, and Flier JS (2002) Leptin, nutrition and reproduction: Timing is everything. *Journal of Clinical Endocrinology and Metabolism* **85**: 804–807.

Strobel A, Issad T, Camoin L *et al.* (1998) A Leptin missense mutation associated with hypogonadism and morbid obesity. *Nature Genetics* **18**: 213–215.

Vigersky RA, Andersen AE, Thompson RH *et al.* (1977) Hypothalamic dysfunction in secondary amenorrhea associated with simple weight loss. *New England Journal of Medicine* **297**: 1141–1145.

Wyshak G and Frisch RE (1982) Evidence for a secular trend in age of menarche. *New England Journal of Medicine* **306**: 1033–1035.

Fetal Origins of Disease *see* **Early Origins of Disease**: Fetal; Non-Fetal

Fiber *see* **Dietary Fiber**: Physiological Effects and Effects on Absorption; Potential Role in Etiology of Disease; Role in Nutritional Management of Disease

FISH

A Ariño, J A Beltrán, A Herrera and P Roncalés, University of Zaragoza, Zaragoza, Spain

Introduction

In discussing the food uses of fishes, the term 'fish' refers to edible species of finfish, molluscs, and crustacea coming from the marine or freshwater bodies of the world, either by capture fisheries or by aquaculture. Accordingly, "fishery products" means any human food product in which fish is a characterizing ingredient, such as dried, salted, and smoked fish, marinated fish, canned seafood, minced fish flesh such as surimi, and miscellaneous products.

Fish is a source of high-quality animal protein, supplying approximately 6% of the world's protein requirements and 16.4% of the total animal protein. According to Food and Agriculture Organization figures, the contribution of fish to the total animal-protein intake is 26.2% in Asia, 17.4% in Africa, 9.2% in Europe, 9% in the former USSR, 8.8% in Oceania, 7.4% in North and Central America, 7.2% in South America, and 21.8% in the low-income food-deficit countries (including China). There are wide differences among countries in fish consumption measured as the average yearly intake per person, ranging from countries with less than 1.0 kg per person to countries with over 100 kg per person.

Edible fish muscle contains 18–20% protein and 1–2% ash; the percentage of lipids varies from less than 1% to more than 20% (in high-fat finfish), and fish has the added advantage of being low in saturated fat. In general, lean fish is not an important source of calories, which are mostly obtained from the staple carbohydrates in the diet. Fatty fish, however, is a significant energy source in many fish-consuming communities in both the developed and the developing worlds. Today it is recognized that fish is probably more important as a source of micronutrients, minerals, and particularly essential fatty acids than for its energy or protein value. The essential micronutrients and minerals in fish include vitamins A and D, calcium, phosphorus, magnesium, iron, zinc, selenium, fluorine, and iodine (in marine fishes).

The protective effect of a small amount of fish against mortality from coronary heart disease (CHD) has been established by numerous epidemiological studies. A diet including two or three servings of fish per week has been recommended on this basis, and researchers have reported a 50% reduction in CHD mortality after 20 years with intakes of as little as 400 g of fish per week.

It has been suggested that the long-chain omega-3 (*n*-3) polyunsaturated fatty acids (PUFAs) (eicosapentanoic acid (EPA; C20:5) and docosahexanoic acid (DHA; C22:6)) in fish offer this protection against CHD. Several recent studies have shown that a large intake of omega-3 fatty acids is beneficial in lowering blood pressure, reducing triacylglycerols, decreasing the risk of arrhythmia, and lowering the tendency of blood platelets to aggregate.

As fish become more popular, the reports of food-borne diseases attributed to fish have increased. Food-borne diseases linked with exposure to fish can result from the fish itself (i.e., toxic species, allergies) or from bacterial (i.e., *Clostridium botulinum*, *Listeria monocytogenes*, *Salmonella*, *Vibrio*, and *Staphylococcus*), viral (i.e., hepatitis, Norwalk gastroenteritis), or parasitic (i.e., *Anisakis* and related worms) contamination. Also, naturally occurring seafood toxins (i.e., scombrotoxin, ciguatoxins, shellfish poisoning from toxic algae) or the presence of additives and chemical residues due to environmental contamination can cause food-borne illnesses. In recent years, reports of contamination of

some fish with methylmercury have raised concerns about the healthfulness of certain fish for some populations.

General Characteristics of Finfish

A very large number of species of finfish are used for food by the world's population. The dressing percentage of finfish (60–70%) is similar to that of beef, pork, or poultry. The percentage of edible tissue in the dressed carcasses of finfish (without head, skin, and viscera) is higher than that of other food animals, because fishes contain less bone, adipose tissue, and connective tissue. There are three main categories of finfish that are widely used as foods. The bony fishes (teleosts) provide two compositional categories: white fishes (or lean fishes) and fatty fish. The third category is the cartilaginous elasmobranch fishes.

White fishes

The flesh of these fishes is very low in fat and consists primarily of muscle and thin layers of connective tissue. The concentrations of most of the B vitamins are similar to those in mammalian lean meats, although fish may contain higher amounts of vitamins B_6 and B_{12}. The mineral levels are also similar, although the very fine bones that are eaten with the fish flesh can raise the calcium content; fish is also a significant source of iodine. These fishes accumulate oils only in their livers, which are a rich source of vitamin A (retinol), vitamin D, and long-chain PUFAs in their triacylglycerols (TAGs).

Fatty Fishes

These fishes have fat in their flesh, which is usually much darker than that of white fishes, with similar blocks of muscle and connective tissue. The amount of fat is related to the breeding cycle of the fish, so that the fat content falls considerably after breeding. The flesh of fatty fishes is generally richer in the B vitamins than that of white fishes, and significant amounts of vitamins A and D are present. The mineral concentrations are not very different, but fatty fish is a better source of iron. The oil of these fishes is particularly rich in very-long-chain PUFA, especially those of the omega-3 (*n*-3) series such as EPA and DHA. These fishes accumulate oils in their muscles, belly flap, and skin (subdermal fat).

Cartilaginous Fishes

The cartilaginous fishes include the sharks and rays, whose flesh is rich in connective tissue and relatively low in fat, although they do accumulate oils in their livers. The concentrations of vitamins and minerals are very similar to those in white fish. These fishes contain urea in relatively large amounts, and so protein values based on total nitrogen are overestimated. The ammonia smell of cooked sharks and rays is not an indication that the fish is spoiled but rather is the result of enzymatic degradation of urea.

General Characteristics of Shellfish

The term 'shellfish' includes any aquatic invertebrate, such as molluscs or crustaceans, that has a shell or shell-like exoskeleton. The cephalopods have an internal shell (as in squids) or no shell (as in octopods). Owing to the presence of the tough exoskeleton, the edible portion in shellfish (around 40%) is less than that in finfish, with the exception of cephalopods, whose dressing percentage is 70–75%. The lipid content of the edible parts of most shellfish is low, as bivalves store their energy surplus as glycogen and not as depot fat, while crustaceans and cephalopods store their fat in their digestive glands (hepatopancreas). In many fish-eating communities, these foods are very highly valued gastronomically.

Molluscs

A wide range of molluscs are eaten by man, including bivalves (such as mussels, oysters, and scallops), gastropods (such as winkles and whelks), and cephalopods (such as squids and octopuses). The flesh is muscular with low levels of fat, although the fat is more saturated and richer in cholesterol than that of finfish. The mineral levels in shellfish are usually somewhat higher than those in finfish, and the vitamin concentrations are low. Bivalves and gastropods are often eaten whole after boiling or sometimes raw; usually, only the muscular mantles of cephalopods are eaten. In some cultures, only selected parts are eaten; for example, only the white adductor muscle of the scallop is eaten in North America.

Crustaceans

Crustaceans include a range of species, both freshwater (such as crayfish) and marine (such as crabs, shrimps, prawns, and lobsters). These animals have a segmented body, a chitinous exoskeleton, and paired jointed limbs. The portions eaten are the muscular parts of the abdomen and the muscles of the claws of crabs and lobsters. The flesh is characteristically low in fat and high in minerals, with vitamin levels similar to those found in finfish.

Nutritional Value of Fish and Shellfish: Introductory Remarks

Fish and shellfish are excellent sources of protein. A 100 g cooked serving of most types of fish and shellfish provides about 18–20 g of protein, or about a third of the average daily recommended protein intake. The fish protein is of high quality, containing an abundance of essential amino-acids, and is very digestible by people of all ages. Seafood is also loaded with minerals such as iron, zinc, and calcium.

The caloric value of fish is related to the fat content and varies with species, size, diet, and season. Seafood is generally lower in fat and calories than beef, poultry, or pork. Most lean or low-fat species of fish, such as cod, hake, flounder, and sole, contain less than 100 kcal (418 kJ) per 100 g portion, and even fatty fish, such as mackerel, herring, and salmon, contain approximately 250 kcal (1045 kJ) or less in a 100 g serving. Most crustaceans contain less than 1% fat in the tail muscle because depot fat is stored in the hepatopancreas, which is in the head region.

Interest in the health benefits of fish and shellfish began decades ago when researchers noted that certain groups of people – including the Inuit and the Japanese, who rely on fish as a dietary staple – have a low rate of ischemic diseases (i.e., heart attack or stroke). Fish, particularly fatty fish, is a good source of the omega-3 fatty acids EPA and DHA. These fats help to lower serum triacylglycerols and cholesterol, help prevent the blood clots that form in heart attacks, and lower the chance of having an irregular heartbeat. In fact, one study found that women who ate fish at least once a week were 30% less likely to die of heart disease than women who ate fish less than once a month. Similar benefits have been found for men. Fish consumption is also related to slower growth of atherosclerotic plaque and lower blood pressure. Especially good sources of omega-3 fats are salmon, tuna, herring, mackerel, and canned tuna and sardines.

When included in the diet of pregnant and breastfeeding women, DHA is thought to be beneficial to infant brain (learning ability) and eye (visual acuity) development. Scientists have found that women who ate fatty fish while pregnant gave birth to children with better visual development. Babies of mothers who had significant levels of DHA in their diet while breastfeeding experienced faster-than-normal eyesight development. Preliminary research also suggests that a diet rich in omega-3 fatty acids – and in DHA in particular – may help to decrease the chance of preterm birth, thus allowing the baby more time for growth and development.

Recent research found that eating just one serving a week of fish decreased the risk of developing dementia by 30%. Eating fatty fish several times a week may also lower the risk of developing prostate cancer by as much as half. A Swedish study of 3500 postmenopausal women eating two servings of fatty fish a week found that they were 40% less likely to develop endometrial cancer than those eating less than one-fourth of a serving a week.

Eating a variety of fish and seafood, rather than concentrating on one species, is highly recommended for both safety and nutrition. It is recommended that pregnant women should avoid certain species of fish and limit their consumption of other fish to an average of 400 g of cooked fish per week. The reason for this recommendation is that, whereas nearly all fish contain trace amounts of methylmercury (an environmental contaminant), large predatory fish, such as swordfish, shark, tilefish, and king mackerel, contain the most. Excess exposure to methylmercury from these species of fish can harm an unborn child's developing nervous system. It is also suggested that nursing mothers and young children should not eat these particular species of fish.

Fish Lipids

In fish, depot fat is liquid at room temperature (oil) and is seldom visible to the consumer; an exception is the belly flaps of salmon steaks. Many species of finfish and almost all shellfish contain less than 2.5% total fat, and less than 20% of the total calories come from fat. Almost all fish has less than 10% total fat, and even the fattiest fish, such as herring, mackerel, and salmon, contains no more than 20% fat (**Table 1**). In order to obtain a good general idea of the fat contents of most finfish species, flesh color might be considered. The leanest species, such as cod and flounder, have a white or lighter color, while fattier fishes, such as salmon, herring, and mackerel, have a much darker color.

The triacylglycerol depot fat in edible fish muscle is subject to seasonal variation in all marine and freshwater fishes from all over the world. Fat levels tend to be higher during times of the year when fishes are feeding heavily (usually during the warmer months) and in older and healthier individual fishes. Fat levels tend to be lower during spawning or reproduction. When comparing fat contents between farmed and wild-caught food fish, it should be remembered that farmed species have a tendency to show a higher proportion of muscle fat than their wild counterparts. Also, the fatty-acid composition of farmed fish depends on the type of dietary fat

Table 1 Fat levels in marine and freshwater fish and shellfish commonly found in the marketplace

Low (<2.5% fat) less than 20% of total calories from fat	*Medium (2.5–5% fat) between 20% and 35% of total calories from fat*	*High (>5% fat) between 35% and 50% of total calories from fat*
Saltwater fish		
Cod	Anchovy	Dogfish
Grouper	Bluefish	Herring[a]
Haddock	Sea bass	Mackerel[a]
Hake	Swordfish	Salmon[a]
Most flatfishes (flounder, sole, plaice)	Tuna (yellowfin)	Sardine
Pollock		Tuna (bluefin)
Shark		
Skate		
Snapper		
Whiting		
Most crustaceans		
Most molluscs		
Freshwater fish		
Pike	Bream	Catfish (farmed)
Perch, bass	Carp	Eel[a]
Tilapia	Trout (various)	Whitefish

[a]More than 10% fat.

Reproduced with permission from Ariño A, Beltran JA, and Roncalés P (2003) Dietary importance of fish and shellfish. In: Caballero B, Trugo L, and Finglas P (eds.) *Encyclopedia of Food Sciences and Nutrition*, 2nd edn. Oxford: Elsevier Science Ltd. pp. 2471–2478.

used in raising the fish. Cholesterol is independent of fat content and is similar in wild and cultivated fishes.

Most protein-rich foods, including red meat and poultry as well as fish, contain cholesterol. However, almost all types of fish and shellfish contain well under 100 mg of cholesterol per 100 g, and many of the leaner types of fish typically have 40–60 mg of cholesterol in each 100 g of edible muscle. It is known that most shellfish also contain less than 100 mg of cholesterol per 100 g. Shrimp contain somewhat higher amounts of cholesterol, over 150 mg per 100 g, and squid is the only fish product with a significantly elevated cholesterol content, which averages 300 mg per 100 g portion. Fish roe, caviar, internal organs of fishes (such as livers), the tomalley of lobsters, and the hepatopancreas of crabs can contain high amounts of cholesterol.

Omega-3 PUFA in Fish and Shellfish

The PUFA of many fish lipids are dominated by two members of the omega-3 (*n*-3) family, C20:5 *n*-3 (EPA) and C22:6 *n*-3 (DHA). They are so named because the first of several double bonds occurs three carbon atoms away from the terminal end of the carbon chain.

All fish and shellfish contain some omega-3, but the amount can vary, as their relative concentrations are species specific (**Table 2**). Generally, the fattier fishes contain more omega-3 fatty acids than the leaner fishes. The amount of omega-3 fatty acids in farm-raised products can also vary greatly, depending on the diet of the fishes or shellfish. Many companies now recognize this fact and provide a source of omega-3 fatty acids in their fish diets. Omega-3 fatty acids can be destroyed by heat, air, and light, so the less processing, heat, air exposure, and storage time the better for preserving omega-3 in fish. Freezing and normal cooking cause minimal omega-3 losses, whereas deep frying and conditions leading to oxidation (rancidity) can destroy some omega-3 fatty acids.

The beneficial effects of eating fish for human health have been well documented. Research has shown that EPA and DHA are beneficial in protecting against cardiovascular and other diseases (**Table 3**). Studies examining the effects of fish consumption on serum lipids indicate a reduction in triacylglycerol and VLDL-cholesterol levels, a factor that may be protective for some individuals. Research also indicates that EPA in particular reduces platelet aggregation, which may help vessels injured by plaque formation. Fish oils also appear to help stabilize the heart rhythm, a factor that may be important in people recovering from heart attacks.

Table 2 Selected fish and shellfish grouped by their omega-3 fatty-acid content

Low-level group (<0.5 g per 100 g)	*Medium-level group (0.5–1 g per 100 g)*	*High-level group (>1 g per 100 g)*
Finfish		
Carp	Bass	Anchovy
Catfish	Bluefish	Herring
Cod, Haddock, Pollock	Halibut	Mackerel
Grouper	Pike	Sablefish
Most flatfishes	Red Snapper	Salmon (most species)
Perch	Swordfish	Tuna (bluefin)
Snapper	Trout	Whitefish
Tilapia	Whiting	
Shellfish		
Most crustaceans	Clams	
Most molluscs	Oysters	

Reproduced with permission from Ariño A, Beltran JA, and Roncalés P (2003) Dietary importance of fish and shellfish. In: Caballero B, Trugo L, and Finglas P (eds.) *Encyclopedia of Food Sciences and Nutrition*, 2nd edn. Oxford: Elsevier Science Ltd. pp. 2471–2478.

Table 3 Summary of the beneficial effects of eating fish for cardiovascular and other diseases

Cardiovascular disease	*Other diseases*
Protects against heart disease	Protects against age-related macular degeneration
Prolongs the lives of people after a heart attack	Alleviates autoimmune diseases such as rheumatoid arthritis
Protects against sudden cardiac arrest caused by arrhythmia	Protects against certain types of cancer
Protects against stroke (thrombosis)	Mitigates inflammation reactions and asthma
Lowers blood lipids such as triacylglycerols and VLDL-cholesterol	
Lowers blood pressure	

VLDL, very low-density lipoprotein.
Reproduced with permission from Ariño A, Beltran JA, and Roncalés P (2003) Dietary importance of fish and shellfish. In: Caballero B, Trugo L, and Finglas P (eds.) *Encyclopedia of Food Sciences and Nutrition*, 2nd edn. Oxford: Elsevier Science Ltd. pp. 2471–2478.

The major PUFA in the adult mammalian brain is DHA. It is among the materials required for development of the fetal brain and central nervous system and for retinal growth in late pregnancy. Brain growth uses 70% of the fetal energy, and 80–90% of cognitive function is determined before birth. However, the placenta depletes the mother of DHA, a situation that is exacerbated by multiple pregnancies. Dietary enhancement or fortification with marine products before and during pregnancy, rather than after the child is born, would be of great benefit to the child and mother. Furthermore, the food sources that are rich in DHA are also rich in zinc, iodine, and vitamin A, so it may be possible to provide several dietary supplements at one time. Deficiencies of the latter micronutrients are established causes of mental retardation and blindness.

The typical Western diet has a ratio of omega-6 to omega-3 essential fatty acids of between 15:1 and 20:1. Several sources of information suggest that a very high omega-6 to omega-3 ratio may promote many diseases, including cardiovascular disease, cancer, and inflammatory and autoimmune diseases. Fish provides an adequate intake of these omega-3 fats, thus improving the omega-6 to omega-3 fatty-acid ratio. Most experts do not advise the routine use of fish-oil supplements: they favor eating fish and shellfish regularly in the context of a healthy diet and a regular pattern of physical activity. Whereas some research shows benefits of fish-oil supplements, research has also shown that people with weakened immune systems should avoid large doses of fish oil. The final conclusion as to whether it is possible to substitute fish consumption with fish oils or omega-3 fatty-acid supplements, and gain the same reduction in mortality from CHD, awaits more studies. However, the protective role of fish consumption is unquestioned.

Fish Proteins

Both finfish and shellfish are highly valuable sources of proteins in human nutrition. The protein content of fish flesh, in contrast to the fat content, is highly constant, independent of seasonal variations caused by the feeding and reproductive cycles, and shows only small differences among species. **Table 4** summarizes the approximate protein contents of the various finfish and shellfish groups. Fatty finfish and crustaceans have slightly higher than average protein concentrations. Bivalves have the lowest values if the whole body mass is considered (most of them are usually eaten whole), whereas values are roughly average if specific muscular parts alone are consumed; this is the case with the scallop, in which only the adductor muscle is usually eaten.

The essential amino-acid compositions of fish and shellfish are given in **Table 5**. Fish proteins, with only slight differences among groups, possess a high nutritive value, similar to that of meat proteins and slightly lower than that of egg. It is worth pointing out the elevated supply, relative to meat, of essential amino-acids such as lysine, methionine, and threonine. In addition, owing in part to the low collagen content, fish proteins are easily digestible, giving rise to a digestibility coefficient of nearly 100.

The recommended dietary allowances (RDA) or dietary reference intakes (DRI) of protein for human male and female adults are in the range of 45–65 g per day. In accordance with this, an intake of 100 g of fish would contribute 15–25% of the

Table 4 Protein content of the different groups of fish and shellfish

Fish group	*g per 100 g*
White finfish	16–19
Fatty finfish	18–21
Crustaceans	18–22
Bivalves	10–12
Cephalopods	16–18

Reproduced with permission from Ariño A, Beltran JA, and Roncalés P (2003) Dietary importance of fish and shellfish. In: Caballero B, Trugo L, and Finglas P (eds.) *Encyclopedia of Food Sciences and Nutrition*, 2nd edn. Oxford: Elsevier Science Ltd. pp. 2471–2478.

Table 5 Content of essential amino-acids in fish and shellfish (g per 100 g of protein)

Fish group	*Isoleucine*	*Leucine*	*Lysine*	*Methionine*	*Phenylalanine*	*Threonine*	*Tryptophan*	*Valine*
Finfish	5.3	8.5	9.8	2.9	4.2	4.8	1.1	5.8
Crustaceans	4.6	8.6	7.8	2.9	4.0	4.6	1.1	4.8
Molluscs	4.8	7.7	8.0	2.7	4.2	4.6	1.3	6.2

Reproduced with permission from Ariño A, Beltran JA, and Roncalés P (2003) Dietary importance of fish and shellfish. In: Caballero B, Trugo L, and Finglas P (eds.) *Encyclopedia of Food Sciences and Nutrition*, 2nd edn. Oxford: Elsevier Science Ltd. pp. 2471–2478.

total daily protein requirement of healthy adults and 70% of that of children. A look at the dietary importance of the Mediterranean diet is convenient: one of its characteristics is the high consumption of all kinds of fish, chiefly fatty fish. In many Mediterranean countries, fish intake averages over 50 g per day (edible flesh); thus, fish protein contributes over 10% of the total daily protein requirements steadily over the whole year in those countries.

Less well known is the fact that the consumption of fish protein, independently of the effect exerted by fish fat, has been related to a decrease in the risk of atherogenic vascular diseases. In fact, it has been demonstrated that diets in which fish is the only source of protein increase the blood levels of high-density lipoprotein relative to those resulting from diets based on milk or soy proteins.

Nonprotein Nitrogen Compounds in Fish

Nonprotein nitrogen (NPN) compounds are found mostly in the fiber sarcoplasm and include free amino-acids, peptides, amines, amine oxides, guanidine compounds, quaternary ammonium molecules, nucleotides, and urea (**Table 6**). NPN compounds account for a relatively high percentage of the total nitrogen in the muscles of some aquatic animals, 10–20% in teleosts, about 20% in crustaceans and molluscs, and 30–40% (and in special cases up to 50%) in elasmobranchs. In contrast, NPN compounds in land animals usually represent no more than 10% of total nitrogen.

Most marine fishes contain trimethylamine oxide (TMAO); this colorless, odorless, and flavorless compound is degraded to trimethylamine, which gives a 'fishy' odor and causes consumer rejection. This compound is not present in land animals and freshwater species (except for Nile perch and tilapia from Lake Victoria). TMAO reductase catalyzes the reaction and is found in several fish species (in the red muscle of scombroid fishes and in the white and red muscle of gadoids) and in certain microorganisms (*Enterobacteriaceae, Shewanella putrefaciens*).

Creatine is quantitatively the main component of the NPN fraction. This molecule plays an important role in fish muscle metabolism in its phosphorylated form; it is absent in crustaceans and molluscs.

Table 6 Nonprotein nitrogen compounds in several commercially important fish species and mammalian muscle (mg per 100 g wet weight)

Compounds	*Cod*	*Herring*	*Shark species*	*Lobster*	*Mammal*
Total NPN	1200	1200	3000	5500	3500
Total free amino-acids:	75	300	100	3000	350
Arginine	<10	<10	<10	750	<10
Glycine	20	20	20	100–1000	<10
Glutamic acid	<10	<10	<10	270	36
Histidine	<1.0	86	<1.0	—	<10
Proline	<1.0	<1.0	<1.0	750	<1.0
Creatine	400	400	300	0	550
Betaine	0	0	150	100	—
Trimethylamine oxide	350	250	500–1000	100	0
Anserine	150	0	0	0	150
Carnosine	0	0	0	0	200
Urea	0	0	2000	—	35

NPN, nonprotein nitrogen.

Reproduced with permission from Ariño A, Beltran JA, and Roncalés P (2003) Dietary importance of fish and shellfish. In: Caballero B, Trugo L, and Finglas P (eds.) *Encyclopedia of Food Sciences and Nutrition*, 2nd edn. Oxford: Elsevier Science Ltd. pp. 2471–2478.

Endogenous and microbial proteases yield some free amino-acids; taurine, alanine, glycine, and imidazole-containing amino-acids seem to be the most frequent. Glycine and taurine contribute to the sweet flavor of some crustaceans. Migratory marine species such as tuna, characterized by a high proportion of red muscle, have a high content (about 1%) of free histidine. A noticeable amount of this amino-acid has been reported in freshwater carp. The presence of free histidine is relevant in several fish species because it can be microbiologically decarboxylated to histamine. Cooking the fish may kill the bacteria and destroy the enzymes, but histamine is not affected by heat, thus becoming a hazard to consumers. The symptoms of the resulting illness (scombroid poisoning) are itching, redness, allergic symptoms, headache, diarrhea, and peppery taste. Scombroid poisoning is most common after ingesting mahi-mahi, tuna, bluefish, mackerel, and skipjack.

Nucleotides and related compounds generally play an important role as coenzymes. They participate actively in muscle metabolism and supply energy to physiological processes. They have a noticeable participation in flavor; moreover, some of them may be used as freshness indices. Adenosine triphosphate (ATP) is degraded to adenosine diphosphate (ADP), adenosine monophosphate (AMP), inosine monophosphate (IMP), inosine, and hypoxantine. This pattern of degradation takes place in finfish, whereas AMP is degraded to adenosine and thereafter to inosine in shellfish. The degradation chain to IMP and AMP in finfish and shellfish, respectively, is very fast. IMP degradation to inosine is generally slow, except in scombroids and flat fishes. Inosine degradation to hypoxantine is slower. IMP is a flavor potentiator, whereas hypoxanthine imparts a sour taste and it is toxic at high levels. ATP, ADP, and AMP decompose quickly leading to a build-up of inosine and hypoxantine. As this corresponds well to a decline in freshness, the ratio of the quantity of inosine and hypoxantine to the total quantity of ATP and related substances is called the *K*-value and used as a freshness index of fish meat.

Guanosine is an insoluble compound that gives fish eyes and skin their characteristic brightness. It is degraded to guanine, which does not have this property; therefore, brightness decreases until it completely disappears.

The NPN fraction contains other interesting compounds, such as small peptides. Most of them contribute to flavor; besides this, they have a powerful antioxidant activity. Betaines are a special group of compounds that contribute to the specific flavors of different aquatic organisms: homarine in lobster and glycine-betaine, butiro-betaine, and arsenic-betaine in crustaceans. Arsenic-betaine has the property of fixing arsenic into the structure, giving a useful method for studying water contamination.

Fish Vitamins

The vitamin content of fish and shellfish is rich and varied in composition, although somewhat variable in concentration. In fact, significant differences are neatly evident among groups, especially regarding fat-soluble vitamins. Furthermore, vitamin content shows large differences among species as a function of feeding regimes.

The approximate vitamin concentration ranges of the various finfish and shellfish groups are summarized in **Table 7**. The RDA for adults is also given, together with the percentage supplied by 100 g of fish. Of the fat-soluble vitamins, vitamin E (tocopherol) is distributed most equally, showing relatively high concentrations in all fish groups, higher than those of meat. However, only a part of the vitamin E content is available as active tocopherol on consumption of fish, since it is oxidized in protecting fatty acids from oxidation. The presence of vitamins A (retinol) and D is closely related to the fat content, and so they are almost absent in most low-fat groups. Appreciable but low concentrations of vitamin A are found in fatty finfish and bivalve molluscs, whereas vitamin D is very abundant in fatty fish. In fact, 100 g of most fatty species supply over 100% of the RDA of this vitamin.

Water-soluble vitamins are well represented in all kinds of fish, with the sole exception of vitamin C (ascorbic acid), which is almost absent in all of them. The concentrations of the rest are highly variable; however, with few exceptions, they constitute a medium-to-good source of such vitamins, comparable with, or even better than, meat. The contents of vitamins B_2 (riboflavin), B_6 (pyridoxine), niacin, biotin, and B_{12} (cobalamin) are relatively high. Indeed, 100 g of fish can contribute up to 38%, 60%, 50%, 33%, and 100%, respectively, of the total daily requirements of those vitamins. Fatty fish also provides a higher supply of many of the water-soluble vitamins (namely pyridoxine, niacin, pantothenic acid, and cobalamin) than does white fish or shellfish. Crustaceans also possess a relatively higher content of pantothenic acid, whereas bivalve molluscs have much higher concentrations of folate and cobalamin.

Table 7 Vitamin content of the different groups of fish and shellfish (mg or μg per 100 g)

	A (μg)	*D (μg)*	*E (mg)*	*B_1 (mg)*	*B_2 (mg)*	*B_6 (mg)*	*Niacin (mg)*	*Biotin (μg)*	*Pantothenic acid (mg)*	*Folate (μg)*	*B_{12} (μg)*	*C (mg)*
White finfish	Tr	Tr	0.3–1.0	0.02–0.2	0.05–0.5	0.15–0.5	1.0–5.0	1.0–10	0.1–0.5	5.0–15	1.0–5.0	Tr
Fatty finfish	20–60	5–20	0.2–3.0	0.01–0.1	0.1–0.5	0.2–0.8	3.0–8.0	1.0–10	0.4–1.0	5.0–15	5.0–20	Tr
Crustaceans	Tr	Tr	0.5–2.0	0.01–0.1	0.02–0.3	0.1–0.3	0.5–3.0	1.0–10	0.5–1.0	1.0–10	1.0–10	Tr
Molluscs	10–100	Tr	0.5–1.0	0.03–0.1	0.05–0.3	0.05–0.2	0.2–2.0	1.0–10	0.1–0.5	20–50	2.0–30	Tr
Cephalopods	Tr	Tr	0.2–1.0	0.02–0.1	0.05–0.5	0.3–0.1	1.0–5.0	1.0–10	0.5–1.0	10–20	1.0–5.0	Tr
RDA	900	5	15	1.2	1.3	1.3	16	30	5.0	400	2.4	90
% RDA/100 g	0–11	0–100	2–20	1–20	2–38	5–60	1–50	3–33	2–20	0.3–12	40–100	0
% RDA/Md	2	50	7	5	15	25	18	5	8	2	100	0

Tr, trace; RDA, recommended dietary allowance; Md, Mediterranean diet.

Reproduced with permission from Ariño A, Beltran JA, and Roncalés P (2003) Dietary importance of fish and shellfish. In: Caballero B, Trugo L, and Finglas P (eds.) *Encyclopedia of Food Sciences and Nutrition*, 2nd edn. Oxford: Elsevier Science Ltd. pp. 2471–2478.

Table 8 Selected mineral content of the different groups of fish and shellfish (mg per 100 g)

	Na	*K*	*Ca*	*Mg*	*P*	*Fe*	*Zn*	*Mn*	*Cu*	*Se*	*Cr*	*Mo*	*I*
White finfish	50–150	200–500	10–50	15–30	100–300	0.2–0.6	0.2–1.0	0.01–0.05	0.01–0.05	0.02–0.1	0.005–0.02	0.005–0.02	0.01–0.5
Fatty finfish	50–200	200–500	10–200	20–50	200–500	1.0–5.0	0.2–1.0	0.01–0.05	0.01–0.05	0.02–0.1	0.005–0.02	0.005–0.02	0.01–0.5
Crustaceans	100–500	100–500	20–200	20–200	100–700	0.2–2.0	1.0–5.0	0.02–0.2	0.1–2.0	0.05–0.1	0.005–0.02	0.01–0.05	0.01–0.2
Molluscs	50–300	100–500	50–200	20–200	100–300	0.5–10	2.0–10	0.02–0.2	0.02–10	0.05–0.1	0.005–0.02	0.01–0.2	0.05–0.5
Cephalopods	100–200	200–300	10–100	20–100	100–300	0.2–1.0	1.0–5.0	0.01–0.1	0.02–0.1	0.02–0.1	0.005–0.02	0.01–0.2	0.01–0.1
RDA			1000	420	700	8	11	2.3	0.9	0.055	0.035	0.045	0.15
% RDA/100 g			1–20	4–50	15–100	2–50	1–90	0–10	1–100	25–100	15–60	10–100	8–100
% RDA/Md			6	5	30	18	2		2	100			100

RDA, recommended dietary allowance; Md, Mediterranean diet.

Reproduced with permission from Ariño A, Beltran JA, and Roncalés P (2003) Dietary importance of fish and shellfish. In: Caballero B, Trugo L, and Finglas P (eds.) *Encyclopedia of Food Sciences and Nutrition*, 2nd edn. Oxford: Elsevier Science Ltd. pp. 2471–2478.

A Mediterranean diet rich in fish – and especially in fatty finfish – contributes steadily over the year to an overall balanced vitamin supply. The last row of **Table 7** illustrates this; the supply of vitamins D, B_2, B_6, B_{12}, and niacin from this particular diet is more than 15% of the daily requirements; all other vitamins, except ascorbic acid, are supplied to a lesser, but significant, extent.

Fish Minerals

The approximate amounts of selected minerals contained in fish are given in **Table 8**. The first point to note is that all kinds of finfish and shellfish present a well-balanced content of most minerals, either macroelements or oligoelements, with only a few exceptions. Sodium content is low, as in other muscle and animal origin foods. However, it must be remembered that sodium is usually added to fish in most cooking practices in the form of common salt; also, surimi-based and other manufactured foods contain high amounts of added sodium. Potassium and calcium levels are also relatively low, though the latter are higher in fish than in meat; in addition, small fish bones are frequently eaten with fish flesh, thus increasing the calcium intake. Fish is a good source of magnesium and phosphorus, at least as good as meat. These elements are particularly abundant in crustaceans; fatty finfish show elevated levels of phosphorus, and bivalve molluscs have high amounts of magnesium.

Fish is a highly valuable source of most oligoelements. Fatty fish provides a notable contribution to iron supply, similar to that of meat, whereas shellfish have higher concentrations of most dietary minerals. In particular, crustaceans and bivalve molluscs supply zinc, manganese, and copper concentrations well above those of finfish. Worth mentioning is the extraordinary dietary supply of iodine in all kinds of finfish and shellfish; however, this depends on the concentration present in feed, particularly in planktonic organisms.

In summary, 100 g of fish affords low levels of sodium and medium-to-high levels of all the remaining dietary minerals. In fact, it can contribute 50–100% of the total daily requirements of magnesium, phosphorus, iron, copper, selenium, and iodine. A Mediterranean diet, rich in fatty fish and all kinds of shellfish, can lead to an overall balanced mineral supply, which may well reach over 20% of daily requirements of phosphorus, iron, selenium, and iodine.

See also: **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Coronary Heart Disease**: Prevention. **Dietary Guidelines, International Perspectives**. **Fatty Acids**: Omega-3 Polyunsaturated. **Food Composition Data**. **Food Safety**: Bacterial Contamination; Other Contaminants; Heavy Metals. **Hyperlipidemia**: Nutritional Management. **Iodine**: Physiology, Dietary Sources and Requirements. **Protein**: Quality and Sources. **Stroke, Nutritional Management**. **Supplementation**: Dietary Supplements.

Further Reading

Ackman RG (1995) Composition and nutritive value of fish and shellfish lipids. In: Ruiter A (ed.) *Fish and Fishery Products: Composition, Nutritive Properties and Stability*, pp. 117–156. Wallingford: CAB International.

Ariño A, Beltran JA, and Roncalés P (2003) Dietary importance of fish and shellfish. In: Caballero B, Trugo L, and Finglas P (eds.) *Encyclopedia of Food Sciences and Nutrition*, 2nd edn., pp. 2471–2478. Oxford: Elsevier Science Ltd.

Exler J (1987, updated 1992) *Composition of Foods: Finfish and Shellfish Products, Human Nutrition Information Service Agriculture Handbook 8-15* Washington DC: US Department of Agriculture.

Food and Agriculture Organization of the United Nations (1989) *Yield and Nutritional Value of the Commercially More Important Species. FAO Fisheries Technical Paper 309. Rome: Food and Agriculture Organization.*

Food and Drug Administration (1989) *The Fish List, FDA Guide to Acceptable Market Names for Food Fish Sold in Interstate Commerce 1988*. Washington DC: US Government Printing Office.

Haard NF (1995) Composition and nutritive value of fish proteins and other nitrogen compounds. In: Ruiter A (ed.) *Fish and Fishery Products: Composition, Nutritive Properties and Stability*, pp. 77–115. Wallingford: CAB International.

Holland B, Brown J, and Buss DH (1993) Fish and fish products. In: *Supplement to the 5th Edition of McCance and Widdowson's The Composition of Foods.* London: The Royal Society of Chemistry and Ministry of Agriculture, Fisheries and Food.

Huss HH (1995) *Quality and Quality Changes in Fresh Fish*, FAO Fisheries Technical Paper 348. Rome: Food and Agriculture Organization.

Lands WEM (1988) *Fish and Human Health* Orlando, FL: Academic Press.

Lovell RT (1989) *Nutrition and Feeding of Fish* New York: Van Nostrand Reinhold.

National Fisheries Institute. http://www.nfi.org.

Nettleton JA (1993) *Omega-3 Fatty Acids and Health* New York: Chapman & Hall.

Southgate DAT (2000) Meat, fish, eggs and novel protein. In: Garrow JS, James WPT, and Ralph A (eds.) *Human Nutrition and Dietetics*, 10th edn. Edinburgh: Churchill Livingstone.

United States Department of Agriculture. Composition of foods. http://www.nal.usda.gov/fnic/foodcomp.

Valdimarsson G and James D (2001) World fisheries – utilisation of catches. *Ocean and Coastal Management* **44**: 619–633.

Flavonoids *see* **Phytochemicals**: Classification and Occurrence; Epidemiological Factors

Folate *see* **Folic Acid**

FOLIC ACID

J McPartlin, Trinity College, Dublin, Ireland

Introduction

Folic acid was initially distinguished from vitamin B_{12} as a dietary anti-anemia factor by Wills in the 1930s. The subsequent chemical isolation of folic acid and the identification of its role as a cofactor in one-carbon metabolism led to the elucidation of deficiency diseases at the molecular level. The term 'folate' encompasses the entire group of folate vitamin forms, comprising the naturally occurring folylpolyglutamates found in food and folic acid (pteroylglutamic acid), the synthetic form of the vitamin added as a dietary supplement to foodstuffs. 'Folate' is thus the general term used for any form of the vitamin irrespective of the state of reduction, type of substitution, or degree of polyglutamylation.

Folate functions metabolically as an enzyme cofactor in the synthesis of nucleic acids and amino-acids. Deficiency of the vitamin leads to impaired cell replication and other metabolic alterations, particularly related to methionine synthesis. The similar clinical manifestations of cobalamin deficiency and folate deficiency underline the metabolic interrelationship between the two vitamins. Folate deficiency, manifested clinically as megaloblastic anemia, is the most common vitamin deficiency in developed countries. Much attention has focused recently on a number of diseases for which the risks are inversely related to folate status even within the range of blood indicators previously considered 'normal.' Food-fortification programs introduced to prevent neural-tube defects (NTD) have proved effective in increasing folate intakes in populations and may be shown potentially to reduce the risk of cardiovascular disease.

Physiology and Biochemistry

Chemistry and Biochemical Functions

Folic acid (**Figure 1**) consists of a pterin moiety linked via a methylene group to a para-aminobenzoylglutamate moiety. Folic acid is the synthetic form of the vitamin; its metabolic activity requires reduction to the tetrahydrofolic acid (THF) derivative, addition of a chain of glutamate residues in γ-peptide linkage, and acquisition of one-carbon units.

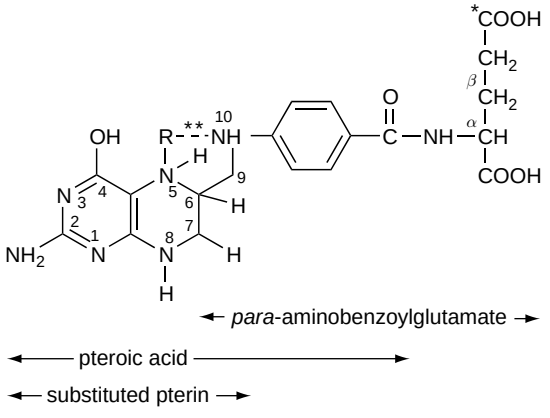


Figure 1 Structural formula of tetrahydrofolate (THF) compounds. In tetrahydrofolic acid R = H; other substituents are listed in **Table 1**. The asterisk indicates the site of attachment of extra glutamate residues; the hatched line and double asterisk indicates the N5 and/or N10 site of attachment of one-carbon units.

Table 1 Structure and nomenclature of folate compounds (see **Figure 1**)

Compound	*R*	*Oxidation state*
5-formylTHF	—CHO	Formate
10-formylTHF	—CHO	Formate
5-formiminoTHF	—CH=NH	Formate
5,10-methenylTHF	—CH=	Formate
5,10-methyleneTHF	$—CH_2—$	Formaldehyde
5-methylTHF	$—CH_3$	Methanol

One-carbon units at various levels of oxidation are generated metabolically and are reactive only as moieties attached to the N5 and/or N10 positions of the folate molecule (**Table 1**).

The range of oxidation states for folate one-carbon units extends from methanol to formate as methyl, methylene, methenyl, formyl, or formimino moieties. When one-carbon units are incorporated into folate derivatives, they may be converted from one oxidation state to another by the gain or loss of electrons.

The source of one-carbon units for folate One-carbon units at the oxidation level of formate can enter directly into the folate pool as formic acid in a reaction catalyzed by 10-formylTHF synthase (**Figure 2**). Entry at the formate level of oxidation can also take place via a catabolic product of histidine, formaminoglutamic acid. The third mode of entry at the formate level of oxidation involves the formation of 5-formylTHF from 5,10-methenylTHF by the enzyme serine hydroxymethyl transferase (SHMT). The 5-formylTHF may be rapidly converted to other forms of folate.

The enzyme SHMT is involved in the entry of one-carbon units at the formaldehyde level of oxidation by catalyzing the transfer of the β-carbon of serine to form glycine and 5,10-methyleneTHF. Other sources of one-carbon entry at this level of oxidation include the glycine cleavage system and the choline-dependent pathway; both enzyme systems generate 5,10-methylene in the mitochondria of the cell.

The removal and use of one-carbon units from folate Single-carbon units are removed from folate by a number of reactions. The enzyme 10-formylTHF dehydrogenase provides a mechanism for disposing of excess one-carbon units as carbon dioxide. (Folate administration to animals enhances the conversion of ingested methanol and formate to carbon dioxide, diminishing methanol toxicity.) Additionally, single-carbon units from 10-formylTHF are used for the biosynthesis of purines (**Figure 2**).

The one-carbon unit of 5,10-methyleneTHF is transferred in two ways. Reversal of the SHMT reaction produces serine from glycine, but since serine is also produced from glycolysis via phosphoglycerate this reaction is unlikely to be important. However, one-carbon transfer from 5,10-methyleneTHF to deoxyuridylate to form thymidylic acid, a precursor of DNA, is of crucial importance to the cell. While the source of the one-carbon unit, namely 5,10-methyleneTHF, is at the formaldehyde level of oxidation, the one-carbon unit transferred to form thymidylic acid appears at the methanol level of oxidation. Electrons for this reduction come from THF itself to generate dihydrofolate as a product. The dihydrofolate must in turn be reduced back to THF in order to accept further one-carbon units.

A solitary transfer of one-carbon units takes place at the methanol level of oxidation. It involves the transfer of the methyl group from 5-methylTHF to homocysteine to form methionine and THF. This reaction is catalyzed by the enzyme methionine synthase and requires vitamin B_{12} as a cofactor. The substance 5-methylTHF is the dominant folate in the body, and it remains metabolically inactive until it is demethylated to THF, whereupon polyglutamylation takes place to allow subsequent folate-dependent reactions to proceed efficiently.

Clinical implications of methionine synthase inhibition The inhibition of methionine synthase due to vitamin B_{12} deficiency induces megaloblastic anemia that is clinically indistinguishable from that caused by folate deficiency. The hematological effect in both cases results in levels of 5,10-methyleneTHF that are inadequate to sustain thymidylate biosynthesis. Clinically, it is essential to ascertain whether the anemia is the result of folate deficiency or vitamin B_{12} deficiency by differential diagnostic techniques. Vitamin B_{12} is essential for the synthesis of myelin in nerve tissue, a function probably related to methionine production from the methionine synthase reaction and the subsequent formation of S-adenosyl-methionine. Hence, vitamin B_{12} deficiency probably leads to nervous disorders in addition to the hematological effects. While the latter respond to treatment with folic acid, the neurological effects do not. Thus, inappropriate administration of folic acid in patients with vitamin B_{12} deficiency may treat the anemia but mask the progression of the neurological defects. Where possible, vitamin B_{12} and folate statuses should be checked before giving folate supplements to treat megaloblastic anemia. The main objection to fortifying food with folate is the potential to mask

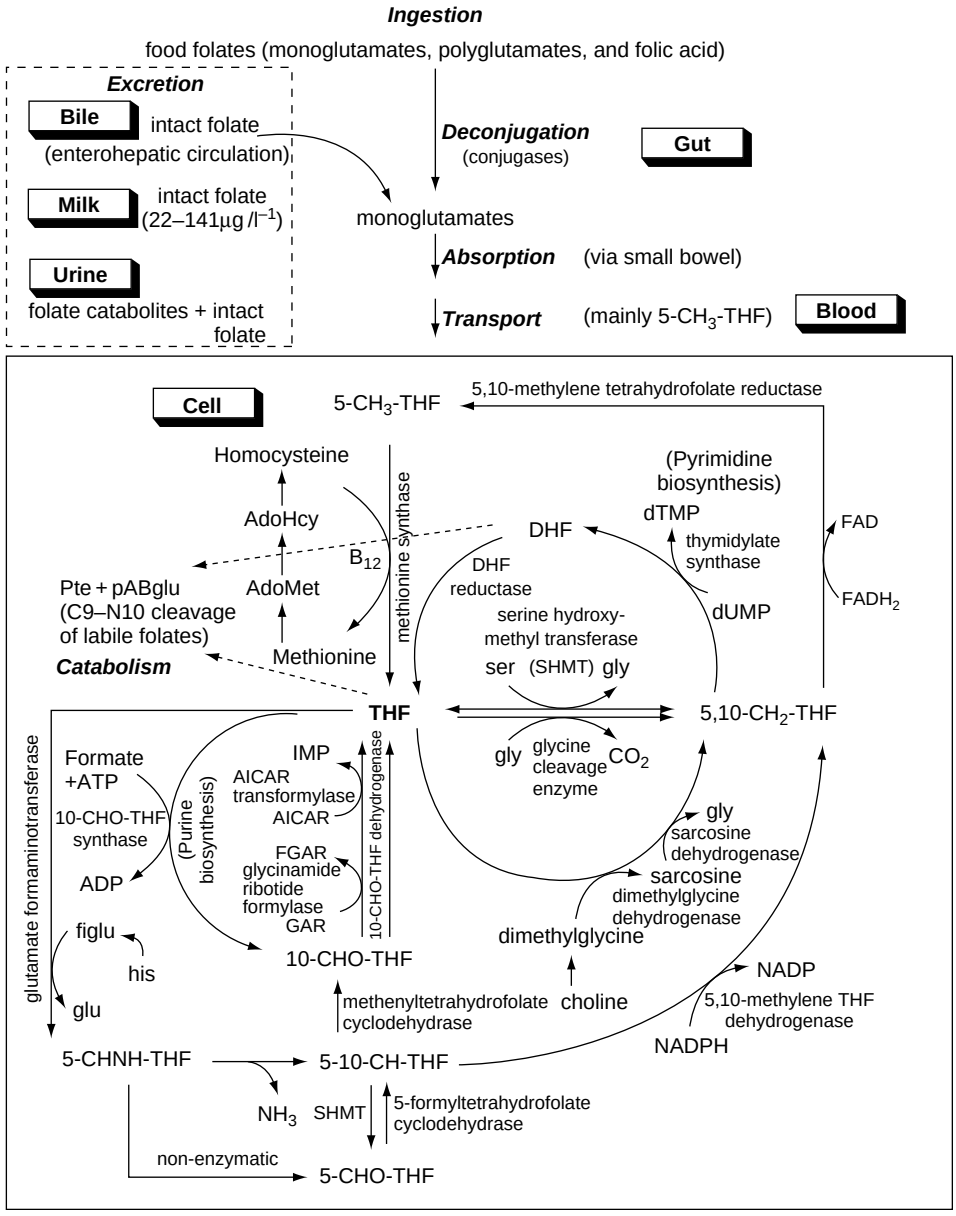


Figure 2 Physiology and metabolism of folate. GAR, glycinamide ribonucleotide; FGAR, formylglycinamide ribonucleotide; AICAR: aminoimidazolecarboxamide ribonucleotide; figlu, formiminoglutamic acid; IMP, inosine monophosphate.

vitamin B_{12} deficiency in the elderly, who are most prone to it.

In summary, the biochemical function of folate coenzymes is to transfer and use these one-carbon units in a variety of essential reactions (**Figure 2**), including *de novo* purine biosynthesis (formylation of glycinamide ribonucleotide and 5-amino-4-imidazole carboxamide ribonucleotide), pyrimidine nucleotide biosynthesis (methylation of deoxyuridylic acid to thymidylic acid), amino-acid interconversions (the interconversion of serine to glycine, catabolism of histidine to glutamic acid, and conversion of homocysteine to methionine (which also requires vitamin B_{12})), and the generation and use of formate.

Many of the enzymes involved in these reactions are multifunctional and are capable of channelling substrates and one-carbon units from reaction to reaction within a protein matrix. Another feature of intracellular folate metabolism is the compartmentation of folate coenzymes between the cytosol and the mitochondria. For instance, 5-methylTHF is associated with the cytosolic fraction of the cell, whereas most of 10-formylTHF is located in the mitochondria. Similarly, some folate-dependent

enzymes are associated with one or other compartment, though some are found in both. Metabolic products of folate-dependent reactions, such as serine and glycine, are readily transported between the two locations, but the folate coenzymes are not.

Folate Deficiency and Hyperhomocysteinemia An important consequence of folate deficiency is the inability to remethylate homocysteine (**Figure 2**). Indeed, there is an inverse correlation between the levels of folate and those of homocysteine in the blood of humans. Many clinical studies, beginning with the observations of children with homocysteinuria presenting with vascular abnormalities and thromboembolism, have demonstrated an association between hyperhomocysteinemia and an increased risk of premature atherosclerosis in the coronary, carotid, and peripheral vasculatures. Even mild hyperhomocysteinemia is recognized to be an independent risk factor for cardiovascular disease. The risk of heart disease was found to increase proportionately in most, but not all, studies, throughout the full of range of blood homocysteine concentrations. An increase in plasma homocysteine of $5\,\mu mol\,l^{-1}$ is associated with a combined odds ratio of 1.3 for cardiovascular disease. Plasma homocysteine is usually shown to be a greater risk factor for cardiovascular disease in prospective studies than in retrospective studies, probably because the populations in the former studies are older. Metabolically, homocysteine may be disposed of by the methionine synthase reaction (dependent on folate and vitamin B_{12}), the transsulfuration pathway (dependent on vitamin B_6), and the choline degradation pathway. Marginal deficiencies of these three vitamins are associated with hyperhomocysteinemia. Of the three vitamins, however, folic acid has been shown to be the most effective in lowering levels of homocysteine in the blood. Convincing evidence of the potential role of folate intake in the prevention of vascular disease has come from a significant inverse relationship between serum folate levels and fatal coronary heart disease. While most studies have focused on the homocysteine-lowering effects of folate, other benefits have also been reported. Potential mechanisms include antioxidant actions and interactions with the enzyme endothelial nitric oxide synthase.

Absorption of Folates

Food folates mainly consist of reduced polyglutamates, which are hydrolyzed to monoglutamates in the gut prior to absorption across the intestinal mucosa. The conjugase enzyme that hydrolyzes dietary folates has been found on the luminal brush border membrane in the human jejunum and has equal affinity for polyglutamates of various chain lengths. Transport is facilitated by a saturable carrier-mediated uptake system, although changes in luminal pH and the presence of conjugase inhibitors, folate binders, or other food components can adversely affect the rate of hydrolysis and intestinal absorption. Such factors account for the wide variation in the bioavailability of the vitamin from foods of plant and animal origins. Some metabolism of the resultant monoglutamate, mainly to 5-methylTHF, appears to occur during the absorption process, though this may not be necessary for transport across the basolateral membrane of the intestinal mucosa into the portal circulation. The degree of metabolic conversion of dietary folic acid depends on the dose; pharmacological amounts are transported unaltered into the circulation.

Transport in the Circulation, Cellular Uptake, and Turnover

Folate circulates in the blood predominantly as 5-methylTHF. A variable proportion circulates freely or bound either to low-affinity protein binders such as albumin, which accounts for about 50% of bound folate, or to a high-affinity folate binder in serum, which carries less than 5% of circulating folate. The physiological importance of serum binders is unclear, but they may control folate distribution and excretion during deficiency.

Though most folate is initially taken up by the liver following absorption, it is delivered to a wide variety of tissues in which many types of folate transporters have been described. Because these transporters have affinities for folate in the micromolar range, they would not be saturated by normal ambient concentrations of folate. Therefore, folate uptake into tissues should be responsive to any increases in serum folate levels arising from folate supplementation. An important determinant of folate uptake into cells is their mitotic activity, as would be expected given the dependence of DNA biosynthesis on folate coenzyme function. Folate accumulation is more rapid in actively dividing cells than in quiescent cells, a factor that is probably related to the induction and activity of folylpoly-γ-glutamate synthase. This enzyme catalyzes the addition of glutamate by γ-peptide linkage to the initial glutamate moiety of the folate molecule. Although

polyglutamate derivatization may be considered a storage strategem, this elongation is the most efficient coenzyme form for normal one-carbon metabolism. The activity of folylpoly-γ-glutamate synthase is highest in the liver, the folate stores of which account for half of the estimated 5–10 mg adult complement. Retention within the cell is facilitated by the high proportion of folate associated with proteins, and this is likely to be increased in folate deficiency.

The mobilization of liver and other stores in the body is not well understood, particularly in deficiency states, though some accounts describe poor turnover rates in folate-depleted rats. Transport across cell membranes during redistribution requires deconjugation of the large negatively charged polyglutamates. Mammalian γ-glutamylhydrolases that hydrolyze glutamate moieties residue by residue and transpeptidases that can hydrolyze folylpolyglutamates directly to mono- or di-glutamate forms of the vitamin have been described for a number of tissues. Thus, mammalian cells possess two types of enzyme that can play a key role in folate homeostasis and regulation of one-carbon metabolism: the folylpolyglutamate synthetase that catalyzes the synthesis of retentive and active folate, and a number of deconjugating enzymes that promote the release of folate from the cell. Polyglutamate forms released into the circulation either through cell death or by a possible exocytotic mechanism would be hydrolyzed rapidly by plasma γ-glutamyl-hydrolase to the monoglutamate form.

Catabolism and Excretion

Folate is concentrated in bile, and enterohepatic recirculation from the intestine accounts for considerable re-absorption and reuse of folate (about 100 $\mu g\,day^{-1}$). Fecal folates mostly arise through biosynthesis of the vitamin by the gut microflora, with only a small contribution from unabsorbed dietary folate. Urinary excretion of intact folates accounts for only a small fraction of ingested folate under normal physiological conditions. The greater amount of excretion in urine is accounted for by products that arise from cleavage of the folate molecule at the C9–N10 bond, consisting of one or more pteridines and *p*-acetamidobenzoylglutamate. The rate of scission of the folate molecule increases during rapid-mitotic conditions such as pregnancy and rapid growth. Scission of folate is perhaps the major mechanism of folate turnover in the body.

Human Folate Requirements

The folate requirement is the minimum amount necessary to prevent deficiency. Dietary recommendations for populations, however, must allow a margin of safety to cover the needs of the vast majority of the population. As is the case with most nutrients, the margin of safety for folate requirement corresponds to two standard deviations above the mean requirement for a population and should therefore meet the needs of 97.5% of the population. Thus, international dietary recommendations contain allowances for individual variability, the bioavailability of folate from different foodstuffs, and periods of low intake and increased use. Current international folate recommendations for FAO/WHO, USA/Canada, and the European Union are listed in **Table 2**.

Table 2 Recommended dietary folate allowances for various population groups ($\mu g\,day^{-1}$)

Category	*Age*	*FAO/WHO (1998)*	*USA/Canada RDA (1998)*	*EU (1993)*
Infants	Up to 6 months	80	65	100
	6 months–1 year	80	80	100
	1–3 years	160	150	100
Children	4–6 years	200	200	130
	7–10 years	300 (age 9–13 years)	300	150
Males	11–14 years	400	300	180
	15–18 years	400	400	200
	19–24 years	400	400	200
	25–50 years	400	400	200
	Over 50 years	400	400	200
Females	11–14 years	400	300	180
	15–18 years	400	400	200
	19–24 years	400	400	200
	25–50 years	400	400	200
	Over 50 years	400	400	200
Pregnant women		600	600	400
Lactating women		500	500	350

The 1998 recommendations for folate are expressed using a term called the dietary folate equivalent (DFE). The DFE was developed to help account for the difference in bioavailability between naturally occurring dietary folate and synthetic folic acid. The Food and Nutrition Board of the US National Academy of Sciences reasoned that, since folic acid in supplements or in fortified food is 85% bioavailable, but food folate is only about 50% bioavailable, folic acid taken as supplements or in fortified food is 85/50 (i.e., 1.7) times more available. Thus, the calculation of the DFE for a mixture of synthetic folic acid and food is μg of DFE = μg food folate + (1.7 × μg synthetic folate).

International recommendation tables are constantly subject to review, particularly in view of the relationship between folate status and the risk of NTD and specific chronic diseases including coronary artery disease and colorectal cancer.

Pregnancy

The crucial role of folate in the biosynthesis of precursors for DNA suggests that folate requirements may vary with age, though folate use is most obviously increased during pregnancy and lactation. Maintaining adequate folate status in women in their child-bearing years is particularly important since a large proportion of pregnancies are unplanned and many women are likely to be unaware of their pregnancy during the first crucial weeks of fetal development. Pregnancy requires an increase in folate supply that is large enough to fulfil considerable mitotic requirements related to fetal growth, uterine expansion, placental maturation, and expanded blood volume. The highest prevalence of poor folate status in pregnant women occurs among the lowest socioeconomic groups and is often exacerbated by the higher parity rate of these women. Indeed, the megaloblastic anemia commonly found amongst the malnourished poor during pregnancy probably reflects the depletion of maternal stores to the advantage of the fetal–placental unit, as indicated by the several-fold higher serum folate levels in the newborn compared with the mother. Considerable evidence indicates that maternal folate deficiency leads to fetal growth retardation and low birth weight. The higher incidence of low-birth-weight infants among teenage mothers compared with their adult counterparts is probably related to the additional burden that adolescent growth places on folate resources.

The lack of hard evidence about the extent of supplementation required in pregnancy prompted the development of a laboratory-based assessment of metabolic turnover, which involved the assay of total daily folate catabolites (along with intact folate) in the urine of pregnant women. The rationale of the procedure was that this catabolic product represents an ineluctable daily loss of folate, the replacement of which should constitute the daily requirement. Correcting for individual variation in catabolite excretion and the bioavailability of dietary folate, the recommended allowances based on this mode of assessment are in close agreement with the latest recommendations of the USA/Canada and FAO/WHO. The data produced by the catabolite-excretion method may provide a useful adjunct to current methods based on intakes, clinical examination, and blood folate measurements to provide a more accurate assessment of requirement.

Folate and Neural-Tube Defects

Much attention has focused over the past 15 years on a number of diseases for which the risks are inversely related to folate status even within the range of serum folate levels previously considered 'normal.' Foremost among these is NTD, a malformation in the developing embryo that is related to a failure of the neural tube to close properly during the fourth week of embryonic life. Incomplete closure of the spinal cord results in spina bifida, while incomplete closure of the cranium results in anencephaly. The risk of NTD was found to be 10-fold higher (6 affected pregnancies per 1000) in people with poor folate status (i.e., less than 150 μg red cell folate per litre) than in those with good folate status (400 $\mu g\,l^{-1}$). International agencies have published folic acid recommendations for the prevention of NTD. To prevent recurrence, 5 mg of folic acid daily in tablet form is recommended, while 400 μg daily is recommended for the prevention of occurrence, to be commenced prior to conception and continued until the 12th week of pregnancy. Given the high proportion of unplanned pregnancies, the latter recommendations are applicable to all fertile women. This amount, however, could not be introduced through fortification because high intakes of folic acid by people consuming fortified flour products would risk preventing the diagnosis of pernicious anemia in the general population and of vitamin B_{12} deficiency in the elderly.

The introduction of 140 μg of folic acid per 100 g of flour in the USA, calculated to increase individual consumption of folic acid by 100 $\mu g\,day^{-1}$, has reduced the incidence of abnormally low plasma folate from 21% to less than 2%, the incidence of mild hyperhomocysteinemia from 21% to 10%,

and, most importantly, the incidence of NTD by about 20% over the first years of universal fortification. Because 30% of the population takes vitamin supplements and presumably would not be expected to derive significant benefit from fortification, the actual effect may be closer to a 30% decrease due to fortification. Recent calculations suggest that, for a variety of reasons, the overall fortification amount was about twice the mandatory amount.

On balance, the introduction of food fortification with folate is regarded as beneficial not only in preventing NTD but also in reducing the incidences of hyperhomocysteinemia (mentioned earlier), colorectal cancer, and a number of neurological and neuropsychiatric diseases in which folate is postulated to play a protective role.

Lactation

Unlike during pregnancy, in which the bulk of folate expenditure arises through catabolism, during lactation the increased requirement is chiefly due to milk secretion. Several observations indicate that mammary tissue takes precedence over other maternal tissues for folate resources. For instance, maternal folate status deteriorates in both early and late lactation, but milk folate concentration is maintained or increased. Moreover, supplemental folate appears to be taken up by mammary epithelial cells preferentially over hematopoietic cells in lactating women with folate deficiency, indicating that maternal reserves are depleted to maintain milk folate content in lactating women. Recommendations are based on the maintenance requirement of nonpregnant nonlactating women and the estimated folate intake required to replace the quantity lost in milk. This increment of between 60 μg and 100 μg daily is based on a milk secretion rate of 40–60 $\mu g\,l^{-1}$ and an absorption rate from dietary sources of between 50% and 70%. The official recommendations might be underestimated, however. On the one hand, a less efficient absorption rate of 50% from a mixed diet is more likely, and, on the other hand, the most recent estimations of milk folate secretion are as high as 100 μg daily. Therefore, an additional 200 μg of folate daily or a total of 500 μg daily seems a more realistic recommended amount for lactating women.

Infants and Children

The high concentration of circulating folate in newborn infants coincides with the rapid rate of cell division in the first few months of life and is reflected in the higher folate requirement for infants on a weight basis than for adults. Though the recommendation standards (see **Table 2**) may underestimate the quantities consumed by many breastfed infants, intake is generally sufficiently above the recommendations that folate deficiency is unlikely.

Data on requirements for older children are sparse, so recommendations for up to adolescence are based on interpolations between the values for very young children and those for adults. Daily recommended levels are above 3.6 μg per kg of body weight, an amount associated with no overt folate deficiency in children and shown to maintain plasma folate concentrations at a low but acceptable level.

Adolescents and Adults

Folate recommendations for adolescents are set at a similar level to that for adults, the smaller weights of adolescents being compensated for by higher rates of growth.

The Elderly

Although folate deficiency occurs more frequently in the elderly than in young adults, recommendations are set at the same level for both groups. Reference recommendations apply to healthy subjects. However, a significant proportion of the elderly are likely to suffer from clinical conditions and to be exposed to a range of factors such as chronic smoking, alcohol, and prescription drugs that may have a detrimental effect on folate status.

Food Sources of Folate

Folate is synthesized by microorganisms and higher plants but not by mammals, for which it is an essential vitamin. The most concentrated food folate sources include liver, yeast extract, green leafy vegetables, legumes, certain fruits, and fortified breakfast cereals. Folate content is likely to depend on the maturity and variety of particular sources. Foods that contain a high concentration of folate are not necessarily those that contribute most to the overall intakes of the vitamin in a population. For example, liver is a particularly concentrated source, providing 320 μg of folate per 100 g, but it is not eaten by a sufficient proportion of the population to make any major contribution to total dietary folate intakes. The potato, on the other hand, although not particularly rich in folate, is considered a major contributor to folate in the UK diet, accounting for 14% of total folate intake because of its high consumption. Prolonged exposure to heat, air, or ultraviolet light is known to inactivate the vitamin; thus, food

Table 3 Contributions of the main food groups to the average daily intake of folate in British and US adults (%)

Food group	*USA (1994)*	*UK (1998)*
Dairy products	8.1	9.8
Meat, poultry, fish	8.5	6.5
Grain products	21.2[a]	31.8
Fruit, fruit juices	10.2	6.9
Vegetables	26.4	31.8
Legumes, nuts, soy	18.5	—
Eggs	5.1	2.4
Tea	—	4.1
Other food	2.1	5.7

[a]Prior to mandatory fortification of flour-based products.

preparation and cooking can make a difference to the amount of folate ingested; boiling in particular results in substantial food losses. The major source of folate loss from vegetables during boiling may be leaching as opposed to folate degradation. Broccoli and spinach are particularly susceptible to loss through leaching during boiling compared with potatoes because of their larger surface areas. The retention of folate during cooking depends on the food in question as well as the method of cooking. Folates of animal origin are stable during cooking by frying or grilling. In addition to highlighting good food sources, public-health measures promoting higher folate intake should include practical advice on cooking. For example, steaming in preference to boiling is likely to double the amount of folate consumed from green vegetables.

While cultural differences and local eating habits determine the contribution of different foodstuffs to folate intake (**Table 3**), as with other nutrients, globalization and the integration of the international food industry may lead to more predictable 'Westernized' diets in the developed and developing world. Internationally, much of the dietary folate in the 'Western' diet currently comes from fortified breakfast cereals, though this foodstuff is likely to be joined shortly in this regard by fortified flour products in the light of the experience of the US food fortification program. In the main, though, adherence to dietary recommendations to increase the consumption of folate-rich foods is likely to enhance the intake not only of folate but also of other nutrients essential to health.

See also: **Adolescents**: Nutritional Requirements. **Amino Acids**: Chemistry and Classification; Metabolism; Specific Functions. **Anemia**: Megaloblastic Anemia. **Breast Feeding**. **Cobalamins**. **Food Fortification**: Developed Countries; Developing Countries. **Fruits and Vegetables**. **Infants**: Nutritional Requirements. **Lactation**: Physiology; Dietary Requirements. **Pregnancy**: Safe Diet for Pregnancy; Prevention of Neural Tube Defects.

Further Reading

Bailey LB (ed.) (1995) *Folate in Health and Disease*. New York: Marcel Dekker.

Blakley R (1969) The biochemistry of folic acid and related pteridines. In: Newbergen H and Taton EL (eds.) *North Holland Research Monographs Frontiers of Biology*, vol. 13, Amsterdam: North Holland Publishing Company.

Boushey CJ, Beresford SA, Omenn GS, and Motulsky AG (1995) A quantitative assessment of plasma homocysteine as a risk factor for vascular disease. *JAMA* **274**: 1049–1057.

Chanarin I (1979) *The Megaloblastic Anaemias*, 2nd edn. Oxford: Blackwell Scientific Publications.

Duthie SJ (1999) Folic acid deficiency and cancer: mechanisms of DNA instability. *British Medical Bulletin* **55**: 578–592.

Homocysteine Lowering Trialists' Collaboration (1998) Lowering blood homocysteine with folic acid based supplements: meta-analysis of randomised trials. *BMJ* **316**: 894–898.

National Academy of Sciences (1998) *Dietary Reference Intakes: Folate, other B Vitamins and Choline*. Wasington, DC: National Academy Press.

Refsum H, Ueland PM, Bygard MD, and Vollset SE (1998) Homocysteine and cardiovascular disease. *Annual Review of Medicine* **49**: 31–62.

Reynolds EH (2002) Folic acid, ageing, depression, and dementia. *BMJ* **324**: 1512–1515.

Scott JM and Weir DG (1994) Folate/vitamin B_{12} interrelationships. *Essays in Biochemistry* **28**: 63–72.

UK Department of Health (2000) *Folic Acid and the Prevention of Disease. Report of the Committee on Medical Aspects of Food and Nutritional Policy. Norwich: Her Majesty's Stationary Office.*

FOOD ALLERGIES

Contents
Etiology
Diagnosis and Management

Etiology

T J David, University of Manchester, Manchester, UK

The concept that certain foods can produce adverse reactions in susceptible individuals has a long history. Hippocrates (460–370 BC) reported that cow's milk could cause gastric upset and urticaria. Later, Galen (131–210 BC) described a case of intolerance to goat's milk. It was Lucretius (96–55 BC) who said, "What is food to one man may be fierce poison to others." In the 1920s and 1930s, a fashion developed of blaming food intolerance for a large number of hitherto unexplained disorders. The uncritical and overenthusiastic nature of the claims, in addition to the anecdotal evidence on which they were based, generally discredited the whole subject. Indeed, the field of food intolerance has been described as "a model of obstruction to the advancement of learning." The whole area has provoked much controversy. The introduction of double-blind provocation tests has placed studies on a more scientific footing, but they are impractical in routine management. The lack of objective and reproducible diagnostic laboratory tests that could eliminate bias has ensured that controversy about food intolerance continues.

Definitions

The word 'allergy' is frequently misused and applied indiscriminately to any adverse reaction, regardless of the mechanism. An allergic response is a reproducible adverse reaction to a substance mediated by an immunological response. The substance provoking the reaction may have been ingested, injected, inhaled, or merely have come into contact with the skin or mucous membranes. Food allergy is a form of adverse reaction to food in which the cause is an immunological response to a food. The much broader term of 'food intolerance' does not imply any specific type of mechanism and is simply defined as a reproducible adverse reaction to a specific food or food ingredient. Outside the United Kingdom, the terminology used sometimes differs. It has been suggested that the term 'food hypersensitivity' should be used to cover all adverse reactions to food, which are then subdivided into food allergy (i.e., immunologically mediated) and food intolerance, which implies a nonimmunologically mediated event.

The term 'food aversion' comprises food avoidance, where the subject avoids a food for psychological reasons such as distaste or a desire to lose weight, and psychological intolerance. The latter is an unpleasant bodily reaction caused by emotions associated with the food rather than the food itself. Psychological intolerance will normally be observable under open conditions but will not occur when the food is given in an unrecognizable form. Psychological intolerance may be reproduced by suggesting (falsely) that the food has been administered.

The term 'anaphylaxis' or 'anaphylactic shock' is taken to mean a severe and potentially life-threatening reaction of rapid onset with circulatory collapse. The term anaphylaxis has also been used to describe any allergic reaction, however mild, that results from specific IgE antibodies, but such usage fails to distinguish between a trivial reaction (e.g., a sneeze) and a dangerous event.

An antigen is a substance that is capable of provoking an immune response. An antibody is an immunoglobulin that is capable of combining specifically with certain antigens. An allergen is a substance that provokes a harmful (allergic) immune response.

Immunological tolerance is a process that results in the immunological system becoming specifically unreactive to an antigen that is capable in other circumstances of provoking antibody production or cell-mediated immunity. The immunological system nevertheless reacts to unrelated antigens given simultaneously and via the same route.

Atopy is the ability to produce a weal-and-flare response to skin prick testing with a common antigen, such as house dust mite or grass pollen. The atopic diseases are asthma (all childhood cases but not all adult cases), atopic eczema, allergic rhinitis, allergic conjunctivitis, and some cases of urticaria.

Mechanisms of Food Allergy

Understanding of the mechanisms of food allergy is poor, and in many cases the precise mechanism is obscure.

Sensitization

The following are possible factors that contribute to immunological sensitization leading to food intolerance:

1. Genetic predisposition: food allergy is commonly familial, suggesting the importance of genetic factors.
2. Immaturity of the immune system or the gastrointestinal mucosal barrier in newborn infants may predispose to sensitisation. The numerous studies that have been performed to determine if food allergy or atopic disease can be prevented by interventions during pregnancy or lactation are based on the idea that there is a critical period during which sensitization can occur.
3. Dosage of antigen: It may be that high dosage leads to the development of tolerance, and low dosage leads to sensitization. This might help to explain the well-documented phenomenon of infants who become allergic to traces of foods that reach them through their mother's breast milk.
4. Certain food antigens are especially likely to lead to sensitization, such as egg, cow's milk, and peanut. The reason why certain foods are more likely to provoke an allergic reaction than others is poorly understood.
5. A triggering event, such as a viral infection: The evidence is anecdotal, but there is a suggestion that food allergy may develop in a previously nonallergic subject after a viral infection such as infectious mononucleosis (glandular fever).
6. Alteration in the permeability of the gastrointestinal tract, permitting abnormal antigen access: The best example of this is the suggestion that acute viral gastroenteritis may damage the small intestinal mucosa, allowing abnormal absorption of food proteins, leading to sensitization. Thus, some data suggest that in a few cases the onset of cow's milk protein allergy follows soon after an episode of gastroenteritis.

Immunological and Molecular Mechanisms

Despite the gastrointestinal barrier, small amounts of immunologically intact proteins enter the circulation and are distributed throughout the body. In normal individuals, the gut-associated lymphoid tissue (GALT), although capable of mounting a rapid and potent response against foreign substances, develop tolerance to ingested food antigens. The means by which tolerance develops is poorly understood, but it is believed that failure to develop tolerance leads to food allergy. The relatively low salivary secretory IgA concentrations, together with the large amount of ingested protein, contributes to the large amount of food antigens confronting the immature GALT. In genetically predisposed infants, these food antigens may stimulate the excessive production of IgE antibodies or other abnormal immune responses.

Heat Treatment

Heat treatment clearly affects certain (but not all) foods, most commonly rendering them less likely to provoke an allergic reaction in a subject who is allergic. Occasionally, the reverse occurs, as in the celebrated case of Professor Heinz Küstner, who was allergic to cooked and not raw fish.

In cow's milk, whey proteins are easily denatured by heat but casein is highly resistant. This observation led to the suggestion that the heat treatment of whey proteins may be a simple and logical strategy for producing a hypoallergenic infant milk formula. However, double-blind, placebo-controlled oral challenges gave rise to immediate hypersensitivity reactions to heat-treated whey protein in four of five children with cow's milk protein intolerance. The reason for these reactions is not known, but one possibility is a reaction to residual casein, which is often present in trace amounts in commercial whey preparations. The small proportion of patients with cow's milk protein intolerance likely to tolerate heat-treated cow's milk, such as evaporated milk, means that heat-treated milk is unlikely to be suitable as a substitute for a cow's milk infant formula.

Cooking reduces the allergenicity of eggs by 70%. However, one of the major allergens in eggs, ovomucoid, a heat-resistant glycoprotein that contributes to the gel-like structure of thick white, is resistant to heating. Heat appears to render a large number of fruits and vegetables less likely to provoke adverse reactions in subjects who are intolerant. Thus, for example, it is not uncommon to see children who are allergic to raw potatoes or fresh pineapple, but almost all such children can tolerate cooked potatoes or tinned pineapples. In some situations, it appears that heat can accelerate a process of denaturation that can in time occur on its own. For example, there have been studies of patients who reacted to fresh melon, pear, peach, pineapple, grape, and banana.

In each case, stewed or tinned fruit caused no reaction. Studies of fresh extracts of these fruits showed that when stored in a refrigerator, the extracts lost their ability to provoke a positive skin test after approximately 3 days.

Heating can increase the allergenicity of certain proteins through the induction of covalent modifications that lead to new antigens or increased stability. In peanuts, for example, the roasting process produces end products with greater resistance to digestion and heightened allergenicity compared with those produced by frying or boiling. This finding may partly account for the low prevalence of peanut allergy in China, where peanut is widely consumed but not roasted.

Prevalence

Unreliability of Self-Reported Food Allergy

Reports of food allergy from individuals or parents of children are notoriously unreliable. Such reports have to be treated with scepticism. It is common for parents to believe that foods are responsible for a variety of childhood symptoms. Double-blind provocation tests in children with histories of reactions to food only confirm the story in one-third of all cases. In the case of purely behavioral symptoms, the proportion that could be reproduced under blind conditions was zero. The same is true of adults' beliefs about their own symptoms. If unnecessary dietary restrictions are to be avoided, one has to be sceptical, and it may be necessary in some cases to seek objective confirmation of food intolerance. The gross overreporting of food allergy has to be borne in mind when examining data on prevalence that are based on unconfirmed subjective reports.

Population Studies

In the European Community Respiratory Health Survey administered to 17 280 adults in 15 countries covering the period 1991–1994, 12% of respondents reported a food allergy or intolerance, ranging from 4.6% in Spain to 19.1% in Australia. The foods most commonly reported to cause shortness of breath were peanut in the United States; fruit in Iceland, Belgium, Ireland, and Italy; and hazelnut in Norway, Sweden, and Germany. The reason for the variation in the reported food triggers is unknown.

A population-based study of 33 110 people in France defined food allergy on the basis of self-reported typical allergic symptoms only and found a rate of 3.5%. In adults, the main foods reported to trigger allergic reactions were seafood, fruit, and vegetables, whereas in children the main foods were egg and milk.

In a questionnaire offered to approximately 30 000 people in 11 388 households in the High Wycombe area of Britain, 3188 of the 18 582 responders (17%) thought that they had some sort of reaction to foods or food additives. A check on the nonresponders showed that they had almost no food-provoked symptoms. Particular attention was then paid to food additives, and it was found that 1372 of the 18 582 responders (7.4%) believed they had adverse reactions to food additives. Of the 1372, 649 attended for a detailed interview, and 132 gave a history of reproducible clinical symptoms after ingestion of food additives. Eighty-one of these completed a trial of double-blind, placebo-controlled challenges with 11 food additives, but a consistent adverse reaction was found in only 2 subjects. One was a 50-year-old atopic man who reported headaches after ingesting coloring agents and who reacted to challenge with annatto, which reproduced his headache at both low (1 mg) and high (10 mg) dose after 4 and 5 h, respectively. He also reacted to placebo on one occasion. The second was a 31-year-old nonatopic woman who reported abdominal pain after ingestion of foods. She had related this to ingestion of preservatives and antioxidants. Her symptoms were reproduced on challenge with annatto at low and high dose.

The parents of 866 children from Finland were asked to provide a detailed history of food allergy, and for certain foods the diagnosis was further investigated by elimination and open challenge at home. Food allergy was reported in 19% by the age of 1 year, 22% by 2 years, 27% by 3 years, and 8% by 6 years. In a prospective study of 480 children in the United States of America up to their third birthday, 16% were reported to have had reactions to fruit or fruit juice and 28% to other food. However, open challenge confirmed reactions in only 12% of the former and 8% of the latter.

Estimates of the prevalence of cow's milk protein allergy are reported to range from 0.3 to 7.5% of subjects.

Natural History

The natural history of food allergy has been little studied. It is well-known that a high proportion of children with food intolerance in the first year of life lose their intolerance in time. The proportion of children to which this happens varies with the food and probably with type of symptoms that are produced. Thus, it is common for allergy to cow's milk or egg to spontaneously disappear with time,

whereas peanut allergy is usually lifelong. In the North American study referred to previously, it was found that the offending food or fruit was back in the diet after only 9 months in half the cases, and virtually all the offending foods were back in the diet by the third birthday. A further study of nine children with very severe adverse reactions to food showed that despite the severity, three were later able to tolerate normal amounts of the offending food and four became able to tolerate small amounts.

Although it is clear that the majority of children with food intolerance spontaneously improve, it remains to be established to what extent this depends on the age of onset, the nature of the symptoms, the food itself, and other factors such as associated atopic disease.

In adults with food allergy, the problem is far more likely to be lifelong. Nevertheless, some adults do become tolerant to foods to which they were allergic. In one adult follow-up study, approximately one-third of adults were found to lose their allergy after maintaining an elimination diet for 1 year.

Cross-Reactions

This term refers to cross-reaction between different species and between different foodstuffs that may or may not belong to the same botanical family.

Animal Milk

There is a marked antigenic similarity between the proteins that cause food allergy in the milk of cows, goats, and sheep. It is often not appreciated that almost all subjects who are allergic to cow's milk protein are allergic to the milks of these other animals. This is one of the many reasons why goat's milk is not an appropriate milk substitute for an infant with cow's milk allergy.

Eggs

The eggs from turkeys, duck, goose, and seagull all contain ovalbumen, ovomucoid, and ovotransferrin, the major allergens in hens' eggs. The eggs of hens and turkeys have a similar relative potency of allergenicity. The immunochemical identity of proteins in the egg white of ducks and geese differs somewhat from that of hens, and they may have less potency as allergens. Of all the bird's eggs listed previously, the eggs of the seagull are the least allergenic and bear the least immunochemical similarity with hen's eggs.

Legumes

It is not always obvious which plants belong to the same family. The Leguminosae include beans, peas, soya, lentils, peanuts, liquorice, carob, and gum arabic. Clinical cross-reactivity is uncommon, and the degree of genetic relationship may be of little relevance. Thus, for example, patients with soya allergy are not uncommonly allergic also to peanuts, although the two legumes are not closely related.

Seafood

The taxonomic diversity (fish, molluscs, and crustaceans) suggests that complete cross-reactivity for all seafood is unlikely to be common. In one study, of 20 children with a history of allergy to cod, there was a history of allergy to sole in 11 (55%), to tuna in 7 (35%), and to mackerel, anchovy, sardine, red mullet, and salmon each in 1 (5%). Most studies of cross-reactivity are based on skin prick and IgE antibody test results, which are of little relevance to clinical sensitivity.

Food and Pollen

Cross-reactions can occur between inhaled pollen and ingested food allergens. There is a well-documented association between allergy to birch tree pollen and allergy to apple, carrot, celery, potato, orange, and tomato. This and other similar types of associations are explained by the conservation of protein across species, and a number of so-called panallergens have been described:

1. Profilins in birch pollen, hazelnut, and apple
2. Class 1 chitinases in avocado, banana, and latex
3. Lipid transfer proteins in apple and peach
4. Tropomyosin in insects and shellfish

Cross-reacting IgE antibodies reactive with related foods can often be detected in people who are allergic to a member of these food groups, but clinical reactions are uncommon.

Special Requirements for the Occurrence of Allergic Reaction to Food

In some individuals, there is a clear one-to-one relationship between the ingestion of a food and a reaction. An example is an individual with allergy to cod. Every time the subject eats cod, there is an immediate allergic reaction. In other individuals, the relationship between the food and an allergic reaction is less precise. There are a number of possible reasons for this.

Timing of Reaction and Delayed Reactions

Most allergic reactions to foods occur within minutes of ingestion of the food. However, sometimes a reaction may be delayed. This is best documented in cow's milk protein allergy, in which three types of reaction are recognized: early skin reaction, early gut reaction, and late reaction. An affected individual usually exhibits only one of these types of reaction. In the early skin reaction group, symptoms begin to develop within 45 min of cow's milk challenge. Almost all patients in this group have a positive skin prick test to cow's milk. In the early gut reaction group, symptoms begin to develop between 45 min and 20 h after cow's milk challenge. Approximately one-third of patients in this group have a positive skin prick test to cow's milk. In the late reaction group, symptoms begin to develop approximately 20 h after cow's milk protein challenge. Only approximately 20% of this late reaction group have a positive skin prick test to cow's milk, and these are mostly children with atopic eczema. Almost all children in the late reaction group present over the age of 6 months, and as a group their age at presentation is significantly higher than that of the two other groups.

Quantity of Food

The quantity of cow's milk, for example, required to produce an allergic reaction varies from patient to patient. Some patients are highly sensitive and develop anaphylaxis after ingestion of less than 1 μg of casein, β-lactoglobulin, or α-lactalbumin. In contrast, there are children and adults who do not react to 100 ml of milk but who do react to 200 ml or more. There is a relationship between the quantity of milk required and the time of onset of symptoms. In one study, the median reaction onset time in those who reacted to 100-ml milk challenges was 2 h, but the median reaction onset time in those who required larger amounts of milk to elicit reactions was 24 h.

Food-Dependent Exercise-Induced Anaphylaxis

In this unusual condition, attacks only occur when the exercise follows within a couple of hours of the ingestion of specific foods, such as celery, shellfish, squid, peaches, or wheat. The mechanisms that result in food-dependent exercise-induced anaphylaxis are obscure. This disorder, although rare, is important in the interpretation of dietary challenge studies of food intolerance because in these patients a simple double-blind food challenge without exercise will fail to validate a history of food intolerance.

Drug-Dependent Food Allergy

There are individuals who only react to specific foods while taking a drug. The best recognised examples of this are individuals who only react to foods while taking salicylate (aspirin).

Effect of Disease Activity

It is a common but poorly understood observation that children with eczema and food allergy can often tolerate some or all food triggers when the skin disease clears (usually when the child is vacationing in a sunny location).

Other Possibilities

It is not known whether food allergy can be confined to occasions when the pollen count is high or when the individual consumes certain other foods. There are no objective studies that address the complex issue of the possible additive effect of orally ingested and possibly inhaled antigens. There are subjects with allergy to foods in whom the severity of adverse reactions clearly varies from time to time, but the reasons for this variability are not known.

See also: **Dairy Products**. **Eggs**. **Food Allergies**: Diagnosis and Management. **Food Intolerance**. **Fruits and Vegetables**. **Immunity**: Physiological Aspects. **Nuts and Seeds**.

Further Reading

Bentley SJ, Pearson DJ, and Rix KJB (1983) Food hypersensitivity in irritable bowel syndrome. *Lancet* **2**: 295–297.

Bernhisel-Broadbent J and Sampson HA (1989) Cross-allergenicity in the legume botanical family in children with food hypersensitivity. *Journal of Allergy and Clinical Immunology* **83**: 435–440.

Bock SA (1987) Prospective appraisal of complaints of adverse reactions to foods in children during the first three years of life. *Pediatrics* **79**: 683–688.

Bush RK, Taylor SL, Nordlee JA, and Busse WW (1985) Soybean oil is not allergenic to soybean-sensitive individuals. *Journal of Allergy and Clinical Immunology* **76**: 242–245.

David TJ (1987) Reactions to dietary tartrazine. *Archives of Disease in Childhood* **62**: 119–122.

David TJ (1993) *Food and Food Additive Intolerance in Childhood*. Oxford: Blackwell Scientific.

De Martino M, Novembre E, Galli L *et al.* (1990) Allergy to different fish species in cod-allergic children: *In vivo* and *in vitro* studies. *Journal of Allergy and Clinical Immunology* **86**: 909–914.

Dreborg S (1988) Food allergy in pollen-sensitive patients. *Annals of Allergy* **61**: 41–46.

Herian AM, Taylor SL, and Bush RK (1990) Identification of soybean allergens by immunoblotting with sera from soy-allergic adults. *International Archives of Allergy and Immunology* **92**: 193–198.

Johansson SGO, Bieber T, Dahl R *et al.* (2004) Revised nomenclature for allergy for global use: Report of the Nomenclature

Review Committee of the World Allergy Organisation, October 2003. *Journal of Allergy and Clinical Immunology* **113**: 832–836.

Pastorello EA, Stocchi L, Pravettoni V *et al.* (1989) Role of the elimination diet in adults with food allergy. *Journal of Allergy and Clinical Immunology* **84**: 475–483.

Pauli G, Bessot JC, Dietemann-Molard A, Braun PA, and Thierry R (1985) Celery sensitivity: Clinical and immunological correlations with pollen allergy. *Clinical Allergy* **15**: 273–279.

Sampson HA (2004) Update on food allergy. *Journal of Allergy and Clinical Immunology* **113**: 805–819.

Sicherer SH (2002) Food allergy. *Lancet* **360**: 701–710.

Young E, Patel S, Stoneham M, Rona R, and Wilkinson JD (1987) The prevalence of reaction to food additives in a survey population. *Journal of the Royal College of Physicians of London* **21**: 241–247.

Diagnosis and Management

T J David, University of Manchester, Manchester, UK

Documenting Possible Food Allergies

The diagnosis of food allergy is made from the history, supported by investigations and by responses to avoidance of specific food triggers. Since the value of investigations is limited, it is especially important to obtain a clear history. There are a number of practical points to be made:

- **Speed of onset.** In general, the quicker the onset of the allergic reaction, the more reliable is the history. If a child develops a violent allergic reaction within a minute or two after ingesting a food, it is much easier to link the reaction to a specific food than if a reaction only occurs hours or days after eating a food.
- **Coincidences need to be excluded.** If a child becomes unwell (e.g., starts wheezing) an hour after eating a specific food, the wheezing could be caused by the food, or it could just be a coincidence. The more times that such a sequence has been observed, the more likely it is that there is a cause and effect relationship.
- **Observations need to be tested for internal consistency.** Someone may believe that he or she is allergic to a food if a symptom (e.g., urticaria) occurs on (say) three occasions after eating a specific food. It is important to find out:
  1. Whether the subject has had the same symptoms on other occasions when the suspect food trigger was not taken.
  2. Whether the subject has taken the suspect food on one or more other occasions without any adverse effects.

Failure to seek inconsistencies such as these is one factor that is responsible for the overdiagnosis of food allergy.

Documenting a Diagnosis of Food Allergy

If it is reported that someone is allergic to an item, it is important to probe further and find out on what basis the person has been deemed allergic. It is common to find children and adults who are believed to be allergic to a food solely on the basis of tests such as skin tests or blood tests, which are in fact almost wholly unreliable (see below). It is also common for people to believe that they are allergic to something because a health professional said so one day, which on further enquiry turns out to be on flimsy or nonexistent grounds.

Another common problem is the misinterpretation of a sequence of events. For example, a child with an ear infection is given an antibiotic, and 3 days later gets diarrhea, so the parents come to believe the child is allergic to the antibiotic. In fact the cause of the diarrhea is far more likely to be either an underlying viral infection, or a disturbance of the gut flora. Another example is the report of a child who is believed to be allergic to sesame seeds because of reactions occurring after eating buns coated with sesame seeds; many such children are in fact not allergic to sesame seeds but are reacting to the egg glaze that has been used as an adhesive for the seed coating. Another common example is the child with asthma who coughs and wheezes after drinking a diluted orange squash drink, with the result that it is believed that the child is reacting to the yellow-orange coloring agent tartrazine. If fact such reactions are more likely to be due to sulfite preservatives in the squash; sulfites trigger symptoms in 60% or more of children with asthma.

Practical Diagnostic Difficulties

Multiple Mechanisms

Reactions to foods are a heterogeneous group of disorders caused by a variety of different immunological and pharmacological mechanisms. In any individual case, the precise mechanism is often not known. No single type of laboratory test could possibly cover all the different types of possible mechanisms of reactions to foods. Even if one

focuses on food allergy, there are a number of different possible immunological mechanisms, including IgE-antibody mediated, cell mediated, and circulating immune complexes.

Inability to Predict Outcome

In many situations (e.g., atopic disease), the subject wants to know whether there will be any benefit from food avoidance (e.g., not drinking cows' milk or not eating apples). Even if there were valid tests for the diagnosis of food intolerance, the outcome of avoidance measures depends on a number of other variables. Allergen avoidance may succeed for the following reasons:

1. the patient was intolerant to the item;
2. coincidental improvement; and
3. placebo response.

The reasons why a trial of food avoidance may fail to help can be summarized as:

1. The subject is not allergic to the food.
2. The period of elimination was too short. For example, where a child has an enteropathy (damage to the small intestine) due to food allergy, it may take a week or more for improvement in symptoms to occur.
3. The food has been incompletely avoided. This may happen in a subject supposedly on a cows' milk protein-free diet who still continues to receive food that contains cows' milk proteins such as casein or whey.
4. The subject is allergic to other items, which have not been avoided. For example, a child with an allergy to cows' milk protein who fails to improve when given a soy-based milk to which they also have an allergy.
5. Coexisting or intercurrent disease, for example, gastroenteritis in a child with loose stools who is trying a cows' milk-free diet.
6. The patient's symptoms are trivial and have been exaggerated or do not exist at all and have either been imagined or made up by the parents.

It is unrealistic to expect there to be a simple test that can overcome all these problems.

Diagnostic Tests

Skin Prick Tests

The principle of skin prick tests is that the skin weal and flare reaction to an allergen demonstrates the presence of mast-cell-fixed antibody, which is mainly IgE antibody. IgE antibody is produced in plasma cells, and is distributed in the circulation to all parts of the body, so that sensitization is generalized and therefore can be demonstrated by skin testing. In the presence of specific IgE antibody, mast cells in the skin release histamine, which in turn causes a visible weal and flare reaction in the skin.

The procedure involves a drop of allergen solution being placed on the skin, which is then pricked with a hypodermic needle. Two control solutions should also be used: the diluent, in order to detect false-positive reactions; and a positive control (e.g., a histamine solution), to enable comparison with a positive result of an allergen solution. The skin prick test induces a response that reaches a peak in 8–9 min for histamine and 12–15 min for allergens. The size of the weal reaction (and not the larger red flare) is measured.

There are numerous problems with skin prick tests, including:

1. There is no agreed definition about what constitutes a positive reaction.
2. The size of the weal depends to some extent on the potency of the extract.
3. Antihistamines and tricyclic antidepressants suppress the histamine-induced weal and flare response of a skin test. The suppressive effect of antihistamines may last from a week up to several months for some of the more recently introduced nonsedating antihistamines.
4. False-positive tests: skin prick test reactivity may be present in subjects with no clinical evidence of allergy or intolerance. This is sometimes described as 'asymptomatic hypersensitivity' or 'subclinical sensitization.' Whilst many with positive skin prick tests will never develop the allergy, some subjects with positive skin prick tests do develop symptoms later. However, since the test cannot identify those who are going to develop symptoms, the skin test information is of no practical value.
5. False-positive results: skin prick test reactivity may persist after clinical evidence of intolerance has subsided. For example, in a study of children with egg allergy, it was noted that 5 out of 11 who grew out of egg allergy had persistently positive skin prick tests after the allergy had disappeared.
6. False-negative tests: skin prick tests are negative in some subjects with genuine food allergies.
7. Skin prick tests mainly detect IgE antibody. However, many adverse reactions to food are not IgE mediated, in which case skin prick tests can be expected to be negative. Taking cows'

milk protein intolerance as an example, patients with quick reactions often have positive skin prick tests to cows' milk protein, but those with delayed reactions usually have negative skin prick tests.

8. False-negative results are a problem in infants and toddlers, when the weal size is much smaller than later in life.
9. There is a poor correlation between the results of provocation tests (e.g., double-blind food challenges) and skin prick tests. For example, in one study of 31 children with a strongly positive (weal >3 mm in diameter) skin prick test to peanut, only 16 (56%) had symptoms when peanuts were administered.
10. Commercial food extracts (sometimes heat treated) and fresh or frozen raw extracts may give different results (more positives with raw foods), reflecting the fact that some patients are allergic to certain foods only when taken in a raw state. In others the reverse is the case.

Skin prick tests are mainly used in research studies. The results of skin tests cannot be taken alone, and standard textbooks on allergy acknowledge that "the proper interpretation of results requires a thorough knowledge of the history and physical findings." The problems in clinical practice are, for example, whether or not a subject with atopic disease (eczema, asthma, or hay fever) or symptoms suggestive of food intolerance will benefit from attempts to avoid certain foods or food additives. However, skin prick test results are unreliable predictors of response to such measures.

Skin test results are known to be misleading in cases of inhalant allergy (e.g., allergy to dust mites or grass pollen) but skin prick tests for food allergy are especially unreliable because of the large number of false-positive and false-negative reactions.

Intradermal Testing

Intradermal testing comprises the intradermal injection of 0.01–0.05 ml of an allergen extract. It can cause fatal generalized allergic reaction (anaphylaxis), and is only performed if a preliminary skin prick test is negative. Intradermal tests are more sensitive than skin prick testing, and hence also produce even more false-positive reactions, making the interpretation of the results of intradermal testing even more difficult than that for skin prick testing. The difficulty in interpretation of the results, the pain of intradermal injections, and the risk of anaphylaxis mean that intradermal testing has no place in the routine investigation of food allergy.

Skin Application of Food Prior to Food Challenges

There is one situation where direct application of food to the skin may be of practical value, and that is prior to a food challenge in a child in whom one fears an anaphylactic reaction. An example might be a 6-month-old infant with a history of a severe allergic reaction to egg. If the parents wish to see if the child has outgrown the allergy without directly administering egg and risking a violent reaction, a simple approach is to rub some raw egg white into the skin and observe the skin for a few minutes. If the skin application of egg in this way causes an urticarial reaction, then a gradual diminution and disappearance of this response during the succeeding months and years can probably be taken to indicate the development of tolerance, and a continuing brisk response to skin contact would constitute a deterrent to an oral challenge. However, this is only an approximate guide, and there are a number of possible reasons why such testing may give false-positive (e.g., using a raw food when the food is usually eaten cooked, such as egg or potato) or false-negative (e.g., the child is receiving antihistamine drug) results.

Tests for Circulating IgE Antibodies: the Radioallergosorbent (RAST) Test

The radioallergosorbent (RAST) test is the best known of a number of laboratory procedures for the detection and measurement of circulating IgE antibody. Unfortunately, the clinical interpretation of RAST test results is subject to most of the same pitfalls as that for skin prick testing. Additional problems with RAST tests are the cost, and the fact that a very high level of total circulating IgE (e.g., in children with severe atopic eczema) may cause a false-positive result. Depending upon the criteria used for positivity, there is a fair degree of correlation between the RAST test and skin prick test results.

Provocation Tests

A provocation test may be useful to confirm a history of allergy. An example might be a child who developed wheezing and urticaria minutes after eating a rusk that contained, as its main ingredients, wheat and cows' milk protein. To determine which component, if any, caused the reaction, oral challenges with individual components can be conducted.

However, the results of provocation tests cannot prove that improvement in a disease has been caused by food avoidance. For example, a child with atopic eczema is put on a diet avoiding many foods, and

the eczema improves. This improvement could be a coincidence, a placebo effect, or due to the diet. Just because the child is shown to react to a single food does not prove that avoidance of that food was the cause of the improvement.

Open and blind challenges Where the subject and the observer knows the identity of the administered material at the time of the challenge, the procedure is said to be an 'open' challenge. In a 'single-blind challenge' the observer but not the patient or family know the identity of the test material. To avoid bias on the part of the observer, a double-blind challenge is required. A 'double blind' challenge involves exposing the subject to a challenge substance, which is either the item under investigation or an indistinguishable inactive (placebo) substance. Neither the subject nor the observer knows the identity of the administered material at the time of the challenge or during the subsequent period of observation.

The purpose of provocation tests The aim of a food challenge is to study the consequences of food or food additive ingestion. Provocation tests are helpful:

1. to confirm a history (parents' observations of alleged food allergy are notoriously unreliable, as are adults' beliefs about their own allergies);
2. to confirm the diagnosis, for example, of cows' milk protein allergy in infancy, where the diagnostic criteria include improvement on elimination diet and relapse on reintroduction;
3. to see if a subject has grown out of a food intolerance; and
4. as a research procedure.

The food challenge should replicate normal food consumption in terms of dose, route, and state of food. It should also be performed in such a way that the history can be verified. Thus, for example, there is no point solely looking for an immediate reaction if the parents report a delayed reaction.

Open food challenges are the simplest approach, but open food challenges run the risk of bias influencing the parents' (or doctors') observations. Often this is unimportant. But in some cases belief in food intolerance may be disproportionate, and where this is suspected there is no substitute for a double-blind placebo-controlled challenge. An open challenge may be an open invitation to the overdiagnosis of food intolerance. For example, in the UK parents widely believe that there is an association between food additives and bad behavior, but in one series, double-blind challenges with tartrazine and benzoic acid were negative in all cases in a study of 24 children with a clear parental description of adverse reaction.

The double-blind placebo-controlled challenge is regarded as the state-of-the-art technique to confirm or refute histories of adverse reactions to foods. The ability to unravel food-related problems is said to be limited only by the imagination of the physician and a clever dietitian. In fact, the technique is subject to a number of potential limitations, not all of which can be overcome.

Effect of dose In some cases of food intolerance, minute quantities of food (e.g., traces of cows' milk protein) are sufficient to provoke florid and immediate symptoms. In other cases, much larger quantities of food are required to provoke a response. Hill *et al.* demonstrated that whereas 8–10 g of cows' milk powder (corresponding to 60–70 ml of milk) was adequate to provoke an adverse reaction in some patients with cows' milk protein allergy, others (with late onset symptoms and particularly atopic eczema) required up to 10 times this volume of milk daily for more than 48 h before symptoms developed.

Concealing large doses is difficult Standard capsules that contain up to 500 mg of food are suitable for validation of immediate reactions to tiny quantities of food, but concealing much larger quantities of certain foods (especially those with a strong smell, flavor, or color) can be very difficult.

Route of administration Reactions to food occurring within the mouth are likely to be missed if the challenge by-passes the oral route, e.g., administration of foods in a capsule or via a nasogastric tube. In practice, patients whose symptoms are exclusively confined to the mouth are unusual, and where there is a history of purely oral reactions an alternative challenge procedure can be employed. In subjects who are intolerant to sulfites, it is well recognized that the administration of sulfites in capsules or directly into the stomach via a nasogastric tube usually fails to provoke an adverse reaction, whereas the oral administration of solution will succeed in doing so.

Problems with capsules Capsules are unsuitable for use in children who cannot swallow large capsules, and this is a major limitation as most cases of suspected food allergy are in infants and toddlers. Furthermore, it is unsatisfactory to allow patients or parents to break open capsules and mix the contents with food or drink, as the color (e.g., tartrazine) or

smell (e.g., fish) will be difficult or impossible to conceal and the challenge will no longer be blind.

Anaphylactic shock danger There is a danger of producing anaphylactic shock, even if it had not occurred on previous exposure to the food. For example, in Goldman's classic study of cows' milk protein intolerance, anaphylactic shock had been noted prior to cows' milk challenge in 5 out of 89 children, but another 3 developed anaphylactic shock as a new symptom after cows' milk challenge. In a study of 80 children with atopic eczema treated with elimination diets, anaphylactic shock occurred in 4 out of 1862 food challenges. The risk appears to be greatest for those who have received elemental diets.

Effect of disease activity A food challenge performed during a quiescent phase of the disease (e.g., urticaria, eczema, or asthma) may fail to provoke an adverse reaction.

Additive effect of triggers Although some patients react repeatedly to challenges with single foods, it is possible (but unproven) that some patients only react adversely when multiple allergens are given together. There certainly are some subjects who only react in the presence of a nonfood trigger, such as exercise or taking aspirin.

Special types of provocation testing Other than giving a suspect food by mouth, and asking the subject to swallow it, there are some alternative approaches, which are outlined below.

Oral mucosal challenge A small portion of food is applied to the mucosa inside the mouth, and one looks for reactions such as swelling of the lips, and tingling or irritation of the mouth or tongue, possibly followed by other more generalized symptoms such as urticaria, asthma, vomiting, abdominal pain, or anaphylactic shock. Patients with food intolerance commonly make use of these oral symptoms, spitting out and avoiding further consumption of a food that provokes the symptom.

Gastric mucosal challenge In this procedure, an allergen is applied directly to the gastric mucosa via an endoscope, and the mucosa is then observed for signs of a reaction. In addition, it is possible to take biopsies of the gastric mucosa to study the histological changes and measure the tissue concentration of mediators of inflammation such as histamine.

Rectal challenges The standard test to confirm a diagnosis of celiac disease is the jejunal biopsy, in which a small portion of jejunal mucosa is obtained with the aid of a special capsule that is swallowed, and which passes into the small intestine. When in the correct location, the capsule is triggered and withdrawn; it contains a portion of intestinal mucosa, which can then be examined under the microscope. Alternatively, gluten can be instilled into the rectum, in order to look for a reaction that would signify celiac disease. This procedure requires multiple biopsies from the rectum, and it is uncertain whether the results are reliable.

Management

Dietary Elimination

The management of food allergy consists largely of elimination from the diet of the trigger food or foods. Elimination diets are used either for the diagnosis or the treatment of food intolerance, or for both. A diet may be associated with an improvement in symptoms because of intolerance to the food, a placebo effect, or the improvement may have been a coincidence. The degree of avoidance that is necessary to prevent symptoms is highly variable. Some patients are intolerant to minute traces of food, but others may be able to tolerate varying amounts. Strict avoidance and prevention of symptoms are the aims in certain instances, but in many cases it is unknown whether allowing small amounts of a food trigger could lead to either enhanced sensitivity or to the reverse, increasing tolerance. The duration required for dietary avoidance varies. For example, intolerance to food additives may last only a few years, whereas intolerance to peanuts is usually lifelong. Although food allergy is common in children, most have grown out of the problem by the age of 5 years; an important exception is those with nut allergy.

Malnutrition

Malnutrition is a major risk of unsupervised diets.

Calcium Cows' milk is an important source of calcium, and avoidance of cows' milk and its products carries the risk of an inadequate intake of calcium. Unfortunately, it is far from clear what constitutes an adequate intake for various different age groups.

Protein, energy Milk, eggs, fish, meat, wheat, and their respective manufactured food products are important sources of protein and energy. Avoidance

of these without the provision of alternative sources of protein and energy runs the risk of an inadequate intake, and growth failure, serious malnutrition, and weight loss are well documented sequelae of unsupervised and inappropriate dietary elimination.

Iodine Cows' milk and dairy products are important sources of dietary iodine. Exclusion of cows' milk products and a number of other items from the diet, coupled with the consumption of large amounts of soy milk, which has been reported to cause hypothyroidism by increasing fecal loss of thyroxine, have resulted in hypothyroidism and growth failure due to dietary iodine deficiency.

High-risk factors The risk of malnutrition from an elimination diet is particularly high in the following situations:

1. The diet is not supervised by a dietitian.
2. There is chronic disease prior to diagnosis, or concurrent chronic disease such as severe atopic eczema. The subject's nutrient requirements may be increased.
3. Malabsorption or enteropathy increases the risk of malabsorption of nutrients.
4. The subject is avoiding sunlight. The risk of vitamin D deficiency may compound the effects of a low calcium intake.
5. The subject is already on a diet that excludes multiple foods, e.g., vegan or macrobiotic diet.

The Role of the Dietitian

The dietitian has three roles in the management of elimination diets. One is to ensure that the resulting diet is nutritionally adequate, and to prevent potential deficiency states by recommending (in an infant) appropriate amounts of infant milk formula, and (in older children or adults) supplements of calcium, vitamins, and so on. Another role is to advise how to avoid specific foods, particularly those contained in manufactured foods. Third, the dietitian makes suggestions as to how to make the diet practical and palatable, and suggests recipes for use with a limited range of foods (e.g., how to make biscuits with potato flour).

Cows' Milk Protein Avoidance

Any form of cows' milk, whether fresh, skimmed, condensed, or evaporated, needs to be avoided. Also forbidden are milk products that contain casein, whey, and nonfat milk solids. Where milk substitutes are required, the choice lies between formulas based on soy protein, casein hydrolysate, or whey hydrolysate. Soya formulas are cheaper, but unsuitable for those who are also intolerant to soya.

Butter, margarine, cream, cheese, ice cream, and yogurt all need to be avoided. Fats that can be used instead include margarines made from pure vegetable fat (e.g., Tomor) and lard. Caution is required with baby foods, as a large number of manufactured products, e.g., rusks, contain milk protein. A common trap is so-called 'vegetarian' cheese, often wrongly believed to be safe for subjects with cows' milk allergy. In fact, it differs from ordinary cheese only in the use of nonanimal rennet and is unsuitable for people with cows' milk allergy. Meat, game, and poultry are all allowed, but sausages and pies should be avoided unless it is known that they are milk free. Intolerance to cows' milk protein is not a reason to avoid beef. Eggs are allowed, but not custard or scrambled egg which may contain milk. Fish is permitted, unless it is cooked in batter (which unless otherwise stated should be assumed to contain milk) or milk. Lemon curd, chocolate spread, chocolate (unless stated to be milk-free), toffee, fudge, caramels, and butterscotch are all unsuitable. All ordinary cereals (e.g., oats) are allowed, but caution is required with manufactured breakfast cereals, some of which contain milk powder.

It is essential to check the list of ingredients on the label of any manufactured foods. There is a special problem with unwrapped foods, because there is no label of ingredients. Examples include bread, sausages, or confectionery.

Egg Avoidance

Eggs (both the white and the yolk) and all products that contain egg or albumen must be avoided. As well as hen's eggs, eggs of other birds such as geese, turkeys, and quails must be avoided. Eggs are widely used to make cakes and are sometimes used in the manufacture of bread. Egg wash or glaze is commonly brushed on to the surface of rolls, buns, or baps, and also bread, cakes, and pastry used in puddings (e.g., apple pie). Sweets can be a hazard because they are usually sold without information about ingredients, and egg is included in several products.

Mayonnaise normally contains egg; custard usually does not, with the exception of egg custard and egg custard tarts. Eggs are an essential ingredient of souffles and certain sauces, such as Bearnaise or Hollandaise sauce.

Egg allergy is not a reason to avoid eating chicken.

Soy Avoidance

The major difficulty is mass-produced bread, because in the UK soy is often included as an

ingredient in flour. Soy is also found in manufactured products that contain hydrolyzed or textured vegetable protein, and minced beef, which unless described as 'pure beef' has been known to include quantities of soy protein.

Wheat-Free and Gluten-Free

These terms cause confusion; they are not interchangeable. Subjects who are allergic to wheat cannot tolerate foods that contain any type of wheat. Subjects with celiac disease can tolerate all wheat proteins other than the gluten fraction.

Peanut Avoidance

Peanut is also known as groundnut or arachis, so these three names need to be sought on labels of manufactured foods as well as some pharmaceutical products. The difficulty comes with 'vegetable oil,' which may include peanut oil; only by writing to the manufacturer of individual products can the composition of the vegetable oil be determined. It is not known to what extent subjects with peanut allergy should avoid peanut oil. Most peanut oil used in food manufacture is highly refined, and contains only very minute quantities of peanut protein. In a number of small-scale studies, subjects with peanut allergy were found not to react when given highly refined peanut oil. However, it remains possible that such oil contains traces of protein sufficient to result in enhanced reactivity, such that when the subject does ingest peanut accidentally the reaction is worse than previously. On this basis, subjects with peanut allergy should really be advised to avoid peanut oil.

Drug Treatment in the Management of Food Allergy

At present, drug treatment has little part to play in the management of food allergies. There are two exceptions. First, there are a very small number of cases in which the reaction to a food is exclusively gastrointestinal, and in whom the reaction can be blocked by taking the drug sodium cromoglycate by mouth 20 min before the trigger food is swallowed. Second, there are a small number of individuals who develop the life-threatening reaction, of anaphylactic shock when exposed to a trigger food. There are three ways in which anaphylactic shock may prove fatal. First, rapid swelling of the soft tissues in the pharynx may completely obstruct the airway; the treatment is to bypass the obstruction, either by passing an endotracheal tube, or by performing a tracheostomy. Another mechanism is severe shock, with a profound drop in blood pressure; the life-saving treatment is to restore the circulating volume with intravenous fluids and to give oxygen. The third mechanism is severe bronchoconstriction (asthma); here, the life-saving treatment is with bronchodilator drugs and artificial ventilation. If patients with life-threatening anaphylactic shock are to be saved, they must be given urgent (within minutes) medical attention. For individuals who have already experienced a life-threatening allergic reaction to a food, it is common practice to provide them with a syringe preloaded with adrenaline (epinephrine), with the aim that this should be administered while waiting for medical help. Unfortunately, self-administered adrenaline is not without its hazards (e.g., inadvertent intravenous administration causing fatal cardiac arrest), and there is no proof that it is life saving; indeed, there are many cases in which the subject died despite the use of epinephrine. Nevertheless, it is the best one can do when faced with someone who is experiencing a life-threatening allergic reaction to a food. The need for urgent medical help cannot be overemphasized.

There is little evidence that antihistamine drugs are of any value. It would be reasonable to take a nonsedating fast-acting antihistamine such as terfenadine if experiencing an allergic reaction to a food, but it is questionable whether it will have much effect.

A number of new approaches to the treatment of IgE-mediated food allergy are being examined. In a double-blind placebo-controlled study of monthly injections of a preparation of anti-IgE antibodies, treated patients with peanut allergy required significantly greater amounts of peanut protein to elicit allergic symptoms compared with control subjects. Another anti-IgE preparation has been used in the treatment of asthma but has not been evaluated in peanut allergy. Theoretically, anti-IgE antibody treatment should be protective against multiple food allergens, although it would have to be administered indefinitely. Other experimental approaches include a concoction of traditional Chinese herbs, injection of heat-killed *Escherichia coli* containing mutated recombinant peanut proteins Ara h 1 to Ara h 3, the use of immunostimulatory sequences, and the use of chimeric protein that could form complexes with allergen-specific IgE bound to mast cells and basophils.

Desensitization

In theory it ought to be possible to desensitize subjects with food allergy by giving injections of gradually increasing quantities of an appropriate extract

of the food trigger. In practice, such treatment is not available. One at present insurmountable difficulty is that desensitization (also known as hyposensitization) treatment carries a small risk of death from the treatment itself. A subject has a series of injections without any major problem, but then without warning drops dead from anaphylaxis after the next injection. There is some data to show that desensitization performed in this way can work, but such subjects would probably require maintenance injections on a permanent basis, and the very subjects most at risk of fatal anaphylaxis from accidental injection are quite probably also the ones most at risk from fatal anaphylaxis resulting from desensitization treatment.

See also: **Celiac Disease**. **Eggs**. **Food Allergies**: Etiology. **Food Intolerance**. **Lactose Intolerance**. **Malnutrition**: Secondary, Diagnosis and Management.

Further Reading

Acciai MC, Brusi C, Francalanci S, Gola M, and Sertoli A (1991) Skin tests with fresh foods. *Contact Dermatitis* **24**: 67–68.

Ancona GR and Schumacher IC (1950) The use of raw foods as skin testing material in allergic disorders. *California Medicine* **73**: 473–475.

Bernstein IL (1988) Proceedings of the task force guidelines for standardizing old and new techniques used for the diagnosis and treatment of allergic diseases. *Journal of Allergy and Clinical Immunology* **82**: 487–526.

Bock SA, Buckley J, Holst A, and May CD (1977) Proper use of skin tests with food extracts in diagnosis of hypersensitivity to food in children. *Clinical Allergy* **7**: 375–383.

Bock SA, Sampson HA, Atkins FM *et al.* (1988) Double-blind, placebo-controlled food challenge as an office procedure: a manual. *Journal of Allergy and Clinical Immunology* **82**: 986–997.

Buttriss J (ed.) (2002) *Adverse Reactions to Food*. The Report of a British Nutrition Foundation Task Force. Oxford: Blackwell.

Curran WS and Goldman G (1961) The incidence of immediately reacting allergy skin tests in a "normal" adult population. *Annals of Internal Medicine* **55**: 777–783.

David TJ (1984) Anaphylactic shock during elimination diets for severe atopic eczema. *Archives of Disease in Childhood* **59**: 983–986.

David TJ (1987) Reactions to dietary tartrazine. *Archives of Disease in Childhood* **62**: 119–122.

David TJ (1989) Hazards of challenge tests in atopic dermatitis. *Allergy* **44**(suppl. 9): 101–107.

David TJ (1993) In *Food and Food Additive Intolerance in Childhood*. Oxford: Blackwell Scientific Publications.

Demoly P, Piette V, and Bousquet J (1998) In vivo methods for study of allergy. Skin tests, techniques, and interpretation. In Adkinson NF, Yunginger JW, Bisse WW *et al.* (eds.) *Middleton's Allergy Principles & Practice*, 6th edn, pp. 631–643. St. Louis Mosby.

Fontana VJ, Wittig H, and Holt LM (1963) Observations on the specificity of the skin test. The incidence of positive skin tests in allergic and nonallergic children. *Journal of Allergy* **34**: 348–353.

Ford RPK and Taylor B (1982) Natural history of egg hypersensitivity. *Archives of Disease in Childhood* **57**: 649–652.

Fries JH and Glazer I (1950) Studies on the antigenicity of banana, raw and dehydrated. *Journal of Allergy* **21**: 169–175.

Goldman AS, Anderson DW, Sellers WA *et al.* (1963) 1. Oral challenge with milk and isolated milk proteins in allergic children. *Pediatrics* **32**: 425–443.

Hill DJ, Duke AM, Hosking CS, and Hudson IL (1988) Clinical manifestations of cows' milk allergy in childhood. II. The diagnostic value of skin tests and RAST. *Clinical Allergy* **18**: 481–490.

Josephson BM and Glaser J (1963) A comparison of skin-testing with natural foods and commercial extracts. *Annals of Allergy* **21**: 33–40.

Lessof MH, Buisseret PD, Merrett J, Merrett TG, and Wraith DG (1980) Assessing the value of skin tests. *Clinical Allergy* **10**: 115–120.

Meglio P, Farinella F, Trogolo E, and Giampietro PG (1988) Immediate reactions following challenge-tests in children with atopic dermatitis. *Allergie Immunologie* **20**: 57–62.

Metcalfe DD, Sampson HA, and Simon RA (eds.) (1997) *Food Allergy: Adverse Reactions to Foods and Food Additives*, 2nd edn. Oxford: Blackwell.

Nater JP and Zwartz JA (1967) Atopic allergic reactions due to raw potato. *Journal of Allergy* **40**: 202–206.

Patel L, Radivan FS, and David TJ (1994) Management of anaphylactic reactions to food. *Archives of Disease in Childhood* **71**: 370–375.

Patterson R, Grammer LC, and Greenberger PA (eds.) (1997) *Allergic Diseases. Diagnosis and Management*, 5th edn. Philadelphia: Lippincott-Raven.

Simons FER (2004) First-aid treatment of anaphylaxis to food: focus on epinephrine. *Journal of Allergy and Clinical Immunology* **113**: 837–844.

Voorhorst R (1980) Perfection of skin testing technique. *Allergy* **35**: 247–261.

FOOD CHOICE, INFLUENCING FACTORS

A K Draper, University of Westminster, London, UK

Food choice is about why we eat the foods we do. This would appear to be a simple and relatively straightforward matter, but human food choice is a complex phenomenon, difficult to predict, and its analysis is not a simple affair. Many factors influence our food choice, and these encompass biological, psychological, economic, social, and cultural influences. These all operate on different aspects of food choice and vary in terms of their relative

strength and influence from person to person and context to context. That said, it is a topic of public health importance for two broad reasons. First, it is important to understand the etiology of those nutritional disorders that can be attributed to dysfunctional patterns of dietary intake. This applies to both deficiency disorders, such as vitamin A deficiency, that are due to inadequate nutrient intakes and diet-related chronic diseases, such as coronary heart disease, that are partly attributable to overconsumption or an imbalanced intake of certain foods and/or nutrients. Second, it is important to reduce morbidity and mortality from these disorders via interventions, such as health education, that are designed to either increase or decrease consumption of specific foods and nutrients. The successful achievement of these objectives relies on an understanding and manipulation of the factors that influence food choice, and there is currently much emphasis on the need for health promotion to be more 'evidence-based.' Food choice, like other so-called health behaviors, is thus an important determinant of health and nutritional status and has been the focus of much research.

One of the significant features of this research is that many different theoretical frameworks drawn from both the natural and the social sciences have been used to study the phenomenon of food choice. These derive from many academic disciplines that include biology and physiology, psychology, social psychology, geography, economics, history, sociology, anthropology, political science, and even philosophy. These all approach the study of food choice in very different ways and identify different factors that influence it. This in part derives from the differing ways in which the phenomenon of food choice and the act of eating have been conceptualized. As noted previously, the question of why we eat what we do is deceptively straightforward, but eating, like sex, presents an ontological or analytical problem: Is eating a purely biological act founded on natural need and determined by physiological mechanisms and whose primary function is the meeting of nutritional requirements or is it a form of intentional social behaviour driven by social, psychological, or economic factors and that may serve nonnutritional ends, such as maintaining social relationships or expressing identity? The particular way in which the nature of food and eating are conceptualized inevitably leads to the identification of different types of factors that are seen to influence it. These methodological differences are compounded in that food choice can also be seen as the outcome of a number of different processes or phases ranging from food production and processing to shopping and finally the act of eating or consumption. Again, different academic disciplines have tended to address different steps within this chain and the way in which they impact on the choice of individuals, households, or wider social groups. Finally, the word 'choice' is somewhat problematic; it is rarely explicitly defined, but it is variously used to refer to acts of food selection that range from unconscious behavioral responses to physiological cues and conscious acts of decision making.

This lack of commensurability in research methods, definition of terms, and the actual object of study thus makes it difficult to review and summarize research findings from across different disciplines regarding the different types of factors that influence food choice. However, the main categories of influence on human food choice can be crudely divided into biological and behavioral factors, psychological factors, economic factors, and social and cultural factors. Within each of these categories, particular factors may be either positive or negative influences in that they may act to encourage or discourage choice of particular food items. These categories of influence are discussed in relation to the theoretical frameworks with which they have been studied and with a brief examination of their explanatory value is given.

Biological and Behavioral Influences

There has been an enormous amount of experimental research conducted on both humans and animals that has sought to identify biological and behavioral mechanisms that regulate food intake. This work is grouped together here with the influences on food choice identified by it because most of this work is based on the assumption that food selection or food choice can be explained by internal physiological processes, and that the principal function served by food and eating is to satisfy nutritional requirements. The notion of choice implicit in such models is therefore not one of conscious decision making but, rather, of choice as the outcome of a physiological or behavioral response to a cue or stimulus that may be either internal or external. The notion of choice thus tends to be elided with that which is eaten, and the unit of analysis is generally that of the individual organism or species.

Many different biological models of food selection have been developed and most of these are based on homeostatic models (i.e., systems that work to maintain a balance between the intake of energy and/or specific nutrients and requirements for them) and are seen to operate via stimulus–response mechanisms of various kinds. Food choice is thus seen as driven either via some kind of internal biofeedback

loop that responds to internal physiological states or stimuli, such as states of physiological need, feelings of hunger or satiety, and the energy and nutrient composition of meals, or via detection and response to external cues or influences, such as taste, smell, or the palatability of foods. Various mechanisms have been identified that act to regulate intake of energy as well as individual nutrients, although the evidence to support innate biological mechanisms governing food choice is probably strongest for overall energy intake. Some mechanisms also only appear to act in fairly extreme physiological circumstances, such as the craving for carbohydrates that follows administration of large doses of insulin. With perhaps the exception of salt and water, the evidence linking physiological needs states or cravings with the choice of specific foods is also equivocal.

The regulation of energy balance and appetite in particular has been the subject of a large amount of research. Much of this work has been carried out in relation to obesity and whether this can be linked to a faulty mechanism or genetic defect of some kind. This work is reviewed in detail elsewhere in this encyclopedia, but a number of different mechanisms have been proposed whereby energy intake and balance might be regulated. These include the adaptive thermogenesis theory (now largely discounted, this proposed that energy expenditure was flexible in some individuals and increased to expend excessive energy intakes); nutrient-based models of feeding in which the energy and/or nutrient composition of the diet is considered to lead to appetite suppression via complex gut-fill cues (e.g., the effect of carbohydrates on neurotransmitters and the central nervous system); and the glucostat, lipostat, and leptin theories, which are considered to operate via satiety effects. However, although experimental studies have shown that complex physiological changes are indeed associated with eating, that these vary with what is eaten, and that humans can respond to the covert manipulation of the energy and nutrient composition of our diets, these studies cannot explain the wide variation and flexibility in human food selection or the development of dysfunctional food habits. Humans appear to be more efficient at regulating up (i.e., eating more) when the energy content of the diet is reduced than regulating down. It has been hypothesized that this is the result of evolutionary adaptation to environments in which food supplies were scarce and/or unreliable and in which the ability to deposit energy stores would carry significant advantages. Thus, although there is evidence to support the idea that humans have some kind of appetite control knob (or even knobs) to regulate energy intake, this knob (or knobs) can be overridden, as evidenced by the increasing rates of obesity in many areas of the world.

In relation to external or behavioral cues influencing food choice, there has also been much work investigating relationships between the organoleptic or sensory properties of food, such as taste and palatability, and food preferences and choice. The palatability of food is a complex construct that combines both the sensory qualities of food (taste, smell, and texture) and our hedonic or pleasure response to that food. There has been much work on what makes certain foods more or less palatable. A high fat content, for instance, enhances the palatability of foods, and a liking for high-fat foods appears to be a universal biological disposition. A liking for sweetness and a dislike of bitter flavors also appear to be universal. Again, it has been argued that these carry evolutionary advantages because sweetness is often linked with good dietary energy sources and bitter tastes with foods containing alkaloids and other poisons. Beyond this, however, most likes and dislikes for specific tastes, flavors, textures, and food appear to be learned, often at an early age. Many food aversions, for instance, are very culturally specific.

In summary, although various stimuli, both internal and external, have been shown to influence certain aspects of human food choice, innate biological or behavioral mechanisms alone cannot explain the enormous diversity in human food choice that we see over time and place. As Rozin, a key researcher in the field of food choice, has pointed out, we need to remember that as a species humans are not specialized feeders; we can and do eat an enormous range of foods and diets that satisfy our nutritional requirements. Thus, although there is undoubtedly a physiological base to human nutrition, biology alone cannot explain the complexity of human food choice.

Psychological influences

Much research has been conducted by psychologists and social psychologists on eating behavior, and they have developed a number of different theoretical models within which food choice is conceptualised as a function of specific psychological characteristics or factors. Psychological factors that have been shown to influence food choice include our knowledge, attitudes, emotions, beliefs, intentions, social norms, feelings of self-control or efficacy, cues to act, our early and ongoing experiences, and our relationships with others, including our mothers and other caretakers. These cannot all be reviewed here, but two models that have been

particularly influential in relation to the study and attempted manipulation of food choice are discussed. Note that only psychological influences on normative food choice are discussed here and not the etiology of complex psychological disorders, such as anorexia nervosa.

The knowledge–attitudes–practice model (also known as knowledge–attitudes–behavior) is perhaps one of the most enduring models of food choice. This is a knowledge-based theory in which it is assumed that a particular practice or behavior is derived from certain attitudes, which in turn flow from our knowledge base. Although attractive in its simplicity, any causal link between nutritional knowledge alone and subsequent food choice is far from proven, and increasing an individual's level of knowledge does not necessarily lead the individual to alter his or her diet. This model has been largely discounted as providing a valid account of human behavior, but it still remains the model of food choice implicit in health education activities that seek to change our behavior via the provision of information and enhancement of knowledge about nutrition.

A number of more complex and sophisticated models of health behavior have been developed that draw on social psychology and social learning theory. These are called social cognition models because in addition to our attitudes, they incorporate the beliefs and norms that we hold about our social environment. A social cognition model that has been widely applied to the study of food choice is the theory of planned behavior developed by Fishbein and Azjen (originally called the theory of reasoned action). The theory of planned behavior has been very influential and, in brief, it posits that the principal determinant of behavior, including food choice, is our intention to perform it (i.e., our behavioral intention), which in turn is determined by our attitude toward that behavior, our social and subjective norms (these are our beliefs about what we think other people want us to do and our motivation to comply or not with their wishes), and our perceived behavioral control ('Do I think that I can do it?'—a construct similar to that of self-efficacy). Studies using this model have shown correlations between the various components of the model and consumption differences; for instance, women tend to consume less high-fat foods and also tend to have more negative attitudes toward them. They thus illustrate that in terms of psychological influences on our food choice, we tend to eat what we like to eat and intend to eat, although our likes and intentions are modified to some extent by what we think we should eat and feel we can eat.

Economic Influences

Economics has also been called the 'science of choice,' and there has been much important work by economists on the analysis of economic influences on food choice. Within economic approaches, food choice is conceptualized as an act of consumption broadly equivalent to the consumption of other goods, whether they be clothes, tractors, or dishwashers. Food choice is thus largely conflated with purchase by economists and, as with other commodities, a range of economic factors have been identified that act to both constrain and encourage our purchase of particular foods, mostly centering on the interplay between price, demand, and income.

The simplest economic model of food choice is the demand curve, which describes an individual's choice of foods as a function of his or her income and the price of the food. Thus, as the price of a particular food commodity decreases, we tend to buy more of it. As income increases, we also tend to buy more of the food as well as spend more money on food overall (although it constitutes a smaller proportion of overall expenditure, a phenomenon known as Engel's law). As consumers, however, we do not just buy for cheapness or price, and various more sophisticated economic explanations have developed, such as utility theory and indifference theory. These incorporate the notion of demand or satisfaction in addition to price or income. Utility theory states that when making choices to purchase a particular good or product from a bundle of goods, as consumers we seek to achieve maximum utility. The term 'utility' refers to the satisfaction of needs, wants, tastes, and aspirations, and choice is conceptualized in terms of maximizing these. Thus, the purchase of a food with high prestige value, such as caviar, makes 'sense' in terms of utility theory if not in terms of nutrition. Indifference theory takes the concept of satisfaction further to explain how we make choices between different combinations of goods or foods for maximum satisfaction; as we eat less of one food, such as vegetables, and more of another, such as fruit, the less willing we become to give up some vegetables to get even more fruit. Economic explanations thus move from accounts of choice on the basis of cheapness to combinations of cost-consciousness with maximizing satisfaction or utility within the budgetary constraints of income.

A number of elegant economic laws have been developed to explain and predict food consumption and patterns of expenditure, such as the wonderfully named law of starchy staples. This predicts that as income increases, traditional staple foods are replaced with more refined staples (e.g., the replacement of sorghum or millet with white rice or wheat),

and this is a pattern of change that is being seen today in countries in transition. Such 'laws' and economic analyses demonstrate how factors such as price and income influence food choice and set constraints to it.

Social and Cultural Influences

As Douglas, a famous anthropologist, stated, food is not feed, and food and eating serve many nonnutritional social functions, such as the expression of identity whether individual or national, the maintenance or rupture of social relationships, and religious and symbolic functions. These subtly influence not only what we eat but also how and where we eat and with whom we eat. Food and drink are not taken at random, but within the context of social exchanges between people and within social contexts, such as the family. Therefore, argue the social scientists, to understand the food choices of individuals we must understand the social contexts within which they occur and also the social rules that influence many aspects of our food choice. These can be seen to work at a number of different levels or in a number of ways on the food choice of individuals.

At the most basic level, it is arguably cultural rules of food use that specify the edible versus the inedible. Although the definition of 'food' might appear to be self-evident, there is huge variation cross-culturally in what is classified as edible. This is perhaps illustrated best by examining differences in what are considered edible animal species in different societies (guinea pigs—a tasty snack or domestic pet?). Beyond this basic definition of the edible versus inedible, in all cultures foods are further classified into complex subgroupings, such as hot and cold foods, foods appropriate for meals, snacks, and special events, foods to be eaten/avoided during pregnancy, and so forth. Cultural rules may also specify the food needs of different categories of people (e.g., the young or the pregnant), meal formats and their patterning over time, the way in which food is eaten and prepared, and the allocation of food within an household.

Taken as a whole, sociological studies of food can be read as a form of lay epidemiology that illustrates that not only are the attributes and values that we give to food largely culturally derived but also beliefs about food are culturally constructed and not the product of ignorance or irrational prejudices. Social science approaches to food choice thus study how choices about food are constructed and negotiated in the context of everyday life and how social rules and contexts influence this, and they can offer insights into what otherwise might appear to be irrational or dysfunctional food choices.

Table 1 A hierarchical model of food choice

Edible substances (as defined by culture and physiology)	
Excluded by	**Subsets of**
Culture and rules of allocation	Permissible foods
Availability and cost	Available and affordable foods
Palatability preferences, attitudes, experience	Preferred and chosen foods

A Hierarchy of Constraints?

There is thus a large range of factors that influence human food choice. These are a complex blend of both positive and negative influences that either encourage or constrain our choice of particular foods as well as a combination of biological, psychological, economic, and social factors. These all operate in very different ways and on different aspects of this phenomenon called food choice. How can these be sorted into any kind of scheme or hierarchy to assess their relative importance in any particular context? Various composite models of food choice have been developed, but these often become so global as to be almost meaningless. Following Wheeler, one solution is to turn the issue on its head and start with an analysis of the factors that constrain or limit choice and so identify the range of choice open to an individual in any particular context and what determines this (**Table 1**).

The relative importance of particular influences or constraints will vary from one context to another and also with how they interact, but such a hierarchical framework allows identification of what the relevant factors may be in any given situation and their relative importance for any given individual and thus what degree of choice is actually available to them.

See also: **Appetite**: Physiological and Neurobiological Aspects; Psychobiological and Behavioral Aspects. **Eating Disorders**: Anorexia Nervosa; Bulimia Nervosa; Binge Eating. **Energy**: Balance; Adaptation. **Hunger**. **Obesity**: Definition, Etiology and Assessment. **Religious Customs, Influence on Diet**. **Socio-economic Status**.

Further Reading

Beardsworth A and Keil T (1997) *Sociology on the Menu: An Invitation to the Study of Food and Society*. London: Routledge.

Conner M and Norman P (eds.) (1995) *Predicting Health Behaviour*. Buckingham, UK: Open University Press.

Fine B, Heasman M, and Wright J (1996) *Consumption in the Age of Affluence: The World of Food*. London: Routledge.

French SJ (1999) The effects of specific nutrients on the regulation of feeding behaviour in human subjects. *Proceedings of the Nutrition Society* **58**: 533–540.

Germov J and Williams L (eds.) (1999) *A Sociology of Food and Nutrition: The Social Appetite*. Oxford: Oxford University Press.
Messer E (1984) Anthropological perspectives on diet. *Annual Review of Anthropology* **13**: 205–249.
Murcott A (1998) Food choice, the social sciences and the 'Nation's Diet' research programme. In: Murcott A (ed.) *The Nation's Diet: The Social Science of Food Choice*. New York: Longman.
Rozin P (1996) Towards a psychology of food and eating: From motivation to model to meaning, morality and metaphor. *Current Directions in Psychological Science* **5**: 1–7.
Rozin P and Schulkin J (1990) Food selection. In: Stricker EM (ed.) *Handbook of Behavioral Neurobiology*, vol. 10, pp. 297–328. New York: Plenum.
Wheeler EF (1992) What determines food choice and what does food choice determine? *BNF Bulletin* **17**: 65–73.

FOOD COMPOSITION DATA

S P Murphy, University of Hawaii, Honolulu, HI, USA

Overview: Why Compile Food Composition Tables?

Food composition data are an integral component of evaluating and planning nutrient intakes. Without information on the nutrient content of foods, it is not possible to convert dietary intake data, based on foods consumed, into nutrient intake data. The science of developing accurate food composition data has advanced substantially with the advent of sophisticated laboratory equipment and methods for food analyses as well as increasingly powerful computers that are used to compile and store the results. The International Network of Food Data Systems (INFOODS) at the Food and Agriculture Organization has provided guidelines and training to help countries improve their food composition tables. However, comprehensive analyses of the many nutrients and other bioactive components of foods remain both challenging and expensive, and given the enormous variety of foods consumed throughout the world, food composition tables are often incomplete. Often, the intake calculations that are based on these tables must be regarded as estimates of true nutrient intakes. Nonetheless, for many purposes related to promoting health, nutrient intake estimates are essential and can lead to actions that improve the health of both individuals and populations.

Procedures for Compiling Food Composition Data

Nutrients to Include

As the number of recognized biologically active components of foods increases, compilers of food composition tables are faced with an ever-expanding list of possible nutrients and other components to include. Some of these are given in **Table 1**. The current version of the US Department of Agriculture's Standard Reference Database (release 16-1) contains up to 125 components for each of more than 6600 foods. Because a wide variety of analytic methods is available for determining and reporting nutrient levels in foods, it is useful to have a common convention for naming the nutrients. Many compilers are using standard nutrient names, called tag names, that have been proposed by INFOODS.

Nutrient values in a food composition table normally reflect the level in 100 g of the food item. Thus, the intake of a nutrient from a specific food can be calculated if the amount consumed is recorded in gram weights (e.g., if 100 g of whole milk has 119 mg of calcium, and a person drank a cup of milk weighing 244 g, then the intake of calcium from the cup of milk would be 291 mg). Many food composition tables also contain the weight of typical portions of each food item, and thus the nutrient profile for these portions can readily be calculated.

A related issue is whether to show nutrient profiles per 100 g of the food as consumed or 100 g of the food as purchased. Because some parts of a food may be discarded as inedible, the nutrients per 100 g as purchased will be lower for these foods. For example, a banana skin is approximately one-third of the weight of a banana. If 100 g of a banana without peel has an energy content of approximately 90 kcal, then the energy content of 100 g of banana with peel is only 60 kcal. Composition tables may simply carry a variable for the average percentage of the food that is edible, but it is obviously important to match the method used to measure the food intake (with or without inedible portions) with the way the composition of the food is given in the table.

Foods to Include

There is also a constantly expanding number of foods available in most regions of the world due to changing

Table 1 Nutrients and other food components that are often included in food composition tables

Macronutrients	*Typical units (usually per 100 g)*	*Related components that may be present in a table*
Energy	kcal and/or kJ	
Protein	g	Individual amino acids; nitrogen
Fat	g	
Carbohydrate	g	May be calculated by difference (100 minus grams of the other macronutrients)
Alcohol	g	
Water	g	
Ash	g	
Carbohydrates and fiber		
Sugars	g	Individual monosaccharides and disaccharides
Starch	g	
Dietary fiber	g	May be divided into soluble and insoluble fiber
Nonstarch polysaccharides	g	
Lignans	μg	
Glycemic load	g	Glycemic index
Fats		
Saturated fatty acids	g	Individual fatty acids
Monounsaturated fatty acids	g	Individual fatty acids
Polyunsaturated fatty acids	g	Individual fatty acids
Omega-3 fatty acids	g	
Omega-6 fatty acids	g	
Trans fatty acids	g	
Conjugated linoleic acid	mg	
Cholesterol	mg	
Minerals		
Calcium	mg	
Phosphorus	mg	
Magnesium	mg	
Iron	mg	Heme iron, non-heme iron
Zinc	mg	
Sodium	mg	
Potassium	mg	
Selenium	μg	
Copper	μg	
Chromium	μg	
Molybdenum	μg	
Manganese	mg	
Fluoride	mg	
Iodine	μg	
Vitamins		
Vitamin A	IU, μg RE, μg RAE	
Carotenoids	mg	Individual carotenoids
Retinol	μg	
Vitamin E	mg α-tocopherol	mg α-tocopherol equivalents, synthetic α-tocopherol
Tocopherols	mg	Individual tocopherols
Vitamin C	mg	
Vitamin D	μg	
Thiamin	mg	
Riboflavin	mg	
Niacin	mg	Niacin equivalents
Folate	μg, μg dietary folate equivalents	Synthetic folic acid
Vitamin B_6	mg, μg	
Vitamin B_{12}	μg	
Pantothenic acid	mg	
Biotin	μg	
Vitamin K	μg	
Other food components		
Isoflavonoids	mg	Individual isoflavonoids
Flavonoids	mg	Individual flavonoids

agricultural practices, increases in imported foods, and new commercial product formulations. Including all these foods in a single composition database has not been attempted, and instead, regions and countries have focused on compiling food composition tables that are specific for their populations. Several types of foods are usually found in such composition tables.

Basic agricultural commodities are considered essential in most tables. These include both plant and animal foods that are typically consumed by the population of interest. Frequently, composite values are given in composition tables and reflect an average of multiple samples collected from different regions of the country. For example, nutrient profiles of oranges in the United States are an average of different species of oranges grown primarily in California and Florida; the average is weighted to reflect the production of different types of oranges. Composition tables may contain both cooked and raw values for a food item, which can be helpful if a food is consumed both ways (e.g., tomatoes). Because nutrients may be lost during cooking, and also because the water and fat contents may change, it is important to have nutrient values that correspond to the form of the food that is actually consumed. In addition to cooked and uncooked forms, basic foods may also be available in processed forms, such as canned, frozen, or dried. Many of these processing procedures can alter the nutrients in foods, and thus it is sometimes desirable to have composition data for the differently processed forms of the food.

In addition to basic foods and ingredients, food composition tables usually also contain values for mixed dishes. Some of these mixtures may reflect common recipes that are used in the home, and others may represent commercially available foods, either in food stores or in restaurants. Because recipes may vary greatly, it is particularly useful if the software that accesses the composition table allows the user to alter the recipe ingredients.

Food Descriptors to Use

It is a challenging task to clearly and completely describe the foods that are in the food composition table. Food names for most tables are devised by the compilers using common names plus appropriate descriptors (e.g., cooked, raw, and canned). Ideally, the food descriptors should fully define the food item so there is no ambiguity about the scientific name, the part of the plant or animal that is consumed, and any cooking or processing that has been applied. Several schemes for describing foods have been proposed, including guidelines from INFOODS and the LanguaL system that is used by several European countries.

Sources of Composition Data

Food composition data come from a variety of sources (**Table 2**). Those that are based on laboratory analyses of foods are considered the gold standard for composition tables. Appropriate methods are often specified by the Association of Official Analytical Chemists. Accurate analytic procedures also should incorporate quality control methods, including the proper use of internal standards, and the analysis of duplicate samples to determine intersample variability.

The scheme that is used to obtain and prepare the food samples for analysis is also important. Ideally, the sample collection scheme would match the foods that are reported by the population of interest. For example, if the purpose of the analysis is to determine the nutrient content of a specific person's diet, then the analyses should be performed for a composite of the foods actually consumed or, in the case of feeding studies, for a composite of the foods to be fed. Because such analyses are usually not feasible, more general composition data are often used. Most food composition tables are intended for use across a broad population, and thus the sampling scheme should reflect the types of foods typically consumed. Often, this is an

Table 2 Sources of data for food composition tables

Source	*Comments on accuracy*
Analytic values	
By the table compilers	Generally the most accurate type of data if sampling and analyses are appropriate.
From published literature	May not be correct if the food items differ on important characteristics.
From the food industry	Values from food labels may be underestimates for nutrients added to foods.
From another composition table	May not be correct if the food items differ on important characteristics.
Imputed values	
Based on a similar food	The accuracy of this process depends on how closely the foods can be matched.
Assumed zero	Can be very accurate for some nutrients (fiber in animal products; vitamin B_{12} in plant products)
Calculated values	
From another form of the same food	Usually requires assumptions about changes such as losses due to cooking.
From a recipe	Typical recipe ingredients and proportions may be difficult to collect.
From a product formulation	Useful method for obtaining nutrient values that are not on the product label.

expensive and challenging task, particularly for national and regional tables. Once the sampling plan is devised, it is also necessary to decide on the protocol for storing and preparing the samples for analysis. Considerable nutrient losses can occur if samples are handled improperly because many nutrients are labile to heat, light, and exposure to oxygen. Methods of indicating the quality of analytic data for foods have been proposed, including attaching a confidence code to each data point so that users can decide if the composition data are appropriate for their purposes.

Analytic data are published in various forms. Many countries or regions publish tables, either printed or in electronic form. For example, large tables are compiled by the US Department of Agriculture and also by the Royal Society of Chemistry and the Department for Environment, Food and Rural Affairs in the United Kingdom. Other sources of analytic data include journal articles and books. A particularly useful journal for food composition values is the *Journal of Food Composition and Analysis*, edited by the INFOODS secretariat.

However, analytic data may not be available for all foods and nutrients of interest, and time and cost constraints may prohibit chemical analyses of these foods. In some cases, these values are left blank, and such missing values are assumed to be the same as zero values by most programs that calculate nutrient intakes. Because an appropriately estimated value for a nutrient is usually superior to a value of zero, several methods are used to derive such estimations. A frequent approach is to obtain data from the food composition table of another region or country. If the foods are of the same genus and species, then the nutrient profiles should be similar. Although variations can occur due to different cultivars within a species, as well as different conditions during growing, storage, and processing, such borrowed composition values are considered preferable to a missing value. Another approach to estimating nutrient profiles is to impute a value from a similar food that does have analytic data. If the known nutrients are similar for two foods (e.g., the macronutrient profiles), and the type of food is similar (e.g., dark green vegetables), then the missing value may be replaced with an imputed value from the similar food. Sometimes, calculations are performed to adjust for differences between the foods. Values for a cooked food can be imputed from a raw food by applying factors for nutrient losses during cooking and adjusting for differences in moisture (and sometimes also fat) content.

Another common method of obtaining composition data is to calculate the values from the ingredients in a mixture. For home-prepared foods, such calculations involve determining the proportions of each ingredient (a recipe) and any changes in moisture content during preparation (the yield). There are many challenges in determining recipes that are appropriate for a large group of individuals, but it is equally challenging to try to collect appropriate samples of these mixtures for chemical analysis. For some mixtures, multiple recipes, and thus multiple entries on the food composition table, may be needed (e.g., home-prepared beef stew, commercially canned beef stew, and beef stew from a restaurant).

Composition data may also be obtained from the nutritional labels on commercial food products, if they are available. Most countries require a list of ingredients on the label, and if the proportions of each can be estimated, then a recipe can be devised. It is more useful, however, if the label gives information on the nutrient profile, for at least some of the main nutrients. These values can be incorporated directly into the composition table and also are useful in estimating the proportions of each ingredient (e.g., the amount of wheat flour may be estimated from the carbohydrate content). Because even the most comprehensive nutrition labels seldom give values for all nutrients of interest for the users of food composition tables, recipes will be needed to estimate values for nutrients not shown on the label. Caution should be used with label values for nutrient-fortified products. Good manufacturing practice dictates that the label underestimate the levels of any nutrient, particularly vitamins, that may degrade with time. This ensures that nutrient levels are at least as high as those stated on the label, even after a substantial time on the shelf. Thus, it is always preferable to obtain average nutrient values directly from the product manufacturer if possible.

Compilations of Composition Data for Dietary Supplements

As the use of dietary supplements increases worldwide, there is an increasing need to quantify intakes of nutrients and botanical products from these sources. Compiling nutrient profiles of such products into tables can be very time-consuming because the number of products continues to grow and formulations of existing products often change over time. Furthermore, average analytic data are seldom available from the supplement manufacturers, and thus database compilers must rely on whatever information is available from the product label. In many countries, a label showing the amount of each nutrient in the product is required.

Uses of Food Composition Data

Evaluate or Plan Nutrient Intakes

Uses of food composition data are varied, and the method of compiling the data may need to be

tailored to the application of interest. Perhaps the most common use of composition data is to estimate intakes of individuals. Dietitians and other health professionals may wish to evaluate the quality of a person's current diet or to plan for changes in a diet to meet specific nutrient goals. For example, a person with elevated serum cholesterol may be counseled to reduce saturated fat and cholesterol intakes and given specific menus of diets low in these nutrients. In order to compile these menus, a nutritionist would require access to composition data for saturated fat and cholesterol in a variety of commonly consumed foods. Although the composition data are often averages across many samples of a food, this level of precision is usually acceptable for counseling applications, where long-term compliance with dietary recommendations is being examined.

Similarly, researchers often evaluate or plan diets for individuals as part of nutrition studies. However, for these applications, the required level of precision of the data may be higher. In a feeding study, it may be crucial that the composition of the menus be tightly controlled, and thus average values across many samples are not appropriate. Indeed, it may be necessary to conduct laboratory analyses of the diets that are used in feeding studies rather than rely on more general composition data.

Food composition data are also used to plan and evaluate intakes of population groups, as in dietary surveys, or in choosing menus for institutions such as schools and hospitals. When intakes are to be evaluated and averaged across a large number of people, the use of aggregated food composition data is appropriate and would lead to less error in the estimates than relying on only a small number of samples.

Some users of food composition data may wish to evaluate the nutrient content of foods as purchased at stores or markets. For example, food consumption data may be evaluated for households rather than for individuals, and these data are usually recorded as foods that are purchased for the household. In this case, the composition data must also be given per quantity of food as purchased, before any inedible portions are removed, and before cooking. Likewise, nutrition education for families may focus on making shopping lists of nutritious foods for the household, and composition data for foods as purchased will be helpful.

Food composition data may be used at an even more aggregated level in estimating food use for a region or country. For example, the US Department of Agriculture estimates the nutrient content of the US food supply annually in order to track trends. These data, often called disappearance data, assign nutrient composition values to the major commodities that are produced (minus any exports) or imported for use as food (e.g., flour, sugar, and butter). The amount of each commodity that is available for consumption is multiplied by the corresponding nutrient composition to give an estimate of nutrient consumption per capita.

Estimate Nutrient Profiles for Product Labels

The food industry also uses food composition data to justify health claims for their products (e.g., to indicate that a food product is low in fat) and, in many countries, to obtain information for printing nutrient information on the product labels. Nutrition labels are often required and are considered an important consumer guide to selecting healthy diets. Large manufacturers of processed foods usually obtain laboratory analyses of the nutrients in their products, but smaller companies may rely on calculating nutrient profiles from the product's ingredients. Restaurant chains are also increasingly likely to provide nutrient composition data for the items on their menus.

Evaluate or Plan Food Intakes

Another use of food composition data is to examine intakes from food groups—at the level of the individual, the population group, or the country. Such analyses are facilitated if each of the food items in a food composition table is assigned to a food group using a predetermined food grouping scheme. Once foods are categorized into groups, it is possible to examine intakes from each group (as grams per day) as well as nutrient intakes from each group (e.g., dietary fiber from grains). A further refinement of the food group assignments includes an indication of the number of servings that each food contains (usually per 100 g of the food). Thus, 100 g of orange juice contains approximately one-half of a serving of fruit (assuming $\frac{3}{4}$ cup, or 188 g, of juice is considered a serving). Using such a scheme, it is possible to calculate the number of servings consumed from each food group in a day and compare these intakes to dietary guidance for a country. The Food Guide Pyramid is used for such guidance in the United States, and the US Department of Agriculture has developed a Pyramid Servings Database that may be used to calculate intakes of 30 food groups (**Table 3**).

Limitations of Food Composition Data

Poor Analytic Procedures

Accurate chemical analysis of the nutrient content of foods is a challenging process and may yield

Table 3 Food groups in the USDA Pyramid Servings Database

Grain group	Total grain
	Whole grain
	Nonwhole grain
Vegetable group	Total vegetables
	Dark-green leafy vegetables
	Deep-yellow vegetables
	White potatoes
	Other starchy vegetables
	Tomatoes
	Other vegetables
Fruit group	Total fruits
	Citrus fruits, melons and berries
	Other fruits
Dairy group	Total dairy
	Milk
	Yogurt
	Cheese
Meat group	Meat, poultry, fish
	Meat (beef, pork, veal, lamb, game)
	Organ meats
	Frankfurters, sausage, luncheon meats
	Poultry (chicken, turkey, other)
	Fish (fish, shellfish, other)
	Eggs
	Soybean products (tofu, meat analogs)
	Nuts and seeds
	Cooked dry beans and peas
Pyramid tip	Discretionary fat
	Added sugars
	Alcohol

inaccurate results for a variety of reasons. For some nutrients (and other food components of interest), accurate procedures may not be available. For example, the usual procedures for analyzing the folate content of foods are known to underestimate the actual levels, and thus estimates of folate intakes are likely to be low. Both the extraction procedures and the enzyme digestion treatments may be less than optimal for food folate, and although recent procedures solve some of these problems, folate values on most food composition tables are probably underestimated. Dietary fiber in foods provides another example of possibly incorrect methods. Many older food composition tables contain a variable named fiber, but the values are for crude fiber. Crude fiber is measured using procedures that destroy some of the physiologically important fibers, and thus it is an underestimate of the true dietary fiber content. Recent methods measure either total dietary fiber (defined as all fibers that are not digested in the human gut, including lignans) or nonstarch polysaccharides (excluding lignans).

Inaccurate analyses may also occur when access to the best laboratory equipment is not available, either because the costs are too high or because the technical expertise to use the equipment is not available. For many of the antioxidant compounds, such as carotenoids and tocopherols, quantification by mass spectrometry (MS) yields the most sensitive detection limits, although analysis using high-performance liquid chromatography (HPLC) is adequate in most cases. However, because the equipment, maintenance, and reagents are often too expensive, laboratories (particularly in developing countries) may use older methods, such as spectrophotometry combined with open-column chromatography. Nutrient values derived using such methods are less accurate than those resulting from HPLC and MS methods.

Users of food composition tables should ask when and how analytic values were obtained. Likewise, compilers of these tables should clearly document the analytic procedures used to obtain all values and ensure that such information is readily available to users.

Inappropriate Sampling Procedures

The way foods are sampled and collected can also impact the quality of the composition data. Many nutrient values vary substantially across multiple samples of the same food. Nutrient composition can be affected not only by the species and cultivar of a plant but also by the growing conditions, time of harvest, and length of storage. Because it is seldom feasible to match all these factors with the diets to be analyzed, composite values, based on the average of multiple samples, are usually given in food composition tables.

Inappropriate Nutrient Forms and Expressions

An important limitation for some food composition tables is the method of expressing the activity of the nutrient. The estimation of nutrient activity is a large and expanding field of research and includes studies of both the absorption of the nutrient and its bioavailability for metabolic processes. For example, iron bioavailability has been debated extensively, and many algorithms for calculations have been proposed. Virtually all these require separating iron that is found as heme-iron in animal products from nonheme sources of iron. If these two variables are not carried on the composition table, it will not be possible to calculate iron bioavailability for specific intakes.

Vitamin A also illustrates the complexity of properly expressing the physiologically meaningful form of a nutrient. Until 1967, the vitamin A value of foods was expressed in international units (IUs), which was equivalent to 0.3 μg of retinol and 0.6 μg of β-carotene. This form of expression is still used in many

composition tables and also on nutrition labels for both foods and dietary supplements. A more relevant unit of activity, micrograms of retinol equivalents (REs), was adopted in 1967 and has been used to set recommended nutrient intake levels. A lower relative pro-vitamin A activity of carotenoids was assumed, and thus it is not possible to directly convert IUs into REs, unless both the retinol and the carotenoid levels of a food are given. Recently, the estimated pro-vitamin activity of carotenoids has been further reduced and a newer unit proposed: micrograms of retinol activity equivalents (RAEs). Again, it is not possible to convert between REs and RAEs (or between IUs and RAEs), unless the retinol and carotenoid components of a food are available. Increasingly, food composition tables are carrying separate variables for the specific forms of nutrients such as vitamin A and iron, but this is not the case for many of the older tables. Such disaggregation is an obvious advantage because it allows for recalculation of nutrient activity when there is a scientific consensus that new availability factors are needed. Tables that cannot be easily updated to reflect new information will lag behind the current knowledge and thus will have more limited usefulness.

Lack of Internal Consistency and Integrity

Compiling food composition tables involves recording nutrient profiles for many foods and many nutrients, and errors can easily occur during this process. Quality control is important in this field, just as it is in the development of any product. Developers of the most accurate composition tables always include procedures to ensure that the numbers are correct. In addition to having several people review any new data before they are added to a table, several automated types of integrity checks are possible. For example, the energy value of a food item should approximate the value calculated from the main components (4 kcal/g times the grams of protein and carbohydrate, plus 9 kcal/g times the grams of fat, plus 7 kcal/g times the grams of alcohol), and any deviations should be investigated. Likewise, the sum of all the macronutrients (water, protein, fat, carbohydrate, alcohol, and ash) should be approximately 100 g if the nutrient profiles are given per 100 g of the food. Quality control procedures such as these should be an integral part of the compilation of food composition data.

Conferences on Food Composition Issues

Most of the issues discussed in this article have been addressed at conferences specifically convened to present advances in food composition data. In the United States, the National Nutrient Databank Conference has been held annually for the past 28 years. In addition, the International Food Data Conference has been held biannually for the past 10 years. The proceedings from several recent conferences have been published in the *Journal of Food Composition and Analysis*.

See also: **Dietary Intake Measurement**: Methodology; Validation. **Iron**. **Supplementation**: Dietary Supplements; Role of Micronutrient Supplementation; Developing Countries; Developed Countries. **Vitamin A**: Biochemistry and Physiological Role.

Further Reading

AOAC International (2002) *Official Methods of Analysis of AOAC International, 17th edn* Gaithersburg, MD: Association of Official Analytical Chemists.

Braithwaite E and Selley B (2004) *International Nutrient Databank Directory*. Available at www.medicine.uiowa.edu/gcrc/nndc/survey.html.

Burlingame B (2004) Fostering quality data in food composition databases: Visions for the future. *Journal of Food Composition and Analysis* **17**: 251–258.

Food Standards Agency (2002) *McCance and Widdowson's The Composition of Foods, 6th edn* Cambridge: Royal Society of Chemistry.

Greenfield H and Southgate DAT (2003) *Food Composition Data. Production, Management and Use*. Rome: Food and Agriculture Organization.

Harrison GG (2004) Fostering data quality in food composition databases: Applications and implications for public health. *Journal of Food Composition and Analysis* **17**: 259–265.

INFOODS (2004) *Directory of International Food Composition Tables,* Available at www.fao.org/infoods/directory.

Klensin JC, Feskanich D, Lin V, Truswell AS, and Southgate DAT (1989) *Identification of Food Components for Data Interchange* Tokyo: United Nations University.

Møller A and Ireland J (2000) *LanguaL 2000—The LanguaL Thesaurus*. Report by the COST Action 99–EUROFOODS Working Group on Food Description, Terminology and Nomenclature, Report No. EUR 19540, European Commission.

Rand WM, Pennington JAT, Murphy SP, and Klensin JC (1991) *Compiling Data for Food Composition Data Bases* Tokyo: United Nations University Press.

Truswell AS, Bateson DJ, Madifiglio KC *et al.* (1991) INFOODS guidelines: A systematic approach to describing foods to facilitate international exchange of food composition data. *Journal of Food Composition and Analysis* **4**: 18–38.

US Department of Agriculture (2004) *Pyramid Servings Database*. Available at www.barc.usda.gov/bhnrc/cnrg.

US Department of Agriculture (2004) *Nutrient Database for Standard Reference*, Release 16-1. Available at www.nal.usda.gov/fnic/foodcomp.

US Department of Agriculture (2004) *Food and Nutrient Database for Dietary Studies,* Release 1.0. Available at www.barc.usda.gov/bhnrc/foodsurvey.

FOOD FOLKLORE

J Dwyer and J Freitas, Tufts University, Boston, MA, USA

Introduction

Food folklore consists of traditional beliefs, legends, and customs about food that have been transferred from one generation to the next by word of mouth. For thousands of years, folklore espousing food's nutritional and medicinal benefits has influenced dietary practices. In this chapter, the history and evolution of food folklore is reviewed, and a guide is provided for determining the scientific validity of food folk beliefs.

History of Food Folklore

In addition to sustaining life, food has come to play a symbolic role in both religious ceremonies and cultural traditions. For example, rice has been associated with fertility in many cultures for millennia and continues to be thrown on newly married couples today. Similarly, bread has been regarded as a symbol of divinity and has played an important role in religious services and observances.

Curative properties have also been ascribed to many foods for thousands of years. In ancient Rome, cabbage was considered the perfect medicinal plant and was prescribed frequently for a wide range of ailments including warts, deafness, and drunkenness. Apples, herbs, garlic, honey, milk, peppers, and many other foods were also highly regarded in ancient cultures for their therapeutic qualities. The prescription of foods as medicines was not necessarily based on scientific fact but instead was often based on early medical theories or magic. The ancient Greeks believed that the body was composed of four humors: blood (hot and moist), phlegm (cold and moist), yellow bile (hot and dry), and black bile (cold and dry). Health was thought to result from a balance of the humors, and illness resulted from an imbalance. To counteract imbalances and restore health, physicians often prescribed specific foods, based on their perceived degree of 'heat' and 'moisture'. For example, fever, a 'hot' 'dry' condition, was attributed to an excess of yellow bile, and 'cool' 'moist' foods, such as cucumbers, were prescribed to treat it. In contrast, oedema, a 'cool' 'moist' condition, was treated with foods that were viewed as 'warm' and 'dry'. The hot, cold, moist, and dry properties of food were also regarded as important in other ancient societies, including China, where achieving a balance between the opposing forces of yin (cold/moist) and yang (hot/dry) has guided traditional medical practice for centuries and continues to be popular today.

The Doctrine of Signatures, based on the notion that 'like cures like', was also popular in the nineteenth century. Therapies were chosen on the basis of similarities of color, aroma, shape, and other characteristics. For example, beet juice, which is deep red, was thought to be an effective cure for blood diseases, while yellow plants were believed to alleviate jaundice and other liver ailments. The pungent odours of onions and garlic were thought to ward off disease, stimulate strength and bravery, arouse libido, and banish evil spirits. Walnuts resemble the brain and so were eaten to improve intellect. The ginseng root, with its resemblance to the human torso, was used by the Chinese as a panacea (**Figure 1**).

The common names of many foods reflect folklore about their curative properties, as shown in **Table 1**. For example, the word ginseng is derived from the root words *gin*, meaning man, and *sing*, meaning essence.

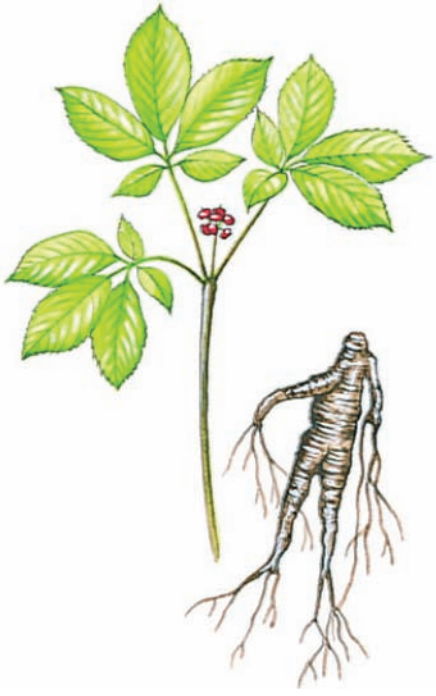


Figure 1 The universal healing properties of ginseng were attributed to the resemblance of its root to the human body. Reproduced with permission from The American Heritage Dictionary of the English Language, Boston, 4th edn, 2000.

Table 1 Food names related to food folklore

Herb (botanical name)	*Folklore*
Blackeye root (*Tamus communis*)	Heals bruises, removes discoloration
Bloodroot (*Sanguinaria candensis*)	Cures blood disorders and heart disease
Birthwort (*Aristolochia longa*)	Alleviates complications associated with childbirth
Eyebright (*Euphrasia officinalis*)	Cures disorders of the eyes
Ginseng (*Panax quinquefolium*)	General human panacea
Heartsease (*Viola tricolor*)	Relieves heart ailments
Liverwort (*Anemone hepatica*)	Relieves liver disorders
Lungwort (*Sticta pulmonaria*)	Cures pulmonary diseases
Maidenhair fern (*Asplenium trichomanes*)	Prevents balding, promotes hair growth
Snakeroot (*Aristolochia serpentaria*)	Antidote for snake bites
Spleenwort (*Asplenium*)	Remedy for disorders of the spleen

Food Folklore Today

While some food folk beliefs continue to be passed down from generation to generation, others have been discarded over the years, and new ones have been introduced. Today, food folklore is spread not only by word of mouth from person to person, but also to large numbers of people simultaneously via the mass media and the Internet. The growing popularity of alternative therapies, organic products, and functional foods has led to the development of new food folklore and increased the popularity of some traditional notions. Several examples of commonly held food folk beliefs of both the past and present are provided in **Table 2**.

Food and nutrition-science concepts that have developed over the past two centuries are newcomers to human thinking about the relationships between food and health. Although some food beliefs, such as the association of carrots with eyesight, have some scientific basis, many others remain unsupported by, or in opposition to, recent scientific findings. Pseudoscience, rather than sound evidence, provides the basis for much of today's food folklore. It is important for food and nutrition professionals to be knowledgeable about current food folk beliefs, because these ideas influence popular views about diet–health relationships. In order to identify beliefs that are of major health significance, it is useful to consider the strength of the belief in folklore and the strength of the scientific evidence surrounding it, as shown in **Table 3**. When folkloric belief and scientific evidence are both strong and in agreement with each other, the folklore is unlikely to pose a major health threat. Similarly, when food folklore and the scientific evidence surrounding it are both weak, few practical problems exist. However, when scientific findings refute, or fail to support, popular food folk beliefs, public health may be threatened. For example, despite folklore that ephedra promotes rapid weight loss, a recent evidence-based review suggests that it does not and that it may even be harmful. Often, though, it is not that the scientific evidence regarding food folklore is weak but that it is unproven or undetermined, indicating that more research needs to be done.

Folklore and Evidence: Fact or Fiction?

Totality of the Evidence

Food folklore cannot be taken as fact without evidence. Single studies are usually inadequate for demonstrating cause–effect relationships, and no single study alone is enough to prove that something is fact or folklore. It is important to consider the totality of evidence and the type and quality of the available research. When many different types of evidence are all supportive of a relationship, the weaknesses of individual studies are mitigated and causal inference is strengthened.

Ideally, the best way to conduct a scientific evaluation of a question is to perform an evidence-based expert review of many randomized double-blind placebo-controlled clinical trials, meta-analyses, and other studies. An evidence-based review often entails the use of statistical techniques to reanalyze the results of many small studies, as well as expert judgment. If all systematic evidence-based reviews produce comparable conclusions, the scientific evidence is good that the folklore belief is justified.

Comprehensive reviews are especially important for far-reaching questions that have implications for large populations. They are also necessary for questions regarding important issues, including life and death, and those that involve very large costs or imply large reimbursements. However, because such reviews are significantly time consuming and require much expertise and money, they can be done only for a few very important questions. For example, the National Institutes of Health (NIH) has sponsored such reviews of obesity treatment, the American Institutes of Cancer Research has done reviews to substantiate their population-based recommendations, and the health

Table 2 Food folkore: current and historic beliefs

Food	*Folklore*
Fruits	
Apple	'An apple a day keeps the doctor away'; prevents cavities and tooth decay
Blueberries	Cure for kidney and urinary-tract ailments; improves vision and memory
Cherries	Cherry gum dissolved in wine relieves a cough; ensure continued fertility
Citrus fruits	Prevent scurvy; Cause low blood pressure; cure the common cold
Cranberries	Prevent scurvy; prevent cure or urinary-tract infections
Currant	Relieves sore throat
Fig	Relieves toothache; mild laxative
Grapefruit	Should be avoided completely when taking medication; burns calories, dissolves fat, aids in weight loss; is 'good for you'
Raspberry	Raspberry leaf tea promotes contractions and aids in labor
Vegetables	
Beets	An effective cure for anemia; help build iron-rich blood
Carrots	Important for good eyesight
Celery	Promotes weight loss
Garlic	Stimulates digestion; inhibits germs; cleanses the blood and intestines; lowers cholesterol and blood pressure
Lettuce	Induces sterility
Onions	Cooked onions cure the common cold; good for the heart
Peppers	Cure for a cold
Potatoes	A historic cure for impotence; prevent scurvy; soothe and soften the skin; are fattening
Spinach	Builds strong muscles
Grains	
Bread	Is the body of the supernatural; cures disease and protects against evil; is a fattening; food brown bread has more fiber than white bread
Flaxseed	Effective cure for constipation; prevents cancer; lowers cholesterol
Oats	Oatmeal and oat bran can prevent heart disease
Dairy	
Milk	Prevents scurvy; helps to heal ulcers; causes constipation; unpasteurized milk has more nutrients than pasteurized; a glass of milk before bed causes drowsiness; mothers who drink a lot of milk have colicky babies; milk and other dairy products are fattening and should be avoided on a low-fat diet; the calcium in milk and other foods causes kidney stones
Yoghurt	Prevents vaginal yeast infections; cures vaginitis, constipation, and diarrhea; yoghurt applied topically heals a sunburn
Meat	
Beef	Eating beef and other red foods causes high blood pressure; extra protein from beef makes muscles stronger
Chicken soup	Cure for the common cold
Eggs	Drinking raw eggs helps build muscle; brown eggs are healthier than white eggs; people with high cholesterol should not eat eggs
Legumes	Beans are a natural laxative
Seafood	Fish is a brain food; fish is good for the heart and prevents heart attacks; pregnant women should avoid eating fish; oysters increase sexual potency
Fat, sweets, and alcohol	
Olive oil	Protects against breast cancer
Cod liver oil	Relieves rheumatism, aching muscles, and stiff joints; prevents rickets
Sugar	Causes tooth decay; causes hyperactivity; eating too much causes diabetes and heart disease
Honey	Is natural and will not raise blood-sugar levels; a mix of honey and water is a good cure for colic
Chocolate	Causes acne; eating chocolate helps to prevent heart disease
Salt	A no-salt diet protects against high blood pressure; sea salt is healthier than table salt; salt tablets prevent muscle cramps
Alcohol	Helps to warm the body in cold weather; acts as a sleep aid if consumed before bedtime; red wine is good for the heart; a nip of brandy cures a cold; drinking alcohol with raw oysters ensures they are safe

effects of ephedrine were recently summarized in an evidence-based review by the Office of Dietary Supplements. Unfortunately, for much food folklore, such studies have never been done.

Type of Evidence

In determining the validity of food folklore, not only the totality of evidence but also the type and quality of available evidence are important. As shown in

Table 3 Folklore: separating fact from fiction

Folklore	Scientific evidence: Strong	Scientific evidence: Weak
Strong	Fact[a]	Fiction/ undetermined[b]
Weak	Emerging science[a]	Fiction[a]

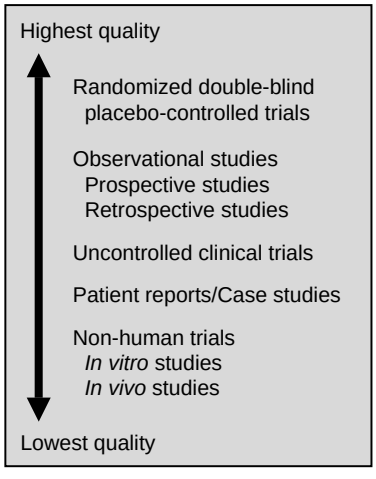
[a]Unlikely to be a major threat
[b]Potential threat to public health

Table 4, the strength of the association between eating a food (cause) and a health outcome (effect) can be ranked according to the type of evidence presented. The best evidence comes from studies that have the most control over the claim or treatment being evaluated and eliminate other factors that may suggest an effect was present, when really it was not. Although randomized double-blind placebo-controlled clinical trials are considered the 'gold standard' for determining diet–health relationships, such studies are rarely available. Lesser levels of evidence must usually be used.

Randomized double-blind placebo-controlled trials When randomized double-blind placebo-controlled trials show a relationship between a specific food and a health effect, the evidence is considered to be very good. These studies exert rigorous control over the claim or treatment being evaluated and over the people who are subjected to it (by randomization) and the assumptions of both the experimenters and the participants (by placebos and blinding). Multiple studies of this type, with an expert review of all other types of data, are considered to be the 'gold standard' for establishing cause–effect relationships. Other types of evidence and studies are lower in the hierarchy, because they are not as definitive in identifying true cause and effect.

Although single randomized trials and non-randomized clinical studies are somewhat less definitive, they are still valuable, because they also permit control over the treatment being evaluated. Often, however, these studies are not large enough, or the study sample is not representative, so the results cannot be generalized to the problem at hand. Other factors that may weaken these studies are not counting of dropouts, lacking or unconvincing placebos, and inappropriate points or biomarkers serving as surrogate end points.

Table 4 Ranking the quality of the evidence



Observational studies Human studies that involve observation rather than direct intervention provide evidence that is satisfactory but less conclusive. These studies are designed to test a relationship between an exposure of interest (folk belief) and a health outcome. Observational studies include both cohort studies (prospective) and case–control studies (retrospective). In a prospective study, a group exposed to the treatment of interest and an unexposed group are followed forward in time. The health outcomes in both groups are observed and evaluated after controlling for confounding factors with the use of statistics. In contrast, retrospective studies compare individuals who have already developed an outcome of interest (case) against those who have not (control). Factors contributing to the development of the outcome are then determined by looking backward in time. Because observational studies cannot be precisely controlled, it is more difficult to establish cause and effect. However, when confounding factors can be adequately controlled for, these studies provide suitable evidence to support diet–health relationships.

Uncontrolled clinical trials Clinical studies in which everyone is treated, in which only those who ask for the treatment are treated, or in which some are treated based on unsubstantiated clinical convictions are suspect. In such studies, no randomization occurs and neither the researcher nor the participant is blinded to the treatment. Therefore, it cannot be determined whether the treatment is actually the cause of the observed results or whether biased

convictions of both experimenters and study participants are falsely contributing to the results. Better evidence is needed before it can be stated with assurance that the folklore based on such observations is true.

Patient reports, case studies, and folklore Even weaker human evidence of cause and effect comes from single medical case reports and anecdotal evidence. These types of evidence are also biased since those who experience success from the treatment are much more likely to report their stories than those who do not.

Animal studies and laboratory experiments Non-human studies involving living animals (*in vivo* studies) or tissue cultures (*in vitro* studies) are useful in providing information on the possible biological plausibility, dose response, and action of a treatment. However, their ability to predict outcomes in humans is poor. Therefore, these studies are unconvincing and should be used only to support other types of evidence.

Guide for Separating Food Folklore Facts from Fiction in Clinical Situations, and A Practical Example

For summarizing and evaluating food folklore involving diet–health relationships, health professionals need to not only evaluate the evidence but also use their clinical judgment and communications skills to relate this to clients or patients. How can food folklore be evaluated in discussions with laypeople and in counselling situations? The strategies are similar to those employed in research and in more formal evidence-based reviews, but contextual realities require tailoring of the approach. One method of evaluation is provided, as shown in **Table 5**, and an actual clinical example follows.

The problem One example of currently popular food folklore is the notion that people on medication should not drink grapefruit juice. Although there is scientific evidence that grapefruit juice interacts with certain medications, making them incompatible, the facts do not suggest either that grapefruit must be eliminated from the diet of those taking medications or that all drugs exhibit these interactions. The process of reviewing this food folklore with the patient to arrive at this conclusion is outlined in detail below.

Table 5 Steps to evaluating food folklore in clinical situations

1. Report
2. Review
3. Recall
4. Relate
5. Recommend
6. Revise

In the late 1980s, in a study examining the interaction between alcohol and felodopine (a calcium antagonist used to lower blood pressure), it was accidentally discovered that grapefruit juice, which was being used as a placebo, dramatically altered the drug's metabolism. The drug was a common one, the juice dose – about 6 ounces (180 ml) – was within the range many people drink, and the effects were large (similar to a doubling of the drug dose). Therefore, it was of potential clinical importance. Since then, more than 200 scientific papers have been published in peer-reviewed journals on the issue of drug interactions with grapefruit, confirming the original observations. By the mid-1990s, the finding had received a great deal of media coverage and the notion that grapefruit juice was dangerous for those on prescription drugs had become a subject of food folklore. This particular bit of folklore is an example of a strongly held belief for which there is some scientific evidence. Under the circumstances, how should clinicians advise patients?

Report In counselling, it is important for the clinician to relate to the patient and establish two-way communication to learn about the folk belief. It may be useful to determine the strength of the individual's conviction about this folklore as well as the source of the belief and whether there is a potential conflict of interest. When the health professional actively listens, it is more likely that the patient will listen, understand, accept, and follow recommendations.

Review In clinical situations, it is also important for health professionals to review all the evidence surrounding the patient's professed food folk belief. A vital piece of information to consider is safety. Although many prescription drugs have side-effects, they are taken under the supervision of a physician, who can monitor adverse effects and take steps to control them. Because folk remedies are often self-administered, such safeguards are lacking. If there is evidence that the implementation of food folklore is likely to be hazardous to health, it must be discouraged. Folk remedies that are effective for one purpose but have negative side-effects must also be cautioned against.

For example, when the patient's prescription drug is one that is metabolized by cytochrome 3A (CYP 3A), a dramatic effect can occur if grapefruit juice or other forms of the fruit are consumed. Grapefruit juice enhances the effects of these drugs over time by decreasing their oral clearance. However, the effects of the interaction depend on the nature of the drug (for some drugs there is little or no effect) and the size of the interaction. Interaction occurs only if the drug is metabolized by CYP 3A, if it normally undergoes pre-systemic extraction with CYP 3A, and if it is given orally. The interactions vary. For example, with the statins – drugs commonly used to lower serum cholesterol – they are strong for simvastatin and lovastatin, moderate for atorvastatin and cervastatin, and low for fuvastatin and pravastatin. Similarly, sedatives, hypnotics, and other drugs vary as to whether they induce interactions or not.

Recall The CYP P450 superfamily consists of many enzymes. They are labelled as follows: CYP (family 1, 2, 3, etc.) (subfamily A, B, C, etc.) (isoform 1, 2, 3, 4, etc.). There is individual variability in the expression of these enzymes, which is probably in part genetic.

CYP 3A is an enzyme that is involved in the metabolism of many drugs. It is present in the liver and gut mucosa and is induced or inhibited by drugs and other chemicals. Under certain conditions, components of foods can also affect it. The CYP 3A enzyme in the gut mucosa (enteric CYP 3A) is affected by grapefruit juice, but CYP 3A in the liver is not. The phytochemicals that are thought to have these effects are furanocoumarin derivatives in the juice, which reversibly and irreversibly inhibit the CYP 3A in the gut mucosa. When grapefruit juice and certain drugs are taken together orally, this leads to effects on their pre-systemic extraction (first-pass metabolism). In consequence, pre-systemic extraction of the drug is reduced, and, because of this, more of the drug reaches the circulation over time. With this increased systemic exposure to the drug, there is an increased drug effect. The duration of the juice's effect depends on the dose and on the time it takes for the enzyme to regenerate. The liver CYP 3A is not affected by grapefruit juice (although it may be affected by some drugs that also affect gut CYP 3A). Thus, the grapefruit–drug interaction does not happen with drugs administered intravenously, since they bypass the gut.

About one-fifth of American households consume grapefruit juice, and it is considered a 'good food' since it has the American Heart Association heart check and the American Cancer Society endorsement. Many older people take their medications and juice together at breakfast. Many people who are elderly take medications, and some of these drugs may pose problems. Therefore, this folklore is highly relevant from the clinical standpoint.

For the grapefruit interaction to take place, the patient must be taking a drug that affects the gut CYP 3A and he or she must express a significant amount of CYP 3A. People differ in these respects, and there may be racial as well as other genetic differences.

Relate The clinician or nutrition scientist must build on his or her own knowledge and that available from expert reviews or sources and place it in the clinical context. Common sense is needed to fit the information to the patient's realities. The facts, which are that most drugs do not interact with grapefruit, must be related to the folklore and the patient's actual condition. Many grapefruit–drug interactions are modest and not clinically important. Fortunately, for every therapeutic class of drugs there is an option that is not affected by the grapefruit interaction.

Recommend In responding and making recommendations, considerations include their importance, feasibility, and effectiveness for the patient. The information is then particularized to fit the patient or questioner's problem and needs.

This is the time to particularize for the patient and lead him or her to the next level of understanding. With many drugs there is no interaction, and the patient can be told to drink the juice if he or she wishes but to alert the physician if adverse reactions occur. It may be possible to change the medication if a major interaction exists, or, for modest interactions, it may be enough to avoid taking drugs and grapefruit juice at the same time and to avoid consuming large quantities (four or more glasses) of juice.

The patient should be praised for asking about possible food–drug interactions and told that these reactions sometimes occur, but do not usually exist. Is the information relevant – that is, if the patient is on a statin, is it the type that is involved in interactions? Does the patient *want* to drink grapefruit juice? If not, the issue is moot. What are patient reactions, and is reassurance needed?

Revise Fortunately, in counselling, although a relatively rapid response is usually required, there is an ongoing relationship with the individual that permits follow-up. The clinician can follow up and revisit the issue later if necessary when more information becomes available or when additional questions arise.

For example, a patient might ask whether, since the drug effect is enhanced with the consumption of grapefruit juice, he or she could save money by taking more grapefruit juice and continuing with lovastatin. The response to this legitimate question is no, not because it is theoretically impossible but because it is difficult to titre the drug, individual differences exist, and reactions are unpredictable, so such a strategy is not recommended.

Conclusions

Food folklore is alive and well and is likely to continue. For summarizing and evaluating food folklore involving diet–health relationships, health professionals need to not only evaluate the evidence but also use their clinical judgment and communications skills to relate this to patients. The six Rs (report, review, recall, relate, respond, recommend, and revise if necessary) provide a guide for evaluating food folklore in clinical situations.

See also: **Dietetics**. **Drug–Nutrient Interactions**. **Food Choice, Influencing Factors**. **Fruits and Vegetables**. **Functional Foods**: Regulatory Aspects.

Further Reading

Duyff RL American Dietetic Association (1999) *Food Folklore: Tales and Truth About What We Eat*. Minneapolis: Chronimed Publishing.

Anderson H, Blundell J, and Chiva M (eds.) (2002) *Food Selection: From Genes to Culture*. Brussels Belgium: Chauvehid, Stavelot, Juillet.

Andrews T (2000) In *Nectar and Ambrosia: An Encyclopedia of Food in World Mythology*. Santa Barbara: ABC-CLIO.

August DA (2000) Clinical guidelines: an evidence based tool to lead nutrition practice into the new millennium. *Nutrition in Clinical Practice* **15**: 211–212.

Cavendish R (1983) *Man Myth and Magic: The Illustrated Encyclopedia of Mythology, Religion and the Unknown*. New York: Marshall Cavendish Corporation.

Green TA (ed.) (1997) *Folklore: An Encyclopedia of Beliefs, Customs, Tales, Music and Art*. Santa Barbara: ABC-CLIO.

Kennett F (1976) *Folk Medicine: Fact and Fiction*. New York: Crescent Books.

Kiple KF and Orneals KC (2000) *The Cambridge World History of Food*. Cambridge: Cambridge University Press.

Koretz RL (2000) Doing the right thing: the utilization of evidence based medicine. *Nutrition in Clinical Practice* **15**: 213–217.

Leach M and Fried J (eds.) (1949) *Funk & Wagnall's Standard Dictionary of Folklore, Mythology, and Legend*. New York: Funk & Wagnall Company.

Lehner E and Lehner J (1962) In *Folklore and Odysseys of Food and Medicinal Plants*. New York: Tudor Publishing Company.

Rinzler CA (1979) *The Dictionary of Medical Folklore*. New York: Thomas Y Crowell Publishers.

Visser M (1991) *The Rituals of Dinner: The Origins, Evolution, Eccentricities, and Meaning of Table Manners*. New York: Penguin Books.

Walker ARP (1998) Food folklore overview. In: Sadler MJ, Strain JJ, and Caballero J (eds.) *Encyclopedia of Human Nutrition*, vol. 2, pp. 875–880. San Diego: Academic Press.

Wilson DS and Gillespie AK (eds.) (1999) *Rooted in America: Foodlore of Popular Fruits and Vegetables*. Knoxville: The University of Tennessee Press.

FOOD FORTIFICATION

Contents

Developed Countries

R Nalubola, Center for Food Safety and Applied Nutrition, US Food and Drug Administration, MD, USA

Published by Elsevier Ltd.

Introduction

Food fortification is increasingly recognized as an effective public health intervention to alleviate nutritional deficiencies. Nutrients may be added to foods to either restore nutrients lost during processing (restoration or enrichment), introduce nutrients not naturally found in the food, or increase the levels of nutrients to above those naturally present in the food (fortification). The terms 'enrichment' and 'fortification' are often used interchangeably to simply refer to the addition of nutrients to food; however in some countries these terms have specific regulatory definitions. In this article, food fortification refers to all such nutrient additions.

History and Current Market Place

Food fortification has been practiced for several decades in developed countries. Historically, certain staple foods have been fortified to alleviate deficiency diseases in the general population, such as the iodization of salt to alleviate goiter, addition of vitamin D to milk to prevent rickets, and enrichment of cereal products with thiamin and niacin to alleviate beriberi and pellagra. For example, in the US, fortification began with the iodization of salt in the State of Michigan in 1924. The practice of salt iodization had spread rapidly throughout the rest of the country by 1928. With the revolution of nutritional science and the elucidation of chemical structures of several nutrients in the 1930s, the production of synthetic preparations of nutrients became possible. By 1934, vitamin D was routinely being added to milk. Although an effort to enrich cereal flours and products with B vitamins and iron started in the 1930s, the enrichment program began in earnest only in 1941 with about 40% of the flour in the country being fortified by 1942. Furthermore, in 1951, the Committee of Food and Nutrition of the National Research Council recommended the addition of vitamin A to table fats (margarine). Several other developed countries experienced similar initiation and implementation of food fortification as a strategy to alleviate nutrient deficiencies in the general population. More recently, cereals fortified with folic acid have been introduced in several countries to increase the folic acid intake of women of childbearing age and reduce the rates of neural tube defects, and the results are already measurable. Foods commonly fortified in developed countries are listed in **Table 1**.

Although the original intent of food fortification was to correct or prevent widespread nutrient deficiencies, the focus now has shifted in developed countries to optimal intakes of nutrients for the prevention of chronic diseases and for overall health and well-being. Classic nutrient deficiency diseases such as goiter and pellagra are no longer common in developed countries at least, in part, due to food fortification programs. However, suboptimal intakes of nutrients such as calcium and vitamin D may occur. In addition, there is heightened consumer awareness and demand for more healthful foods. Consequently, food industries perceived a market for foods with improved nutritional profiles, resulting in a proliferation of fortified foods in the current market place. Ready-to-eat breakfast cereals are a prime example of industry-initiated food fortification. Published reports indicate that in 1969, only about 16% of the breakfast cereals in the US market place were fortified; that number rose to about 92% in 1979 following certain legislative actions providing more flexibility to industry for the addition of nutrients to foods. A market survey in Germany highlights the heterogeneity of current fortification practices: a total of 288 fortified foods in six different food categories (beverages, sweets, cereals, milk products, powdered instant beverages, and ready-to-eat meals) were found to be fortified with a wide range of nutrients.

Table 1 Foods commonly fortified in developed countries

Foods	*Nutrients*
Cereal grain flours and products	Thiamin, riboflavin, niacin iron, and folic acid
Milk and milk products	Vitamins A and D
Fats and spreads	Vitamins A and D
Salt	Iodine
Ready-to-eat breakfast cereals	Many vitamins and minerals
Juices and other beverages	Vitamin D and calcium
Infant foods	Many vitamins and minerals

Characteristics of a Fortification Program

Food fortification can be mandatory (required by government regulations) or voluntary (permitted by government regulations and policies). Some characteristics of both types of fortification programs are presented in **Table 2**. As with any intervention, each of these has certain advantages and limitations that have to be considered to determine the regulatory course that is most appropriate for the country and for the public health concern at hand. For example, it is reported that voluntary fortification of foods with folic acid in Australia and New Zealand, while demonstrating benefits, has not reached the target population to the extent intended. Therefore, policymakers are considering mandatory fortification as an option to ensure sufficient folate intake among all women in Australia and New Zealand.

At the international level, the Codex Alimentarius Commission has adopted a set of criteria which allow for the rational addition of nutrients to improve the nutritional quality of the overall food supply and prevent indiscriminate fortification of foods, while acknowledging that the need for, and appropriateness of, addition of nutrients to foods will depend on the nutritional problems of a country, the characteristics of the target populations, and

Table 2 Characteristics of mandatory and voluntary food fortification

Mandatory fortification	*Voluntary fortification*
• Initiated by government[a]	• May be initiated by government or by industry
• Regulations require addition of nutrients	• Regulations provide for optional addition of nutrients
• Food vehicle(s) are staple foods consumed in significant and relatively stable amounts	• Both staple and nonstaple foods are used as food vehicles
• Driven primarily by a documented need for a public health intervention	• Driven primarily by consumer demand and market forces
• Resource-intensive, coordinated multi-sector effort	• Relatively fewer resources are needed, particularly from the public sector
• Targeted to reach populations at risk of deficiency	• Can be introduced quickly
• Excessive intakes of fortified nutrients are minimized	• In the case of breakfast cereals, proven to make substantial contributions to nutrient intakes
• Adverse effects on other nutrient intakes or health conditions due to fortification are curtailed	• Random fortification and overfortification of the food supply is a potential concern although this can be minimized by appropriate regulatory restrictions
• Fortification levels and food vehicles are tightly controlled	
• A cost-effective nutrition intervention for governments	
• Proven to be an effective strategy to eliminate micronutrient deficiencies in the general population	

[a]While government-initiated fortification is usually mandatory, it can be 'voluntary' in that food companies have the option of selling unfortified versions of the food provided they are appropriately labeled. For example, in Canada, 'flour' or 'enriched flour' is required to contain specified amounts of thiamin, riboflavin, niacin, folic acid, and iron (B.13.001, Food and Drug Regulations) and 'milk' is required to be fortified with vitamin D (B.08.003, Food and Drug Regulations). In the US, however, 'enriched flour' is required to contain these nutrients, but 'flour' is not (Title 21 of the Code of Federal Regulations, sections 137.105 and 137.165) and 'milk, vitamins A and D added' is required to contain specified amounts of these vitamins, but 'milk' is not (21 CFR 131.110).

their food consumption patterns. The Codex general principles for the addition of essential nutrients to foods (CAC/GL 09-1987) state the following:

- The nutrient should be present at a level that will not result in an excessive or an insignificant intake of the added nutrient considering the amounts from other sources in the diet.
- The nutrient should not result in an adverse effect on the metabolism of any other nutrient.
- The nutrient should be sufficiently stable in the food during packaging, storage, distribution, and use.
- The nutrient should be biologically available from the food.
- The nutrient should not impart undesirable characteristics to the food, or unduly shorten shelf-life.
- The additional cost should be reasonable for the intended consumers, and the addition of nutrients should not be used to mislead the consumer concerning the nutritional quality of the food.
- Adequate technology and processing facilities should be available, as well as methods of measuring and/or enforcing the levels of added nutrients.

The Codex principles mirror the criteria proposed in the US (in 1974) as conditions to be met to support food fortification, including the following:

- There should be a demonstrated need for increasing the intake of an essential nutrient in one or more population groups.
- The food selected as a vehicle should be consumed by the population at risk and intake of this food should be stable and uniform.
- The amount of nutrient added should be sufficient to correct or prevent the deficiency when the food is consumed in normal amounts by the population at risk.
- The addition of the nutrient should not result in excessive intakes.

Many developed countries have adopted various regulations to either require or permit manufacturers to fortify their food products in a safe and appropriate manner, and to market such products in a manner that is truthful and not misleading to consumers. In some countries, such as Norway, Finland, and Denmark, voluntary fortification has been considered unnecessary and potentially harmful and, therefore, is generally restricted. However, in many other countries, such as the US, UK, Canada, Switzerland, and Belgium, regulations are less restrictive allowing foods to be fortified voluntarily as long as fortification is safe and harmful levels of nutrients are avoided.

Contribution of Food Fortification to Nutrient Intakes

Individual nutrient intakes in developed countries are strongly influenced by food fortification. The amounts of thiamin, riboflavin, niacin, and iron in

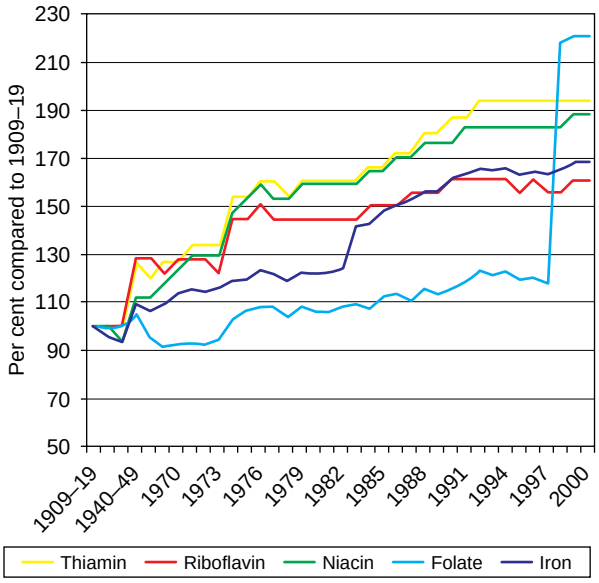


Figure 1 Trends in the amounts of thiamin, riboflavin, niacin, folic acid, and iron in the US food supply between 1909 and 2000. Fortification of cereal grain products with thiamin, riboflavin, niacin, and iron was initiated in 1941 and fortification levels were adjusted in 1973. The levels of iron were reversed in 1978 and then increased in 1981. Folate fortification began in 1998. Source of data: Gerrior S, Bente L, and Hiza H (2004) *Nutrient Content of the United States Food Supply, 1909–2000*. Home Economics Research Report No. 56. United States Department of Agriculture, Center for Nutrition Policy and Promotion, Alexandria, VA, USA.

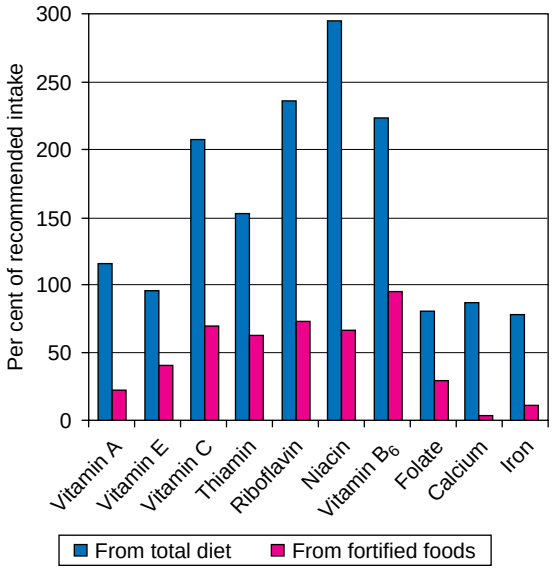


Figure 2 Nutrient intakes from the total diet and fortified foods among 2–13-year-old German children in 1996. Source of data: Sichert-Hellert W, Kersting M, Alexy U, and Manz F (2000) Ten-year trends in vitamin and mineral intake from fortified food in German children and adolescents. *European Journal of Clinical Nutrition* **54**(1): 81–86.

the US food supply markedly improved following fortification of cereal grain products in the 1940s (**Figure 1**). Cereal grain products provided about 31% of the thiamin in the food supply in 1909–19 compared to about 59% in 2000; 14% of riboflavin in 1909–19 to 39% in 2000; 30% of niacin in 1909–19 to 45% in 2000; and 33% of iron in 1909–19 to 52% in 2000. Similarly, fortification sharply increased the amounts of folate in the food supply since its inception in 1998. Cereal grain products accounted for 23% of folate in the food supply in 1909–19, which rose to 35% in 1997 and, following fortification, jumped to 70% in 1998 and remained there in the most recent data of 2000.

An analysis of food consumption patterns between 1985 and 2000 among German children and adolescents showed that food fortification increased the intakes of several vitamins and minerals by 20–50%. In the case of vitamins A, C, thiamin, riboflavin, niacin, and B_6, fortification raised already adequate intakes from nonfortified foods. In the case of vitamin E and folate, inadequate intakes from nonfortified foods were increased to 100% and 80%, respectively, of reference intake levels (**Figure 2**). Fortified foods accounted for less than 5% of total calcium intake, 10–20% of iron, vitamin A, and folate intakes, 40–50% of vitamin C, E, thiamin, riboflavin, and niacin intakes, and up to 80% of vitamin B_6 intake among this population group.

Calcium fortification of flour is estimated to have supplied up to 30% of the total calcium intake in Danish adults and 13% of total calcium intake in adolescents in Britain. A Swedish multicenter study, 1980–81, on food habits and nutrient intake in children aged 1–15 years revealed that enriched low-fat milk and margarine are important food sources of vitamin A in this population group. Vitamin D-fortified margarine is estimated to provide up to 48% of the total vitamin D intake among Australian men and women.

While government-initiated fortification of foods can greatly change the nutritional profile of the food supply, the contribution of foods voluntarily fortified by industry should also be considered. In the US, such fortified foods are reported to increase intakes of several nutrients, some of which are already adequately provided by nonfortified foods and standardized enriched foods (**Table 3**). Among voluntarily fortified foods, ready-to-eat breakfast cereals are the most significant food category, accounting for the top food source of many vitamins and minerals in several countries, including the US, UK, France, and Spain. Fortified juices and beverages appear to be another substantial contributor, primarily for intakes of vitamins C and D. Fortified breakfast cereals are associated with higher intakes

Table 3 Contribution of voluntary food fortification to nutrient intakes in the US population ≥ 1 year of age[a]

Nutrient	*Median intake (% RDA)*		*Contribution of voluntarily fortified food[b] sources to median intake (% RDA)*
	Excluding voluntarily fortified food[b] sources	*Including voluntarily fortified food[b] sources*	
Vitamin A	70	85	15
Vitamin C	98	122	24
Thiamin	101	113	14
Riboflavin	107	119	12
Niacin	104	117	13
Folate	99	122	23
Calcium	76	76	0
Iron	84	96	12
Zinc	66	70	4

[a]Source of data: Berner LA, Clydesdale FM, and Douglass JS (2001) Fortification contributed greatly to vitamin and mineral intakes in the United States, 1989–91. *Journal of Nutrition* **131**: 2177–2183.
[b]In this analysis, researchers determined the contribution of voluntarily fortified foods, namely the food categories ready-to-eat cereals and fortified cooked cereals, vitamin-and mineral-fortified drinks, meal replacements and supplements, and calcium-fortified milk beverages and juices. Standardized, enriched foods (i.e., foods governed by federal food standards of identity), including enriched flour, enriched bread, enriched rice, enriched pasta, and vitamin-fortified milk, were excluded in this analysis.

of vitamin A among children in Spain and Northern Ireland. In the UK, which has a long tradition of fortification, fortified foods contribute positively to nutrient intakes. For example, in the UK, in 1950, red meat and vegetables were the primary sources of iron and vitamin C, respectively, among 4-year-olds. However, by 1992, most dietary iron came from fortified breakfast cereals and most vitamin C was provided by fortified beverages. Similarly, fortified breakfast cereals were found to be the principal source of folic acid and the second most important food source of vitamins B_6 and D among children in Spain.

Effectiveness of Food Fortification as a Public Health Intervention

The benefits, at a national level, of historical fortification efforts are generally not well documented; however, epidemiological evaluations of pilot-scale programs are considered to have played an important role in the widespread implementation of these programs. Iodization of salt was found to decrease the incidence of goiter by 74–90% in different counties in the State of Michigan in the US in the first 10 years of the program (1924–35). Salt iodization is credited with the elimination of endemic cretinism and endemic goiter in Switzerland. Since its inception in 1922, periodic evaluations of the program triggered increases in fortification of levels of iodine, most recently in 1998. In 2001, an evaluation of a national sample of Swiss school children and pregnant women showed adequate iodine status, underscoring the importance of iodized salt in that nation's food supply and the value for periodic monitoring in the success of a fortification program.

A successful intervention with vitamin A-fortified margarine initiated in Newfoundland in 1944–45 led to a marked improvement in vitamin A status, as indicated by serum retinol levels in a sample of the population. Similarly, observations on the curative effects of milk fat, but not of margarine, eventually led to the enrichment of margarine with vitamin A in Denmark.

An evaluation of the possible health impact of niacin fortification of cereal grains in the US showed that fortification played a significant role in the decline of pellagra-attributed mortality in the 1930s and 1940s and, finally, in the elimination of pellagra in the country. Fortification was particularly significant during a period when food availability and variety were considerably less than are evident today.

Effectiveness of iron fortification is less clear owing primarily to the complex etiology of anemia. Several cereal grain products and other foods, especially breakfast foods, are commonly fortified with iron in developed countries and iron fortification is generally assumed to be responsible, at least in part, for the marked reduction in the prevalence of iron deficiency anemia in these countries. However, many other factors, such as improved socioeconomic conditions, increased meat intake, and iron supplementation may have played important roles. Furthermore, the most common iron source used in cereal fortification in Western countries is reduced iron, which has been found to be poorly bioavailable. Nevertheless, the effectiveness of iron

fortification is apparent in some cases. For example, in Sweden, fortification of flour with iron was withdrawn in January, 1995, because the benefits of such fortification were considered uncertain. However, recent investigations suggest that, after accounting for possible confounding factors, iron intake decreased by 39% and iron deficiency anemia increased by 28% among adolescent girls in the 6 years following withdrawal of iron fortification. The effectiveness of iron fortification is also clear in the case of targeted fortification of infant foods. The use of iron-fortified infant formulas and cereals is credited with the virtual elimination of iron deficiency among American infants.

The effectiveness of food fortification as a public health strategy is evident in the case of recent folate fortification efforts. Since its inception in November, 1998, folic acid fortification (150 μg per 100 g of food) in Canada has produced measurable benefits. In Newfoundland, the average rates of neural tube defects, which remained unchanged between 1991 and 1997, fell by 78% with concurrent increases in blood folate levels of women after the implementation of folic acid fortification. This survey did not find evidence of improved folate status masking hematological manifestations of vitamin B_{12} deficiency, which was a concern carefully considered in setting the fortification levels of folic acid. Studies in other Canadian provinces also report significant reductions in neural tube defects: up to 32% in Quebec, 48% in Ontario, and 54% in Nova Scotia. Folate fortification is also reported to be associated with a 60% reduction in neuroblastoma, an embryonic tumor, among Canadian children.

In the US, enriched cereal grain products have been required to be fortified with 140 μg of folic acid per 100 g of food since January, 1998. Since then, folate levels of baked products, cereal grains, and pasta have doubled or tripled and breakfast cereals are one of the most highly fortified food sources of folate. Consequently, typical folic acid consumption in the country is estimated to have increased by more than 200 $\mu g\,day^{-1}$ due to fortification, along with a substantial improvement in folate status of different population groups. According to the Centers for Disease Control and Prevention, the rates of neural tube defects fell by about 26% from 1995–96 to 1999–2000 in the US although there is debate that this figure may be an underestimate. Careful monitoring and surveillance of the long-term effects of these fortification programs is needed to ensure desired benefits without unintended consequences.

Several factors are considered essential for the success of fortification as a public health intervention. Key among them are:

- a documented need for food fortification, i.e., assessing the gap between current and desired intakes;
- choosing an appropriate food vehicle(s) that is consumed by most of the population in relatively constant amounts;
- setting a fortification level that is not only efficacious in the target population but also safe for the general population;
- resolving any concerns of adverse nutrient interactions;
- establishing clear fortification regulations and policies;
- a public education campaign to increase consumer awareness; and
- periodic assessments of the impact of the intervention to determine any necessary adjustments to the fortification policy to ensure that the desired benefits are achieved and that excessive intakes are minimized.

Relevant government bodies, food industry, professional health organizations, consumer associations, and trade organizations play important roles in formulating a coordinated and concerted effort in the successful implementation of the fortification program.

Emerging Issues

As noted above, both mandatory and voluntary fortifications have played important roles in the overall nutritional adequacy of food supply in developed countries. While classic nutrient deficiency diseases have been alleviated, there is increasing awareness that suboptimal intakes of nutrients may occur in some population subgroups. For example, calcium and vitamin D intakes may be inadequate, particularly among women and the elderly. In the US, median calcium intakes among adolescent girls and adult women are below the recommended Adequate Intake levels. Similarly, the lowest dietary intakes of vitamin D in the US are reported by female teenagers and female adults, with only about 30% of adolescent girls and 20–25% of women cosuming sufficient dietary vitamin D to meet the Adequate Intake levels. Dietary patterns are changing among teenagers with a preference for soft drinks over milk and other dairy beverages. Up to 90% of older adults in the US are reported to consume inadequate amounts of vitamin D. In Australia and New Zealand, women and the elderly are also considered high-risk groups for marginal vitamin D status. In Europe, a substantial percentage of the elderly are reported to have low vitamin D status, ranging from about 10% in the Nordic countries to about 40% in France. Current

scenarios of food fortification do not appear to be reaching the population groups most at risk of these deficiencies.

In contrast, overfortification and random fortification of the food supply is also a growing concern. Where the regulatory system in the country permits, a wide variety of foods are being fortified with a broad spectrum of nutrients. The need for such a vast number of fortified foods for generally nutritionally adequate populations, particularly given the wide consumption of vitamin and mineral supplements, is questioned. Recently, advances in scientific knowledge have enabled estimations of upper (safe) levels of intake and various government agencies have established tolerable upper intake levels for different micronutrients. These upper levels of nutrient intake are being carefully considered as countries continue to monitor their existing fortification programs and policies to appropriately balance risks of deficiency and toxicity in well-nourished populations.

Another emerging issue relates to the so-called 'functional foods' as consumers look for foods associated with health benefits beyond basic nutrition. Nutrition research in the recent past has shifted from evaluating the benefits of the whole food itself to the benefits of bioactive components isolated from such foods. Substances such as lycopene, lutein, and probiotics are being added to foods with claimed health benefits, although the efficacy of such fortification is largely unknown.

Conclusion

Fortification of foods commonly consumed by populations at risk of micronutrient deficiencies has been demonstrated to be an effective public health intervention. Iodization of salt, fortification of milk with vitamin D, and the addition of thiamin, riboflavin, niacin, and iron to cereal grain products have a long history in several developed countries, some starting as early as the 1920s. More recently, several countries have either required or permitted the fortification of cereal grain products with folic acid. In addition, where the regulatory environment permits, the food industry has voluntarily fortified a variety of foods with a wide array of nutrients. The importance of food fortification in the diets of developed countries is clear. Surveys indicate that fortified foods make significant contributions to the intakes of nutrients among different population groups, resulting in improved nutrient status and/or related health conditions. The success of food fortification in developed countries can be attributed to several factors, notably cooperation among different sectors to raise consumer awareness and demand for more healthful foods. However, as the popularity of fortified foods grows and trends toward random fortification and overfortification of the food supply continue, careful monitoring of existing fortification programs and policies become more critical to ensure not only the adequacy but, perhaps more importantly, the safety of food fortification in generally well-nourished populations.

See also: **Ascorbic Acid**: Deficiency States. **Calcium**. **Cobalamins**. **Folic Acid**. **Food Fortification**: Developing Countries. **Functional Foods**: Regulatory Aspects. **Iodine**: Deficiency Disorders. **Niacin**. **Riboflavin**. **Thiamin**: Beriberi. **Vitamin A**: Deficiency and Interventions. **Vitamin D**: Rickets and Osteomalacia. **Vitamin B_6**.

Further Reading

Allen L, de Benoist B, Dary O, and Hurrell R (2005) *Guidelines on Food Fortification with Micronutrients*. World Health Organization, Geneva.

Backstrand JR (2002) The history and future of food fortification in the United States: a public health perspective. *Nutrition Reviews* **60**(1): 15–26.

Bauerfeind JC and Lachance PA (eds.) (1991) *Nutrient Additions to Foods*. Trumbull, CT: Food and Nutrition Press.

Food and Agriculture Organization (1996) *Food Fortification: Technology and Quality Control*. Report of an FAO technical meeting held in Rome, 20–23 November, 1995. FAO Food and Nutrition Paper 60. Food and Agriculture Organization of the United Nations, Rome.

Gerrior S, Bente L, and Hiza H (2004) *Nutrient Content of the United States Food Supply, 1909–2000*. Home Economics Research Report No. 56. United States Department of Agriculture, Center for Nutrition Policy and Promotion Alexandria, UA, USA.

Liu S, West R, Randerll E *et al.* (2004) A comprehensive evaluation of food fortification with folic acid for the primary prevention of neural tube defects. *BMC Pregnancy and Childbirth* **4**: 20.

Meltzer HM, Aro A, Andersen NL, Bente K, and Alexander J (2003) Risk analysis applied to food fortification. *Public Health Nutrition* **6**(3): 281–290.

Prynne CJ, Paul AA, Price GM *et al.* (1999) Food and nutrient intake of a national sample of 4-year-old children in 1950: comparison with the 1990s. *Public Health Nutrition* **2**(4): 537–547.

Sichert-Hellert W, Kersting M, Alexy U, and Manz F (2000) Ten-year trends in vitamin and mineral intake from fortified food in German children and adolescents. *European Journal of Clinical Nutrition* **54**(1): 81–86.

Yetley EA and Rader JI (2004) Modeling the level of fortification and post-fortification assessments: U.S. experience. *Nutrition Reviews* **62**(6): S50–S59.

Developing Countries

O Dary and J O Mora, The MOST Project, Arlington, VA, USA

Introduction

Micronutrient deficiencies are a consequence of diets with low variety, which are common in developing countries where the diet is based on starchy foods. Fortification of commonly consumed edible products is a potential solution to increase the intake of micronutrients. However, feasibility and effectiveness of fortification depends on production in the industrial setting, reliable food control mechanisms, and frequent and sufficient consumption of the fortified products. These criteria are not common in developing countries, and this limits the potential of this intervention to improve the health of their inhabitants. Nevertheless, some examples already exist that show that food fortification is feasible and useful in developing countries. Increasing, urbanization in these countries makes fortification an essential food technology for their immediate future.

The Need to Improve Micronutrient Intakes in Developing Countries

Micronutrient deficiency is a consequence of a lack of variety in the diet. This condition is aggravated in developing countries, where the diets consist primarily of low-cost starchy foods. **Table 1** shows the micronutrient density of wheat, wheat flour, maize-meal, corn-masa flour, and rice. In all cases, the cereal-based derivatives are good sources of energy, but are very poor in most micronutrients, in particular zinc, iron, calcium, and vitamins associated with foods of animal origin, such as vitamins A, B_2, and B_{12}. These products are also poor in folate and vitamin C. Extraction of flour from cereal grains increases the energy density by two times in the case of corn, and by three times in the case of wheat (**Table 1**), but reduces the original micronutrient content of the grain by 30–85%. **Table 1** illustrates that white wheat flour (extraction rate lower than 80%) is not a good source of any micronutrient, whereas rice and, especially, corn derivatives maintain satisfactory quantities of vitamin B_6 and niacin. Corn sub products also retain some zinc, and adequate amounts of vitamin B_1. Iron content is also higher, but the presence of strong iron-absorption inhibitors, especially in corn-masa flour, makes corn derivatives poor sources of this nutrient. Processing of the corn grains with lime (calcium oxide) to produce masa-flour increases the content of zinc, calcium, and niacin. Nevertheless, considering that 50% or more of energy is satisfied through the consumption of these foods in poor populations of developing countries, it is easy to explain why they are at very high risk of suffering the consequences of vitamin and mineral deficiencies.

Table 2 illustrates that, in general, diets in developing countries are not only poor in energy (less than 77% adequacy) but also in zinc, iron, calcium, vitamin B_{12}, folate, vitamin A, and vitamin B_2. Adequacies of these micronutrients are from 35% to 70% of the estimated average requirement (EAR). Vitamin B_1, niacin, and vitamin C have better adequacies, although these are still unsatisfactory, being in the order of 70–100% of the EAR. Iodine deficiency, as in most human societies in the world, is

Table 1 Micronutrient density of main energy sources in developing countries (per 100 g of product)

Nutrient	*EAR*[a]	*Wheat*	*White wheat flour*	*Maize-meal*	*Corn-masa flour*	*Rice*
Energy (kcal)		112	364	361 (% EAR)	365	360
Zinc	11.7 mg	26	6	15	21	10
Iron	21.6 mg[b]	17	8	16	8	4
Calcium	833 mg	4	2	2	17	1
B_{12}	2.0 μg	0	0	0	0	0
Folate	320 μg	14	6	8	0	3
Vitamin A	429 μg	0	0	0	0	0
B_1	1.0 mg	40	6	25	31	7
B_2	1.1 mg	16	5	11	5	5
B_6	1.1 mg	9	4	34	31	14
Niacin	12 mg	50	8	14	20	13
Vitamin C	37 mg	0	0	0	0	0

[a]EAR, estimated average requirement is the average (median) daily nutrient intake level estimated to meet the needs of half the healthy individuals in a particular age and gender group.
[b]EAR for white wheat flour (low extraction flour) = 10.8 mg; and for corn-masa flour = 43.2 mg.

Table 2 Nutrient adequacy in some developing countries (approximate % EAR in diet)[a]

Nutrient	*EAR*	*S. Africa (1999)*	*Bangladesh (1995–96)*	*Nicaragua (1993)*	*Mexico (2000)*	*Philippines (1998)*[b]
Energy (kcal)		77	75	75	67	70
Zinc	11.7 mg	56	–	77	51	–
Iron	21.6 mg	65	53	48	38	61
Calcium	833 mg	61	41	82	77	58
B_{12}	2.0 μg	280	–	90	–	–
Folate	320 μg	89	–	35	66	–
Vitamin A[c]	429 μg	106	66	50	60	49
B_1	1.0 mg	140	118	70	–	76
B_2	1.1 mg	157	45	82	–	73
B_6	1.1 mg	118	–	128	–	–
Niacin	12 mg	127	155	71	–	89
Vitamin C	37 mg	193	87	140	83	119

[a]Almost all individuals in the population should have an intake equal or greater than the corresponding EAR. In this table, adequacy rates smaller than 100% means that half or more of the population is not fulfilling the nutritional requirement.
[b]Data for children 4–8 years old.
[c]Conversion factor for vitamin A from plant sources was changed from 1:6 to 1:12 (retinol:β-carotene).

also widespread due to the absence of foods containing this nutrient. It seems that only vitamin B_6 may be the exception for micronutrient deficiencies in developing countries. Improvement in micronutrient adequacy cannot be expected with a simple increase of the consumption of food (in order to satisfy the daily energy requirements) without changes in the quality of the diet, because inadequacies originate from the low-micronutrient content of the usual foods consumed.

In South Africa, nutrient adequacy for vitamin B_{12}, vitamin A, and vitamin B_2 was better than in the other countries included in **Table 2**; this may be a consequence of the greater consumption of eggs and milk in this country. Similarly, Nicaragua and Mexico presented lower deficiency for calcium, which is obviously associated with the consumption of corn-masa flour.

It is generally recognized that developing countries require not only an increase in intake of iodine, vitamin A, and iron, but also most vitamins and minerals.

Conditions Limiting the Implementation of Effective Food Fortification Programs in Developing Countries

Many of the micronutrient deficiencies in developing countries might be solved by raising the consumption of foods of animal origin (eggs, milk, fish, poultry, and meat). However, this strategy is difficult to implement in the short term because of its high cost, as well as cultural and religious beliefs. Thus, food fortification seems an attractive alternative. However, fortification is not a simple task in developing countries, as explained below.

Small Proportion of Centrally Produced Foods

Food fortification is a technology, and hence it necessitates the existence of processing centers (industrial plants or fortification facilities), which must have a minimum technological development that is compatible with the advantages of large market economies and quality control procedures. If this requirement is not satisfied, it would be wasteful, worthless, and risky to try to implement a food fortification program, because of prohibitive cost and unsafe practices. The salt iodization program is occasionally mentioned as an example against this statement, but this cannot be used as a model of other food fortification programs, as explained later. Factories for processed foods (breakfast cereals, pastas, vegetable fats, beverages, and the like) exist in most developing countries but, unlike the situation in developed countries, their products are generally out of the reach of the populations at risk. Food staples such as corn, rice and, in some situations wheat, are produced and processed locally in very small facilities, and hence they are very difficult or impossible to fortify. Fortification of rice still faces cost and technological constraints, especially in regions where rice is the main staple.

The difficulty of finding suitable vehicles to be fortified in the developing world has led to investigation of the fortification of oil, sugar, salt, fish and soy sauces, curry powder, and, in the past, other condiments such as monosodium glutamate. The main limitation of these food vehicles is that they allow the addition of only one or very few micronutrients.

Low Consumption of Potential Food Vehicles

The effectiveness of food fortification in developing countries is hampered by the fact that those edible products that are amenable to fortification, are not consumed frequently enough or in sufficient amounts by the most at risk individuals. In the past, it was assumed that increasing the level of micronutrients in the food vehicle may overcome this constraint. It is now realized that the content of micronutrients in widely consumed products depends on keeping their consumption safe for everyone (especially for individuals with very large consumption), on maintaining the technological compatibility with the properties and uses of the food vehicle, and on maintaining the price of the product within an acceptable range. As a consequence, the fortification formulations are more or less fixed by maximum possible values, which in turn determine that a minimum daily consumption of the fortified vehicles is necessary in order to provide useful amounts of micronutrients. This condition reduces the potential benefits of food fortification for very poor families, as well as for small children, because they usually consume low amounts of industrially manufactured products.

Table 3 illustrates the minimum consumption amounts of industrially produced foods to supply from 20% to 70% of the EAR of iodine, iron, and vitamin A through fortification. This range was selected under the assumption that a source of a nutrient is one that provides at least 20% of the EAR of that nutrient, and if a food supplies 70% EAR it means that that food might be sufficient to supply all of the amount needed to complement the diet with that nutrient. Daily consumption amounts should be in the following ranges: oil, 5–20 ml for vitamin A; sugar, 10–35 g for vitamin A; low-extraction wheat flour, 40–140 g for iron, and 50–150 g for vitamin A; salt, 4–15 g for iron, and slightly less for vitamin A and iodine; and fish/soy sauce, 6–20 ml for iron. Ideally, consumption patterns should be equal to or higher than the uppermost figure of the range, but never below the lowest value to have public health significance. If consumption is within the range, then probably more than one product should be fortified and consumed.

Weak Enforcement and Monitoring Capabilities

Many developing countries have regulations and laws making food fortification mandatory. However, those legal instruments often remain without compliance because of the impossibility of enforcing them. There are also many examples of false, misleading, and exaggerated claims regarding the presence of micronutrients in the foods. These claims are used as simple advertizing tools but without checking that the products really contain the vitamins and minerals in the amounts that are declared on the labels. The lack of reliable food control mechanisms is a barrier to the effectiveness of fortification in developing countries.

Low Bioavailability of Iron and Other Minerals

Iron is a relatively abundant mineral in plant sources, but its bioavailability is very low. Iron in cereal-based diets – without fermentation or elimination of the bran – might be $\leq 5\%$ absorbed.

Table 3 Consumption range of fortified foods to supply 20–70% EAR[a] (range of consumption/(minimum fortification level in ppm[b]))

Nutrient	*Oil*	*Sugar*	*Wheat flour*	*Salt*	*Fish/soy sauce*
Iodine	–	–	–	1.5–5.0 g[c] / (25 ppm)	–
Iron[d]	–	–	45–165 / (45 ppm)	4.0–15.0 g / (500 ppm)	6–20 ml / (350 ppm)
Vitamin A	5–20 ml / (20 ppm)	10–35 g / (10 ppm)	50–175 g / (2.0 ppm)	3.0–10.0 g / (35 ppm)	–

[a]Approximately equivalent to 15–50% Recommended Nutrient Intake (RNI). It is assumed that a food should supply at least 20% EAR of a nutrient to be considered as a source of that nutrient; and that a supply of 70% would mean that that food would be sufficient to complement the whole recommendations of that nutrient.

[b]Higher levels cannot be used because risk of excessive intakes for individuals who consume large amounts of these foods; high cost; or technological incompatibility. Supply of EAR was estimated considering losses during storage, transportation and distribution, and after food preparation.

[c]In the case of iodine, because iodized salt is almost its only source, the nutritional goal should be to supply 100%EAR, which means that the usual salt consumption should be 7 grams/day.

[d]In all cases, assuming a diet with iron bioavailability of 10%.

Many developing countries assume that their diets have 10% bioavailability for iron, but in fact this is not the case. This fact has caused iron deficiency to be underestimated, and the estimation of the potential impact of iron fortification overestimated. Perhaps a similar situation may be happening with zinc and calcium.

In a diet with abundant iron-absorption inhibitors (rice, high extraction wheat and corn flours, and cornmasa flour), even the most absorbable iron compounds, such as NaFeEDTA and ferrous bisglycinate, have absorption rates not higher than 10%. Water-soluble or acid-soluble iron salts, such as ferrous sulfate and ferrous fumarate, respectively, are absorbed at rates of from 5% to 10%, depending on the content of iron absorption inhibitors in the diet. Electrolytic iron and other difficult to dissolve iron compounds are absorbed in even lower proportions. This condition combined with the fact that the most absorbable iron compounds change the sensory characteristics of foods (color, flavor, and odor) makes the improvement of iron status in cereal-based diets through food fortification a very difficult challenge.

Use of electrolytic iron deserves a special mention because, once in the fortified food, it cannot be distinguished from other types of elemental iron, such as reduced iron and atomized iron, whose bioavailability rates are even lower. The impossibility of enforcement makes the use of electrolytic iron an unattractive option under the conditions in developing countries.

Price of Fortified versus Nonfortified Product

When fortification is being considered as a public health strategy, feasibility analyses regarding cost implications and trade practices are usually neglected in favor of biological efficacy trials. This approach has produced reports and publications showing biological impact but it has not supported the creation of true and effective programs. This was the case when monosodium glutamate was fortified with vitamin A in the Philippines and Indonesia in the 1980s. The nutrient was bioavailable and metabolically efficacious, but the fortified product had a very short life span in the market. This was principally due to the 17% reduction in weight of the sachets in order to keep the price similar to the unfortified product. The consumers noticed the difference and preferred to choose the unfortified alternative. The fortified product also had problems with color and micronutrient segregation, but the uncompetitive price was the main reason for the failure of the program.

Table 4 shows the annual cost per person of providing 70% of the EAR through food fortification. In the case of iodine, the annual cost is only US$0.003 per person. Iodine is the only micronutrient to combine several properties that permit us to recommend a 'universal' measure to provide iodine in most countries of the world. These characteristics are: its mineral nature (resistant to environmental factors, mainly in the form of potassium iodate even when combined with raw salt); very low cost; very small interactions with the food matrix; very low nutritional requirements (microgram range); and absence of foods that are good sources of this nutrient. These properties make it possible to use only one carrier, i.e., salt, without the need to consider other ways to complement the amounts of iodine supplied by this means. No other nutrient has all these attributes. Therefore, iodization of salt is the exception and not the model for food fortification.

Table 4 also shows that independently of the food vehicles that can be used to increase the dietary iron, the annual investment would depend on the price of the type of iron that must be used to be compatible with the food matrix. Thus, to supply 70% EAR throughout the year using ferrous fumarate would cost US$0.045 per person' using encapsulated ferrous sulfate US$0.257, and using NaFeEDTA US$0.405. These iron compounds are compatible with white wheat flour, salt and fish/soy sauce, respectively.

Table 4 Annual cost per person of providing 70% EAR through food fortification[a] (US$)

Nutrient	*Oil*	*Sugar*	*Wheat flour*	*Salt*	*Fish/soy sauce*
Iodine	–	–	–	0.003	–
Iron	–	–	0.045[b]	0.257[c]	0.405[d]
Vitamin A[e]	0.023	0.150	0.130	0.175	–
Total	0.023	0.150	0.175	0.435	0.405

[a]At the levels of fortification and consumption indicated in **Table 3**.
[b]Ferrous fumarate.
[c]Encapsulated ferrous sulfate.
[d]NaFeEDTA.
[e]All cases, except oil, use microencapsulated vitamin A. Oil uses fat-soluble vitamin A.

The nature of the two latter vehicles does not allow use of other iron alternatives at this time.

Similarly, the cost of fortification with vitamin A would change if the nutrient is added to an oily matrix (US$0.023/year per person), or to a dry or water-soluble matrix (US$0.130 to US$0.150/year per person) such as wheat flour, sugar, or salt.

In principle, the less costly alternatives should be used preferentially, but if these food matrixes are not consumed regularly nor in sufficient amounts, then the more expensive options are necessary. **Table 4** illustrates that the total annual costs, even in the most costly cases of food fortification, are inexpensive enough compared with the large public health benefits that are expected from the prevention of the consequences of these micronutrient deficiencies.

However, the viability of the fortification program depends not only on the total annual cost, but also on the relative price increase of the food vehicle, because the difference in price compared to the unfortified choice will determine the feasibility of production, trade, and enforcement. Thus, using the case of salt as an example (**Table 5**), adding iodine increases the price by 0.6%, iron by 16.7%, and vitamin A by 11.7%. Then, it is logical to conclude that considering salt as a vehicle of iron and vitamin A, although it might be technologically compatible and biologically efficacious, would face a lot of opposition from the producers with high risk of noncompliance in the absence of strong enforcement capabilities. In theory, this type of fortification is possible and has potential biological impact, but in practice the substantial challenges might impede their implementation.

Table 5 also includes the production cost and the price increase for other fortification examples. It is easy to see that double or triple fortification of salt, or iron fortification of fish/soy sauce, would need very strict and reliable enforcement systems and permanent financial mechanisms, in order to be sustainable and efficient interventions.

Table 5 and **Table 6** show that flours are suitable for the addition of several micronutrients simultaneously, and that the price increase is relatively low. Therefore, flours should be used to their maximum potential wherever and whenever it is possible.

Examples of Food Fortification Programs in Developing Countries

Despite the multiple limitations that affect food fortification in the developing world, there are several examples that confirm its feasibility and benefits.

Vegetable Fats and Oil Fortified with Vitamin A and D

Addition of vitamin A to margarine and other vegetable fats started in 1918 in Denmark, when cases of xerophthalmia were associated with the replacement of butter by margarine. Then, the practice of nutritional equivalence, that is emulating the nutritional composition of butter, was adopted by the industry. Thus, vitamin A and D started to be added to margarine and other vegetable fats. Despite the fact that the bioavailability and utilization of vitamin A has been confirmed experimentally in the Philippines and other countries, the compliance and effectiveness of this practice at the national level has not been documented. Similarly, the addition of vitamin A in oil has been proven to be efficacious in

Table 5 Production cost of food fortification[a]

Food (price)	*Oil (US$0.50/kg)*	*Sugar (US$0.40/kg)*	*Wheat flour (US$0.40/kg)*	*Salt (US$0.30/kg)*	*Fish/soy sauce (US$0.50/kg)*
Nutrient	(US$ per Metric Ton/(% price)				
Iodine	–	–	–	1.75 (0.6%)	–
Iron	–	–	0.90[b] (0.2%)	50.00[c] (16.7%)	55.00[d] (11.0%)
Vitamin A[e]	3.50 (0.7%)	12.00 (2.7%)	2.02 (0.5%)	35.00 (11.7%)	–

[a]At the levels of fortification indicated in **Table 3**.
[b]Ferrous fumarate.
[c]Encapsulated ferrous sulfate.
[d]NaFeEDTA.
[e]All cases, except oil, microencapsulated vitamin A.

Table 6 Characteristics of a suggested fortification formulation for white wheat flour[a]

Nutrient	*Level (mg kg^{-1})*	*Cost (US$ per Metric Ton)*	*% Price*[b]	*% EAR in 100 g consumption*[c]
Zinc	30.0	0.16	0.04	52
Iron	45.0	0.90	0.22	48
B_{12}	0.01	0.59	0.15	45
Folic acid	2.0	0.32	0.08	97[d]
Vitamin A	2.0	2.02	0.51	39
B_1	6.0	0.30	0.08	45
B_2	4.0	0.24	0.06	39
B_6	5.0	0.25	0.06	47
Niacin	45.0	0.54	0.14	38
Total	–	5.32	1.34%	–

[a]Formulation adequate for daily consumption of 50–200 g day^{-1}. If consumption is greater, micronutrient levels should be reduced. Countries might reduce the levels of some micronutrients based on their particular nutritional profile and presence of other fortification programs with large public health coverage.
[b]Assuming US$0.40/kg.
[c]Considering losses during food preparation.
[d]Dietary folate = 1.7 × folic acid.

Brazil, but national effectiveness studies are still pending. Oil is currently fortified with vitamin A in Coted'Ivoire, Mali, Morocco, Nigeria, Oman, Philippines, Uganda, and Yemen. These programs have been designed to provide at least 100 μg vitamin A (23% EAR), which may have some biological consequence in a portion of the population. The main restriction of this practice is the destruction of vitamin A when oil is exposed to light inside transparent bottles.

Sugar Fortified with Vitamin A

In the 1970s, the Central American countries suffered from vitamin A deficiency. The only food that was identified as a good vehicle for fortification (centrally produced, affordable fortification cost, and widely consumed) was sugar. The technology of the addition of vitamin A was developed and, together with the introduction of the intervention, its biological effectiveness was evaluated. Vitamin A level increased in serum and breast milk of individuals in poor communities. Sugar is now the most important source of this nutrient in El Salvador, Guatemala, Honduras, and Nicaragua. It supplies 200–1000 μg of vitamin A (50–200% EAR), depending on the daily consumption pattern (30–150 g day^{-1}). Nowadays, vitamin A deficiency is practically nonexistent in those countries. Zambia started this program in 1998 but, differentiate from, Central America, the impact has been modest due to the lower sugar consumption (20 g day^{-1}) and sugar use (50% of population compared with nearly 100% in Central America). Nevertheless, sugar is the main source of vitamin A for those who consume it in Zambia. Nigeria has already started to implement this program for all sugar refined in the country. The vitamin A added to sugar is microencapsulated, and hence the type of packaging has little influence on its stability.

Wheat Fortified with Multiple Micronutrients

The practice of adding micronutrients to restore the nutritional quality of wheat grain with iron, vitamins B_1 and B_2, niacin, and sometimes calcium has been followed by many wheat flour mills in developing countries since the 1950s, especially in those industries with links to companies in the US and Canada. Nevertheless, attention to fortification of wheat flour was raised in the 1990s as a measure advocated by international organizations to reduce iron deficiency. Generally, the iron source is elemental iron, either electrolytic or reduced iron; exceptions are Central America where ferrous fumarate is used, and Peru, Cuba, and Chile, which use ferrous sulfate. A study in Sri Lanka, where an additional 12 mg day^{-1} of reduced (28% EAR) or electrolytic (56% EAR) iron was provided, did not find any improvement in hemoglobin levels after 2 years of treatment. Similarly, a study in Bangladesh, which supplied 3.3 mg day^{-1} of reduced iron (35% EAR) to 6–15-year-old children failed to modify any parameter associated with iron status. Only Chile has indirect evidence that iron from fortified wheat flour contributes to maintain good nutritional status of this nutrient. Wheat flour consumption in that country (250 g day^{-1}) provides 7.5 mg day^{-1} iron from ferrous sulfate (69% EAR). Other studies are on-going with the purpose of assessing if it is legitimate to use electrolytic iron for the fortification of wheat flour.

Most recently, attention has been given to folic acid, after evidence that neural tube defects in the US and Canada were reduced by the intake of an

additional 200 μg day^{-1} of folic acid (106% EAR), by means of the consumption of cereals fortified with this nutrient. These results have been confirmed in Chile, where bioavailability and biological utilization of folic acid was also clearly demonstrated.

Experimental efficacy trials of wheat flour fortified with vitamin A in the Philippines and Bangladesh have established that biological impact can be found with intake levels of 100–200 μg day^{-1} (25–50% EAR). Despite this finding, vitamin A has not been widely considered as a micronutrient to be added to wheat flour, although South Africa, Nigeria, and the Philippines have regulations including this micronutrient as part of the formulation. In comparison with sugar, addition of vitamin A to wheat flour has a lower cost and slightly better stability.

In addition to iron and folic acid, many countries also include vitamins B_1, B_2, and niacin in the fortification formulations. This is the practice followed by most Latin American countries. Now, Latin America is also considering the incorporation of vitamin B_{12} and zinc. Biological effects of the presence of other nutrients apart from iron and vitamin A have not been systematically evaluated in developing countries. However, when wheat flour fortification started in the US, it was documented that cases of beriberi, ariboflavinosis, and pellagra, associated with vitamins B_1, B_2, and niacin deficiencies, respectively, decreased drastically.

Wheat flour fortification has now extended to Bahrain, Jordan, Morocco, Saudi Arabia, Pakistan, Iran, Indonesia, and some regions of China, Central Asian countries and a few African countries.

Corn Products Fortified with Multiple Micronutrients

In Venezuela, precooked and degermed corn flour is fortified with iron (reduced iron and ferrous fumarate, in a mixture 2:3), vitamins B_1, B_2, niacin, and vitamin A. The usual consumption of this food provides 2.4 mg of iron from ferrous fumarate (22% EAR), and 1.6 mg from reduced iron (3.7% EAR). It also supplies 200 μg of vitamin A (47% EAR). A formal evaluation of this program has not been carried out, although it has been associated with maintenance of the iron status of the population despite the economical deterioration of the country. This argument is controversial because the additional supply of iron is not high, and the prevalence of anemia was low only in the year when fortification was introduced, whereas anemia rates after fortification were similar to those existent before fortification. Without a study under controlled conditions it is difficult to assign the mentioned effect to this program. Fortified precooked corn flour is theoretically more important as a source of vitamin A than iron, but no experimental evidence has been obtained in this regard. Nevertheless, it is interesting to note that a national nutritional survey, carried out in 2002, did not find vitamin A deficiency in Venezuelan preschoolers.

In Africa, maize-meal is currently being used as a vehicle for iron, zinc, vitamin A, and B complex vitamins, including in some cases vitamin B_{12}. Biological impact has not been evaluated.

Masa-corn flour is fortified with iron (reduced, or ferrous fumarate, or ferrous bisglycinate), and vitamins B_1, B_2, and niacin, in some Central American countries, and in Mexico also with zinc. Efficacy trials have shown the better bioavailability of NaFeEDTA over reduced iron in this matrix, although NaFeEDTA has not been used as yet in an industrial setting.

Salt Fortified with Iodine and Fluoride

Most developing countries have joined the initiative for universal salt iodization; evidence obtained worldwide confirms that this provides sufficient iodine to human populations. However, some countries have started to reduce the content of iodine, because of concerns of unnecessary or excessive supply. A few countries, such as Colombia, Costa Rica, Jamaica, Mexico and Uruguay, have also utilized salt as a vehicle for delivering fluoride with the purpose of reducing tooth decay. In Costa Rica, tooth decay was reduced 65% after 12 years of initiating the program. As in the case of iodine, fluoride content has recently been lowered after epidemiological monitoring determined that excessive fluoride was being supplied.

Soy/Fish Sauces and Curry Powder Fortified with Iron

Efficacy trials carried out with soy sauce in China and fish sauce in Vietnam that were fortified with NaFeEDTA have shown an impact on reduction of anemia and improvement in the iron status indicators. The amount of iron supplied was 10 mg day^{-1}, which is about 93% of the EAR assuming a diet with 10% iron bioavailability. The plan is now to make these experiences national programs. A similar study with curry powder in South Africa provided similar conclusions. In these cases, effectiveness and technical feasibility have been proven, and the existence of the program depends now on assuring industrial acceptance, permanent financing, and continuous enforcement.

Implications and Conclusion

The examples described in the previous section demonstrate that food fortification is possible and that it can be effective in developing countries. However, it is important to point out that the biological impact depends on the chemical properties and amounts of the micronutrients that are supplied and not on the fortified food itself. It is not a surprise that biological impact has been found in efficacy trials using many edible products as the fortification vehicles, such as biscuits, milk, sugar-based beverages, and "sprinkles." If these products contain the proper form and amount of micronutrients, it is highly probable that they will produce the expected biological outcomes.

Although food fortification has limitations in developing countries, mainly due to low accessibility and affordability of industrially produced foods by poor sectors of the population, the increasing trend of urbanization makes this strategy very important. Quality of the diets will be improved, and hence the health and well being of millions of persons. The aim is to supply sufficient micronutrients to populations having poor-quality diets, and fortification is one of the most efficient mechanisms to achieve this goal.

See also: **Ascorbic Acid**: Physiology, Dietary Sources and Requirements; Deficiency States. **Calcium**. **Folic Acid**. **Food Fortification**: Developed Countries. **Iodine**: Physiology, Dietary Sources and Requirements; Deficiency Disorders. **Iron**. **Niacin**. **Nutrition Policies In Developing and Developed Countries**. **Riboflavin**. **Sodium**: Salt Intake and Health. **Supplementation**: Developing Countries. **Vitamin A**: Biochemistry and Physiological Role; Deficiency and Interventions. **Vitamin B_6**.

Further Reading

Allen L, de Benoist B, Dary O, and Hurrell R (2004) *Guidelines on Food Fortification with Micronutrients*. Geneva: World Health Organization.

Bauerfeind JC and Lachance PA (eds.) (1991) *Nutrient Additions to Foods*. Trumbull, CT: Food and Nutrition Press.

Hetzel BS, Dunn JT, and Stanbury JB (eds.) (1987) *The Prevention and Control of Iodine Deficiency Disorders*. Amsterdam: Elsevier Press.

Hurrell R (2002) How to ensure adequate iron absorption from iron-fortified food. *Nutrition Reviews* **60**(II): S7–S15.

Hurrell R, Bothwell T, Cook JD, Dary O, Davidsson L, Fairweather-Tait S, Hallberg L, Lynch S, Rosado J, Walter T, and Whittaker P (2002) The usefulness of elemental iron for cereal flour fortification: A SUSTAIN Task Force Report. *Nutrition Review* **60**: 391–406.

Micronutrient Initiative (ed.) (1998) *Food Fortification to End Micronutrient Malnutrition*. Arlington, UA: State of the Art.

Mora JO, Dary O, Chinchilla D, and Arroyave G (2000) *Vitamin A Sugar Fortification in Central America. Experience and Lessons Learned* Arlington, VA: MOST/USAID Micronutrient Program.

Nestel P (1993) *Food Fortification in Developing Countries*. Arlington, UA: VITAL, USAID, Micronutrient Initiative.

Pan American Health Organization (2002) Iron compounds for food fortification: Guidelines for Latin America and the Caribbean 2002. *Nutrition Reviews* **60**(II): S50–S61.

Roche (2003) Vitamins. Nutriview (special issue 12). http://www.nutrivit.org/vic/staple/N2003_spec.pdf.

Food Intake *see* **Dietary Intake Measurement**: Methodology; Validation. **Dietary Surveys**. **Meal Size and Frequency**

FOOD INTOLERANCE

T J David, University of Manchester, Manchester, UK

Definition of Food Intolerance

Food intolerance can be defined as a reproducible adverse reaction to a specific food or food ingredient, and it is not psychologically based. Although this appears straightforward, there are a number of difficulties with this definition, and these are discussed below.

Lack of Definition of 'Adverse Reaction'

One problem with our definition of food intolerance is the lack of definition of what constitutes an adverse reaction. All eating causes reactions, which include satiety, feeling warm, the urge to defecate (due to the gastrocolic reflex), and weight gain.

Variation in Tolerance

People vary in their tolerance of events. For some, flatus is an unacceptable and embarrassing problem, whereas for others it is the normal effect of eating baked beans.

Any Food taken in Excess may be Harmful

The definition above does not take into account dosage. Large quantities of certain foods may result in disease in certain individuals, although such disorders are not usually included in the category of food intolerance. Any food, however harmless, can be harmful if taken in excess. Notable examples of this are:

1. Apples, pears, and honey are rich sources of fructose, a sugar which in early childhood is not well absorbed if taken in large quantities. Thus, if a child takes a quantity of fructose in excess of that which can be absorbed in the gastrointestinal tract, the result will be loose stools (diarrhea) due to the osmotic effect of unabsorbed fructose. It should be noted that whereas this applies to normal children, there is in addition a small number of children who are especially poor at handling ingested fructose, and in these children even small quantities of fructose-containing foods will cause florid diarrhea.
2. Chicken liver is a rich source of vitamin A. There are reported cases of infants who were fed large quantities of chicken liver, and who developed raised intracranial pressure as a consequence of vitamin A toxicity.
3. In those who are genetically predisposed, ingestion of an excess of purine-rich foods contributes to hyperuricemia, leading to gout, a disorder which is not usually regarded as a form of food intolerance.

Principal Mechanisms and Pathophysiology of Food Intolerance

The principal mechanisms resulting in food intolerance and the pathophysiology (where this is understood) are discussed below.

Food Allergy

The term 'allergy' implies a definite immunological mechanism. This could be antibody mediated, cell mediated, or due to circulating immune complexes. The clinical features of an allergic reaction include urticaria (nettle rash), angioedema, rhinitis (sneezing, nasal discharge, blocked nose), worsening of pre-existing atopic eczema, asthma (wheezing, coughing, tightness of the chest, shortness of breath), vomiting, abdominal pain, diarrhea, and anaphylactic shock.

Enzyme Defects

Inborn errors of metabolism may affect the digestion and absorption of carbohydrate, fat, or protein. In some subjects the enzyme defect is primarily gastrointestinal, causing defects in digestion or absorption. An example is lactase deficiency (see below). In other subjects, the enzyme defect is systemic. An example is the rare disorders of hereditary fructose intolerance, described below.

Lactase deficiency An example of an enzyme defect causing food intolerance is lactase deficiency. In this condition, which is primarily a disorder that affects infants and young children, there is a reduced or absent concentration of the enzyme lactase in the small intestinal mucosa. Affected individuals are unable to break down ingested lactose, the main sugar found in milk, and which if unabsorbed passes into the large intestine, where there are two consequences. One is an osmotic diarrhea. The other is that some of the unabsorbed lactose is broken down by intestinal bacteria, accompanied by the production of gas (hydrogen) leading to abdominal distension and flatus and the production of organic acids that cause perianal soreness or excoriation. The production of hydrogen, its absorption into the bloodstream, and its excretion in the breath lead to a very simple and elegant test for sugar intolerance: the breath hydrogen test. In this test, the subject swallows a portion of the sugar which one suspects the subject cannot absorb. Breath is collected every half an hour and the hydrogen content is measured. In the normal individual, the sugar is absorbed and hydrogen is not produced. In the intolerant individual, the sugar is not absorbed, hydrogen is produced, and a steep rise in hydrogen concentration is found in the exhaled air.

The management of lactose intolerance is to avoid foods that contain lactose (mainly cows' milk and its products). For infants it is worth noting that the soya-based infant formulas are lactose free. In theory, an alternative is to add microbial β-galactosidase to cows' milk, which can produce a lactose-free milk, with the inconvenience that it has a sweeter flavor and requires a 24-h incubation period at 4 °C.

In infants and young children, lactase deficiency is usually a transient problem occurring after an episode of gastroenteritis, but it is commonly a feature of any disease that causes damage to the intestinal mucosa (e.g., celiac disease). Levels of lactase tend to fall during mid to later childhood, and in a

number of populations (e.g., African, Mexican, Greenland Eskimo) a high proportion of adults have very little lactase activity. This adult deficiency is believed to have a genetic basis. Man is the only animal apart from the domestic cat that drinks milk after weaning, and deficiency of lactase in adults could in certain populations be considered the normal state.

Hereditary fructose intolerance In this condition, which has autosomal recessive inheritance, there is deficiency of the liver enzyme fructose 1,6-bisphosphate aldolase. As a result, fructose-1-phosphate accumulates in liver cells, and acts as a competitive inhibitor for phosphorylase. The resulting transient inhibition of the conversion of glycogen to glucose leads to severe hypoglycemia (low blood glucose concentration). Affected infants are symptom free as long as their diet is limited to human milk. If they receive milk formulas or any food that contains fructose they develop attacks of hypoglycemia, shock, coma and convulsions. There may be jaundice, an enlarged liver, and sometimes progressive liver disease. The treatment requires the complete elimination of fructose from the diet, which may be difficult as fructose is a widely used food additive and sweetener. A trivial but interesting feature of the condition in survivors is a notable reduction in dental caries, a beneficial result from the need to avoid many types of confectionery.

Pharmacological Mechanisms

Caffeine A good example of a pharmacological agent found in food with the ability to cause adverse reactions is caffeine. The stimulant effect, which may be welcome at times but unwelcome at others, of 60 mg caffeine in a cup of tea or 100 mg caffeine in a cup of coffee are well recognized. What is less well recognized is that heavy coffee or tea drinkers can suffer a number of other side effects of caffeine, which stimulates gastric secretion and can cause heartburn, nausea, vomiting, diarrhea, and intestinal colic. Also common are irregular heartbeats, episodes of rapid pulse, sweating, tremor, anxiety, and sleeplessness. Caffeine also has a diuretic effect.

Sodium nitrite Another pharmacological effect occurs when unusually large quantities of sodium nitrite are ingested. Sodium nitrite is an antioxidant used as an antibacterial agent, and in quantities of 20 mg or more it can cause dilatation of blood vessels causing flushing and headache, and urticaria.

Tyramine, histamine, and other vasoactive amines A further example of a pharmacological mechanism is the adverse effect of various vasoactive amines such as tyramine, serotonin, tryptamine, phenylethylamine, and histamine, which are found in a range of foods such as tuna, pickled herring, sardines, anchovy fillets, bananas, cheese, yeast extracts (such as Marmite), chocolate, wine, spinach, tomato, and sausages. There appear to be three main mechanisms in operation:

1. An abnormally high intake of vasoactive amines, such as histamine or tyramine, either because of a high content in food or because of synthesis of these chemicals in the gut by bacteria.
2. An abnormal effect whereby drugs or chemicals in food interfere with the enzymes that break down vasoactive amines.
3. An abnormal release from mast cells of histamine and other mediators of inflammation, triggered by eating certain foods such as strawberries, shellfish, and alcohol.

Vasoactive amines are the normal constituents of many foods. They arise mainly from the decarboxylation of amino acids, but they may also develop during normal food cooking and during the storage of food. **Table 1** shows the histamine level of various sausages. The term 'semi-dry' when applied to sausages (**Table 1**) means sausages that are fermented for varying periods. During this sausage ripening process, the histamine concentration increases, depending upon the length of the ripening process. It is estimated that 70–1000 mg of histamine ingested in a single meal is necessary for the

Table 1 Histamine levels in sausages

Type of sausage	*Histamine level (mg/100g)*	
	Mean	*Range*
Cooked sausages[a]		
Bologna	0.55	0.19–0.84
Cooked salami	0.83	0.47–5.86
Kosher salami	0.50	0.33–0.97
Semidry sausages[a]		
Thuringer cervelat	2.35	1.03–3.63
Thuringer	1.19	0.31–2.56
Dry sausages[a]		
Italian dry salami	2.14–24.5[b]	0.42–36.4[b]
Pepperoni	1.03–38.1[b]	0.72–55.0[b]
Chorizo	2.29	0.60–8.08

[a]The sausages were obtained from retail markets in the San Francisco Bay area.
[b]Depending upon the brand tested.
Source: Taylor SL, Leatherwood M, Lieber ER (1978) A survey of histamine levels in sausages. *Journal of Food Protection* **41**: 634–637.

onset of toxicity, depending on individual sensitivity. Thus, 130 g of the pepperoni sample that contained 55.0 mg histamine per 100 g (see **Table 1**) would be necessary to cause symptoms in the most sensitive individuals.

The largest amounts of histamine and tyramine are found in fermented foods such as cheese, alcoholic drinks, sausage, sauerkraut, and tinned fish. Badly stored food (see below) such as mackerel and tuna can also contain large amounts of histamine.

The effects of large doses of tyramine, histamine, and other vasoactive amines are extremely variable. Histamine causes flushing (by dilatation of blood vessels), constriction of smooth muscle in the intestine and the bronchi, increased heart rate, headache, fall in blood pressure, and asthma. Tyramine causes constriction of blood vessels, and it stimulates the release of noradrenaline from nerve endings. It can also cause the release of histamine and prostaglandins from mast cells. Dietary tyramine is known to induce hypertension and headache in patients who are taking monoamine oxidase inhibitor drugs. This effect has been shown to be due to inhibition, by these drugs, of intestinal and hepatic metabolism of tyramine, so that the amine accumulates.

The variable effect of histamine taken by mouth is in part due to the varying degree of inactivation in the gastrointestinal tract. Histamine is inactivated by mucoproteins that are produced in the gastrointestinal tract mucosa, but this inactivation can be blocked by other amines such as cadaverine and putresceine, which also bind strongly to mucoproteins. Thus, when food is taken that contains cadaverine and putresceine, more histamine can be absorbed. In fact, most of the histamine that is absorbed is degraded as it is transported across the mucosa by the intestinal enzyme diamine oxidase. Cadaverine and putresceine also have a high affinity for diamine oxidase, and can also interfere with the inactivation of histamine by this enzyme. Another barrier to the absorption of histamine is provided by the liver enzyme methyl transferase.

Thus, the effect of histamine and other vasoactive amines on an individual will depend on a number of factors, which include:

1. The amount of vasoactive amine that is present in food.
2. The amount of histamine released (as a result of an allergic process).
3. The permeability of the gastrointestinal tract, including inactivation by mucus and by mechanisms in the gut mucosa.
4. Interference with the synthesis or release of enzymes involved in amine breakdown (e.g., liver damage causing reduced activity of methyl transferase).

Tyramine and migraine There has been interest in a possible relationship between dietary tyramine and migraine. One hypothesis is that some patients with migraine have defective metabolism of ingested tyramine in the intestinal wall, which leads to increased absorption, apparently explaining why foods that contain tyramine can provoke attacks in susceptible individuals. However, there is no evidence that the activity of monoamine oxidase, the main tyramine metabolizing enzyme, is lower in patients with food-induced migraine than in other individuals prone to migraine, although levels of monoamine oxidase in platelets are generally lower in patients with migraine.

Set against these theoretical arguments, in fact most attempts to induce migraine by tyramine challenge in children and adults have been unsuccessful. Furthermore, a controlled study of exclusion of dietary vasoactive amines in children with migraine failed to demonstrate benefit. In the latter study, patients were randomly allocated to either a high-fiber diet low in dietary amines or a high-fiber diet alone. Although there was no significant difference in the results for the two groups, both groups showed a highly significant decrease in the number of headaches, emphasizing the need for a control diet in studies designed to show that dietary manipulation improves disease.

Of the foods reported to be common triggers of attacks of migraine, only cheese is rich in tyramine. Chocolate is low in this and other vasoactive amines, and red wine usually contains no more tyramine than white wine. Alcoholic drinks, particularly red wine, are commonly reported to provoke attacks of migraine. Whether these attacks are due to the alcohol itself or some other compound is a matter of debate. The major chemical difference between red and white wine is the former's high concentration of phenolic flavonoids such as anthocyanins and catechins, which as well as having direct effects on blood vessels may also inhibit the enzyme phenolsulfotransferase. Patients with food-induced migraine were shown to have significantly lower levels of platelet phenolsulfotransferase activity, and it has been hypothesized (but not proven) that low activity of this enzyme could lead to an accumulation of phenolic or monoamine substrates, which in turn might directly or indirectly provoke attacks of migraine.

Regardless of the possible mechanism, there are a number of subjects with migraine who are made worse by specific dietary triggers such as cheese or wine, for whatever reason, and avoidance of specific food triggers in susceptible subjects may prove helpful in reducing the frequency of attacks.

11β-hydroxysteroid dehydrogenase and liquorice Liquorice contains an enzyme that inhibits 11 β-hydroxysteroid dehydrogenase, resulting in sodium and water retention, hypertension, hypokalemia, and suppression of the renin–aldosterone system.

Irritant Mechanisms

Certain foods have a direct irritant effect on the mucous membranes of the mouth or gut, such as the irritant effect of coffee or curry. In certain individuals, food intolerance only occurs in the presence of a coexisting medical disorder. For example, the ingestion of spicy food, coffee, or orange juice provoke esophageal pain in some patients with reflux esophagitis. This effect is unconnected to the temperature or acidity of the food, or to any effect on the lower esophageal sphincter. The treatment in susceptible individuals is to avoid the trigger food item.

Specific Drug–Food Combinations

One example of drug-induced food intolerance is potentiation of the pressor effects of tyramine-containing foods (e.g., cheese, yeast extracts, and fermented soya bean products) by monoamine oxidase inhibitor drugs. Another is the effect of taking alcohol in patients with alcohol dependence during treatment with disulfiram (Antabuse). The reaction, which can occur within 10 min of alcohol and may last for several hours, consists of flushing and nausea.

Toxic Mechanisms

Nature has endowed plants with the capacity to synthesize substances that are toxic, and thus serve to protect them from predators whether they be fungi, insects, animals, or humans. Thus, many plant foods contain naturally occurring toxins. On a worldwide scale, reactions to naturally occurring toxins may outnumber allergic reactions, although it is currently fashionable to pay more attention to the latter.

Protease inhibitors Soya beans were originally introduced into the US as a source of oil, the extracted meal being used as a by-product that could provide animals with a source of protein. However, it was recognized that it was necessary to subject soya beans to heat treatment if they were to support the growth of animals. It was later found that the substance responsible for growth inhibition in raw soya beans was a protease (trypsin) inhibitor and it is now known that protease inhibitors are widely distributed throughout the plant kingdom, particularly in legumes, and to a lesser extent in cereal grains and tubers. In addition to inhibition of growth, one of the most characteristic responses of most animals to trypsin inhibitor is enlargement of the pancreas. The depression of growth is believed to result from endogenous loss of protein (i.e., loss into the gastrointestinal tract) due to hypersecretion by the pancreas. Soya beans products that have been adequately heat treated to inactivate trypsin inhibitor are safe for consumption.

Lectins There is a protein present in most legumes and cereals that has the property of being able to agglutinate the red blood cells of various species of animals: the so-called phytohemagglutinins or lectins. Some of these lectins, such as ricin from the castor bean, are extremely toxic. Others, such as those in the soya bean, are nontoxic. Lectins appear to be responsible for the fact that many other legumes, unless properly cooked, not only fail to support the growth of animals but can lead to death. Lectins are found in many food items commonly consumed in the human diet including tomatoes, bean sprouts, raw vegetables, fruits, spices, dry cereals, and nuts, and it is not known whether these are harmful in any way. However, it is well recognized that inadequate cooking of red kidney beans can cause severe gastrointestinal upset, with vomiting and diarrhea. It is for this reason that it is recommended that raw red kidney beans should be cooked by initially boiling hard for 10 min.

Lathyrogens Lathyrism is a paralytic disease that is associated with the consumption of chickling pea or vetch, *Lathyrus sativus*. The causative factor is believed to be an amino acid derivative, β-*N*-oxalyl-, -diaminopropionic acid; this is a metabolic antagonist of glutamic acid, a substance that is involved in the transmission of nerve impulses in the brain.

Mimosine Mimosine is an amino acid that comprises 1–4% of the dry weight of the legume *Leucaena leucocephala*, and consumption of its leaves, pods, and seeds leads to hair loss in animals. Mimosine is also a goitrogen (see below).

Djenkolic acid In parts of Sumatra the djenkol bean is a popular food item. The bean is a seed of the leguminous tree, *Pithecolobium lobatum*, and resembles the horse chestnut in size and color. Consumption of this seed leads to kidney failure that is accompanied by blood and needle-like clusters in the urine, which have been identified as containing the amino acid djenkolic acid.

Goitrogens Substances capable of producing goiter are present in plants belonging to the cabbage family, including cabbage, turnip, broccoli, cauliflower, brussel sprouts, kale, rape seed, and mustard seed. Cows' milk is a vector for the transmission of goitrogens from animals fed kale and turnips, and may have been responsible for endemic goiter in countries such as Australia and Finland.

Cyanogens A number of plants are potentially toxic because they contain glycosides from which hydrogen cyanide may be released by enzymatic hydrolysis. The most common plants eaten by humans, in order of their potential cyanide content, are: lima beans (*Phaseolus lunatus*), sorghum, cassava, linseed meal, black-eyed pea (*Vigna sinensis*), garden pea (*Pisum sativum*), kidney bean (*Phaseolus vulgaris*), Bengal gram (*Cicer arietinum*), and red gram (*Cajanus cajans*).

Vicine and convicine These are β-glucosides that are present in broad beans (*Vicia faba*). When consumed by individuals with deficiency of the enzyme glucose-6-phosphate dehydrogenase, these substances precipitate the condition of favism, which is characterized by anemia caused by hemolysis of red blood cells. The enzyme deficiency is a genetic disorder that is confined largely to inhabitants of countries surrounding the Mediterranean basin (Italy, Sicily, Lebanon, Israel, and north Africa) although individuals of the same ethnic background residing in other countries may also suffer from favism.

Cycasin Cycad seeds or nuts are obtained from *Cycad circinalis*, a palm-like tree that grows throughout the tropics and subtropics. The seeds, unless thoroughly washed, are extremely toxic, causing poisoning in humans and tumors in experimental animals. The toxic ingredient methyl-azoxymethanol, the aglycone of cycasin, is released on hydrolysis of cycasin by intestinal bacteria.

Pyrrolizidine derivatives Pyrrolizidine alkaloids are found in a wide variety of plant species. The toxic ingredient belongs to a class of compounds that are derivatives of pyrrolizidine. Large numbers of people have been poisoned through consumption of cereal and grain crops contaminated with pyrrolizidine-containing plants. It is also possible that milk from cows grazing on pastures that contain such plants could act as a vector for the transmission of pyrrolizidine to humans. In one part of western USA one such plant, the tansy ragwort (*Senecio jacobea*) is readily consumed by cows and goats, and the milk from such animals has been shown to contain significant amounts of a pyrrolizidine derivative, jacoline.

Lupin alkaloid Milk from animals that have eaten plants from the lupin family, notably *Lupinus latifolius*, may contain quinolizidine alkaloids such as anagyrine. There is strong evidence that these alkaloids are teratogenic in animals, causing severe bony deformities, and there is some evidence that similar defects may occur in the offspring of human mothers who drink alkaloid-containing milk in pregnancy.

Other examples There are numerous other examples of toxic substances present in foodstuffs. These include solanidine in potatoes, cyanide in tapioca, mycotoxins in mushrooms and cereal grains, and phototoxic furocoumarins in angelica, parsley, dill and celeriac, which in sufficient quantities can give rise to a wide variety of toxic reactions (**Table 2** and **Table 3**).

Table 2 Examples of toxic constituents of plant foodstuffs and their role in plant physiology

Toxic constituent	*Type of food containing toxic constituent*	*Physiological role of toxic constituent*	*Role in plant defense: mechanism of toxic constituent*
Protease inhibitors	Legumes, cereals, potatoes, pineapple	?Prevents degradation of storage protein during seed maturation	Part of defense against invading microbes following mechanical damage to leaves
Hemagglutinins	Legumes, cereals, potatoes	(a) Attach glycoprotein enzymes (b) Role in embryonic development/ differentiation (c) Role in sugar transport or store (d) ?Involved in root nodule nitrogen-fixing bacteria symbiosis	(a) Counteract soil bacteria (b) Antifungal (c) Protect against seed predators
Glucosinolates	Radish, horseradish, turnip, cabbage, rape seed	?Disease & insect resistance role	
Cyanogens	Almonds, cassava, corn, peas, butter beans, bamboo shoots		
Saponins	Alfalfa, French beans, soya beans		

Adapted from: Leiner IE (ed.) (1980) *Toxic Constituents of Plant Foodstuffs*, 2nd edn. New York: Academic Press.

Table 3 Examples of foodborne toxins or toxin-producing organisms, excluding plant foodstuffs

Pathogen or toxin	*Principal symptoms*	*Common food source*
Bacillus cereus	(a) diarrhea	Proteinaceous food, vegetables, sauces, puddings
	(b) Vomiting	Fried rice
Bacillus subtilis	Vomiting, diarrhea	Meat and pastry
	Flushing, sweating	Meat/seafood with rice
Bacillus licheniformis	Diarrhea	Cooked meat and vegetables
Clostridium botulinum	Neuroparalytic disease (botulism)	Meat, fish, vegetables, hazelnut conserve
Clostridium perfringens	Diarrhea, abdominal pain	Meat, poultry
Salmonella enteridis	Diarrhea, abdominal pain, fever, vomiting	Poultry, eggs
Staphylococcus aureus	Vomiting, abdominal pain, diarrhea	Numerous but specially cooked high-protein foods
Verotoxin-producing *Escherichia coli*	Hemorrhagic colitis	Ground beef
Listeria monocytogenes	Listeriosis	Unpasteurized cheese, undercooked meat
Dioxins and dibenzofurans	Adverse effects uncertain when consumed in quantities found in food	Fish
Cantharidin	Sensitivity to urethra and genitalia; priapism	Frogs that have Meloidae (blister beetles)
Methyl mercury	Brain damage	Fish, bread
Toxic alkaloid (saxitoxin) in dinoflagellates and plankton	Diverse neurological disorders (paralytic shellfish poisoning)	Clams, oysters, scallops, and mussels
Brevetoxins	Paraesthesia, abdominal pain, diarrhea, transient blindness, paralysis, death (neurotoxic shellfish poisoning)	Clams, oysters, scallops, and mussels
Ciguatera toxin	Diverse gastrointestinal and neurological disorders	Fish (especially reef predators)
Tetrodotoxin	Diverse gastrointestinal and neurological disorders	Puffer fish, certain newts
Domoic acid	Vomiting, diarrhea, hyperexcitation, seizures, memory loss (amnesic shellfish poisoning)	Mussels
Okadaic acid, dinophysis toxins, yessotoxin, pectenotoxins	Diarrhea, vomiting, abdominal pain (diarrhetic shellfish poisoning)	Mussels, scallops
Scombrotoxin (usually histamine)	Headache, palpitations, gastrointestinal disturbance	Mackerel, tuna, and related species
Tetramine (red whelk poisoning)	Diplopia, dizziness, leg pains	Whelks
Grayanotoxins (in honey from areas of Turkey where Rhododendrons are grown)	Hypotension, bradycardia, vomiting, sweating	Honey
Unknown (? in algae) (turtle flesh poisoning)	Cardiorespiratory failure, death	Turtles

Food Storage

Chemical changes in food during storage can produce substances that cause food intolerance. An example is intolerance to ripe or stored tomatoes in subjects who can safely eat green tomatoes, where ripening of the fruit produces a new active glycoprotein. Some adverse reactions resulting from food storage come into the category of toxic reactions, such as the rise in levels of histamine and tyramine in certain foods during storage as a result of bacterial decarboxylation. An example of this is the production of histamine in badly stored mackerel and other fish: scombroid fish poisoning. Contamination of food by antigens such as storage mites or microbial spores may give rise to adverse effects, particularly asthma and eczema. Contamination of food by microorganisms may result in adverse effects. For example, celery, parsnip, and parsley may become infected with the fungus *Sclerotinia scleriotiorum* ('pink rot'), resulting in the production of the photosensitizing chemicals psoralen, 5-methoxypsoralen, and 8-methoxypsoralen.

Practical Applications

Food arouses not only the appetite but also the emotions. The passion for food that is natural (i.e., free from extraneous ingredients) is not new; in 1857, a survey of adulterants in food showed that childrens' sweets were commonly colored by red lead (lead oxide), lead chromate, mercuric sulfide, and copper arsenite. By the late 1850s, 'pure and unadulterated' had become the stock advertising slogan of those anxious to cash in on the then

newly awakened fears of the public. The current scale of the use of additives in food comes as a surprise to most people, and it is understandable that many should find these substances vaguely menacing. Nonetheless, the current phobia of food additives and food processing, and the obsession for so-called natural or health food arises largely out of misinformation and ignorance. Obsession with so-called natural or health food ignores the wide range of naturally occurring toxins in foods. The concept of health food is wholly misleading. For example, a survey of 'crunchy' peanut butter showed that 11 out of 59 samples from health food producers contained over 100 $\mu g\,kg^{-1}$ of aflatoxins, over 10 times the proposed maximum permitted level for total aflatoxins. Only one of the 26 samples from other producers contained aflatoxins in excess of 10 $\mu g\,kg^{-1}$, and none contained more than 50 $\mu g\,kg^{-1}$.

See also: **Caffeine**. **Food Allergies**: Etiology; Diagnosis and Management. **Food Safety**: Mycotoxins. **Fructose**. **Lactose Intolerance**. **Vitamin A**: Biochemistry and Physiological Role; Deficiency and Interventions.

Further Reading

Ashwood-Smith MJ, Ceska O, and Chaudhary SK (1985) Mechanism of photosensitivity reactions to diseased celery. *British Medical Journal*, vol. **290**: 1249.

Bjarnason I, Levi S, Smethurst P, Menzies IS, and Levi AJ (1988) Vindaloo and you. *British Medical Journal* **297**: 1629–1631.

Bleumink E, Berrens L, and Young E (1967) Studies on the atopic allergen in ripe tomato fruits. *International Archives of Allergy* **31**: 25–37.

Ciegler A (1975) Mycotoxins: occurrence, chemistry, biological activity. *Lloydia* **38**: 21–35.

Conning DM and Lansdown ABG (eds.) (1983) *Toxic Hazards in Food*. London: Croom Helm.

Edwards CRW (1991) Lessons from licorice. *New England Journal of Medicine* **325**: 1242–1243.

Farese RV, Bigieri EG, Shackleton CHL, Irony I, and Gomez-Fontes R (1991) Licorice induced hypermineralocorticoidism. *New England Journal of Medicine* **325**: 1223–1227.

Forsythe WI and Redmond A (1974) Two controlled trials of tyramine in children with migraine. *Developmental Medicine and Child Neurology* **16**: 794–799.

Gibson GG and Walker R (eds.) (1985) *Food Toxicology – Real or Imaginary Problems?* London: Taylor & Francis.

Gumbmann MR, Spangler WL, Dugan GM, and Rackis JJ (1986) Safety of trypsin inhibitors in the diet: effects on the rat pancreas of long-term feeding of soy flour and soy protein isolate. In: Friedman M (ed.) *Nutritional and Toxicological Significance of Enzyme Inhibitors in Foods*, pp. 33–79. New York: Plenum Press.

Hall MJ (1987) The dangers of cassava (tapioca) consumption. *Bristol Medico-Chirurgical Journal* **102**: 37–39.

Harris JB (ed.) (1986) *Natural Toxins. Animal, Plant, and Microbial*. Oxford: Clarendon Press.

Horwitz D, Lovenberg W, Engelman K, and Sjoerdsma A (1964) Monoamine oxidase inhibitors, tyramine, and cheese. *Journal of the American Medical Association* **188**: 90–92.

Kaufman HS (1986) The red wine headache: A pilot study of a specific syndrome. Immunology and Allergy Practice **8**: 279–284.

Knudson EA and Kroon S (1988) In vitro and in vivo phototoxicity of furocoumarin-containing plants. *Clinical Experimental Dermatology* **13**: 92–96.

Leiner IE (ed.) (1980) *Toxic Constituents of Plant Foodstuffs*, 2nd edn. New York: Academic Press.

Lessof MH (1992) *Food Intolerance* London: Chapman & Hall.

Littlewood JT, Glover V, Davies PTG, Gibb C, Sandler M, and Rose FC (1988) Red wine as a cause of migraine. *Lancet* **1**: 558–559.

Mahoney CP, Margolis MT, Knauss TA, and Labbe RF (1980) Chronic vitamin A intoxication in infants fed chicken liver. Pediatrics 1980; **65**: 893–896.

Masyczek R and Ough CS (1983) The "Red Wine Reaction" syndrome. *American Journal of Enology and Viticulture* **34**: 260–264.

Moffett A, Swash M, and Scott DF (1972) Effect of tyramine in migraine: a double-blind study. *Journal of Neurology, Neurosurgery and Psychiatry* **35**: 496–499.

Moffett AM, Swash M, and Scott DF (1974) Effect of chocolate in migraine: a double-blind study. *Journal of Neurology, Neurosurgery and Psychiatry* **37**: 445–448.

Morgan RGH, Crass RA, and Oates PS (1986) Dose effects of raw soyabean flour on pancreatic growth. In: Friedman M (ed.) *Nutritional and Toxicological Significance of Enzyme Inhibitors in Foods*, pp. 81–89. New York: Plenum Press.

Noah ND, Bender AE, Reaidi GB, and Gilbert RJ (1980) Food poisoning from raw red kidney beans. *British Medical Journal* **281**: 236–237.

Price SF, Smithson KW, and Castell D (1978) Food sensitivity in reflux esophagitis. *Gastroenterology* **75**: 240–243.

Rackis JJ, Wolf WJ, and Baker EC (1986) Protease inhibitors in plant foods: content and inactivation. In: Friedman M (ed.) *Nutritional and Toxicological Significance of Enzyme Inhibitors in Foods*, pp. 299–347. New York: Plenum Press.

Salfield SAW, Wardley BL, Houlsby WT, Turner SL, Spalton AP, Beckles-Wilson NR, and Herber SM (1987) Controlled study of exclusion of dietary vasoactive amines in migraine. *Archives of Disease in Childhood* **62**: 458–460.

Sandler M, Youdim MBH, and Hanington E (1974) A phenylethylamine oxidising defect in migraine. *Nature* **250**: 335–337.

Taylor SL (1986) Histamine food poisoning: toxicology and clinical aspects. *CRC Critical Reviews in Toxicology* **17**: 90–128.

Wuthrich B and Ortolani C (eds.) (1996) *Highlights in Food Allergy*. Basel: Karger.

FOOD SAFETY

Contents

Mycotoxins

J D Groopman and T W Kensler, Johns Hopkins University, Baltimore, MD, USA

Mycotoxins are toxic fungal metabolites of enormous chemical diversity that naturally contaminate the human food supply. These compounds induce an array of toxicologic effects when consumed in sufficient quantities. The three major genera of mycotoxin-producing fungi are *Aspergillus*, *Fusarium*, and *Penicillium*. This field has been comprehensively reviewed by the Council for Agricultural Science and Technology (CAST) and concisely by Etzel. The potential production of mycotoxins is insidious since fungal growth can occur both prior to and after grain harvest. Ecological conditions such as drought or damage to seeds by insects or mechanical harvesting can enhance mycotoxin production during both growth and storage. Mycotoxin production also occurs over a wide range of moisture content, relative humidity, and temperature. The major crops affected throughout the world are corn, peanuts, cotton, wheat, rice, and the processed food derived from these commodities.

Following the discovery of the carcinogenic aflatoxins 40 years ago, the search for mycotoxins has led to the identification of more than 100 toxigenic fungi and more than 300 mycotoxins worldwide. Most of these mycotoxins have not been linked to any toxic syndromes in animals or people, but some, such as aflatoxins, certain trichothecenes, fumonisins, and ochratoxins, have been implicated in highly lethal episodic outbreaks of mold poisoning in exposed animals and/or human populations. Mycotoxins with carcinogenic potency in experimental animal models include aflatoxins, sterigmatocystin, ochratoxin, fumonisin, patulin, and penicillic acid. Of these agents, aflatoxin B_1 has been classified as a category I human carcinogen by the International Agency for Research on Cancer (IARC). This article briefly describes the occurrence, biological effects, mechanistic studies, and where available, epidemiological associations of dietary exposure to major mycotoxins with human disease outcomes.

Aflatoxins

Chemistry and Occurrence

The aflatoxins (AFs) were discovered as the causative agent of turkey X disease, which resulted in the death of thousands of turkey poults, ducklings, and chicks fed a contaminated peanut meal. Chemically, the AFs are a highly substituted coumarin moiety containing a fused dihydrofurofuran moiety. Four major AFs designated B_1, B_2, G_1, and G_2 are produced by *A. flavus* and *A. parasiticus*. AFB_1 and AFB_2 were named because of their strong blue fluorescence under ultraviolet light, whereas AFG_1 and AFG_2 fluoresced greenish-yellow.

Commodities most often shown to contain AFs are peanuts, various other nuts, cotton seed, corn, and rice. Human exposure can occur from consumption of AFs from these sources and the products derived from them, as well as from tissues, eggs, and milk (AFM_1) from animals that have consumed contaminated feeds. When contamination occurs, AFB_1 generally predominates. Although contamination by the molds may be universal within a given geographical area, the levels or final concentrations of AFs in the grain product can vary from less than 1 ppb to greater than 12,000 ppb. It is important to note that obvious contamination of a commodity with *A. flavus* or *A. parasiticus* does not necessarily indicate the presence of AFs, and the appearance of a sound, uninfected sample of commodity does not preclude the existence of significant quantities of AFs.

Widespread concern regarding the toxic effects of AFs in humans and animals and possible transfer of residues from animal tissues and milk to humans has led to regulatory actions governing the interstate as

well as global transport and consumption of AF-contaminated food and feed commodities. The US Food and Drug Administration has set the action levels of AF in commodities. For feeding mature nonlactating animals, the action level is 100 ppb; for commodities destined for human consumption and interstate commerce, it is 20 ppb; and for milk it is 0.5 ppb.

Toxic Effects

AFs are potent liver toxins, and their effects in animals vary with dose, length of exposure, species, breed, and diet or nutritional status. These toxins may be lethal when consumed in large doses; sublethal doses produce a chronic toxicity, and low levels of chronic exposure can result in cancer, primarily liver cancer, in many animal species. AFB_1, the most potent and most commonly occurring AF, is acutely toxic to all species of animals, birds, and fishes tested. Sheep and mice are the most resistant, whereas cats, dogs, and rabbits are the most sensitive species. Chronic aflatoxicosis is characterized by bile duct proliferation, periportal fibrosis, icterus, and cirrhosis of liver. Prolonged exposure to low levels of AFB_1 leads to hepatoma, cholangiocarcinoma, or hepatocellular carcinoma and other tumors. The molecular basis for the toxicology of aflatoxin has been reviewed by Wild and Turner.

Some cases of acute aflatoxicosis in humans have been reported in the literature, especially in the subpopulations of developing countries. Clinical manifestations were characterized by vomiting, abdominal pain, pulmonary edema, and fatty infiltration and necrosis of the liver. There was a putative aflatoxin poisoning in western India when there was consumption of heavily molded corn. There were at least 97 fatalities, and these deaths occurred only in households where the contaminated corn was consumed. Histopathology of liver specimens revealed extensive bile duct proliferation, a lesion often noted in experimental animals after acute AF exposure. An incident of acute aflatoxicosis in Kenya was also associated with consumption of maize highly contaminated with AF. There were 20 hospital admissions with a 20% mortality. Also, the consumption of AF-contaminated noodles resulted in acute hepatic encephalopathy in children in Malaysia. Up to 3 mg of AF was suspected to be present in a single serving of contaminated noodles.

Carcinogenicity in Animals

AFB_1 is a potent liver carcinogen in many species of animals, including rodents, nonhuman primates, and fish. In appropriate circumstances, dependent on such variables as animal species and strain, dose, route of administration, and dietary factors, significant incidences of tumors have been induced at sites other than the liver, such as kidney and colon. AFB_1 has been demonstrated to induce liver tumors in two species of lower primates: the tree shrew (*Tupaia glis*) and the marmoset (*Saguinus oedipomidas*). All liver tumors of the tree shrew were classified as hepatocellular carcinoma (HCC) and developed in a manner similar to those of the rat. Unlike the case with rats, in the marmoset histologic observation revealed the association of cirrhotic changes with liver tumor development. Rhesus monkeys have also proven to be susceptible to AFB_1 carcinogenicity. Data from 47 monkeys, representing three species (rhesus, cynomolgus, and African green), that had received AFB_1 have been published. Primary liver tumor incidence was 19% (5/26) in animals surviving for longer than 6 months, and total tumor incidence in these animals was 50% (13/26).

Metabolism

Metabolism plays a critical role in the biological activity and disposition of AF. To produce a DNA damage product, AFB_1 undergoes an initial two-electron oxidation by the cytochrome P450 family members CYP1A2 and CYP3A4, yielding aflatoxin B1-8,9-oxide. This epoxide reacts with the *N*7 atom of guanine to form a pro-mutagenic DNA adduct (aflatoxin–*N*7–guanine). The aflatoxin–DNA adduct is unstable and undergoes depurination, leading to its urinary excretion. Aflatoxin B1-8,9-oxide is also a substrate for several isoforms of human glutathione *S*-transferases (GSTs), which yield a stable, nontoxic, polar product that is excreted in the bile. The aflatoxin–glutathione product also undergoes sequential metabolism in the liver and kidneys to be excreted as a mercapturic acid (aflatoxin–*N*-acetylcysteine) in the urine. Aflatoxin B1 also undergoes extensive oxidation, which is catalyzed by cytochrome P450s. In addition to formation of the 8,9-oxide, oxidation by CYP1A2 yields a stable urinary metabolite, aflatoxin M1, that is excreted in milk. Aflatoxin M1 is less carcinogenic or mutagenic than aflatoxin B1, but it is equally toxic. The oxidation products of aflatoxin can be excreted without further biotransformation or can be conjugated by UDP-glucuronosyl transferases. Collectively, these end products of aflatoxin biotransformation are biomarkers of exposure to aflatoxin and risk of hepatocellular carcinoma.

Aflatoxin and Human Cancer

HCC is the fifth leading cause of cancer mortality throughout the world, and in areas of Asia and Africa it accounts for nearly 70% of all cancer deaths. Furthermore, due to the lack of symptoms in the early stages and rapid growth rates of tumors, most HCCs are discovered in very advanced stages. The 5-year mortality rate for individuals diagnosed with HCC is greater than 95%. In the People's Republic of China, HCC is the third leading cause of cancer mortality and accounts for at least 250,000 deaths per year, with an incidence in some counties approaching 100 cases per 100,000 per year. Moreover, in high-risk regions of the world the median age of onset of HCC is decades earlier than in the United States.

In the early 1990s, nested case–control studies conducted in Shanghai utilized these biomarkers to establish a significant association between aflatoxin exposure and HCC. They showed that the risk of HCC increased dramatically (60-fold) in individuals who had been exposed to aflatoxin and had chronic hepatitis infection compared to those with neither the chemical nor viral exposures. Additional studies in Qidong and Taiwan have confirmed this striking chemical–viral interaction. The underlying mechanism for this interaction remains poorly understood.

The relationship between aflatoxin exposure and development of human HCC is further highlighted by molecular biological studies on the *p53* tumor suppressor gene, the most common mutated gene detected in many human cancers. The initial results came from three independent studies of *p53* mutations in HCCs occurring in populations exposed to high levels of dietary AF and found high frequencies of G → T transversions, with clustering at codon 249. On the other hand, studies of *p53* mutations in HCCs from Japan and other areas where there is little exposure to AF revealed no mutations at codon 249.

A positive correlation has been observed between population estimates of aflatoxin exposure and the proportion of HCC cases with a *p53* 249^{ser} mutation detected in plasma. Kirk and colleagues analyzed restriction length fragment polymorphisms to detect this mutation in the plasma of liver cancer patients in The Gambia, West Africa. Jackson *et al.* subsequently used mass spectrometry to detect the same mutation in matched plasma and tumor samples from cancer patients in Qidong. Continuing validation of this biomarker in a prospective cohort has shown that it can be detected 1 or more years in advance of HCC diagnosis. Collectively, these genetic serum markers reveal a new paradigm for early identification of at-risk individuals and HCC diagnosis.

Fumonisins

Occurrence

The fumonisins are a class of mycotoxins produced by *Fusarium moniliforme*, a fungus that is a ubiquitous contaminant of corn, and are also found at high levels in milk and cereal products. Six fumonisins have been isolated and characterized from *F. moniliforme*. They are designated as fumonisin B_1 (FB_1), B_2, B_3, B_4, A_1, and A_2. Only FB_1 and FB_2 appear to be toxicologically significant and have been studied to any extent. FB_1 and FB_2 were first isolated in 1988 and invariably occur together, with FB_2 at levels of 15–35% of FB_1. Levels of FB_1 have annual variation but are consistently in the 0.5 to 2 ppm range in US cornmeal and have been reported as high as 150 ppm in corn destined for human consumption in South Africa. Regulatory limits for fumonisins in commodities are currently being promolugated worldwide.

Toxicity

Fusarium moniliforme contaminated corn has been associated with several human and animal diseases, including luekoencephalomalacia (LEM) in horses, pulmonary edema in swine, and hepatoxicity in horses, swine, and rats. Both culture material from *F. moniliforme*-inoculated corn and pure FB_1 are capable of producing similar effects in animals. Neurotoxic signs and symptoms, including loss of feed consumption, lameness, ataxia, oral and facial paralysis, and recumbency, begin within days after initial consumption of moldy corn or by direct administration of FB_1 and may be rapidly followed by seizures and morbidity. Focal malacia and liquefaction of cerebral white matter with peripheral hemorrhage is the pathognomonic finding.

Studies have provided possible insights into the mechanisms of toxicity. The fumonisins bear considerable structural similarity to the long-chain (sphingoid) base backbones of sphingolipids. It was demonstrated that incubation of rat hepatocytes with fumonisins inhibited sphingosine biosynthesis. FB_1 increased the amount of the biosynthetic intermediate sphinganine, which suggests that fumonisins inhibit the conversion of sphinganine to *N*-acyl-sphinganines. It was subsequently shown, using mouse cerebellar neurons in culture, that FB_1 inhibited ceramide synthase in mouse brain microsomes with a competitive-like kinetic behavior with respect to both sphinganine and stearoyl-CoA. Thus, disruption of the *de novo* pathway of sphingolipid biosynthesis may be a critical event in

the diseases that have been associated with consumption of fumonisins.

Carcinogenicity in Animals

Rats fed a diet supplemented with maize contaminated with the *F. moniliforme* that had caused an outbreak of LEM in horses all developed hepatic nodules, cholantiofibrosis, or cholangiocarcinomas within 6 months. Lifetime studies in rats fed diets containing maize inoculated with *F. moniliforme* yielded a high incidence of liver tumors. The carcinogenicity of FB_1 has been directly assessed in a study in which a semipurified diet containing 50 mg/kg of pure (>90%) FB_1 was fed to rats. Ten out of 15 FB_1-treated rats (66%) developed primary HCC.

Human Health Effects

In an initial study in high-risk and low-risk regions of Transkei (South Africa) for esophageal cancer, cancer rates were correlated with the proportion of maize samples infected by *F. moniliforme*. In a follow-up study, the mean proportions of maize kernel infected with *F. moniliforme* in both healthy and moldy maize samples from households in the high-incidence esophageal cancer area were significantly higher than those in the low incidence area. FB_1 and FB_2 levels in healthy maize samples from the low-risk area were approximately 20 times lower than those in healthy samples from high-risk areas. One study estimated that naturally poisoned horses consumed levels of fuminisins equivalent to those shown to be toxic experimentally, and that humans in high esophageal cancer risk areas can potentially consume levels higher than those shown to be carcinogenic in rats.

A number of surveys have been conducted in Henan Province in northern China. Fungal strains from samples of wheat, corn, dried sweet potato, rice, and soya beans were cultured and isolated in five counties with a high incidence of esophageal cancer and three with a low incidence. The frequency of contamination by *F. moniliforme* was significantly higher in food samples from high-risk areas, although the frequency of contamination by all other fungi analyzed was also significantly higher in samples from the high-risk counties. Although these studies, as well as those conducted in South Africa, demonstrate correlations between high esophageal cancer rates and contamination of foods, primarily corn, with *F. moniliforme*, the specific role of the fumonisins or other related toxins in the etiology of this cancer remains to be firmly established.

Ochratoxins

Occurrence

Ochratoxins are a group of structurally related metabolites that are produced by *A. ochraceus* and related species, as well as *P. viridicatum* and certain other *Penecillium* species. The major mycotoxin in this group is ochratoxin A (OA), which appears to be the only one of major toxicological significance. Chemically, OA contains an isocoumarin moiety linked by a peptide bond to phenylalanine. OA has been detected in many food commodities throughout the world but is found primarily in grains grown in northern temperate areas resulting in contamination of breads and cereal products. In addition to cereals, animal products such as sausage can be significant human dietary sources of OA. Although OA has been found in many foodstuffs in many countries, the highest frequency of OA contamination in foods (~10%) was encountered in Croatia, where Balkan endemic nephropathy (BEN) is highly prevalent. Moreover, average concentrations are higher in foods from nephropathic regions. Many countries have set regulatory limits for OA ranging from 1 to 50 ppb for food and from 100 to 1000 ppb for animal feeds.

Toxicity

The toxicity of OA varies considerably with dose and between species. Dogs and pigs are the most sensitive species (0.2 and 1 mg/kg body weight, respectively). Synergistic effects of OA with other mycotoxins, such as citrinin and penicillic acid, on the LD_{50} were seen in mice following intraperitoneal injection. OA is nephrotoxic to a number of animal species, and the presence of OA in feed is believed to be the most important cause of spontaneous mycotoxic porcine and poultry nephropathy. OA also produces hepatic toxicity at high doses. OA is teratogenic in mice, rats, and hamster, and the major target in the fetus is the developing central nervous system. OA is also immunosuppressive at low doses, affecting immune function at both the level of antibody synthesis and natural killer cell activity. The toxic mechanism of OA has been shown to be inhibition of protein synthesis by competition with phenylalanine in the phenylalanyl-tRNA synthetase-catalyzed reaction. OA also inhibits other enzymes that use phenylalanine as a substrate, such as phenylalanine hydroxylase. The effect of OA on protein synthesis is followed by an inhibition of RNA synthesis, which may affect proteins with a high turnover. OA has also been found to enhance lipid peroxidation *in vivo*.

Carcinogenicity

OA has been tested for carcinogenicity by oral administration in mice and rats. The kidney, and in particular the tubular epithelial cells, was the major target organ for OA-induced lesions. In male ddY and DDD mice, atypical hyperplasia, cystadenomas, and carcinomas of the renal tubular cells were induced, as were neoplastic nodules and hepatocyte tumors of the liver. In B6C3F1 mice, tubular-cell adenomas and carcinomas of the kidneys were induced in male mice, and the incidences of hepatocellular adenomas and carcinomas were increased in male and female mice. In male and female F344 rats, OA induced neoplastic effects in the kidneys.

Human Health Effects

It has been suggested for several decades that excessive exposure to OA plays a substantive role in the development of BEN. BEN is a bilateral, noninflammatory, chronic nephropathy in which the kidneys are extremely reduced in size and weight and show diffuse cortical fibrosis. Functional impairments are characterized by progressive hypercreatininemia, hyperuremia, and hypochromic anemia. In an endemic area of Croatia, an extremely high incidence of urinary tract tumors in the endemic areas for BEN, particularly urothelial tumors of the pelvis and ureter, has been reported. In Bulgaria, 16 cases of urinary tract tumors were reported among 33 autopsied patients with BEN. A causal relationship between exposure to OA and these human diseases is suggested by (i) similarities in the morphological and functional renal impairments induced by OA in animals and those observed in BEN and (ii) the finding that foods from the endemic areas are more heavily contaminated with OA than foods from disease-free areas. Analyses of serum samples in European countries from nearly a dozen studies revealed that blood from healthy humans was contaminated with OA at concentrations of 0.1–40 ng/ml. The frequency of contamination of human sera, which ranged from 4 to 57%, seems to indicate continuous, widespread exposure of humans to OA.

An association between BEN and/or urinary tract tumors and OA content in blood samples has been reported. Among 61 patients with BEN and/or urinary tract tumors, 14.8% had levels of 1 or 2 ng/ml and 11.5% had more than 2 ng/ml OA in their blood. This proportion was significantly higher than that in a control group of 63 individuals from unaffected families in the endemic villages (7.9 and 3.2%, respectively). A case–control study provided molecular evidence for the possible role of ochratoxin in the development of urinary tract tumors in Bulgaria.

Trichothecenes

The trichothecenes are a family of more than 150 structurally related compounds produced by several fungal genera (*Fusarium*, *Cephalosporium*, *Myrothecium*, *Stachybotrys*, and *Trichoderma*). Chemically, they are sesquiterpenes characterized by a double bond at position C-9, an epoxide ring at C-12, and various patterns of hydroxy and acetoxy substitutions at positions C-3, C-4, C-15, C-7, and C-8. There are four naturally occurring trichothecene mycotoxins (deoxynivalenol, nivalenol, T-2 toxin, and diacetoxyscirpenol) produced in food and feed by *Fusarium* species.

Deoxynivalenol

Deoxynivalenol (DON) is probably the most widely distributed *Fusarium* mycotoxin. Its occurrence in foods in North America, Japan, and Europe is common, but the concentration is relatively low; however, its contamination in cereals in some developing countries, particularly in southern China and areas of South America and Africa, is usually high during some years. In 1980 and 1981 in Canada and 1982 in the United States, DON was found in wheat as the result of severe infestations with the wheat scab fungus, *F. graminearum*. In both countries, the soft winter wheats were the most severely affected. In Canada, dried corn was found to contain higher levels of DON. Several countries have set guidelines or official tolerance levels for DON in food and feed. The range varies from 0.005 to 4 mg/kg. Acute mycotoxicoses affecting fairly large numbers of people and caused by ingestion of DON-contaminated food have been reported in China, India, and other countries.

T-2 Toxin

T-2 toxin is produced primarily by *F. sporotrichioides* and has been reported in many areas of the world. It is formed in large quantities in the unusual circumstance of prolonged wet weather at harvest. Natural contamination of foods and feeds by T-2 toxin in the United States has been reported in only one incident involving heavily molded corn. An official tolerance level of 0.1 mg/kg was established for T-2 toxin in grains in Russia.

T-2 toxin, as the representative trichothecene, has been well studied for its toxic effects on various animal models and has been reviewed in detail.

General signs of toxicity in animals include weight loss, decreased feed conversion, feed refusal, vomiting, bloody diarrhea, severe dermatitis, hemorrhage, decreased egg production, abortion, and death. Histologic lesions consist of necrosis and hemorrhage in proliferating tissues of the intestinal mucosa, bone marrow, spleen, testis, and ovary. T-2 toxin can alter hemostasis and affect cellular immune response in animals, and it is a strong inhibitor of protein and DNA synthesis. T-2 toxin is also teratogenic in mice and rats. As the major trichothecene mycotoxin, T-2 toxin has been implicated in a variety of animal and human toxicosis, such as alimentary toxic aleukia, Msleni joint disease, scabby grain toxicosis, and Kashin–Beck disease.

Other Mycotoxins

Zearalenone

Zearalenone (ZEN) is produced primarily by *F. graminearum* and is among the most widely distributed *Fusarium* mycotoxins. It is associated mainly with maize but occurs in modest concentrations in wheat, barley, sorghum, and other commodities. An official tolerance level of 1 mg/kg ZEN in grains, fats, and oils was established in Russia. Proposed levels in other countries are 0.2 mg/kg ZEN in maize in Brazil and 0.03 mg/kg in all food in Romania. ZEN has estrogenic effects in domestic pigs and experimental animals. F-2 toxicosis and hyperestrogenism are two diseases in pigs caused by ZEN. ZEN is teratogenic to mice and rats and induces chromosomal anomalies in cultured rodent cells. Its carcinogenicity was tested by administration in the diet of mice in one experiment and in that of rats in two experiments. An increased incidence of hepatocellular adenomas was observed in female mice, and an increased incidence of pituitary adenomas was observed in mice of both sexes. No increase in the incidence of tumors was observed in rats.

Sterigmatocystin

Sterigmatocystin is produced by several species of *Aspergillus*, *Penicillium luteum*, and a *Bipolaris* species. Chemically, sterigmatocystin resembles the AFT and is a precursor in the biosynthesis of AFT. It has been detected at low concentrations in green coffee, moldy wheat, and in the rind of hard Dutch cheese. Sterigmatocystin is a hepatotoxin and is less potent than the AFT. It was mutagenic in the Ames test, the *Rec* assay, and the *Bacillus subtilis* assay. It can covalently bind to DNA and form DNA adducts. It has been proven that sterigmatocystin is carcinogenic to rats and mice, mainly inducing liver tumors.

Patulin

Patulin is produced primarily by *P. expansum*. Other *Penicillium* and *Aspergillus* species can also be patulin producers. Commodities found contaminated with patulin are mainly fruits and fruit juices in Europe and North America. Patulin is appreciably stable in apple and grape juices, and it may constitute a potential threat to humans. Currently, 11 countries have set regulatory limits for patulin in fruit juice ranging from 30 to 50 ppb. The toxicity of patulin has been studied in many experimental models, including chicken, quail, cat, cattle, rabbit, mice, and rats. The toxic effects on these animals were found to be edema and hemorrhage in brain and lungs; capillary damage in the liver, spleen, and kidney; paralysis of motor nerves; and convulsions. Patulin is also an immunosuppressive agent that inhibits multiple aspects of macrophage function.

See also: **Cancer**: Effects on Nutritional Status. **Liver Disorders**. **Nuts and Seeds**.

Further Reading

Chen J-G *et al.* (1998) Population-based cancer survival in Qidong, People's Republic of China. *IARC Scientific Publications* **145**: 27–35.

Council for Agricultural Science and Technology (CAST) (2003) *Mycotoxins: Risks in Plants, Animals and Human Systems.* Washington, DC: CAST.

Etzel RA (2002) Mycotoxins. *Journal of the American Medical Association* **287**: 425–427.

Harris CC (1996) *p53* tumor suppressor gene: from the basic research laboratory to the clinic—an abridged historical perspective. *Carcinogenesis* **17**: 1187–1198.

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans (1993) *Some Naturally Occurring Substances: Food Items and Constituents, Heterocyclic Aromatic Amines and Mycotoxins,* vol. 56. Lyon, France: IARC.

Jackson PE, Qian G-S, Friesen MD *et al.* (2001) Specific *p53* mutation detected in plasma and tumors of hepatocellular carcinoma patients by electrospray ionization mass spectrometry. *Cancer Research* **61**: 33–35.

Kirk GD, Camus-Randon AM, Mendy M *et al.* (2000) Ser-249 *p53* mutations in plasma DNA of patients with hepatocellular carcinoma from The Gambia. *Journal of the National Cancer Institute* **92**: 148–153.

Li L and Rao K (eds.) (2001) *Cancer Incidence and Mortality in Cities and Counties of P. R. China 1988–1992*. China Medical Science and Technology Press.

Lye MS, Ghazali AA, Mohan J, Alwin N, and Nair RC (1995) An outbreak of acute hepatic encephalopathy due to several aflatoxicosis in Malaysia. *American Journal of Tropical Medicine and Hygene* **53**: 68–72.

Park D and Troxell T (2002) US perspective on mycotoxin regulatory issues. *Advances in Experimental Medicine and Biology* **504**: 277–285.

Parkin DM, Pisani P, and Ferlay J (1999) Estimates of the worldwide incidence of 25 major cancers in 1990. *International Journal of Cancer* **80**: 827–841.

Sun Z-T *et al.* (1999) Increased risk of hepatocellular carcinoma in male hepatitis B surface antigen carriers with chronic hepatitis who have detectable urinary aflatoxin metabolite M1. *Hepatology* **30**: 379–383.

Wang L-Y, Hatch M, Chen CJ *et al.* (1996) Aflatoxin exposure and the risk of hepatocellular carcinoma in Taiwan. *International Journal of Cancer* **67**: 620–625.

Wild CP and Turner PC (2002) The toxicology of aflatoxins as a basis for public health decisions. *Mutagenesis* **17**: 471–481.

Yu MW, Lien JP, Chiu YH *et al.* (1997) Effect of aflatoxin metabolism and DNA adduct formation on hepatocellular carcinoma among chronic hepatitis B carriers in Taiwan. *Journal of Hepatology* **27**: 320–330.

Pesticides

M Saltmarsh, Alton, UK

What are Pesticides?

Pesticide is a generic term that covers a wide range of natural and synthetic chemicals (over 700 in total) that are used to protect crops from attack from pests, both before and after harvest. There are many different sorts of pests. The term includes insects, slugs and snails, nematode worms, mites, rodents, weeds, molds, bacteria and viruses. The chemicals can be applied before and during growth of the plant or on to the stored crop as, for example, fumigants, which are used to kill pests that have infested stored cocoa or grain. Chemicals used to treat pests on animals are not included; they are considered as veterinary medicines.

The pesticide formulation used by the farmer will include the pesticide chemical itself and a number of other chemicals that enable it to be applied and to work as effectively as possible. These will include solvents, adhesives, and surface-active agents such as emulsifiers. In some cases other chemicals, known as 'safeners,' are applied to minimize the damage done to the crop while maintaining the effectiveness of the spray on the target.

It is estimated that worldwide usage of pesticides is around 2.5 million tons with a cost in 1997 of US$21 billion.

Why Do We Need Pesticides?

Food crops are subject to attack by a multitude of pests and diseases and pesticides are applied to minimize the damage to the crop. It has been estimated that without protection world cereal crop yields would fall by between 46 and 83%. History is littered with records of crop failures and famine caused primarily by rodent, insect or fungus. Some of these events have had a wide-ranging and long-lasting effect, like the 1845–1846 Irish potato famine and the 1917–1918 German 'turnip winter,' the latter so called because the potatoes rotted and turnips were the only stored root crop that was available to feed the population through the winter. Both these events, in which 1.5 million and 700 000 people died, respectively, were caused by potato blight, infection by the fungus *Phytophthora infestans*. Famine caused by massive swarms of locust is still all too common in Northern Africa and Arabia. Less spectacular but as disastrous is the loss of an estimated 30% of harvested crops in India to rodents.

In addition to the loss of the crop, pesticides are used to control agents which make the crop toxic rather than healthy. Two examples are the toxins caused by fungi. When an insect bores into a peanut it allows spores of the fungus *Aspergillus flavus* to enter and grow, producing the aflatoxins, a series of carcinogens. When rye (*Secale cereale*) grows in damp conditions a fungus, *Claviceps purpurea*, can grow on the seed. If this seed is subsequently ground into flour and made into bread it can cause consumers to suffer hallucinations, gangrene, and death. Outbreaks amounting to epidemics were common in the Middle Ages in Europe and one occurred as recently as 1951 in France.

A second reason relates not so much to quantity as to quality. Supermarkets in the developed nations offer a wide range of fresh produce at competitive prices. Consumers do not like holes made by slugs and snails in their fresh lettuce. They do not expect scab marks on their apples, or holes made by small maggots in their carrots. Flour millers do not expect to have to clean the grain from weed seeds before milling. Even small defects can dramatically reduce the value of the crop, or indeed make it unsaleable, and the need for a competitive price requires minimal labor input so that application of pesticide is essential.

Types of Pesticides

There are currently around 600 pesticides, both natural and synthetic. Natural pesticides include both chemicals derived from plant sources and biological agents such as parasitic wasps, mites, bacteria, and chemicals contained within or exuded by plants or bacteria. While there is no inherent reason why natural products should be any safer than synthetic ones (after all, insect

venoms and toxins and poisonous plants are natural), it appears that the risks do lie in their potential impact on the environment rather than on their effect in food. There are also increasing numbers of cases where plants have been given a gene which expresses a natural pesticide (see *Bacillus thuringiensis*, below).

At the time of writing, naturally derived pesticides make up less than 5% of the world pesticide market, but a great deal of work is being devoted to the screening of natural sources and this proportion will certainly increase. The most successful natural product development so far has been that of the pyrethrin insecticides, of which 33 are currently available.

The largest classes of pesticides are pyrethrins, organochlorines, organophosphates, and carbamates, although there are many smaller classes with only one or two members. The chemical structures of the key members of the major groups are given in **Table 1**.

Important Pesticide Groups

This list covers the important pesticide groups and some individual pesticides but does not attempt to be comprehensive.

Pyrethrins

Pyrethrins are chemically related to pyrethrin, which is a secondary metabolite found in the flowers of the pyrethrum plant (*Chrysanthemum cinerariaefolium*). Dried pyrethrum flowers were used as an insecticide in ancient China and in the middle ages in Persia. The dried flowers are still used. Current production is around 20 000 tons per annum centered in Kenya and Tanzania. The pyrethrins are effective insecticides, having very low dose rates and rapid knockdown of insects but being harmless to mammals under all normal conditions. Natural pyrethrins break down rapidly under the influence of oxygen and UV light. This limits their use in agriculture, but recently synthetic analogs have been developed to overcome these problems. Starting from the structure of the natural product a large number of synthetic compounds have been made. It is worth noting how they differ in effectiveness: deltamethrin is a broad range insecticide; allethrin is particularly toxic to house flies (*Musca domestica*) but much less effective with other insects; flumethrin is active against cattle ticks; while others are acaricides or miticides with little or no insecticidal activity.

Bacillus thuringiensis

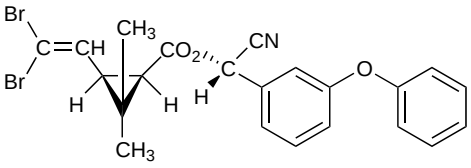
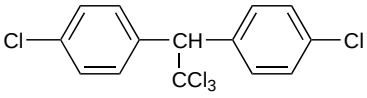
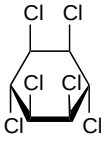
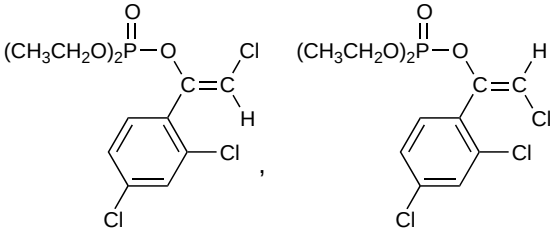
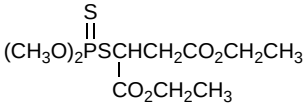
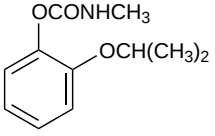
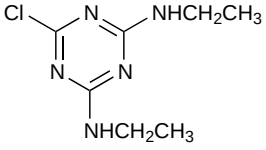
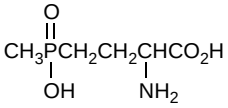
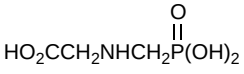
Bacillus thuringiensis is a widely distributed bacterium that during sporulation produces a crystal inclusion which is insecticidal when ingested by the larvae of a number of insect orders. Susceptible orders include Lepidoptera, Diptera, and Colcoptera. The action of *B. thuringiensis* was first observed in 1901 as the cause of a disease of silkworms. Several strains of the bacterium have been identified with activity against a range of insects including cabbage looper, tobacco budworm, mosquito, black fly, and more recently nematodes, ants and fruit flies. While the bacterium appears an ideal insecticide (having a toxicity 300 times greater than synthetic pyrethroids), it requires careful use. It is most effective against neonates and early larval instars so that spraying must be timed for egg hatch. It also has no contact activity and must be ingested so the plant must be well covered to ensure the insect receives a lethal dose. Furthermore it has a half-life in the field as short as 4 h, so careful timing is essential for it to be effective. Despite these limitations, it has been shown to be an important component of crop management programes.

One way of overcoming the problems of application of *B. thuringiensis* is to incorporate the gene responsible for expression of the protein into the crop plant. This has been achieved with maize (*Zea mays*) to protect against the European corn borer, with cotton (*Gossypium hirsutum*) to protect against a range of budworms and bollworms, and with potato (*Solanum tuberosum*) against Colorado beetle. (Cotton may seem irrelevant in a text on food but cottonseed oil is used extensively in cooking oils, margarines, and industrial fats.) This genetic modification has great benefits but care has to be taken that the food product has not changed in some unpredicted way. All genetically modified foods have to be extensively tested and cleared by regulatory agencies before release.

Neem oil

This is an oil obtained from the neem tree, *Azadirachta indica* A. Juss. It has been used as an insecticide in India and Africa but is increasingly being developed as a significant commercial product. It contains a number of compounds, one of the most active being azardirachtin, which is an insect antifeedant but also shows growth inhibitory and endocrine disrupting effects. This product and its individual components is at the beginning of its commercial development, which is likely to result in a series of products as significant as those from pyrethrum.

Table 1 Chemical structure and acceptable daily intake (ADI) of some pesticides

Compound	*Class*	*Structure*	*ADI (mg per kg body weight)*
Deltamethrin	Pyrethrin		0.01
DDT	Organochlorine		0.02
Lindane (HCH)	Organochlorine		0.008
Chlorfenvinphos	Organophosphate (mixture of two isomers)		0.002
Malathion	Organophosphate		0.02
Propoxur	Carbamate		0.02
Simazine	Triazine		0.005
Glufosinate			0.02
Glyphosate			0.3

Microbial phytotoxins

These are herbicides and include the highly commercially successful glufosinate, a synthetic form of phosphinothricin, first isolated from *Streptomyces hygroscopicus*, a soil-borne microbe. This compound is a potent, irreversible inhibitor of glutamine synthetase which is used in plants for photorespiration. Many attempts have been made to make synthetic variants of phosphinothricin

without success. Other members of this group include anisomycin and herboxidiene, derived from other *Streptomyces* strains. The veterinary insecticide, avermectin is derived from *Streptomyces avermitilis*.

Organochlorines

The organochlorines were the first group of synthetic insecticides and without them the dramatic decrease in malaria observed in the 1950s would have been impossible. The best known of this class is DDT (dichlorodiphenyltrichloroethane) but others include 2,4 DD, hexachlorbenzene, and lindane. Of these only lindane (γ-hexachlorocyclohexane, see **Table 1**) is still in use in the developed world.

These compounds are very slow to break down in the environment and one result of this persistance was the decline in bird numbers graphically described by Rachel Carson in the book *Silent Spring*. The problem was that DDT was concentrated through the food chain and predator birds in particular were failing to raise chicks. Since the organochlorine pesticides and other sources of organochlorines in the environment have been largely phased out, numbers of many species of birds are rising again. It is recognized that pesticides are still having an adverse influence on numbers of some birds that inhabit farmland. However, this is not a straightforward effect. In the case of the grey partridge, for example, it is because herbicides have reduced the number of weeds, which in turn has reduced the number of insects that feed on the weeds, resulting in fewer insects for the chicks to eat.

The mechanism of action of the organochlorines is not known in detail although they appear to act on the central nervous system. In humans the organochlorine compounds tend to accumulate in the body fat and in mothers' milk. While there is no direct evidence that they cause mutations or cancers, there is concern that lindane may be a carcinogen and its role in breast cancer is still under review. However, in contrast, DDT and γ-HCH have both been shown to inhibit tumors in mice initiated by aflatoxin B_1.

Although organochlorine pesticides have largely been phased out in Europe, analysis for them continues and low levels of lindane are still being detected in milk in the UK (typically at $0.005\,mg\,kg^{-1}$ compared with the maximum residue limit (see below) of $0.008\,mg\,kg^{-1}$ and an acceptable daily intake (see below) of 0.05 mg per kg body weight).

Organophosphorus compounds

Organophosphorus compounds generally contain both sulfur and phosphorus linked to carbon atoms. Their discovery was a by-product of the development of nerve gases. The group includes parathion, malathion, dimethoate, diazinon, and chlorfenvinphos. They are used as herbicides, insecticides, and fungicides. They break down quickly in the environment and do not concentrate in body fats, although they may be stored for some time. However, their mode of action – inhibition of acetylcholine esterase – means that they affect both insects and mammals and their use depends on the effective dose in the target species being below the sensitivity of other species.

Acute effects of sublethal doses of organophosphates in man include sweating, salivation, abdominal cramps, vomiting, muscular weakness, and breathing difficulties. Concern has also been expressed about long-term effects following acute exposure. Research suggests that some victims may show reductions in some neurobehavioral tests when tested some months after exposure. There are also concerns that people who do not appear to have suffered acute poisoning have subsequently developed debilitating illnesses. Symptoms include extreme exhaustion, mood changes, memory loss, depression, and severe muscle weakness.

Carbamates

Carbamates are derived from carbamic acid and are used against both insects and weeds. They are also acetylcholine esterase inhibitors. They are very reactive and are used up rapidly after application.

Methyl bromide

Methyl bromide was for many years the fumigant of choice for destroying insects in stored crops, but it is now being withdrawn as part of the general restriction on volatile organohalogen compounds because of their damaging effect on the ozone layer. It is being replaced by a number of less environmentally damaging compounds, including phosphine, although none currently available is as effective or as cheap as methyl bromide.

Phosphine

Phosphine has been used as a fumigant for many years. It is highly reactive and leaves no residues but great care has to be taken in its application because it is very toxic to humans.

Control of Pesticides

Control over pesticides is exercised in two ways: stringent testing on new pesticides before they are permitted and measurement of the residue in the crop.

Testing pesticides

There are a number of national and international bodies that approve new pesticides within their areas of responsibility. These include Codex Alimentarius, the European Union, and the US Food and Drug Administration (USFDA). Currently, within the European Union, registration of pesticides is being harmonized under Directive 91/414 EEC. Annexe 1 of this directive will identify all active ingredients permitted in pesticides. As yet this annexe is incomplete and member states are still acting under their national laws.

Within the UK pesticide registration is carried out under the Control of Pesticide Regulations 1986 and is the responsibility of the Ministry of Agriculture, Fisheries, and Food who are advised by the Advisory Committee on Pesticides.

In the USA a new Food Quality Protection Act of 1996 replaced both the Food, Drug, and Cosmetic Act and the Insecticide, Fungicide, and Rodenticide Act to provide a comprehensive regulatory scheme for pesticides.

In order to gain approval for use, pesticides are subjected to an extensive testing program including toxicity tests on mammals, plants, insects, fungi, birds, bees, fish, earthworms, and other soil organisms. The toxicity studies include effects of pesticides on fetuses and infant animals. There are also environmental tests which include laboratory tests on the breakdown and movement of the chemical in plants, soil, water, air, mammals, birds, and fish. These latter tests determine the rate of decay in the various species. Laboratory tests are followed by prolonged field trials to determine the fate of the chemical and its breakdown products in the environment and to estimate how the pesticide is concentrated up the food chain. On average it takes about 10 years to develop a new pesticide at a cost of about £50 million. The complete dossier of results has to be submitted to the approval body who determine whether the tests have been sufficiently rigorous to allow an acceptable daily intake (ADI) of the pesticide to be set. The ADI is defined as the amount of a pesticide that can be taken in each day throughout a person's life with the practical certainty, on the basis of all known facts, that no harm will result. This is determined on the basis of the highest level at which the pesticide has no observable effect in animal tests. This is then reduced by a factor of 10 in case humans are more sensitive than the animals used in the tests, and by a further factor of 10 to allow for cases where some humans may be more sensitive than others. In some cases, where the data show unusual effects, the safety factor can be increased from 100 to 500 or 1000. In practice the amount of pesticides to which the population is exposed is far below this level.

Table 1 includes the ADI for a number of the more common pesticides. There is no evidence that there are any cases where the combined effects of two pesticides are greater than the sum of their individual effects, in other words there is no evidence of synergy in toxicology between the different pesticides. Once maximum residue limits (MRL see below) for foodstuffs have been set on the basis of good agricultural practice, a total dietary intake is determined by considering all commodities in which the pesticide is likely to be used, and assuming the upper range of consumption, all foodstuffs at the MRL and no losses during transport, storage or food preparation. This figure is then compared with the ADI. For all permitted pesticides in the UK the figure is below the ADI.

Maximum residue limits

Maximum residue limits (MRLs) are statutory limits set on individual active ingredient and foodstuff combinations. They are based on residue levels which result when the pesticide is used according to the instructions on the label and in accordance with good agricultural practice (GAP). MRLs may be used to ensure that the pesticides are only being used in accordance with GAP. Many countries have codes of good operating practice with training for farmers and operators to ensure that pesticides are used at optimal levels. Some countries rely on the Codex Alimentarius Committee on Pesticide Residues to establish MRLs, while others set their own. (Codex Alimentarius is an international body which has over 120 countries as members and their standards are increasingly being accepted as the basis of world trade in foodstuffs.)

In the USA the FDA used to set tolerances for pesticide/foodstuff combinations but under the 1996 Act it sets a level for each pesticide in all foods based on the principle of a reasonable certainty of no harm. This is defined as a lifetime cancer risk of less than 1 in a million. There is also a requirement that residue tolerances must be specifically determined as being safe for children.

Within the EU, individual member states have historically set their own MRLs which differ from

state to state. Directive 76/895 established a common MRL setting regime and a series of subsequent directives has fixed the levels for a series of pesticides in fruit, vegetables, cereal products, and products of animal origin. There is an ongoing program to harmonize the levels throughout the Union.

Most industrialized countries have pesticide surveillance programs which cover both home-produced and imported commodities and these report annually. The EU has an annual specific coordinated program to check compliance in nominated combinations of pesticide and foodstuff. MRLs require sophisticated equipment for their determination because the levels are so low and the minimum detectable limit depends on the foodstuff. For example the tolerance for aldrin and dieldrin (two organochlorines) in the USA is between 0.05 and 0.1 mg kg^{-1} (parts per million), depending on the foodstuff. There are over 600 different active ingredients available commercially. Because there are so many, laboratories around the world have developed sophisticated rapid analytical techniques to allow them to screen pesticides by class so that retailers, food manufacturers, and governments can carry out analyses as a matter of routine.

The MAFF 7th Report of the UK Working Party on Pesticide Residues in 1996 showed 68% of samples had no detectable residue, 31% had residues below the MRL, and <1% were over the MRL. Similar results were obtained by the FDA who report results with relation to the tolerance to the pesticide/commodity combination. In 1995, of over 9000 samples analyzed, 64% had no detectable residues, 34% had residues below the tolerance, <1% had residues over the tolerance and <1% had residues for which there is no tolerance in that particular pesticide/commodity combination.

In all cases where MRLs or tolerances are exceeded follow-up action is taken. For home-produced materials, this involves investigation of the grower and prosecution if necessary. For imports, exceeding the level causes the consignment to be refused entry.

Maximum levels of pesticides are also set for drinking water. Pesticides get into water from spraying, runoff, percolation or from treatment of fish in aquaculture. Good practice is increasingly being developed to minimize the levels in raw water and treatment works are developing systems to reduce incoming levels to levels acceptable for drinking water.

Endocrine Disruption

The possibility that a number of chemicals discharged into the environment as a result of human activity may disrupt the endocrine system of a wide range of mammals has recently been given considerable prominence. Among the chemicals cited are the organochlorine pesticides, most of which have now been withdrawn for other reasons. While there is no doubt that there are a significant number of cases of endocrine disruption, the evidence to point to any particular chemical as a cause is lacking. It is also worth noting that deliberate endocrine disruption is a mechanism of a number of natural insecticides which act so as to inhibit development of juvenile larvae to adults. Fortunately these pesticides are reactive and usually have a short life in the field.

It is also true that there are very many naturally occurring endocrine disruptors, including the phytoestrogens present in vegetables, notably soya beans, peas, beans, cabbage, and hops. However, since this issue is very serious a considerable amount of work has now been initiated and its results will have implications for future testing of pesticides.

Future Prospects

In many parts of the world it is recognized that there has been too great a reliance on pesticide use and not enough on improving agricultural practices. There is increasing pressure to move towards minimizing pesticide usage in order to both improve the environment and to reduce cost. This is being done by using newer, more specific pesticides and by adopting improved agricultural practices and integrated pest management (a combination of biological and chemical control).

Biological control is not new. In the 1930s *Macrocentrus homonae* was introduced into Sri Lanka from Indonesia to control the tea small leaf roller (*Adoxophyes*) with such success that no chemical control measures are needed for this pest even today. More recently there have been some impressive results from using predator insects, for example in the control of cassava green mite (*Mononchellus tanajoa*) in West Africa and white fly in European greenhouses.

In terms of agricultural practice, improved crop hygiene, crop rotation, better understanding of optimal timing of application, and varying sowing dates, together with the development of more powerful and more discriminating pesticides has brought about a decrease in pesticide inputs. This is seen dramatically in the case of oil seed rape (canola). Less than 1% of the weight of herbicide applied to this crop in 1983 was applied in 1993.

Unfortunately pests develop resistance to individual pesticides over time and research is continually

needed to develop both new pesticides and resistant varieties of crops to keep the pests in check. There has been some success with new pesticides having new modes of action such as the antifeedants and antimolting agents, but this will be a continuing battle for the foreseeable future.

See also: **Phytochemicals**: Classification and Occurrence; Epidemiological Factors.

Bacterial Contamination

N Noah, London School of Hygiene and Tropical Medicine, London, UK

The burden of gastroenteritis (GE) in the world, in terms of both morbidity and mortality, is enormous. In the developing world (e.g., Southeast Asia), diarrhea vies with acute respiratory tract infection as the leading cause of death in childhood. Even in the more developed world, infectious GE is a significant cause of illness and time lost from work, and death does occur. Infectious intestinal disease in England is estimated to cost the country £743 million p.a. (in 1994–95 prices). GE caused by bacteria was far more costly than that caused by viruses. The more sophisticated surveillance systems become, the more GE they uncover.

Not all GE is caused by food. Probably most GE is caused by poor hygiene leading to direct or indirect transmission of infection without the assistance of food. Nevertheless, a major cause of infectious GE throughout the world is contaminated food. The definition of food poisoning (FP) is not straightforward. In essence, FP is an acute gastroenteritis caused by food. Hepatitis A, typhoid, and brucellosis, however, are not usually considered as FP, whereas botulism is, even though it causes paralysis and not GE, as is listeria, which causes septicemia and meningitis.

Bacteria are the most common known cause of FP and, with the possible exception of the Norwalk-like viruses, of GE also. As one would expect, bacterial FP is more common in summer than winter.

Bacteria produce their effects on the intestinal tract either by direct invasion of the mucosa or by the production of toxin. Some of the toxins are produced outside the intestinal tract—in the food; others are formed in the intestine. Some invasive bacteria also produce a toxin in the intestine. This article provides an overview of the bacterial causes of FP.

Bacterial Toxins

There are three main forms of bacterial enteric toxin:

Enterotoxin producing excess fluid secretion into the gut (cholera and some types of *Escherichia coli*)
Cytotoxin causing inflammation and mucosal damage (shigella and enterohemorrhagic *E. coli*)
Neurotoxin affecting the nervous system (botulism and staphylococcal toxin)

Some *E. coli* strains produce toxin; these are dealt with under Invasive Bacteria. Red kidney bean, scombrotoxin and other fish toxins, and heavy metal poisoning are dealt with elsewhere in this book.

Staphylococcal Food Poisoning

Background Staphylococcal food poisoning (SFP) is one of the few causes of bacterial FP that can commonly be attributed to a food handler. Humans frequently carry staphylococci either in an infected site or asymptomatically. Infected sites include wounds and abscesses, which may be the source of large numbers of staphylococci. Asymptomatic sites include throat, nostrils, fingernails, or hair. In general, only coagulase-positive staphylococci (*Staphylococcus aureus*), and only certain types, produce enterotoxin. Rarely, some coagulase-negative strains may occasionally produce toxin. Because the organism is also carried by many animals, outbreaks attributable to inadequate processing of a precontaminated food can occur also.

Growth and survival Staphylococci are killed by normal cooking temperatures. Any staphylococci that survive because of inadequate heat penetration or, more frequently, by postcooking contamination from a food handler will, if it is an enterotoxigenic strain and given the right conditions of warmth, moisture, pH, and time, produce toxin. Growth of staphylococci and production of toxin are optimum at approximately 20–37 °C, but growth can occur between 8 and 48 °C. This toxin is fairly heat stable; boiling for approximately 30 min is required to destroy it. Canning is usually, but not always, sufficient. The toxin is also resistant to radiation.

Many foods can cause SFP. Because the organism can grow in foods with high salt or sugar content (possibly because there is less competition from other organisms), ham is a common cause of SFP,

as are desserts, especially those containing cream. Other foods implicated include meats and other high-protein foods, salads with mayonnaise, canned mushrooms, cream, cheese, salami, and eggs (and probably any moist food). Cows with mastitis may occasionally infect milk.

Characteristic sequence of events A whole leg of ham is prepared for consumption and cooked. It is then sliced warm by a chef who has no skin lesions. The slices are overlapped on a tray. The tray is covered and left to cool for several hours before being refrigerated. Staphylococci from the nose of the chef are conveyed to the warm ham slices. Because of the large surface area, and the large bulk of covered overlapping slices of meat taking time to cool, staphylococci grow and toxin is formed. Refrigerating or reheating the meat will not destroy toxin. Similarly, ice cream may be contaminated during the preparation process by a food handler. If it is then left at room temperature, it will allow the staphylococci to grow and form toxin. Freezing the dish afterwards does not destroy toxin, and consumers are likely to contract SFP regardless of how long the ice cream is kept frozen.

Clinical features The toxin is an enterotoxin and a receptor in the gut appears to be necessary. There may also be a neurotoxic effect that acts on the vomiting center in the brain. With SFP, onset of symptoms is often dramatic. Vomiting is the most prominent symptom. It occurs between 2 and 4 h but may range from 30 min to 8 h after eating. Nausea, abdominal colic, and diarrhea are also common. Generally, as with most toxins, the higher the concentration (or the greater the amount ingested), the shorter the incubation period and the more severe the symptoms. Individual susceptibility is also a determining factor in severity. The illness is usually over in a day or two, but deaths have occurred, sometimes as a result of acute hypotension (another well-known but rare effect of the toxin).

Diagnosis Because many people carry staphylococci, it is important to ensure that the type causing an outbreak of FP is the same in the carrier and those affected; merely showing that a food handler carries staphylococci is insufficient. The organism can be grown from, or enterotoxin can be detected in, implicated foods, which usually contain $>10^6$ organisms/g. The organism can also be isolated from vomit or stool of patients and from the hands, nose, abscess, or infected wound of the food handler. Phage typing of strains, with detection and typing of enterotoxin, can also be performed. Enterotoxins A–I are recognized, although type A used to be the most common. With the advent of polymerase chain reaction, other types are being seen more frequently. As with all FP, the absence of laboratory-supporting evidence does not necessarily mean that the diagnosis is wrong or the implicated food innocent.

Bacillus cereus

Background *Bacillus cereus* is widely distributed in the environment and is not a contaminant of food. It is found in rice and other natural foods, such as herbs and spices, cream, and dry foods.

Growth and survival Unlike the staphylococci, *B. cereus* is a spore-forming organism that survives prolonged boiling. It causes two fairly distinct types of food poisoning, emetic and diarrheic. The diarrheal toxin is heat labile and, like *Clostridium perfringens*, formed in the gut. The foods commonly associated with it are 'proteinaceous' and, like *C. perfringens*, associated with meats, stews, desserts, and sauces. The emetic type is 'farinaceous,' associated mainly with cooked rice, and produces an illness similar to SFP. Different serotypes of *B. cereus* cause these two different forms of FP, and the toxins are different also. Other members of *Bacillus* sp. are discussed later. Some strains will grow at refrigeration temperatures in milk and other foods.

Clinical features and characteristic sequence of events The emetic type of *B. cereus* FP is caused by preformed toxin (cereulide) in food, usually rice that has cooled slowly. This usually happens when a large bulk of rice, as in Chinese restaurants, is allowed to cool at room temperature for many hours, often overnight. The center of the mass will stay warm for a long enough period for the spores of the bacillus to germinate and form toxin. The toxin is heat stable and will survive the quick frying given to it in a Chinese restaurant. The incubation period is usually short (1–6 h), and the symptoms, predominantly vomiting, tend to be milder than those for SFP, which it otherwise resembles.

The diarrheal form of *B. cereus* FP is similar to that caused by *C. perfringens*. The toxin, unlike the emetic type, is an enterotoxin formed in the intestine and is heat labile. The predominant symptoms are diarrhea and abdominal colic. The incubation period, as expected for an organism that multiplies in the intestine and then produces its toxin, is also longer (8–16 h). This type of *B. cereus* FP can be caused by a wide variety of foods, including meat, vegetables, and dairy products.

Diagnosis The mere presence of *B. cereus* in a food is insufficient because it is a normal contaminant of many natural foods. The diagnosis is confirmed by the finding of the organism in high concentrations [10^6–10^8/g, minimum 10^5] in cooked rice, or other food for the diarrheic type, and obtaining it from the stool or vomit of cases. Alternatively, the same serotype should be present in food and patient specimen. Detection of the toxin in the food may also be sufficient.

Clostridium botulinum

Background *Clostridium botulinum* is an anaerobic spore-forming bacterium widely distributed in soil and mud. The toxin is the most lethal substance known to man, with a LD_{50} of 0.00003 μg/kg body weight. In one incident, an adult was paralysed for more than 6 months after eating less than two teaspoonfuls of a rice salad. Tetanus toxin from *Clostridium tetani* and ricin from the castor bean (the next most toxic substances) have LD_{50} values of 0.0001 and 0.02 μg/kg, respectively. The seven toxin types, A–G, affect the nervous systems of vertebrates, animals, and birds more commonly than man. Birds in aquatic environments seem especially susceptible to mass die-offs caused by botulism. Invertebrates are not susceptible but can harbor the bacteria and toxins in their bodies. Types A, B, and E are the only toxins that affect man. Type E is acquired from fish. Type C is the main bird toxin, although types D and E are also important.

Growth and survival Because the organism grows anaerobically, special conditions have to be present for it to cause FP. Fortunately, although the organism is ubiquitous, FP caused by it is rare. First, the spores have to be present, which is not uncommon because they are widely, although patchily, distributed in soil and aquatic environments. Second, the spores have to survive cooking, which again is not difficult because they can survive heating at 100 °C for 2 h. Third, they have to be allowed to germinate and grow in anaerobic conditions. This accounts for the rarity of botulism, although accidents have occurred, and occasionally still occur, with home canning (once popular in the United States), and it occasionally occurs in preserved meat in Europe (the term botulism is derived from *botulus*, the Latin term for sausage) and in preserved rotting or fermenting food in the Arctic and areas of the Far East. Large bulks of certain types of food (e.g., canned hazelnut puree) may also be susceptible to growth and toxin formation. Commercial canning, except for the occasional accident, destroys spores by the heating processes used. The vegetative forms of *C. botulinum* are as susceptible as most other vegetative bacteria to heat, and the toxin is also destroyed by boiling—the human types A, B, and E in 2 min at 70 °C and all toxins for 5 min at 80 °C. The pH is also important: The lower the pH, the less resistant are the spores to heat, and a low pH (<4.5) affects the ability of the spores to form toxin. Hence, bottled pickled vegetables in vinegar tend to be safe. High concentrations of salt also affect the viability and toxin-forming properties of *C. botulinum*.

Clinical features The incubation period is 12–36 h (range, 6 h–10 days).The toxin destroys the cholinergic nerves in the motor end plates (MEPs). These are the junctions of the nerves with muscle, preventing the release of acetylcholine from the cholinergic nerves in the MEP and paralysing the muscle. Once this has happened, no amount of antitoxin or antibiotic is going to help. The combination of nausea, vomiting, or diarrhea followed by symmetrical descending paralysis of cranial and autonomic nerves is almost diagnostic. Thus, the characteristic neurological symptoms are blurred vision, dry mouth, difficulty swallowing, dysarthria, diplopia, and descending paralysis. Recovery has to wait until new MEPs form. There is a high fatality rate, but with modern technology patients can often be kept alive artificially until new nerve terminals have formed new MEPs, which may take several months.

Baby botulism

Some babies, usually younger than 6 months of age, acquire a form of botulism that is usually mild. It is thought to be caused by ingested spores multiplying in and colonizing the baby's intestine, forming toxin. The initial symptom is often constipation, leading to poor feeding, irritability, neck paralysis, and generalized weakness. Honey is thought to be one cause of baby botulism.

Diagnosis The diagnosis is made by the demonstration of botulinus toxin in food, stool, or serum. Growing the organism from food is suggestive but not diagnostic, whereas fecal isolates are rare except in affected individuals.

Clostridium perfringens

Background Food poisoning caused by *C. perfringens* is also toxin mediated. It differs, however, from those described previously in that toxin is formed in the intestine after ingestion of the bacteria. Like other clostridia, it is anaerobic, gram positive, and spore forming. There are five types, classified A–E

according to the enterotoxin formed; type A is the one that causes FP. Some strains, but not generally those that cause FP, can cause gas gangrene. *Clostridium perfringens* is primarily found in soil and is transmitted to animals and man by ingestion of vegetables and other plants. It is thus commonly found in the intestine of man and animals. When animals are eviscerated, the organism contaminates the inside of the carcass. Flies can transmit the organism to food.

Clostridium perfringens FP is common. Fortunately, it is rarely fatal except in those who are debilitated or immunocompromised.

Growth and survival *Clostridium perfringens* does not tend to multiply on the surface of raw meat. It grows optimally at warm temperatures of approximately 43–47 °C (range, 20–50 °C) and low oxygen levels found in the interior of a cooked dish. The cooking process will have driven off oxygen and thus facilitated sporulation and subsequent growth of the organism. Vegetative cells are not resistant to heat, but spores of the FP strains of *C. perfringens* can survive boiling for several hours. If cooling is slow, vegetative cells reform and grow rapidly. After ingestion, toxin is formed from multiplying cells in the intestine, although both toxin and vegetative cells appear to be necessary to produce symptoms.

Clinical features and characteristic sequence of events A casserole is prepared containing, among other ingredients, cubed meat. It is heated to boiling and allowed to cook for 1 or 2 h until ready. However, it is not needed immediately, and because of its bulk and lack of refrigeration facilities it is left overnight in a warm kitchen. It is warmed the next day before serving. Symptoms of diarrhea with abdominal pain begin 8–24 h later. The illness may last only a few hours, and there are no sequelae, except in those who are already debilitated.

Diagnosis The organism can be cultured from the stools of affected people and should be compared with that isolated from food for toxin production. Enterotoxin detection in stools is important confirmatory evidence. The organism has to be detected in high numbers in food to be significant. Molecular typing methods are available to compare isolates from food and feces.

Vibrio cholerae

Background Cholera appears to have originated in India. It first spread to Asia in 1817–1823, the first pandemic. The second pandemic reached Europe in 1826–1837, and subsequent to this there were five additional pandemics. The most recent began inexplicably in 1961 with a mild strain, the el Tor biotype, which had been endemic in Indonesia since 1937. More recently, it has become endemic in areas of South America. *Vibrio cholerae* 0139 is a new strain that emerged in the Indian subcontinent in 1992.

It is mainly to cholera that we owe the introduction of sanitation and the development of 'public health.' Although not a common cause of FP or GE in developed countries, the vibrios, especially *V. cholerae*, still cause large, mainly waterborne outbreaks in the developing world. It is the only gastrointestinal infection that is internationally notifiable. Because large numbers of organisms are required for infection, case-to-case transmission is uncommon.

Growth and survival The bacteria are aquatic and prefer briny waters. They can be found in many warm plankton-rich coastal waters, including the Mediterranean, Gulf of Mexico, and those of Southeast Asia and South America. Bivalved molluscs concentrate them, as they do many other bacteria. Other fish and shellfish can be contaminated, and inadequate cooking and storage will allow growth sufficient to cause FP. They prefer moist, slightly salty foods. Unfortunately, the el Tor strain is more likely to produce asymptomatic infections, persist longer in the environment, multiply more rapidly in food, and produce less immunity than the classical type. The organism produces an enterotoxin in the intestine.

Clinical features Cholera, in its most dramatic form, is characterized by an acute outpouring of watery diarrhea (rice water stools) and vomiting causing death within 24 h by acute loss of fluid and electrolytes. However, the clinical syndrome ranges from the symptomless to the mild and less dramatic forms. The organism is not invasive, and if the loss of fluid and salts can be counterbalanced by infusion of equal amounts of fluid supplemented by electrolytes, the patient will survive. Patients with an absence of acid in the stomach, and those with blood group O, are especially prone to severe disease. The incubation period is 1–3 days (range, 12 h–5 days).

Characteristic sequence of events Sewage-contaminated seafood or water is by far the most common source of infection. The vibrio can grow successfully in cooked rice and other grains contaminated by food handlers, and salad vegetables can be

contaminated by water. However, the organism often seems to find an 'environmental niche' and may persist in some communities for years.

Diagnosis The organism is usually isolated from the stool using special media. It can also be distinguished by light microscopy, and specific antisera will halt motility of the organisms. Agglutination tests with serum will distinguish 01 from 0139 and other strains. The organism can also be isolated from the environment using enriched media. Toxin production or the presence of the toxin gene can, and should, also be demonstrated. A 4-fold or greater rise in antibody is also helpful in diagnosis.

Invasive Bacteria

Salmonella Infections

Background Salmonellae are the most common known cause of bacterial FP in developed (and possibly less well-developed) countries of the world. Other bacteria such as campylobacter commonly cause GE but not necessarily FP.

There are a large number of serotypes of salmonella. They are typed according to their somatic [O] or flagellar [H phases 1 and 2] antigens according to the Kauffman–White scheme and are generally named after a geographical location. Further typing, usually by phages, can be done to distinguish the more common serotypes. Some serotypes are pathogenic for man, some for animals, others for birds including poultry, and some for humans as well as animals and birds. They are gram-negative bacilli that do not form spores but can survive for remarkably long periods on dried foods.

Salmonellae are widely distributed in and excreted by living creatures so that environmental contamination is inevitable. Protein foods processed in bulk for animals and poultry can cause widespread infection in them and subsequently in humans. In the United Kingdom, for example, fishmeal imported from Peru and fed to poultry caused a large outbreak of *Salmonella agona* infection in humans that lasted for several years through the late 1960s and early 1970s. Since then, there have been outbreaks of *S. hadar* infection in turkeys, and more recently various phage types of *S. enteritidis* infection in poultry and hens' eggs have caused outbreaks in many countries. In eggs, transmission is mainly 'vertical' (i.e., through oviducts). Before this, salmonellae gained entry to the insides of eggs mainly through the shell. If eggs were shelled in bulk, contamination of just one or two shells would be enough to contaminate the whole bulk of eggs and then grow given the right conditions. However, mayonnaise, paradoxically, is best kept at room rather than refrigerator temperatures if contaminated with salmonella. This is because acid, in the form of lemon juice or vinegar, kills salmonella more efficiently at warm than cool temperatures.

Other potent sources of contamination are sewage, polluted water, or direct fecal contamination of foodstuffs. Thus, many foods are bought already contaminated. Recent examples include mung beans, black pepper, dried herbs and spices, chocolate, spent yeast (used as a flavoring vehicle in packet potato crisps), infant dried milk, salamis, and sausages. Indeed, it is important to note that almost any food can be contaminated given the right circumstances. A multistate outbreak of salmonellosis in the United States was traced to tomatoes that had been washed in a contaminated water bath. An extensive outbreak in the United Kingdom was shown to be caused by lettuce. These episodes are particularly worrying because they show that any vegetable eaten raw may cause a salmonella infection. Cross-contamination from raw meat to relishes and dressings in a kitchen may also occur. Direct contamination of a food by a food handler, at any rate enough to cause an outbreak, is rarely documented. Indeed, infected food handlers are nearly always victims of the food they have prepared, not the source of contamination. Cases of human carriers with prolonged carriage occur.

Growth and survival Although salmonellae do not form spores, and are fairly easily destroyed by heat, they survive for a remarkably long period in the environment. An outbreak of *S. virchow* and *S. saint-paul* infection was caused in several countries in Europe by green lentils (mung beans) imported from Queensland. They were used to make bean sprouts, which required overnight growth in a warm waterbath. Slow drying of salmonellae makes them more resistant to dry heat. Moisture is important when using heat to kill bacteria.

Salmonellae grow best at 37 °C, and the danger temperatures are 30–45 °C. Growth stops below approximately 7 °C and above approximately 63 °C. Antibiotic-resistant strains are becoming more common.

The infective dose of salmonellae in humans is quite large—approximately 10^7 organisms. However, in certain circumstances, much lower doses may cause symptoms. Fatty foods such as chocolate, cheese, salami, and mayonnaise seem to require much smaller doses, and patients with immunosuppression, low acid levels in their stomach (achlorhydria), as well as the elderly and debilitated may also

be especially vulnerable. Thus, salmonellae may be transmitted nosocomially, especially in geriatric or psychogeriatric wards, and in such outbreaks it may be a waste of time to pursue a food source. Growing antibiotic resistance in salmonellae is proving to be a problem.

Characteristic sequence of events Various scenarios are described here:

A chicken dish is undercooked and then left in a warm environment for some hours before consumption. Alternatively, the chicken may be thoroughly cooked but is then replaced in an unwashed container or plate, or cut with a knife that was used for raw chicken, and allowed to stand. The contaminated utensil may be used on another dish, thus contaminating it.

A dried herb or spice, such as black pepper, is added to a dish after cooking but while still warm, and the dish is then left to stand.

Raw egg is added to a product without cooking, as with mayonnaise or mousse, or only lightly cooked. When the light cooking involves several hundred eggs, 1 contaminated egg and time in the warmth are enough to create many infective doses. In one outbreak, 800 eggs were used and left for a considerable period before being lightly cooked to make a hollandaise sauce. More than 100 guests at the wedding were affected. Subsequent examination of leftover unused eggs from the same batch suggested that the rate of contamination of eggs was very low, perhaps less than 1 per 1000. Low rates of contamination of a common food that is generally eaten after light cooking or raw can cause a large number of salmonella infections.

Clinical features The incubation period of salmonella infection is 12–36 h but can range from 6 to 72 h. Clinical features of salmonella infection range from asymptomatic to mild and enteric fever. Enteric fever (typhoid) is usually caused by *S. typhi* or *S. paratyphi A*. More commonly, salmonellae cause severe diarrhea with fever and abdominal pain. Vomiting is uncommon. Some salmonellae, such as *S. cholerae-suis*, may cause multiple abscesses, and people with sickle cell disease may have bone abscesses with any salmonella. Septicaemia (blood poisoning), meningitis, and other localised infections are also occasional complications of salmonellosis. Patients with AIDS and other immunosuppressive conditions are particularly vulnerable to severe complications.

Salmonellae and hens eggs In the late 1980s, *S. enteritidis* rapidly became the most common cause of salmonella infections in the United Kingdom. Previously, *S. typhimurium* had been by far the most frequently reported salmonella species. Between 1984 and 1987, the number of human *S. enteritidis* infections increased by approximately 50% per year. In 1988, the number more than doubled and by 1993 it was virtually 10 times that diagnosed in 1984. By 1993, *S. enteritidis* accounted for approximately five times the number of *S. typhimurium* infections. Most of this was due to eggs, although some of it was also attributable to chicken. Many European countries and countries elsewhere experienced similar trends.

Diagnosis The diagnosis of a salmonella infection is usually made by isolation of the organism from stool or food. Some salmonellae, such as *S. typhimurium* and *S. enteritidis*, are so common that further differentiation is necessary for epidemiological purposes. Further characterization of the organism can be undertaken by phage typing and antibiotic resistance profiles. Plasmid analysis may also be useful in differentiating strains of the same phage type.

Campylobacter Infections

Background The importance of campylobacter as a cause of GE was only recognized in the mid-1970s. They are now the most common known bacterial cause of GE in most developed countries. (In less developed countries, asymptomatic infection is more common.) *Campylobacter jejuni* is the most common species, but *C. coli* is common in some areas.

Campylobacter spp. are found in the intestines of many animals and birds, including cattle and horses, household pets, and chickens. Rates of contamination of chicken carcasses vary from >75% in the United Kingdom and The Netherlands to <30% in Sweden and Norway. Some of these differences may be due to the method of isolation used.

The reported incidence of human infection in Western Europe is high. In a survey of 15 countries, the annual incidence varied from 2.9 to 166.8 per 10^5 population, with a mean of 71 (1999 data). Because these are laboratory-confirmed infections, the true incidence will be considerably higher. The wide range of incidences almost certainly reflects rates of laboratory diagnosis and reporting rather than variation in incidence.

Growth and survival The reason for the late recognition of campylobacters is their fastidiousness: They grow best in an O_2 concentration of 5%, in a

special medium, and at a temperature of 42 °C. They are also sensitive to heat, being destroyed readily by cooking, and do not survive for long (probably a few hours only) on the surfaces of foods. They nevertheless are highly successful in causing infection, probably because of their ubiquity in the environment, domestic animals, and birds and the small dose required for infection (possibly no more than 200 organisms may be enough).

Characteristic sequence of events and clinical features Although campylobacters undoubtedly cause FP, the source of infection in most instances, especially sporadic cases, is unknown. It is highly probable that many cases, perhaps even most cases, are caused by direct contact with animals, birds, the environment (both domestic and outside), meat carcasses, and possibly other people. Food-borne outbreaks in the past have been traced to untreated water and milk and also milk from bottles whose tops have been pecked by birds. Undercooked poultry is undoubtedly a risk factor, and meat prepared at barbecues, which includes pork, veal, and beef as well as chicken, has also been implicated. In one study, consuming organic products, both meat and vegetables, and eating in a restaurant were risk factors. In another, eating grapes was found to be a risk factor, and salads have also been implicated, but it is possible that some of these foods were contaminated from another source or directly by a food handler.

Other risk factors include travel to foreign countries; handling and cooking of food, especially raw meat; contact with animals and pets (especially those with diarrhea) and visiting an animal farm; swimming; and sailing.

The incubation period of 3–5 days is long compared to that of most other FP bacteria. As with most gastrointestinal (GI) infections with a long incubation period, symptoms are mostly associated with the lower GI tract. Thus, vomiting is uncommon, and abdominal pain and diarrhea are the main symptoms. An accompanying fever is usual, and the diarrhea is often bloody. The illness may last a few days, and the antibiotic ciprofloxacin is now the treatment of choice for severe or prolonged illnesses.

Septicemia or other localized infections are rare complications. One of the well-known complications of campylobacter infection is Guillain–Barre syndrome, in which a symmetrical paralysis affects the body some weeks after the infection. Recovery is usually spontaneous but may take several months. In the acute phases of the illness, respiratory support may be needed.

Diagnosis The organism can be grown from stools, rectal swabs, and food. Special media and O_2 concentrations of 5–10% are needed for campylobacter. Two typing systems are available—Penner and Lior.

Escherichia coli

Background *Escherichia coli* are a remarkable group of organisms with a wide range of infections, including meningitis, septicaemia, and urinary infections. They are often also nonpathogenic. Those that cause GE also have a wide range of pathogenic mechanisms and are divided into various fairly distinct groups: enteropathogenic (EPEC), enterohemorrhagic (EHEC), enteroinvasive (EIEC), and enterotoxigenic (ETEC) are the main ones, although some groups—diffusely adherent (DAEC) and enteroaggregative (EAEC)—have recently been described. Some EHEC strains produce a shiga (or verocyto-) toxin, STEC, which includes *E. coli* O157:H7 as well as other strains. However, because the O157 strains are much more common, STEC strains are classified as O157 and non-O157. Shiga toxin is produced by other bacteria also, including *S. dysenteriae* type 1. Only these verocytotoxin-producing strains are considered in detail here because they are commonly foodborne and can cause serious illness and death.

Escherichia coli GE is not a notifiable infection, so there are few if any data on its impact on communities. Moreover, ETEC is not identified routinely by the stool culture methods commonly used. Infections with ETEC strains are common worldwide at all ages. These strains are the most common known cause of travellers' diarrhea but can also be spread by food. This toxigenic group includes strains that produce heat-labile and heat-stable enterotoxins. Heat-labile enterotoxin is closely related to cholera toxin and causes profuse watery diarrhea. EPEC strains are common infections in neonates and infants, tend to spread from person to person, and are not commonly known to be associated with food. EIEC and the two newer strains are rare. EIEC outbreaks related to food have been occasionally described, including one caused by French cheese exported to the United States.

In the United States, *E. coli* O157:H7 is estimated to cause 20 000 cases and 250 deaths annually, 67% of outbreaks are foodborne, 8% waterborne, and 22% transmitted case to case. Swimming in contaminated water can also transmit the infection. *Escherichia coli* O157:H7 was recognized as a cause of FP only in 1982. Some strains with the verocytotoxin (VT) gene produce a toxin that causes

the hemolytic–uraemic syndrome (HUS). Other *E. coli* strains also produce VT.

Escherichia coli, including *E. coli* O157:H7, is a normal inhabitant of the intestines of many mammals, including cattle, sheep, and goats. Contamination can occur directly from the intestines to carcass to meat or via the faeces of these animals to raw vegetables and other foods.

Growth and survival The organism can survive for a considerable length of time on contaminated meat and vegetables. In two outbreaks caused by apple juice, the organism was present on the surface of apples that had fallen to the ground. The orchard was frequented by deer that were subsequently found to be carrying the organism. Manure was also used. The apples were not processed adequately to kill most organisms, and waxing may have sealed the organisms onto the surface of the fruit. The cider was not pasteurized. The organism is more resistant than salmonella to acid: *E. coli* O157:H7 has been shown to survive for 21 days in cider at a pH of 3.7–3.9 at 4 °C, with only approximately a 5% kill-off. It can grow very successfully over several weeks in manure slurries. The infectious dose is thought to be small, so case-to-case infection may occur. Like most vegetative organisms, it is destroyed by heat.

Characteristic sequence of events In a town in North Cumbria, England, 61 patients had diarrhea, many with blood, over three weeks. A total of 114 people were found to be infected, ranging in age from 3 months to 85 years. Investigations implicated a farm supplying pasteurized milk. Nine days before the first case, a problem had occurred in the heat-exchanger plates of the pasteurization unit. No tests were undertaken after new plates were fitted, and temperature monitoring was inadequate. The unit was one that a few months before had been the subject of a food hazard warning. *Escherichia coli* O157 was isolated from 66 environmental and animal feces samples on the farm but not from the milk or the pasteurization plant.

In an outbreak in the United States, 501 patients became ill after eating inadequately processed hamburgers from a restaurant chain. HUS developed in 45 cases, and 3 died.

Undercooked hamburgers and ground beef are a common cause of *E. coli* O157:H7 infection. The process of grinding beef can spread the organism from the surface of the meat to the inside. Other vehicles of infection include raw milk, unchlorinated water, apple juice, unwashed fruits and vegetables including alfalfa sprouts and radish tops, or swimming in unchlorinated pools.

Clinical features The infectious dose for *E. coli* O157:H7 is thought to be fewer than 700 organisms. The incubation period of 3–5 days is long compared with that of most other FP bacteria. As with most GI infections with a long incubation, symptoms are mostly associated with the lower GI tract. Vomiting is uncommon, and abdominal pain and diarrhea, often bloody, are the main symptoms. Fever is usual. The illness may last a few days, and the antibiotic ciprofloxacin is now the treatment of choice for severe or prolonged illnesses. HUS is characterized by hemolytic anemia, thrombocytopenic purpura, renal failure, and a death rate of 3–5%.

Diagnosis The usual method of diagnosis is to isolate the organism from stools or food, which is straightforward. However, because most of the *E. coli* in the intestine is part of the normal flora and nonpathogenic, it is necessary to demonstrate virulence by further tests or assigning it to a serotype, which normally requires more sophisticated techniques in specialist laboratories. Serotyping is performed on the somatic cell wall antigens (O antigen) and the flagellar antigen (H). On the basis of the serotyping of the O antigens, the organisms can be classified as EPEC, ETEC, etc. DNA tests are increasingly being used. Thus, *E. coli* O111 is an EPEC strain, O115 with an H antigen is an ETEC strain, O115 without an H antigen is an EIEC strain, and O157:H7 is an EHEC strain. Toxins are now tested for using enzyme-linked immunosorbent assay or DNA probes. For EIEC strains, the conjunctival sac of a guinea pig is used. Serology tests are also used, but they are not reliable indicators of recent infection.

Other Organisms

Shigella

Shigella requires a very low dose to cause infection and does not grow very well in food. Thus, it is more commonly spread case to case, especially among kindergarten and primary school children. Affected patients may excrete the organisms for weeks. Nevertheless, some large and important outbreaks have been caused by food contaminated by sewage-polluted water or food handlers. In 1995, an extensive *S. sonnei* outbreak caused by lettuce imported from Spain affected people in many countries in northern Europe. In another outbreak caused by shrimp, infection was transmitted by a food handler who mixed the shrimp by hand with mayonnaise and tomato sauce. The incubation period is 24–48 h, and although bloody or mucoid diarrhea

is the usual symptom, it is characteristically accompanied by tenesmus—a feeling of wanting to defecate without being able to do so.

Listeria

Listeriosis, caused by the bacterium *Listeria monocytogenes*, is an unpleasant and rare infection that affects the more vulnerable, such as fetuses, infants, pregnant women, the elderly, and the immunocompromised. It causes septicaemia and meningitis, which is unusual for a FP organism. Fatality rates for invasive disease are high, as many as one in three. GI symptoms may be absent or mild. Also unusual, it can grow (albeit slowly) at normal refrigeration temperatures (0–4 °C). It is also very resistant in the environment, both to cold and to heat, so that it can survive for long periods. Normal pasteurization processes will inactivate it, but some organisms can survive the high-temperature/short-time process. The incubation period for invasive disease tends to be long (up to 3 weeks), but for GI symptoms very short periods of 1 day have been recorded. Dairy products, including soft cheese and butter, and pate are the most common sources of listerial FP, and a large and sustained outbreak in England and Wales between 1987 and 1989 was thought to be due to Belgian pate. The list of foods also includes hot dogs and other ready-to-eat delicatessen meats.

Yersinia enterocolitica

Like Listeria, *Y. enterocolitica* can reproduce at refrigeration temperatures. It is often missed in the laboratory because it requires special media and grows best at 25 °C. Many strains are nonpathogenic, but it is difficult to predict which strains are pathogenic and which not using current laboratory methods, although serotypes 3, 8, and 9 are most commonly associated with illness. The organism is found in many pets and other animals, and it may survive pasteurization, especially high-temperature/short-time milk treatment. Outbreaks have been associated with raw milk and dairy products. Undercooked pork and tofu have also caused outbreaks, and in one incident a caregiver who handled pork sausages passed the infection on to some infants. These infections appear to be more common in Scandinavia than elsewhere. The incubation period is 24–48 h to 5 days. Infants and young children are most often affected. Clinical features are characteristically fever and profuse watery diarrhea, but it may mimic acute appendicitis, risking an unnecessary operation. Occasionally, in vulnerable patients, septicaemia may occur.

Vibrio parahaemolyticus

Like *V. cholerae* (and several other aquatic organisms described later), *V. parahaemolyticus* is an aquatic organism that thrives in shallow coastal waters. Deep-sea fish do not tend to harbor the organism and usually become contaminated in fish markets. Precooked frozen shrimp may be contaminated and cause FP if served without further cooking, as in a seafood cocktail. *Vibrio parahaemolyticus* FP is associated with raw, undercooked, or contaminated seafood and is especially common in Japan and probably other countries in which seafood is a staple of the diet. Contamination from raw to cooked seafood is a common cause. The incidence of *V. parahaemolyticus* FP has increased in many Asian countries and the United States since 1996, and this is thought to be caused by a pandemic clone. Diarrhea, abdominal pain, and nausea are the predominant symptoms. The diarrhea can be severe, with blood or mucus in the stool. Vomiting is a less common feature, but fever can occur. The incubation period ranges from 4 h to 4 days, but most cases occur between 12 and 24 h. Death is uncommon.

The diagnosis is made by culture of the organism from feces or food. *Vibrio parahaemolyticus* can be easily isolated from most aquatic environments, but such strains are predominantly Kanagawa negative. Only the Kanagawa-positive strains (i.e., those producing a thermostable hemolysin that can be confirmed in a laboratory) cause GE, and it is thought that they multiply selectively in the human intestine. The infectious dose can be fairly small, 2×10^5 to 3×10^7.

Vibrio vulnificus

Like the other vibrios, infection with this organism is acquired from seafood. The epidemiology is similar but the organism appears to be the most common vibrio causing FP in the United States. It also causes septicaemia by ingress through a skin lesion in the food handler and in people with chronic liver disease through consumption of raw seafood. Gastroenteritis can also occur.

Aeromonas and *Plesiomonas shigelloides*

Aeromonas is another aquatic organism that prefers brackish and fresh water. It is generally accepted as a cause of FP, after initial doubts, in both adults and children. A profuse watery diarrhea is typical, although a dysentery-like syndrome is sometimes associated with it. The incubation period is 18–24 h. Sporadic infections are more common than full-blown outbreaks. Consumption of raw shellfish should be avoided (not only for aeromonas). *Plesiomonas shigelloides* is also an aquatic organism, and FP from it is rare. Its role in FP has not been fully

elucidated. Diarrhea is the usual symptom, and the incubation period is approximately 24 h.

Bacillus subtilis and *Bacillus licheniformis*

Bacillus subtilis is a member of the *Bacillus* genus and is similar to *B. cereus*, except that its natural habitat is slightly different, and so the foods causing illness also differ. It has been recognized increasingly as a cause of GE, characterized mainly by vomiting. The incubation period is short, 2 or 3 h, although many cases occur within 1 h. Foods implicated include meat and vegetable products such as meat pies, sausage rolls, curries with accompanying rice dishes, and even bread, crumpets, and pizza. The organism is present at high levels in implicated food (10^5–10^9 cfu/g) and can be isolated easily from both food and feces. Another member of this genus, *Bacillus licheniformis*, tends to cause diarrhea. Cooked meats and vegetables have been implicated. The median incubation period is approximately 8 h. Other members of this genus can also cause FP.

Bacillus anthracis

Rare in developed countries, anthrax FP is caused by the consumption of severely infected meat that has been insufficiently cooked. Bloody diarrhea is characteristic, accompanied by nausea, vomiting, and acute abdominal pain. The incubation period is from 2 days to many weeks.

Brucellosis

Although not usually considered as FP, brucellosis deserves mention because it is commonly caused by food. *Brucella melitensis* in particular is nearly always foodborne, and cheese, milk, and other dairy products made from unpasteurized milk are common vehicles. Occasionally, contaminated meats may be responsible. *Brucella abortus* is associated with cows and cow products, *B. melitensis* with goats, and *B. suis* with pigs. Brucellosis is a serious and prolonged systemic illness, with fever, night sweats, headache, aches and pains, and, sometimes, profound depression. Many developed countries have eradicated brucella from livestock.

Streptococcal Pharyngitis

Notwithstanding the definition of FP as causing GE, streptococcal sore throat with fever has been well documented to spread in this way. Usually, a food handler has a streptococcus group A in his or her throat, which may or may not be causing symptoms, and transfers this to a food that is then left in a warm environment for some time before consumption. Foods implicated include cheese, milk, eggs, and meat. The incubation period is 24–48 h. To confirm a source, typing of strains is important, as is sound epidemiological evidence, because many people carry these streptococci in their throats.

Prevention of Bacterial Food Poisoning

With the increasing trend toward the manufacturing of foods in large quantities for distribution not only nationally but also internationally, the potential for vast outbreaks of foodborne disease is considerable. Outbreaks of salmonellosis and *E. coli* FP caused by cheese, salami, chocolate, beef jerky, infant dried milk, minced beef, hamburgers, and even potato crisps have all been documented. In one outbreak of *E. coli* O157:H7 infection, 34 lots of 281 000 lb of beef patties were manufactured in one plant, and 7 of 21 lots tested were found to be contaminated. The introduction of HACCP (hazard analysis and critical control point) in food manufacturing processes has been a significant advance in the production of safer food and the prevention of FP. Microbiological guidelines now exist for ready-to-eat foods. The establishment of Enternet in countries in Europe and elsewhere is an important step in the early detection of such outbreaks and the curtailment of their effects. This is a voluntary surveillance system that shares information on organisms causing FP. For example, an outbreak of FP affecting several tourists from different countries after staying at a hotel on the Mediterranean coast will be detected more quickly and enable an early warning to be instituted. If a contaminated food is distributed through several countries, the country that first detected a problem can issue an early warning.

Fortunately, most outbreaks of FP are more localized, although large numbers of people may still be affected. The most common problems in the preparation of food are inadequate cooking, leaving prepared food too long at too high a temperature, and allowing cross-contamination from raw to cooked food.

All foods entering the kitchen should be considered to be potentially hazardous. In any investigation, it is important not to assume that a food cannot be the cause just because it is unlikely or not known to have caused FP in the past. Salads and other vegetables or fruit eaten raw may be contaminated, and outbreaks have been caused by lettuce (*S. sonnei* and *E. coli* VTEC 0157), raspberries and strawberries (*Cyclospora cayetanensis* and hepatitis A), alfalfa sprouts (*S. enteritidis*), and radish sprouts grown hydroponically (*E. coli* VTEC

0157). Some of these foods were contaminated at the source by water or sewage, others during processing by infected food handlers (NLV and hepatitis A), and others by food handlers during preparation. It is difficult to avoid or prevent such infections in the kitchen short of cooking everything, and more stringent codes for hygiene at the growing farms and processing plants are required.

In the kitchen, it is important to keep raw and ready-to-eat foods entirely separate. Salad and fruit fall into the ready-to-eat category. Raw meats especially should have their own utensils, surfaces, and cutting boards. For cooked food entirely different utensils, surfaces, and cutting boards should be used, unless washed thoroughly in very hot water and detergent or a dishwasher and then left to dry. Otherwise, FP organisms will transfer from raw to cooked food and grow.

Cooking food, especially meat, thoroughly will destroy vegetative organisms, including salmonellas, although spores will survive. If the cooling down period is too long—normally approximately 1 h is considered the limit before refrigeration is necessary—*C. perfringens* or *B. cereus* that have survived as spores will grow. So will salmonellas and many other FP bacteria if the food was inadequately cooked. Cooking will not normally destroy preformed toxins of *S. aureus*. Infected food handlers may occasionally cause outbreaks of FP, but with the exception of staphylococci, bacterial FP from this source is rarer than is generally thought. Salmonellae almost always originate in a food, and *B. cereus* and *C. perfringens* always arise from the food. Food handlers whose hands have been contaminated usually transmit shigella, hepatitis A, or norovirus.

Thus, moist food needs to be kept either hot (above approximately 60 °C) or cold (below 8 °C, preferably 4 °C). Cooling food, even freezing it, will not destroy organisms (except for some helminths). Undercooked chicken that has been refrigerated will still need thorough cooking before it is fit to eat. A large bulk of frozen meat or poultry will need to be thawed before cooking. Large frozen turkeys may need several days in a refrigerator to thaw fully. The inside of the meat is the last to thaw and the last to cook.

Grinding meat will disperse organisms through it. Thus, hamburgers and sausages need thorough cooking.

Drinking raw milk is hazardous: A large number of organisms, from tuberculosis to *Streptococcus zooepidemicus*, can be spread in this way. Hens' eggs caused many cases of salmonella FP (mainly *S. enteritidis*) in the 1980s and 1990s throughout much of Europe and the United States. The rate of contamination was low, probably not more than 2 or 3 per 1000 eggs, but the number of cases was large because of the popularity of eggs as a food and because it is normal to eat them less than fully cooked, not only on their own but also in other dishes such as sauces and mousse. Fortunately, screening and vaccination of flocks in recent years have made this a less hazardous source of salmonellosis. Irradiation of food is effective but is not popular with the public.

Education of food handlers may be straightforward, but, especially in countries in which food handlers have low status and pay, compliance is more difficult. Nevertheless, education of the general public has been slow but has progressed: Most people now realize the importance of thawing poultry and meat thoroughly before cooking, and the large outbreaks of salmonella FP that used to occur at Christmas time in England and Wales caused by inadequately defrosted and cooked turkeys are now very rare.

See also: **Dairy Products**. **Eggs**. **Fish**. **Food Safety**: Mycotoxins. **Meat, Poultry and Meat Products**.

Further Reading

Campylobacter Sentinel Surveillance Scheme Collaborators (2003) Point source outbreaks of campylobacter jejuni infection—Are they more common than we think and what might cause them? *Epidemiology and Infection* **130**: 367–375.

Centers for Disease Control and Prevention. Staphylococcal food poisoning. Available at www.cdc.gov/mmwrrhtml/00050415.htm

Eley AR (1996) *Microbial Food Poisoning*, 2nd edn. London: Chapman & Hall.

Evans MR, Salmon RL, Nehaul L *et al.* (1999) An outbreak of *Salmonella typhimurium* DT170 associated with kebab meat and yoghurt relish. *Epidemiology and Infection* **122**: 377–383.

Gilbert RJ and Humphrey TJ (1998) Foodborne bacterial gastroenteritis. In: Collier L, Balows A, and Sussman M (eds.) *Topley and Wilson's Microbiology and Microbial Infections*. London: Edward Arnold.

Hedberg CW, Angulo FJ, White KE *et al.* (1999) Outbreaks of salmonellosis associated with eating uncooked tomatoes: Implications for public health. *Epidemiology and Infection* **122**: 385–393.

Horby PW, O'Brien SJ, Adak GK *et al.* (2003) A national outbreak of multiresistant Salmonella enterica serovar typhimuriom DT 104 associated with consumption of lettuce. *Epidemiology and Infection* **130**: 169–178.

Mintz ED, Tauxe RV, and Levine MM (1998) The global resurgence of cholera. In: Noah ND and O'Mahony M (eds.) *Communicable Disease: Epidemiology and Control*. Chichester, UK: John Wiley.

Neimann J, Engberg J, Molbak K, and Wegener HC (2003) A case–control study of risk factors for sporadic campylobacter infections in Denmark. *Epidemiology and Infection* **130**: 353–366.

Noah ND (1992) Food poisoning. In: Truswell AS (ed.) *ABC of Nutrition*. London: British Medical Journal.

Osterholm MT and Norgan AP (2004) The role of irradiation in food safety. *New England Journal of Medicine* **350**: 1898–1901.

Perales I and Garcia MI (1990) The influence of pH and temperature on the behaviour of S. enteritidis phage type 4 in homemade mayonnaise. *Letters in Applied Microbiology* **10**: 19–22.

Roberts JA, Cumberland P, Sockett PN *et al.* (2002) The study of infectious intestinal disease in England: Socioeconomic impact. *Epidemiology and Infection* **129**: 1–11.

Takkinen J, Ammon A, Robstad O, Breuer T, and the Campylobacter Working Group (2003) European Survey of Campylobacter surveillance and diagnosis 2001. *Eurosurveillance* **8**: 207–213.

Tuttle J, Gomez T, Doyle MP *et al.* (1999) Lessons for a large outbreak of *E. coli* O157:H7infections: Insights into the infectious dose and method of widespread contamination of hamburger patties. *Epidemiology and Infection* **122**: 185–192.

Zaidi AKM, Awasthi S, and deSilva HJ (2004) Burden of infectious diseases in South Asia. *British Medical Journal* **328**: 811–815.

Other Contaminants

C K Winter, University of California at Davis, Davis, CA, USA

Food may be contaminated with many chemicals that pose the potential for toxicological consequences in humans consuming the contaminated food items. In addition to the presence of contaminants such as mycotoxins, pesticide residues, and heavy metals, food may contain numerous organic contaminants that enter the food supply from environmental sources or as a result of chemical reactions that occur during food processing. This article focuses on three types of food contaminants: dioxins (including dibenzofurans and polychlorinated biphenyls), acrylamide, and perchlorate. Each of these classes has been subject to considerable regulatory scrutiny, scientific study, and popular media coverage. It is likely that concerns regarding the presence of these contaminants in the food supply will continue throughout the next decade or longer, and that significant efforts will be made to reduce human exposure to these substances from food. This article discusses how these types of food contaminants enter the food supply, the types of food items in which they are most likely to occur, and the potential toxicological consequences resulting from exposure to these contaminants.

Dioxins

Dioxins are organic chemicals that comprise a family of ubiquitous environmental contaminants. Technically speaking, the dioxins of potential toxicological concern are polychlorinated dibenzo-*p*-dioxins (PCDDs). They are related, both structurally and toxicologically, to polychlorinated dibenzofurans (PCDFs) and polychlorinated biphenyls (PCBs). Structures of generic PCDDs, PCDFs, and PCBs are shown in **Figure 1**. Due to their structural and toxicological similarity and to avoid confusion, all three related groups of chemicals are considered to represent "dioxins" for the purposes of this article. Specific chemicals belonging to this family are referred to as congeners. Collectively, there are more than 200 dioxin-related congeners, and each possesses unique toxicological and chemical properties.

Occurrence in the Environment and in Food

PCDDs and PCDFs are primarily introduced into the environment as by-products of combustion processes. These by-products have been identified in the exhaust gases from sources such as cigarette smoke; industrial and municipal waste incinerators; power plants burning coal, oil, or wood; and automobiles. In addition to these human sources, PCDDs and PCDFs are also produced naturally by combustion in forest fires and from volcanic eruptions.

Historically, PCDDs and PCDFs have also been produced as impurities during organic chemical synthesis. The most notable and most toxic dioxin congener, 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD), has been shown to be produced in the synthesis of the herbicide 2,4,5-T, one of the herbicide components of Agent Orange, notoriously used in the Vietnam War. Although 2,4,5-T is now banned for use in the United States because of TCDD and other dioxin impurities, health concerns over the exposure of military veterans to Agent Orange and to TCDD continue to be raised. PCDDs and PCDFs can also be produced through the use of chlorine

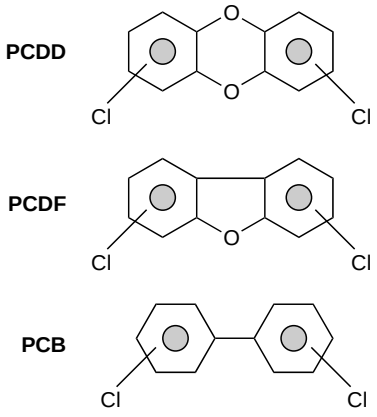


Figure 1 Chemical structures of generic PCDDs, PCDFs, and PCBs.

to bleach wood pulp, although most bleaching processes now use nonchlorine agents such as hydrogen peroxide.

PCBs have been produced synthetically since the 1930s and have been widely used for industrial applications, such as dielectric fluids in transformers (due to their inflammability) and capacitors in electrical machinery. Like their PCDD and PCDF counterparts, PCBs are extremely persistent in the environment and are of toxicological concern. As a result, the synthesis and industrial use of PCBs were significantly curtailed in the 1970s, although environmental residues of PCBs are still commonly detected today.

Although dioxin release into the environment has been known to occur for several decades, data are still limited with respect to the degree to which dioxins contaminate the food supply. Dioxin analysis in the laboratory is extremely expensive because methods must identify hundreds of different congeners, detection limits are required in the sub-part per trillion range, and significant precautions must be taken to minimize exposure of laboratory personnel to the analytical standards used for dioxin congeners.

Dioxins are highly fat soluble and have been shown to accumulate in the fat of birds, fish, and food animals. The US Environmental Protection Agency (EPA) has estimated that more than 95% of human exposure to dioxins results from dietary intake of animal fats. The major food sources for dioxin exposure include fish, poultry, meats, milk, and milk products. Dioxins are excreted in human breast milk and result in exposures to nursing infants.

Historically, it has been shown that human dioxin exposures, as determined by analyzing human tissues and environmental samples, have decreased significantly since 1987 due to engineering controls to limit dioxin emissions during combustion processes and to increased regulatory control over other sources of dioxin exposure. Dietary dioxin exposures to UK consumers were reduced by nearly two-thirds from 1982 to 1992, and subsequent studies showed even lower exposures in 1997. Nevertheless, dioxins are still ubiquitous in the environment and human exposure still occurs.

Toxicological Considerations

Dioxin exposure at significant dose levels has been linked to a large number of adverse health effects. Large acute exposures, resulting from chemical accidents and/or occupational exposure to dioxins, have caused a severe skin condition known as chloracne. A variety of other skin effects, such as rashes and discoloration, have also been attributed to acute dioxin exposures, as has liver damage.

Concerns from chronic exposure to dioxins include cancer, reproductive effects, and developmental effects. The most toxic dioxin congener, TCDD, was classified by the International Agency for Research on Cancer as a human carcinogen.

From a biochemical standpoint, PCDDs, PCDFs, and PCBs appear to cause their toxic effects through chemical binding to a specific cellular receptor known as the Ah receptor. Specific dioxin congeners vary dramatically with respect to their abilities to bind with the Ah receptor; TCDD binds extremely effectively, whereas other congeners are more limited in their binding capabilities. The degree to which various dioxin congeners bind with the Ah receptor seems to be directly related to the number and location of chlorine atoms on the congeners.

Assessing the potential human health risks from exposure to dioxins presents significant challenges. Dioxin levels in specific food items can be quite variable, and, as discussed previously, data concerning dioxin levels on foods are frequently not available.

Another difficulty encountered in assessing dioxin risks is to appropriately account for exposures to the various congeners and to account for the toxicological differences among congeners. This is most appropriately achieved through a toxic equivalency factor (TEF) approach that assigns a potency factor to each of the congeners relative to that of the most toxic dioxin TCDD. For example, the TEF for TCDD is 1 and the TEF for 1,2,3,4,7,8-hexachlorodibenzo-*p*-dioxin (with chlorines added to the 1 and 2 positions and otherwise similar to TCDD) is 0.1 based on findings that 1,2,3,4,7,8-hexachlorodibenzo-*p*-dioxin is 10 times less capable of binding to the Ah receptor than is TCDD. To calculate a total dioxin exposure, the dietary contributions of each of the dioxin congeners are multiplied by their corresponding TEFs and summed to determine a TCDD equivalent exposure.

According to the World Health Organization (WHO), a tolerable daily intake (TDI) for TCDD was established at 10 pg TCDD per kilogram bodyweight per day in 1990, although revisions by WHO reduced the TDI range to 1–4 pg/kg/day in 1999. A 1997 UK survey of dioxin consumer exposure provided an upper bound of 1.8 pg TCDD equivalent/kg/day. Surveys from other countries, using slightly different TEF approaches, yielded exposures of 0.7 pg/kg/day in Italy, 1.4 pg/kg/day in Norway, 2.4–3.5 pg/kg/day in Spain, and 0.2 pg/kg/day in New Zealand.

The US Food and Drug Administration (FDA) has been monitoring finfish, shellfish, and dairy products for dioxins since 1995 and initiated dioxin analysis of foods analyzed in its Total Diet Study in 1999. Specific findings from the FDA's annual Total Diet Study can be obtained from the FDA, although human exposure estimates, in terms of the amount of TCDD equivalent exposure per kilogram of body weight per day, have not been published by the FDA.

The EPA recommends that consumers follow the existing Federal Dietary Guidelines to reduce fat consumption and, subsequently, dioxin exposure. Such guidelines suggest that consumers choose fish, lean meat, poultry, and low- or fat-free dairy products while increasing consumption of fruits, vegetables, and grains. Dioxin exposure can be further minimized by trimming visible fat from meats, removing the skin of fish and poultry, reducing the amount of butter or lard used in cooking, and replacing cooking methods such as frying with methods such as boiling or oven broiling.

Acrylamide

Acrylamide is a widely used and versatile industrial chemical. Its most common use is as a coagulant in water treatment and purification. It is also used as a soil conditioner, in the sizing of paper and textiles, in ore processing, and as a construction aid for the building of tunnels and dam foundations.

Acrylamide is considered by the International Agency for Research on Cancer to be "probably carcinogenic to humans" based on the results of several animal carcinogenicity studies. As a result, there has been widespread concern about the potential risks from exposure to acrylamide among industrial, manufacturing, and laboratory workers. Consumer exposure to acrylamide in treated drinking water has posed a much lower concern since drinking water is subject to special treatment techniques that control the amount of acrylamide in drinking water.

Swedish researchers developed laboratory techniques that allowed for the detection of biological reaction products (hemoglobin adducts) of acrylamide in human blood samples; results from their studies allowed correlations to be made between occupational activities and acrylamide exposures. The findings that acrylamide occurred in tobacco smoke and that smokers had increased levels of hemoglobin adducts relative to nonsmokers provided a suggestion that acrylamide may be formed during incomplete combustion of organic matter or during heating. Interestingly, the researchers found significant levels of hemoglobin adducts in blood samples of nonsmoking humans not exposed occupationally to acrylamide. This led to speculation that the human diet could contain significant quantities of acrylamide. In April 2002, Swedish researchers published results of research that demonstrated the presence of acrylamide in several common foodstuffs, with the highest levels found in fried and baked foods. These findings stimulated worldwide interest in identifying the potential mechanisms for acrylamide formation in foods, in assaying a wide variety of foods for acrylamide levels, and in developing risk assessment and risk mitigation procedures.

Occurrence in Food

The findings from the initial Swedish study indicated that the highest levels (150–4000 μg/kg) of acrylamide were detected in carbohydrate-rich foods such as potatoes and in heated commercial potato products (potato chips) and crispbread. Moderate levels (5–50 μg/kg) were measured in protein-rich foods that were heated, whereas unheated or boiled foods showed no detectable acrylamide (<5 μg/kg).

The governments of several countries throughout the world performed similar analyses of acrylamide in foods and findings were fairly consistent with those reported in the Swedish study. The FDA analyzed dozens of foods for acrylamide levels and concluded that the highest levels were observed in french fries (29 samples; range, 117–1030 μg/kg) and in potato chips (40 samples; range, 117–2762 μg/kg). Multiple samples from different lots of the same commercial food products showed significant variability, with the highest levels often several times greater than the lowest levels. Commercial potato products that could be prepared by baking or by other methods showed much higher levels of acrylamide in the baked products. Acrylamide levels in baby food ranged from below the detection level (<10 μg/kg) to 130 μg/kg. All infant formula samples had levels below 10 μg/kg, and acrylamide levels in dairy products were also low.

The widespread findings of acrylamide in foodstuffs throughout the world provided the basis for numerous studies designed to elucidate the mechanisms for acrylamide formation in foods. It has been demonstrated that acrylamide can be formed from classical Maillard reactions as well as from reaction of the fatty acid oxidation product acrolein with ammonia and subsequent oxidation steps. The most plausible explanation for the relatively high acrylamide levels in fried potato products derives from a mechanism involving the reaction of the amino group of the amino acid asparagine with the

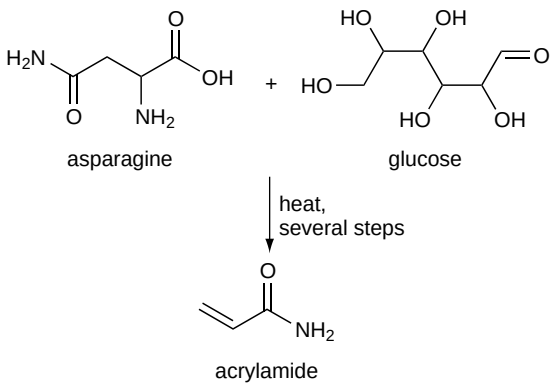


Figure 2 Proposed mechanism for acrylamide formation in foods.

carbonyl group of a reducing sugar such as glucose during baking and frying. This mechanism is shown in **Figure 2**. Potatoes are high in asparagine and in reducing sugars, and they are commonly prepared for consumption by frying or baking; all of these factors help explain the relatively high levels of acrylamide in heated potato products.

Toxicological Considerations

Laboratory toxicology studies have indicated that acrylamide is carcinogenic and also has been associated with the development of reproductive toxicity, genotoxicity, and neurotoxicity. Epidemiological and analytical studies of people exposed to acrylamide in the workplace have indicated that acrylamide does indeed enter the bloodstream of workers and can be detected and quantified as hemoglobin adducts, thus indicating both exposure and absorption of acrylamide. Such studies have not, however, indicated increases in cancer rates among those exposed occupationally to acrylamide. To date, the only documented toxicological effect observed in epidemiological studies of workers exposed to acrylamide is neurotoxicity. This effect is primarily an acute effect caused by large exposures to acrylamide for relatively short periods of time, leading to nervous system damage, weakness, and incoordination of limbs.

From a biochemical standpoint, it is likely that the health effects caused by high levels of exposure in humans and in laboratory animals may result from a Michael-type nucleophilic addition reaction of amino acids (both amino and sulfhydryl groups), peptides, and proteins to acrylamide because of the presence of the α,β-unsaturated conjugated structure in acrylamide. This is a common toxicological pathway for many reactive compounds. It is likely that high doses of acrylamide may overwhelm the defensive mechanisms of the body such as glutathione conjugation and may cause reaction with biologically significant nucleophiles, leading to mutations and possible carcinogenicity.

Although it is clear that humans have been consuming significant amounts of acrylamide in their diets for a long time, the relatively new discovery of acrylamide as a food contaminant has raised several questions. Significant efforts are currently being made to better understand the levels of acrylamide throughout the food chain and to estimate dietary exposure to acrylamide. In addition, there is much emphasis on developing food processing approaches that can reduce acrylamide formation.

Regulatory limits for acrylamide in food have yet to be established since dietary acrylamide risk assessments are still being developed. In the meantime, the FDA recommends that consumers eat a balanced diet that includes a wide variety of foods low in trans fat and saturated fat and rich in high-fiber grains, fruits, and vegetables.

Perchlorate

Perchlorate exists as an anion (ClO_4^-) with a central chlorine atom surrounded by four oxygen atoms arranged in a tetrahedron. Perchlorate is manufactured in the United States and is used as the primary ingredient of solid rocket propellant. Perchlorate wastes from the manufacture and/or improper disposal of perchlorate-containing chemicals are frequently detected in the soil and water. Levels of perchlorate have been detected in 58 California public water systems and in water samples from 18 states.

The widespread water contamination by perchlorate and its potential to cause health effects in those consuming contaminated drinking water have led four US agencies—the EPA, Department of Defense, Department of Energy, and National Aeronautics and Space Administration—to request that the US National Academy of Sciences convene a study on "Toxicological Assessment of Perchlorate Ingestion."

Occurrence in Food

Although the primary concerns from perchlorate contamination result from drinking water consumption, recent evidence has indicated that perchlorate may contaminate food items as well. A small survey of 22 lettuce samples purchased in northern California showed perchlorate contamination in 4 samples. A subsequent study of California lettuce showed detectable perchlorate levels in all 18 samples tested. The toxicological

significance of such findings has not been established, but the studies clearly indicate that perchlorate can enter lettuce, presumably from growing conditions in which perchlorate has contaminated water or soil.

Milk has also been shown to be subject to perchlorate contamination. A small survey of seven milk samples purchased in Lubbock, Texas, indicated that perchlorate was present in all of the samples at levels ranging from 1.12 to 6.30 μg/l. To put such findings in perspective, the State of California has adopted an action level of 4 μg/l for perchlorate in drinking water, whereas the EPA has yet to establish a specific drinking water limit.

Toxicological Considerations

Perchlorate is thought to exert its toxic effects at high doses by interfering with iodide uptake into the thyroid gland. This inhibition of iodide uptake can lead to reductions in the secretion of thyroid hormones that are responsible for the control of growth, development, and metabolism. Disruption of the pituitary–hypothalamic–thyroid axis by perchlorate may lead to serious effects, such as carcinogenicity, neurodevelopmental and developmental changes, reproductive toxicity, and immunotoxicity. Specific concerns relate to the exposures of infants, children, and pregnant women because the thyroid plays a major role in fetal and child development.

The ability of perchlorate to interfere with iodide uptake is due to its structural similarity with iodide. In recognition of this property, perchlorate has been used as a drug in the treatment of hyperthyroidism and for the diagnosis of thyroid or iodine metabolism disorders.

Ammonium perchlorate was found to be nongenotoxic in a number of tests, which is consistent with the fact that perchlorate is relatively inert under physiological conditions and is not metabolized to active metabolites in humans or in test animals.

Workers exposed to airborne levels of perchlorate absorbed between 0.004 and 167 mg perchlorate per day. These workers showed no evidence of thyroid abnormality, and a No Observed Adverse Effect Level was established at 34 mg absorbed perchlorate/day. Perchlorate does not accumulate in the human body, and 85–90% of perchlorate given to humans is excreted in the urine within 24 h.

See also: **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Fish**. **Food Intolerance**. **Food Safety**: Mycotoxins; Pesticides; Bacterial Contamination; Heavy Metals.

Further Reading

Becher G (1998) Dietary exposure and human body burden of dioxins and dioxin-like PCBs in Norway. *Organohalogen Compounds* **38**: 79–82.

Buckland SJ (1998) Concentrations of PCDDs, PCDFs and PCBs in New Zealand retain foods and assessment of dietary exposure. *Organohalogen Compounds* **38**: 71–74.

Environmental Protection Agency (2001) *Dioxin: Scientific Highlights from Draft Reassessment*. Washington, DC: US Environmental Protection Agency, Office of Research and Development.

Food and Drug Administration (2002) *Exploratory Data on Acrylamide in Foods*. Washington, DC: US Food and Drug Administration, Center for Food Safety and Applied Nutrition.

Friedman M (2003) Chemistry, biochemistry, and safety of acrylamide. A review. *Journal of Agricultural and Food Chemistry* **51**: 4504–4526.

Jimenez B (1996) Estimated intake of PCDDs, PCDFs and co-planar PCBs in individuals from Madrid (Spain) eating an average diet. *Chemosphere* **33**: 1465–1474.

Kirk AB, Smith EE, Tian K, Anderson TA, and Dasgupta PK (2003) Perchlorate in milk. *Environmental Science and Technology* **37**: 4979–4981.

Sharp R and Walker B (2003) *Rocket Science: Perchlorate and the Toxic Legacy of the Cold War* Washington, DC: Environmental Working Group.

Tareke E, Rydberg P, Karlsson P, Eriksson S, and Tornqvist M (2002) Analysis of acrylamide, a carcinogen formed in heated foodstuffs. *Journal of Agricultural and Food Chemistry* **50**: 4998–5006.

Urbansky ET (2002) Perchlorate as an environmental contaminant. *Environmental Science and Pollution Research* **9**: 187–192.

Zanotto E (1999) PCDD/Fs in Venetian foods and a quantitative assessment of dietary intake. *Organohalogen Compounds* **44**: 13–17.

Heavy Metals

G L Klein, University of Texas Medical Branch at Galveston, Galveston TX, USA

Food that we are culturally habituated to consume is usually thought to be safe. However, some foods are naturally contaminated with substances, the effects of which are unknown. Crops are sprayed with pesticides while they are being cultivated; some animals are injected with hormones while being raised. Meanwhile, other foods are mechanically processed in ways that could risk contamination. This article discusses food contamination with heavy metals, the heavy metals involved, their toxicities, and their sources in the environment. A brief consideration of medical management is also included. Five metals are considered in this category: lead, mercury, cadmium, nickel, and bismuth.

Lead

How Does Lead Contaminate Food?

Although lead is primarily known as an environmental contaminant that is ingested in paint chips by young children in urban slums or from contaminated soil or inhaled in the form of house dust or automobile exhaust, it may also enter the food and water supply. Ways in which this can occur include fuel exhaust emissions from automobiles that may contaminate crops and be retained by them, especially green leafy vegetables. Animals used for food may graze on contaminated crops and thus may also be a potential source of lead. Moreover, lead from soldered water pipes may contaminate tap water used for drink or for food production.

Permissible Intakes

In the United States, the maximum quantity of lead in the water supply that is permitted by the Environmental Protection Agency is 15 μg (0.07 μmol l^{-1}). The Food and Drug Administration (FDA) Advisory Panel recommends that no more than 100 μg (50 μmol) of lead per day should be ingested from food products.

Dietary Lead: Absorption and Consequences

People with certain macronutrient and micronutrient deficiencies are prone to experience increased absorption of lead in the diet. Thus, depletion of iron, calcium, and zinc may promote lead absorption through the gastrointestinal tract. Whereas adults may normally absorb approximately 15% of their lead intake, pregnant women and children may absorb up to 3.5 times that amount, and the explanation for this difference is not clear.

The effects of the entry of lead into the circulation depend on its concentration. Thus, the inhibition of an enzyme active in hemoglobin synthesis, δ-amino levulinic acid dehydratase (ALAD), occurs at blood lead concentrations of 5–10 μg dl^{-1} (0.25–0.5 μmol l^{-1}). Another enzyme active in heme biosynthesis, erythrocyte ferrochelatase, is inhibited at a blood lead level of 15 μg dl^{-1} (0.75 μmol l^{-1}). Reduction of the renal enzyme 25-hydroxyvitamin D-1-α hydroxylase, which converts circulating 25-hydroxyvitamin D to its biologically active steroid hormone, 1α,25-dihydroxyvitamin D (1,25$(OH)_2$D) or calcitriol, is observed at a blood lead concentration of 25 μg dl^{-1} (1.25 μmol l^{-1}). Behavioral changes and learning problems may begin to occur at blood levels previously thought to be normal, 10–15 μg dl^{-1} (0.5–0.75 μmol l^{-1}).

Manifestations of Lead Toxicity

Perhaps due to their increased absorption of lead from the diet, children appear to be more susceptible to the toxic effects of lead. These involve the nervous system, including cognitive dysfunction; the liver; the composition of circulating blood; kidney function; the vitamin D endocrine system and bone (**Table 1**); and gene function, possibly with resultant teratogenic effects. Chronic exposure results in high blood pressure, stroke, and end-stage kidney disease in adults.

Neurologic Full-blown lead encephalopathy, including delirium, truncal ataxia, hyperirritability, altered vision, lethargy, vomiting, and coma, is not common. Although peripheral nerve damage and paralysis may still be reported in adults, the most common toxicity observed is learning disability and an associated high-frequency hearing loss occurring in children with blood lead levels previously assumed to be safe. At low blood levels of lead (less than 10 μg dl^{-1}), children may lose IQ points, possibly due to the interference of lead in normal calcium signaling in neurons and possibly by blocking the recently reported learning-induced activation of calcium/phospholipid-dependent protein kinase C in the hippocampus.

The physicochemical basis of these changes derives largely from small animal data. Rats exposed to lead from birth develop mitochondrial dysfunction, neuronal swelling, and necrosis in both the cerebrum and the cerebellum. Exposure on day 10 of life elicited only the cerebellar pathology, and lead exposure after $3\frac{1}{2}$ weeks of life failed to produce any of these changes. In combination with manganese, lead has also produced peroxidative damage to rat brains and has been shown to inhibit nitric oxide synthase in the brains of mice. Additionally, an increase in blood arachidonic acid and in the ratio of arachidonic to linoleic acid following lead exposure in several species, including humans, may provide evidence in support of a peroxidative mechanism of damage to neural tissue following lead exposure.

Lead has also produced necrosis of retinal photoreceptor cells and swelling of the endothelial lining of retinal blood vessels in rats. Lead may also damage the auditory nerves in rats, and it may be partially responsible for the high-frequency hearing loss observed in humans. Finally, organic lead compounds may also disturb brain microtubular assembly.

Liver Although there are no outwardly recognizable manifestations of lead toxicity to the liver, studies in

Table 1 Heavy metal toxicities by tissues

Tissue	*Heavy metal*	*Dietary source(s)*	*Toxicity*
Neurologic	Lead	Green, leafy vegetables, canned food with lead solder, water	Learning disability, ataxia, encephalopathy, irritability
	Mercury	Seafood, agricultural crop contamination	Psychomotor retardation, paralysis, microcephaly, convulsions, choreoathetotic movements
	Bismuth	Medications	Paraesthesias, tremors, ataxia, reduced short-term memory
Bone	Lead	See above	Reduced conversion of vitamin D to active form, ?reduced osteoclast function
	Mercury	See above	?Reduced bone formation and bone density
	Cadmium	Seafood, plant roots in contaminated soil	?High bone turnover, secondary hyperparathyroidism
Bone marrow	Lead	See above	Decreased hemoglobin synthesis, decreased erythrocyte survival
	Mercury	See above	Increased hemolysis, alteration of T helper and T suppressor lymphocytes
	Cadmium	See above	?Reduced erythrocyte count
	Nickel	Vegetables, especially legumes, spinach and nuts	Decreased helper T cells and increased suppressor T cells
Gastrointestinal	Lead	See above	Decreased binding of L-tryptophan to hepatocellular nuclei
	Mercury	See above	Anorexia, fetal hepatic cell damage
	Cadmium	See above	Abdominal pain, vomiting, diarrhea
Renal	Lead	See above	Proximal tubular dysfunction: glycosuria, aminoaciduria, hyperphosphaturia, decreased renal conversion of 25-hydroxyvitamin D to 1,25-dihydroxyvitamin D, the biologically active form
	Mercury	See above	Renal tubular dysfunction, proteinuria, autoimmune damage
	Cadmium	See above	Proteinuria, glycosuria

rats indicate that amino acid binding to hepatocyte nuclei may be altered by lead. Thus, liver function may be subtly or subclinically affected and further studies are needed to elucidate this possibility.

Blood composition The major consequences of lead toxicity to the blood are microcytic anemia and decreased erythrocyte survival. The anemia is largely due to the inhibition of ALAD and erythrocyte ferrochelatase, which are critical to heme biosynthesis. Although the pathogenesis of the decreased red blood cell survival is not clear in humans, animal data indicate that the pentose phosphate shunt and glucose-6-phosphate dehydrogenase (G6PD) are inhibited by lead, suggesting that increased hemolysis may also contribute to the reduction in erythrocyte survival.

Kidney function Studies from the US National Institute of Occupational Safety and Health have reported that lead exposure reduced glutathione *S*-transferase expression in the kidneys of rabbits, indicating increased susceptibility to peroxidative damage. Renal proximal tubular dysfunction is described with lead intoxication and can result in glycosuria, aminoaciduria, and hyperphosphaturia as well as a reduced natriuretic response to volume expansion. This latter effect of lead exposure may possibly offer an explanation of how lead accumulation may contribute to hypertension.

Vitamin D endocrine system and bone As previously mentioned, lead can contribute to the reduced conversion of 25-hydroxyvitamin D to $1,25(OH)_2D$. The extent to which this action may contribute to vitamin D deficiency is not known, but there is at least the potential for lower circulating levels of $1,25(OH)_2D$ to play a role in reduced intestinal calcium absorption. This in turn may result in further lead absorption. Additionally, lead accumulating in bone has been reported to cause osteoclasts to develop pyknotic nuclei and manifest inclusion bodies, possibly lead, in the nucleus and cytoplasm. Although it has yet to be proven, these findings suggest a reduction in the resorptive function of osteoclasts. This may be a protective mechanism by the body to prevent the liberation of lead stored in bone, but at the same time lead may prevent the uptake by bone of additional calcium.

Genetic/teratogenic effects Lead has been reported to alter gene transcription by the reduction of DNA

binding to zinc finger proteins. This interruption of transcription has the potential to produce congenital anomalies in animals or humans. Studies have reported that lead crossing the placenta has produced urogenital, vertebral, and rectal malformations in the fetuses of rats, hamsters, and chick.

Management

Chelation therapy with dimercaprol succinic acid is recommended for anyone with a blood lead level higher than $25\,\mu g\,l^{-1}$ ($1.2\,\mu mol\,l^{-1}$), as shown in Table 2.

Mercury

How Does Mercury Contaminate Food?

The primary portal of mercury contamination of food is via its industrial release into water, either fresh or salt water, and its conversion to methyl mercury by methanogenic bacteria. As the marine life takes up the methyl mercury, it works its way into the food chain and is ultimately consumed by humans. This is the scenario that occurred following the release of inorganic mercury from an acetaldehyde plant into Minimata Bay in Japan in 1956 and 1965 and is responsible for the so-called 'Minimata disease.' Furthermore, acid rain has increased the amount of mercury available to be taken up by the tissues of edible sea life and can enhance the toxicity of certain fish. An unfortunate consequence of seafood contamination with methyl mercury is the contamination of fish meal used to feed poultry, resulting in mercury accumulation in the poultry as well as in the eggs. Additionally, mercury-containing pesticides can contaminate agricultural products. In Iraq in 1971 and 1972, wheat used in the baking of bread was contaminated with a fungicide that contained mercury.

Permissible Intakes

Limits of mercury intake set by the UN Food and Agriculture Organization (FAO) and the World Health Organization (WHO) are 0.3 mg per person per week, of which no more than 0.2 mg should be methyl mercury. Furthermore, FAO and WHO have set limits of mercury contamination of foods as not to exceed 50 parts per billion wet weight ($50\,\mu g\,l^{-1}$). Hair mercury content is used as a marker of methyl mercury burden.

Dietary Mercury: Absorption and Consequences

Although the precise mechanism of mercury absorption and transport has not been clarified, one possibility is its use of molecular mimicry. Studies of methyl mercury show that it binds to reduced sulfhydryl groups, including those in the amino acid cysteine and glutathione. Methyl mercury-1-cysteine is similar in conformation to the amino acid methionine and may be taken up by the methionine transport system in the intestine. Also, inasmuch as it has been shown that deep-frying of fish, with or without breading, will increase the mercury content, it has been postulated that mercury may be absorbed with the oil from the frying process.

A Swedish study reported a direct correlation between the amount of seafood consumed by pregnant mothers and the concentration of methyl mercury in their umbilical cord blood. Although fetal tissue mercury concentration is generally lower than the maternal concentration, the exception to this is liver. According to a Japanese study, mercury is stored in the fetal liver, bound to metallothionein. With development, the amount of metallothionein decreases and the mercury in liver is redistributed primarily to brain and kidney. In studies of offspring of animals exposed to mercury vapors, behavioral changes have been detected.

With regard to toxicity, mercury affects the skin, kidneys, nervous system, and marrow, with

Table 2 Recommended management of toxic symptoms caused by heavy metal contaminants in food

Element	*Agent*	*Comments*
Lead	Dimercaptosuccinic acid	Blood lead levels greater than $25\,\mu g$ ($1.2\,\mu mol$) l^{-1}; treatment of children with blood levels exceeding $10\,\mu g$ ($0.5\,\mu mol$) l^{-1} advocated due to learning problems
Mercury	Dimercaptosuccinic acid	Dimercaprol and D-penicillamine have also been used, but dimercaprol complicated by increased amount of mercury in brain
Cadmium	Diethyldithiocarbamate	Also used: dimercaprol, D-penicillamine, and dicalcium disodium EDTA
Nickel	Insufficient studies for recommended agent	Parenteral administration of diethyldithiocarbamide for acute toxicity may be helpful but unproven
Bismuth	Insufficient studies for recommended agent	Dimercaprol has been used anecdotally and reversed the symptoms of myoclonic encephalopathy; many choose to stop bismuth-containing drugs with a gradual resolution of symptoms

consequent effects on the blood cells, immune system, and bone formation.

Manifestations of Mercury Toxicity

Skin Mercury produces a symptom complex called acrodynia. Its main features are redness of the lips and pharynx, a strawberry tongue, tooth loss, skin desquamation, and pink or red fingertips, palms, and soles. The eyes are also affected, and photophobia and conjunctivitis are seen. In addition, there is enlargement of the cervical lymph nodes, loss of appetite, joint pain, and, occasionally, vascular thromboses, possibly by the induction of platelet aggregation, which has been shown in *in vitro* experiments. There is also a neurological component to this symptom complex: irritability, weakness of the proximal muscles, hypotonia, depressed reflexes, apathy, and withdrawal.

Kidneys Mercury has been hypothesized to stimulate T lymphocytes to produce a glomerular antibasement membrane antibody, which produces sufficient damage to lead to the proteinuria observed with mercury toxicity (**Table 1**). The basis for this theory derives from studies in rats in which mercuric chloride injection produced these antibodies, both as IgG and IgM. There was also an observed increase in $CD8^+$ (suppressor) T cells in the glomeruli. In addition, the rats developed proximal tubular necrosis. However, it is not clear that this theory is correct because methyl mercury can induce apoptosis, or programmed cell death, of the T lymphocytes, possibly by damaging mitochondria and inducing oxidative stress.

Nervous system In the large epidemics of methyl mercury ingestion reported in both Japan and Iraq, infants were reported to have psychomotor retardation, flaccid paralysis, microcephaly, ataxia, choreoathetotic motions of the hands, tonic seizures, and narrowing of the visual fields (**Table 1**). Studies of neonatal rats injected with methyl mercuric chloride reported postural and movement changes during the fourth week of life. These were associated with degeneration of cortical interneurons, which produce γ-aminobutyric acid (GABA) as a neurotransmitter. In the caudate nucleus and putamen, these GABAergic and somatostatin immunoreactive interneurons manifested the abnormalities. Pregnant rats given methyl mercury by intraperitoneal injection demonstrated rapid (within 2 h) effects on their fetuses, including mitochondrial degeneration of cerebral capillary endothelial cells, which led to hemorrhage. In turn, the bleeding disrupted normal neuronal migration.

In addition, methyl mercury may disrupt neuronal microtubular assembly and, perhaps by molecular mimicry (as described previously), may bind to the sulfhydryl groups of glutathione, causing peroxidative injury to the neurons. Following intracerebral injection in the rat, methyl mercuric chloride distributes in the Purkinje and Golgi cells of the cerebellum as well as in three different layers of cerebral cortical cells—III, IV, and VI.

Mercury exposure in humans can result in deficits in attention and concentration, especially under pressure of time deadlines. One report suggests that this may be due to mercury damage to the posterior cingulate cortex, where these functions are regulated.

Finally, *in vitro* studies of rat cerebellar granular cells suggested that incubation with methyl mercury caused an increased, although delayed, phosphorylation of certain proteins. The 12- to 24-h time course from mercury exposure to phosphorylation was believed to be consistent with the alteration of gene expression by mercury. Thus, the effects of mercury on the nervous system are multiple.

Bone marrow: Immune cells, blood cells, and bone formation A toxic effect of mercury on bone marrow would explain the abnormalities in red cell production, immune cell production, and bone formation (**Table 1**); all of the cells involved arise from stem cells found in the marrow and are presumably affected by mercury.

With regard to the immune cells, mercury induces an autoimmune response manifested by an increase in $CD4^+$ (helper) and $CD8^+$ (suppressor) T lymphocytes and in B lymphocytes in peripheral lymphoid tissue. This may explain in part the autoimmune nephropathy as well as the enlarged lymph nodes of acrodynia, previously described. Additionally, mercury may impair integrin signaling pathways in neutrophils, which may give rise to neutrophil dysfunction.

Hemolysis of red blood cells resulting from mercury exposure may be at least in part due to peroxidative damage inasmuch as studies on workers chronically exposed to mercury vapors demonstrate a reduction in erythrocyte enzyme activity of glutathione peroxidase and superoxide dismutase, as well as in G6PD.

Finally, although the effects of mercury exposure on bone have not been studied in humans, experiments in mice indicate that the administration of an anti-metallothionein antibody and mercury results in

decreased biochemical markers of bone formation and decreased bone mineral density. The mechanism for this is unknown, but mercury interference with differentiation of osteogenic precursor cells is postulated.

Genetic/teratogenic effects The uptake and redistribution of mercury by fetal hepatic tissue have been previously discussed. Abnormalities described with in utero exposure to mercury during the epidemics in Japan and Iraq have included low birth weight, malformation of the brain (both cerebrum and cerebellum), an abnormal migratory pattern of neurons, mental retardation, and failure to achieve developmental milestones. This remains a problem today for pregnant women who consume seafood. The FDA recommends that intake of large predator fish, such as swordfish and shark, be limited since they contain large amounts of mercury. Even tuna is considered to contain more mercury than most other seafood.

Management

Chelation with dimercaptosuccinic acid is recommended (**Table 2**).

Cadmium

How Does Cadmium Contaminate Food?

Cadmium enters the food chain in much the same way that lead and mercury do—by means of industrial contamination. Cadmium is often used as a covering of other metals or in the manufacture of batteries and semiconductors; it readily transforms into a gas as the metal ores are smelted. The cadmium then condenses to form cadmium oxide, which deposits in soil and water near the source. Cadmium accumulates in lower marine life, such as plankton, mollusks, and shellfish, and continues through the food chain as these organisms are consumed. However, contamination of the human food supply is limited by this route since cadmium is toxic to fish and fish embryos. In contrast to seafood, vegetables are affected differently because cadmium is taken up by the leaves and roots of plants, so those near industrial sources may be very high in cadmium.

Permissible Intakes

A 1991 study of adults consuming rice contaminated with cadmium in the Kakehashi River Basin of Ishikara, Japan, correlated cadmium intake with renal tubular dysfunction and established a maximum allowable intake of 110 μg per day. Canadian studies have estimated daily intake in study populations to be approximately half that, and the French have estimated cadmium exposure in the diet as being only 3 or 4 μg per day. The Provisional Tolerable Weekly Intake (PTWI) established by FAO/WHO is $7\,\mu g\,kg^{-1}$ body weight per week, a slightly more conservative estimate than the Japanese study but still in general agreement with it.

Dietary Cadmium: Absorption and Consequences

Fortunately, only 2–8% of dietary cadmium is absorbed and significant cadmium ingestion is accompanied by vomiting. Therefore, the gastrointestinal route is not as significant as inhalation of dust particles as a source of significant exposure.

Manifestations of Toxicity

Toxic manifestations of cadmium ingestion include renal dysfunction, osteoporosis and bone pain, abdominal pain, vomiting and diarrhea, anemia, and bone marrow involvement (**Table 1**).

Gastrointestinal toxicity The mechanisms for cadmium's effects on the gastrointestinal tract are not certain. Whether these toxicities stem from an irritative effect of the metal or whether there is cellular damage has not been resolved in animal or *in vitro* studies. One possibility is that *in vitro* studies of neural tissue suggest that cadmium blocks adrenergic and cholinergic synapses. Therefore, it is possible that cadmium interferes with autonomic nervous system influence on gastrointestinal motility.

Renal toxicity Renal tubular dysfunction is manifest in patients with itai itai disease as glycosuria and proteinuria, including excessive excretion of α- and β-microglobulin. Approximately 50–75% of cadmium accumulation in the body occurs in the liver and kidneys. Urinary cadmium excretion of $200\,\mu g$ $(1.78\,\mu mol)\,g^{-1}$ of renal cortical tissue has been associated with tubular dysfunction. In the kidney, cadmium is bound to metallothionein. When the amount of intracellular cadmium accumulation exceeds metallothionein binding capacity, this is the point at which renal toxicity is hypothesized to occur.

Bone marrow and bone In short-term accumulation of cadmium in the marrow, there is a proliferation of cells in the myeloid/monocyte category. However, with longer term burden, marrow hypoplasia is reported, including decreased production of

erythropoietin. Although a reduction in marrow cells may indicate that the osteogenic precursors in the marrow may also be reduced (**Table 1**), this is not borne out by studies both in humans and in rats. In these cases, biochemical markers of bone formation (osteocalcin) and resorption (deoxypyridinoline) are both increased, indicating a high turnover state. In rats, circulating parathyroid hormone levels are also elevated, suggesting that the high turnover is due to secondary hyperparathyroidism and subsequent inability of the bone matrix to mature and bind calcium and phosphate. Parenteral administration of 1,25-dihydroxyvitamin D has been reported to decrease circulating parathyroid hormone in the rat and to reduce bone turnover. Moreover, other animal studies report that cadmium interferes with hydroxyapatite nucleation and growth, thus making it difficult for bone matrix to bind to calcium.

Management

Chelation therapy is recommended using calcium, disodium ethylene diaminetetraacetic acid, dimercaprol, D-penicillamine, or diethyldithiocarbamate (**Table 2**).

Nickel and Bismuth

Dietary Contamination

Nickel and bismuth are not considered to be common dietary contaminants. Nickel is mainly inhaled as a dust by workers, whereas bismuth is mainly ingested in bismuth-containing medications such as Pepto-Bismol. Vegetables contain more nickel than other foods, and high levels of nickel can be found in legumes, spinach, lettuce, and nuts. Baking powder and cocoa powder may also contain excess nickel, possibly by leaching during the manufacturing process. Soft drinking water and acid-containing beverages can dissolve nickel from pipes and containers. Daily nickel ingestion can be as high as 1 mg (0.017 mmol) but averages between 200 and 300 μg (3.4 and 5.1 μmol).

Permissible Intakes

The maximum permissible intake of nickel is not known. Bismuth intake is related to whole blood bismuth levels. If these levels exceed $100\,\mu g\,l^{-1}$, bismuth-containing medication should be discontinued.

Toxicity

Nickel ingestion by women resulted in an increase in interleukin-5 levels 4 h after ingestion and a decrease in $CD4^+$ and an increase in $CD8^+$ lymphocytes 24 h following the nickel intake. Thus, alterations in the immune response may be associated with excessive nickel ingestion, consistent with reports of tumor production in animals and humans by inhalation of nickel-containing dust or powders. The mechanism for nickel-associated toxicity is purported to be oxidative. For bismuth, neurotoxicity, including irritability, numbness and tingling of the extremities, insomnia, poor concentration, impairment of short-term memory, tremors, dementia masquerading as Alzheimer's disease, and abnormal electroencephalograms, has been reported. Discontinuation of the bismuth may result in restoration of normal neurological function. Production of these symptoms in animals was associated with a brain bismuth concentration of $8\,\mu g\,g^{-1}$ brain tissue; a brain bismuth concentration of $4\,\mu g\,g^{-1}$ brain tissue was not associated with these neurotoxic manifestations. However, hydrocephalus was reported. At 1 μg bismuth g^{-1} brain tissue, no neurotoxic features were observed in animals. Nephropathy, osteoarthropathy, and thrombocytopenia have also been reported with bismuth toxicity.

Management

Insufficient controlled clinical trials have been performed to make clear-cut recommendations for pharmacotherapy for toxicity from either nickel or bismuth. Diethyl dithiocarbamide chelation therapy when promptly administered intravenously has been reported to be effective in acute nickel carbonyl poisoning. In addition, there have been anecdotal case reports of the reversal of myoclonic encephalopathy caused by bismuth with use of dimercaprol. However, no recommendations can be given at the present time.

See also: **Ascorbic Acid**: Physiology, Dietary Sources and Requirements; Deficiency States. **Food Safety**: Other Contaminants. **Vitamin D**: Physiology, Dietary Sources and Requirements.

Further Reading

Bierer DW (1990) Bismuth subsalicylate: History, chemistry and safety. *Reviews of Infectious Disease* **12**(supplement 1): S3–S8.

Bjomberg KA, Vahter M, Peterson-Grawe K *et al.* (2003) Methyl mercury and inorganic mercury in Swedish pregnant women and in cord blood: Influence of fish consumption. *Environmental Health Perspectives* **111**: 637–641.

Blumenthal NC, Cosma V, Skyler D *et al.* (1995) The effect of cadmium on the formation and properties of hydroxyapatite in vitro and its relation to cadmium toxicity in the skeletal system. *Calcified Tissue International* **56**: 316–322.

Burger J, Dixon C, Boring CS *et al.* (2003) Effect of deep frying fish on risk from mercury. *Journal of Toxicology and Environmental Health* **66**: 817–828.

Jin GB, Inoue S, Urano T *et al.* (2002) Induction of anti-metallothionein antibody and mercury treatment decreases bone mineral density in mice. *Toxicology and Applied Pharmacology* **115**: 98–110.

Knowles S, Donaldson WE, and Andrews JK (1998) Changes in fatty acid composition of lipids in birds, rodents, and preschool children exposed to lead. *Biological Trace Element Research* **61**: 113–125.

Kollmeier H, Seeman JW, Rothe G *et al.* (1990) Age, sex and region adjusted concentrations of chromium and nickel in lung tissue. *British Journal of Industrial Medicine* **47**: 682–687.

Kurata Y, Katsuta O, Hiratsuka H *et al.* (2001) Intravenous 1-α,25 $(OH)_2$ vitamin D_3 (calcitriol) pulse therapy for bone lesions in a murine model of chronic cadmium toxicosis. *International Journal of Experimental Pathology* **82**: 43–53.

Murata K, Weche P, Renzoni A *et al.* (1999) Delayed evoked potential in children exposed to methylmercury from seafood. *Neurotoxicology and Teratology* **21**: 343–348.

Needleman HL, Schell A, Bellinger D *et al.* (1990) The long-term effects of exposure to low doses of lead in childhood. An 11-year follow-up report. *New England Journal of Medicine* **322**: 83–88.

Report of the International Committee on Nickel Carcinogenesis in Man (1990) *Scandinavian Journal of Work and Environmental Health* **49**: 1–648.

Royce SC and Needleman HL (1990) *Agency for Toxic Substances and Disease Registry Case Studies in Environmental Medicine*, pp. 1–20. Atlanta: US Department of Health and Human Services, Public Health Service.

Simon JA and Hudes ES (1999) Relationship of ascorbic acid to blood lead levels. *Journal of the American Medical Association* **281**: 2289–2293.

Watanabe C, Yoshida K, Kasanume Y *et al.* (1999) In utero methylmercury exposure differentially affects the activities of selenoenzymes in the fetal mouse brain. *Environmental Research* **80**: 208–214.

Worth RG, Esper RM, Warra NS *et al.* (2001) Mercury inhibition of neutrophil activity: Evidence of aberrant cell signaling and incoherent cellular metabolism. *Scandinavian Journal of Immunology* **53**: 49–55.

Fortification *see* **Food Fortification**: Developed Countries; Developing Countries

FRUCTOSE

N L Keim, US Department of Agriculture, Davis, CA, USA
P J Havel, University of California at Davis, Davis, CA, USA

Published by Elsevier Ltd.

Fructose, a monosaccharide, is naturally present in fruits and is used in many food products as a sweetener. This article reviews the properties and sources of fructose in the food supply, the estimated intake of fructose in Western diets, the intestinal absorption of fructose, and the metabolism of fructose and its effect on lipid and glucose metabolism. The health implications of increased consumption of fructose are discussed, and inborn errors of fructose metabolism are described.

Properties and Sources of Fructose

Fructose has a fruity taste that is rated sweeter than sucrose. Sweetness ratings of fructose are between 130% and 180% (in part dependent on the serving temperature) compared to the standard, sucrose, rated at 100%. Both sucrose and fructose are used extensively in foods to provide sweetness, texture, and palatability. These sugars also contribute to the appearance, preservation, and energy content of the food product.

Natural sources of dietary fructose are fruits, fruit juices, and some vegetables. In these foods, fructose is found as the monosaccharide and also as a component of the disaccharide, sucrose (**Table 1**). However, the primary source of fructose in Western diets is in sugars added to baked goods, candies, soft drinks, and other beverages sweetened with sucrose and high-fructose corn syrup (HFCS). HFCS is produced by hydrolyzing the starch in corn to glucose using α-amylase and glucoamylase. This is followed by treatment with glucose isomerase to yield a mixture of glucose and fructose. The process typically produces a HFCS composed of 42% fructose, 50% glucose, and 8% other sugars (HFCS-42). By fractionation, a concentrated fructose syrup containing 90% fructose can be isolated (HFCS-90). HFCS-42 and HFCS-90 are blended to produce HFCS-55, which is 55% fructose, 41% glucose, and 4% other sugars. HFCS-55 is the preferred sweetener used by the soft drink industry, although HFCS-42 is also commonly used as a sweetener in many processed food products. Concentrated

Table 1 Sucrose, glucose, and fructose contents of fruits, vegetables, and sweeteners

Food item	*Serving size*	*Sucrose (g)*	*Glucose (g)*	*Fructose (g)*
Apple	1 medium	2.86	3.35	8.14
Apple juice	1 cup	4.22	6.20	13.89
Banana	1 medium	2.82	5.88	5.72
Blueberries	1 cup	0.16	7.08	7.21
Cantaloupe	1/8 melon	3.00	1.06	1.29
Cherries	1 cup	0.18	7.71	6.28
Grapes	1 cup	0.24	11.52	13.01
Oranges	1 medium	5.99	2.76	3.15
Peaches	1 medium	4.66	1.91	1.50
Pears	1 medium	1.29	4.58	10.34
Plums	1 medium	1.04	3.35	2.03
Pineapple	1 cup diced	8.48	2.70	3.18
Raspberries	1 cup	0.25	2.29	2.89
Strawberries	1 cup	0.20	3.39	4.15
Watermelon	1/16 melon	3.46	4.52	9.61
Avocado	1 fruit	0.10	0.14	0.14
Broccoli	1 cup	0.09	0.43	0.60
Carrots, baby	10 small	2.70	1.00	1.00
Corn, sweet	1 ear	1.85	0.45	0.43
Cucumber	1 cup	0.00	0.75	0.89
Onions	1 slice	0.44	0.74	0.44
Peas, green	1 cup	7.24	0.17	0.57
Potatoes	1 medium	0.36	0.70	0.58
Spinach	1 cup	0.02	0.03	0.04
Sweet potato	1 medium	3.28	1.25	0.91
Tomatoes	1 medium	0.00	1.54	1.69
Honey	1 Tbsp	0.19	7.51	8.60
Maple syrup	1 Tbsp	11.26	0.47	0.18
Molasses	1 Tbsp	5.88	2.38	2.56

Values obtained from US Department of Agriculture nutrient database accessed at www.nal.usda.gov/fnic/foodcomp.

fruit juices, such as apple juice and white grape juice, are also used to sweeten beverages. The amount of fructose in fruit juices varies, as does its proportion with glucose, but clearly the use of concentrated apple juice provides more fructose relative to glucose (**Table 1**), compared to either sucrose or HFCS-55. Nevertheless, considering the variety of sweeteners commonly available, it is likely that fructose constitutes approximately 50% of energy from added sweeteners.

As a result of the addition of sweeteners and sugars to so many food products, the consumption of fructose has increased from the mid-1970s to the mid-1990s. Sugars added to the diet are difficult to quantify accurately, but based on food intake survey data, total fructose consumption provides approximately 12% of adult energy intake or ~60 g/day based on a 2000-kcal diet. Individuals who are avid consumers of soft drinks, such as adolescent males, typically consume more than two times the average intake or more than 100 g/day of fructose from added sweeteners. Considering the US population as a whole, and based on both food disappearance data and food survey data, total fructose consumption has increased approximately 25% during the past three decades.

Absorption of Fructose

Dietary fructose is ingested as the simple monosaccharide and also as part of the disaccharide sucrose. Sucrose is hydrolyzed by sucrase at the intestinal brush border to yield one molecule of glucose and one of fructose. Glucose is rapidly absorbed via a sodium-coupled cotransporter and arrives at the liver via the portal circulation. Fructose absorption is accomplished primarily by a fructose-specific hexose transporter, GLUT-5. This transporter is found in the jejunum on both the brush border and the basolateral membranes. Expression of GLUT-5 increases within hours of exposure to a fructose-enriched diet, indicating that the transporter is regulated by luminal signals. However, consumption of a large amount of pure fructose can exceed the capacity of intestinal fructose absorption, resulting in diarrhea. Several studies have shown that when a single dose of 50 g of fructose is consumed by healthy adults, more than half experience malabsorption, and in some studies malabsorption is also observed with a 25-g dose. Fructose malabsorption results in abdominal bloating, flatulence, and diarrhea. However, the intestinal absorptive capacity for fructose increases when

glucose is consumed along with fructose. Thus, coingesting glucose to roughly balance fructose, as occurs when most fruits or sucrose is consumed, largely alleviates problems of fructose malabsorption. In addition, fructose absorption increases during sustained fructose consumption, suggesting adaptation to increased fructose intake.

Fructose Metabolism

The predominant site of fructose metabolism is the liver, where fructose enters the intermediary pathways of carbohydrate metabolism. Fructose is readily extracted by the liver because of the presence of an active hepatic enzyme system for metabolizing fructose, and the majority of ingested fructose is cleared in a single pass through the liver. Thus, the concentration of fructose circulating in blood is low after consumption of moderate amounts of fructose. Other tissues that take up small quantities of fructose include the kidney, skeletal muscle, and adipose tissue. The GLUT-5 transporter is expressed in these tissues but at relatively low levels.

In the liver, fructose is phosphorylated and forms fructose-1-phosphate. This reaction requires ATP and is catalyzed by fructokinase (EC 2.7.1.4), an enzyme with high affinity and specificity for fructose. Fructose-1-phosphate is then cleaved by hepatic aldolase (aldolase B; EC 4.1.2.13) to form dihydroxyacetone phosphate (DHAP) and glyceraldehyde. DHAP is an intermediate metabolite in both the gluconeogenic and glycolytic pathways. Thus, a portion of the original fructose carbon structure forms glucose, and, in fact, a small increase in circulating glucose occurs after ingestion of fructose. The glyceraldehyde intermediate is phosphorylated by triokinase (EC 2.7.1.28) to form glyceraldehyde-3-phosphate, another intermediate in the glycolytic pathway. The triose phosphate compounds provide substrate for glycolysis and oxidative metabolism, formation of glycogen, and synthesis of glucose and fatty acids. With the formation of the triose phosphates, the metabolism of fructose and glucose converges. However, prior to this step, there are important differences in fructose and glucose metabolism that impact both carbohydrate and lipid metabolism. The initial reaction that primes fructose for entry to the glycolytic pathway allows it to bypass the critical rate-limiting step of glycolysis. This critical step precedes the formation of triose phosphates; glucose carbons pass through an intermediate step in which fructose-6-phosphate is converted to fructose-1,6-bis-phosphate. This reaction is catalyzed by the allosterically regulated enzyme phosphofructokinase (PFK; EC 2.7.1.11) and is the most important control point in the glycolytic sequence. Among the multiple effectors of PFK are ATP and citrate; these products of glucose oxidation exert an inhibitory effect on the enzyme (**Figure 1**). The

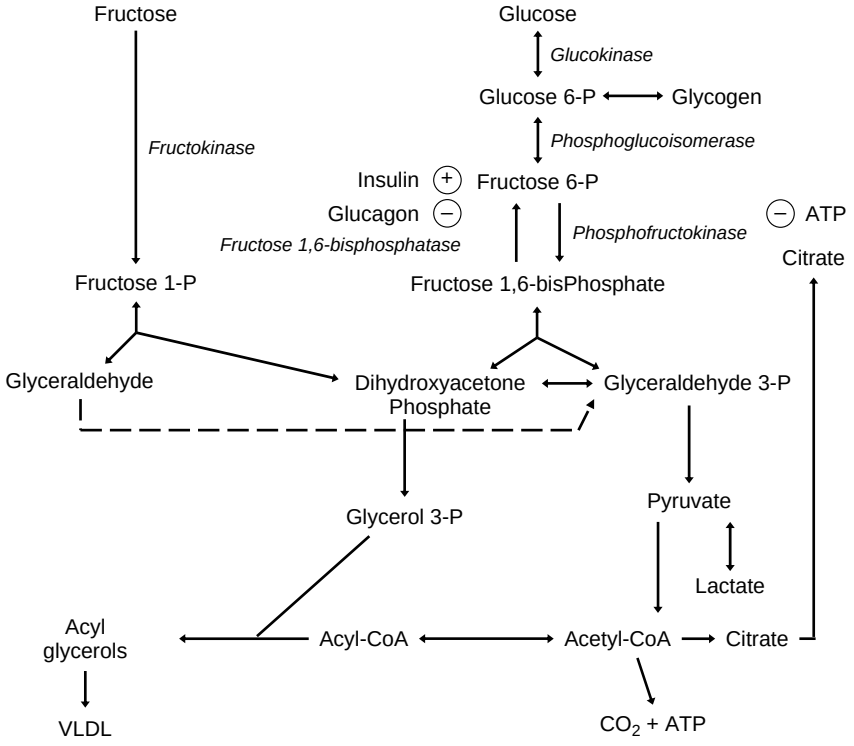


Figure 1 The intermediary pathways and fructose metabolism. Reproduced with permission by the American Journal of Clinical Nutrition. Copyright *Am J Clin Nutr.* American Society for Clinical Nutrition.

allosteric inhibition of PFK effectively reduces the rate of glycolysis and decreases hepatic glucose uptake overall. In contrast, the entry of fructose carbons through the pathway proceeds without this limitation.

Fructose and Lipid Metabolism

When large amounts of fructose are ingested, the glycolytic pathway becomes saturated with intermediates. In these circumstances, the intermediates become substrates for triacylglycerol synthesis: DHAP can be converted to glycerol, and acetyl-CoA can enter the lipogenic pathway to form fatty acids that are then esterified to the glycerol molecule to form triacylglycerols. During the initial step of lipogenesis, malonyl-CoA is formed. This intermediate serves to inhibit the transport of fatty acids into the mitochondria, where they are oxidized. By this regulatory mechanism, esterification of the newly synthesized fatty acids is reinforced. Studies have shown that the ingestion of fructose results in increased synthesis of fatty acids compared to ingestion of a comparable amount of glucose. The increased availability of fatty acids and subsequent triacylglycerol synthesis results in the production and secretion of triacylglycerols from the liver in the form of very low-density lipoproteins. Studies in animals have demonstrated that when large quantities of fructose or sucrose are consumed, an increase in blood triacylglycerol concentration occurs. Similar findings have been observed in humans, although some humans appear to be more susceptible to fructose consumption than others. For example, the lipogenic sequence may be accentuated in humans with preexisting hypertriacylglycerolemia or in those who are insulin resistant. Since high circulating triacylglycerol levels have been identified as a risk factor for coronary heart disease, long-term exposure to high levels of dietary fructose may contribute to a chronic, unfavorable lipid profile and increase risk of coronary heart disease.

Fructose and Glucose Metabolism

With fructose ingestion, there is an increased flux through the glycolytic pathway, with formation of pyruvate and lactate. As fructose-1-phosphate is formed, at the initial priming stage of glycolysis, it feeds forward and enhances the activation of pyruvate kinase (EC 2.7.1.40), thereby facilitating the passage of fructose carbon to pyruvate and lactate. With fructose ingestion, it is common to observe increases in blood lactate concentrations.

In the postprandial state, fructose serves to promote the formation of glycogen, but only when it is consumed along with glucose. This occurs through the activation of glycogen synthase (EC 2.4.1.11) and the inhibition of glycogen phosphorylase (EC 2.4.1.1).

In the starved state, fructose actively serves as a substrate for gluconeogenesis and glucose production. The gluconeogenic pathway and the glycolytic pathway share many common intermediates and enzymes, but the direction of the carbon flux through these pathways is controlled by several allosteric enzymes unique to each pathway. Since fructose enters the glycolytic pathway beyond the major gluconeogenic–glycolytic pivotal point (the interconversion between fructose-6-phosphate and fructose-1,6-bisphosphate), it does not exert an inhibitory effect on the gluconeogenic rate-limiting enzyme, fructose-1,6-bisphosphatase (EC 3.1.3.11). Consequently, there is no inhibition of gluconeogenesis by fructose as fructose carbons proceed through the glycolytic pathway. When a large quantity of fructose is infused intravenously, hepatic glucose production and output increase.

Consumption of large amounts of fructose is also associated with an impairment of glucose disposal. Prolonged feeding of fructose or sucrose to animals impairs insulin signaling and induces insulin resistance. Less is known about the effect of fructose ingestion on glucose tolerance and insulin resistance in humans because the scientific literature contains conflicting results. However, the lipogenic effects of fructose may contribute to insulin resistance indirectly since increased blood levels of triacylglycerols and fatty acids and deposition of lipid in liver and skeletal muscle have been implicated in the etiology of insulin resistance.

Fructose and Diabetes

Historically, in the nutritional management of diabetes mellitus, the ingestion of fructose was recommended as a sweetener for diabetics because it causes smaller increases in blood glucose following ingestion compared to similar amounts of glucose, sucrose, or starches. In fact, fructose, in small quantities, increases the hepatic uptake of glucose and promotes glycogen storage, probably by stimulating the activity of hepatic glucokinase (EC 2.7.1.2). Also, in individuals with type 2 diabetes mellitus, the addition of a small amount of fructose to an oral glucose tolerance test improves the glycemic response, indicating improved glycemic control. It must be emphasized, however, that the consumption of large quantities of fructose is not recommended, particularly for diabetics who, as a group, are at increased risk for cardiovascular disease, because of

potentially adverse effects of fructose on lipid metabolism, body weight regulation, and oxidative stress that may contribute to diabetic complications.

Fructose Consumption, Body Weight, and Obesity

With the increase in fructose intake, primarily as sugar-sweetened beverages, occurring coincidently with the increase in prevalence of overweight and obesity during the past two decades, it is important to examine the evidence that links fructose consumption and body weight gain. In epidemiological studies, consumption of larger amounts of soft drinks and sweetened beverages is associated with greater weight gain in women and increased energy intake and higher body mass index in children. In experimental studies, when fructose- or sucrose-sweetened beverages are added to the diet, subjects do not compensate for the additional energy provided by these beverages by reducing energy intake from other sources, and total energy intake increases. Possibly, this lack of compensation may be explained by the lack of a significant effect of fructose ingestion on the secretion of hormones involved in the long-term regulation of food intake.

Data comparing the effects of ingesting fructose- and glucose-sweetened beverages with meals indicate that fructose ingestion results in smaller increases in blood glucose and insulin concentrations following the meals. In addition, circulating leptin concentrations are lower, and the normal suppressive effect of meal consumption on ghrelin concentrations is attenuated with fructose beverages. Glucose, insulin, leptin, and ghrelin are all involved in the long-term control of food intake and body weight regulation through the central nervous system. Since these key signals are absent or weakened with fructose consumption, chronic consumption of a diet high in fructose could contribute, along with dietary fat and inactivity, to increased energy intake, weight gain, and obesity.

Inborn Errors of Fructose Metabolism

Several genetically based abnormalities in fructose metabolism have been described in humans. Fructokinase deficiency leads to high levels of fructose in the blood and urine. In the absence of fructokinase, fructose can be metabolized to fructose-6-phosphate by hexokinase (EC 2.7.1.1), although at a low rate. Consequently, no serious health problems are associated with this abnormality.

The aldolase A, B, and C enzymes catalyze the reversible conversion of fructose-1-diphosphate into glyceraldehyde-3-phosphate and dihydroxyacetone phosphate, and deficiencies in the A and B enzymes have been identified. Aldolase A is expressed in embryonic tissues and adult muscle. Possibly owing to the importance of this enzyme in fetal glycolysis, its deficiency results in mental retardation, short stature, hemolytic anemia, and abnormal facial appearance. There is no known treatment for aldolase A deficiency. Aldolase B is expressed in liver, kidney, and intestine, and a deficiency of this enzyme is more common and can be exhibited at first exposure to fructose during infancy or can have its onset in adulthood. Upon ingestion of fructose-containing foods, vomiting and failure to thrive are apparent. Hypoglycemia (in some cases severe), increased blood uric acid, and liver dysfunction also occur. Fortunately, this disorder can be treated effectively by completely eliminating fructose from the diet.

Deficiency of fructose-1,6-bisphosphatase is also considered a genetic disorder of fructose metabolism. This enzyme has a critical role in the enzyme complex regulating glycolysis and gluconeogenesis. Deficient individuals exhibit hypoglycemia, acidosis, ketonuria, hyperventilation, and often hypotonia and hepatomegaly. The urinary excretion of many organic acids is altered; notably, urinary glycerol is elevated and is useful in the diagnosis of this disease. The treatment includes avoidance of dietary fructose, sorbitol, and prolonged fasting.

D-Glyceric aciduria is caused by D-glycerate kinase (EC 2.7.1.31) deficiency. Only 10 cases have been documented, with symptoms ranging from none to metabolic acidosis, failure to thrive, psychomotor retardation, spastic tetraparesis, and seizures. The absence of significant symptoms in some suggests that this enzyme deficiency is essentially benign, and other associated enzyme deficiencies may underlie the more severe symptoms.

See also: **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Glucose**: Chemistry and Dietary Sources; Metabolism and Maintenance of Blood Glucose Level. **Inborn Errors of Metabolism**: Classification and Biochemical Aspects. **Obesity**: Definition, Etiology and Assessment.

Further Reading

Bray GA, Nielsen SJ, and Popkin BM (2004) Consumption of high-fructose corn syrup in beverages may play a role in the epidemic of obesity. *American Journal of Clinical Nutrition* **79**: 537–743.

Elliott SS, Keim NL, Stern JS, Teff K, and Havel PJ (2002) Fructose, weight gain, and the insulin resistance syndrome. *American Journal of Clinical Nutrition* **76**: 911–922.

Forbes AL and Bowman BA (eds.) (1993) Health effects of dietary fructose. *American Journal of Clinical Nutrition* **58**(supplement 5): 1S–823S.

Fried SK and Rao SP (2003) Sugars, hypertriacylglycerolmia, and cardiovascular disease. *American Journal of Clinical Nutrition* **78**: 873S–880S.

Havel PJ (2001) Peripheral signals conveying metabolic information to the brain: Short-term and long-term regulation of food intake and energy homeostasis. *Experimental Biology and Medicine* **26**: 963–977.

Havel PJ (2003) Regulation of energy homeostasis and insulin action by gastrointestinal and adipocyte hormones. In: Strasburger CJ (ed.) *Pituitary and Periphery: Communication in and out*, pp. 89–114. Bristol, UK: BioScientifica.

Havel PJ (2004) Update on adipocyte hormones: Regulation of energy balance and carbohydrate/lipid metabolism. *Diabetes* **53**(supplement 1): S143–S151.

Havel PJ (2005) Dietary fructose: Implication for dysregulation of energy homeostasis and lipid/carbohydrate metabolism. Nutrition reviews. (In press).

Joost HG and Thorens B (2001) The extended GLUT-family of sugar/polyol transport facilitators: Nomenclature, sequence characteristics, and potential function of its novel members (Review). *Molecular Membrane Biology* **18**: 247–256.

Keim NL, Levin RJ, and Havel PJ (2005) Carbohydrates. In: Shils ME, Ross AC, Shike M, Caballero B, Weinseir RL, and Cousins RJ (eds.) *Modern Nutrition in Health and Disease*, 10th editon. Baltimore: Lippincott, Williams & Wilkins, Inc. In press.

Kelley DE (2003) Sugars and starch in the nutritional management of diabetes mellitus. *American Journal of Clinical Nutrition* **78**: 858S–864S.

McGuinness OP and Cherrington AD (2003) Effects of fructose on hepatic glucose metabolism. *Current Opinion in Clinical Nutrition and Metabolic Care* **6**: 441–448.

Skoog SM and Bharucha AE (2004) Dietary fructose and gastrointestinal symptoms: A review. *American Journal of Gastroenterology* **99**: 2046–2050.

Teff KL, Elliott SS, Tschoep M, Kieffer TJ, Rader D, Heiman M, Townsend RR, Keim NL, D'Allesio D, and Havel PJ (2004) Dietary fructose reduces circulating insulin and leptin, attenuates postprandial suppression of ghrelin, and increases triglycerides in women. *Journal of Clinical Endocrinology & Metabolism* **89**: 2963–2972.

Wright EM, Martin MG, and Turk E (2003) Intestinal absorption in health and disease—Sugars. *Best Practice & Research Clinical Gastroenterology* **17**: 943–956.

FRUITS AND VEGETABLES

A E Bender, Leatherhead, UK

This article is a revision of the previous edition, article by A E Bender, pp. 902–906, © 1999, Elsevier Ltd.

Fruits and vegetables have considerable potential as a source of nutrients but the amounts eaten vary enormously both within and between countries. Some 3000 species are known to be edible and there are said to be more than 1500 species of wild tropical plants. In the foreword to *Traditional Plant Foods*, published by the Food and Agriculture Organization of the United Nations, it is stated that "rural Africa is rich in nutritious plant foods but in recent decades social and economic changes have militated against their propagation and use." This is because the promotion of major cereals has led to the eclipse of traditional plants. Furthermore, in developing regions many plant foods are regarded as being 'merely' children's food or poor man's food. Indeed, more fruits and vegetables are eaten in industrialized countries where there is an abundance of foods of all kinds than in developing countries, where any addition to the food supply is valuable.

For example, the average daily intake of fruit and vegetables in some underdeveloped regions is only 10–12 g per day compared with the recommendation in Western countries of five (and even up to nine) helpings (at least 375 g per day).

Definition

Plant foods are usually divided into seeds (including cereals), nuts, and a third combined group of fruits and vegetables.

Fruits are the fleshy seed-bearing parts of plants, while stems, roots, shoots, leaves, some seeds (including peas, beans, and lentils), tubers (underground storage organs such as potato, Jerusalem artichoke, sweet potato, yam), underground stems (taro, onion), and flower buds and flowers (cauliflower, broccoli) are all classed as vegetables.

However, through popular usage some fruits such as tomato and cucumber are classed as vegetables, and rhubarb, a stem, as fruit. So the group includes a large number of very diverse foods which differ considerably in nutritional value.

Macronutrients

Fruits and vegetables contain a very high proportion of water (e.g., up to 96% in cucumber) and so, with some exceptions, supply only small amounts of

macronutrients. The exceptions are those few eaten in sufficiently large amounts as to constitute a staple food, such as potato, plantain, cassava, and taro (colocasia, also known as eddo, dasheen, and old cocoyam). These are also lower in water content than many other members of the group and make a significant contribution to the carbohydrate and so the energy intake, and a small contribution of protein. For example, a 1000 kcal (4200 kJ) portion of plantain (*Musa* spp.) supplies 10 g protein. Cassava (manioc, *Manihot utilissima*) is extremely hardy and prolific and is a valuable staple in some communities, although a poor source of protein (1%).

Even in Western countries the potato (*Solanum tuberosum*) makes a contribution to the carbohydrate and protein of the diet, as well as supplying significant amounts of thiamin and ascorbic acid, because several hundred grams are often eaten per day.

One fruit, avocado (*Persea americana*), is a source of fat—17–27%—two-thirds of which is monounsaturated. Olives (*Olea europea*) contain 10–12% fat, two-thirds of which is monounsaturated, and of course are a commercial source of oil. Leafy vegetables are very watery (80–90%) and make an insignificant contribution to the intake of macronutrients. However, such proteins that are supplied are relatively rich in the amino acid lysine and to some extent, depending of course on the amount eaten, complement the relative deficiency of lysine in cereals (see Leaf Protein below).

The legumes—peas, beans, and lentils—play a special role in the diet since, as eaten, they contain more dry matter than most fruits and vegetables, around 30%. Unlike fresh fruits and vegetables they are usually stored for long periods in the dry state. After rehydration and cooking most of them including 'baked beans' (mature haricot beans, *Phaseolus vulgaris*) supply 5–8 g protein, 10–15 g carbohydrate, 300–400 kJ (80–100 kcal), and 4–5 g dietary fiber per 100 g. Legumes are sometimes described as rich sources of protein, but this is based on the dried product, not as eaten. When rehydrated and cooked their protein content is less than that of cereal products.

Garden peas (*Pisum sativum*) are commonly cooked from the fresh or frozen state, i.e., wet, but the macronutrient content is similar to that quoted above.

Green beans/French beans (pods and seeds of *Phaseolus vulgaris*) contain only 10% dry matter and 2 g protein, 100 kJ (25 kcal), and 2 g dietary fiber per 100 g.

Sprouted beans, commonly mung beans (*Vigna radiata*) but also alfalfa (lucerne, *Medicago sativa*) and adzuki beans (*Vigna angularis*), are also low in dry matter content and supply macronutrients in amounts similar to those in French beans.

Micronutrients

The major contribution of fruits and vegetables to the diet lies in their content of vitamins and minerals, but there are enormous variations between different types and between varieties of the same type.

Vitamins

The vitamin C content of different types of fruits ranges between some thousands of milligrams per 100 g in the instance of the West Indian cherry (acerola/Barbados cherry, *Malpighia punificolia*) to a few milligrams in apples (species of *Malus sylvestris*) and pears (varieties of *Pyrus communis*). The yellow–orange colored fruits (e.g., apricots, *Prunus armeniaca*, pawpaw, *Carica papaya*) supply carotene, as also do all green vegetables, while it is absent from white types. Leafy vegetables are rich sources of vitamin K and many fruits and vegetables are significant sources of folate.

Growing conditions including soil, fertilizer (type, amount, and time of application) and the state of maturity influence the vitamin content, particularly of vitamin C. In some foods, such as citrus fruits and tomato, the vitamin C is influenced by exposure to sunshine. One further cause of variation in vitamin content, again applying particularly to vitamin C and to a lesser extent to folate, is the loss after cropping, particularly from leaves that are bruised or wilted. Consequently it is not possible, apart from very broad generalizations, to state the vitamin content of a particular fruit or vegetable. Furthermore, new varieties have been and are being developed that are particularly rich in some vitamins.

Comparisons between food composition tables from various national authorities are unrealistic. Thus among six such tables (Australia, Germany, Great Britain, Spain, Italy, and the USA) the range of carotene in pumpkins is quoted from 0.24 to 19 mg per 100 g; for thiamin between 0.15 and 0.5 mg per 100 g; and for vitamin C between 10 and 50 mg per 100 g. For tomato the carotene ranges between 0.8 and 4 mg per 100 g. Even within one set of tables, in which the same sampling and analytical techniques are presumably used, there is a range of carotene in sweet potatoes (*Ipomea batatas*) from 0.3 to 4.6 mg per 100 g; and for lettuce (*Lactuca sativa*) from 0.16 to 1.6 mg per 100 g.

Mineral Salts

These are stable compared with some of the vitamins but the amounts can vary with different growing conditions, though to a lesser extent than described above for vitamins. Chemical analysis can be misleading since part of the mineral may be present

in the food in a bound, unavailable form which is not liberated during digestion. In addition there may be other substances present in the same food or eaten at the same meal that interfere with absorption. Substances such as oxalates, phytates, tannins, and dietary fiber reduce the amount absorbed.

Generally, vegetables are good sources of potassium with a very high potassium/sodium ratio, but rather poor sources of iron of low availability, although the amount absorbed is a balance between enhancement by the vitamin C present and various factors that reduce absorption. Vegetables are also a minor source of iodine depending on the content in the soil water.

Both fruits and vegetables are often described as sources of calcium but they are minor sources, i.e., many supply around 50 mg per 100 g compared with milk at 120 mg (which is consumed in much larger quantities) and cheese at several hundred mg per 100 g. Some common fruits and vegetables do not contain any calcium and the richer sources such as parsley at 330 mg and watercress at 220 mg per 10 g are eaten in relatively small amounts. Spinach stands out with 600 mg calcium per 100 g, but this is partly unavailable.

Dietary Fiber

All fruits and vegetables are sources of dietary fiber—more precisely nonstarch polysaccharides—but in different amounts. Thus, taken from the same food composition tables, they vary in fruits from around 0.5 g per 100 g in grapes, lychees, melons, and cherries, through 1.5 g per 100 g in oranges, peaches, pineapple, plums, rhubarb, and apples, 2.5 g per 100 g in mangoes, olives, and pears, to nearly 4 g per 100 g guava, blackberries, and blackcurrants (Englyst method of analyis). Similarly vegetables vary from 1–1.5 g (potatoes, cauliflower, celery, lettuce) to 2 g (aubergine, French beans, cabbage), 4 g (baked beans, lentils, Brussels sprouts), and 6–7 g per 100 g (broad beans, kidney beans).

Leaf Protein

Leafy vegetables contain so little protein that excessive amounts would have to be consumed to make a significant contribution to the diet: Such amounts would include unacceptable intakes of dietary fiber (chiefly cellulose). This problem has been overcome by extracting the protein from leaves, including grass; the soluble proteins are separated from the fibrous parts and concentrated by heat coagulation. This product can be added on the domestic scale without further purification to foods such as stews or can be further refined to remove color and dried for storage, which adds to the cost and the technology required. Since grass and many leaves provide a continuing crop leaf protein offers considerable possibilities in developing countries but has been little exploited.

Developing Regions

As quoted in the introduction, there are numerous species of wild plants that could make useful contributions to the diet. The Food and Agriculture Organization has frequently drawn attention to the possibility of collecting wild plant foods, of growing them under protection, or of cultivating them. It has been calculated, for example, that adding 100 g of leafy vegetables to the diet of a 6-year-old child whose staple is cereal or cassava could supply three times the daily need for vitamins A and C, all the folate, calcium, and iron, 15% of the B vitamins, and 15% of the protein. Regular consumption of amaranth leaves and leaves of the drumstick tree (*Moringa olifera*) is recommended for children as a public health measure in many areas of the world, particularly to overcome the widespread problem of vitamin A deficiency.

Vegetarians

Although there is little evidence that the avoidance of animal products results in malnutrition, indeed, in some instances the reverse may be true, the intake of some nutrients may be at risk if all animal foods (including fish, eggs, and milk products) are shunned. Vitamin B_{12} is present only, with very few exceptions (e.g., some yeasts), in animal foods, so supplementation of the diet is virtually essential.

The average intake of carnitine by strict vegetarians is only one-tenth of that of people eating a mixed diet but plasma levels are within 'normal' limits. However, dietary carnitine may be required by premature infants and possibly by full-term infants and may be required by adults taking certain drugs.

There are very few plant sources of taurine and it is not known to what extent this may be a dietary essential. However, plasma levels in strict vegetarians are close to the 'normal' range.

Toxins and Contaminants

Fruits and vegetables contain large numbers of nonnutrients, some of which are toxic, but they are rarely harmful under ordinary conditions. Antinutritional substances in plant foods include antienzymes that interfere with digestion, antivitamins, and substances such as oxalates, phytates, tannins,

and dietary fiber that can interfere with the absorption of some minerals. Glucosinolates, which are responsible for the characteristic flavor of vegetables of the families Cruciferae and Brassicaceae, are goitrogenic but appear to be an insignificant hazard to human health in the amounts usually eaten. In addition some legumes contain lectins in amounts sufficient to have been the cause of occasional cases of food poisoning when incompletely destroyed by cooking.

Most of these substances are destroyed by heat and have been found harmful to animals when included in feed in the raw state. Some are present in salad vegetables that are eaten raw, but in amounts too small to be harmful.

Cassava, especially the bitter variety, contains cyanide. This is usually removed in traditional food processing but not infrequently is a source of harm when the food is not properly treated.

Plantains eaten as a staple provide sufficient 5-hydroxytryptamine to affect central and peripheral nervous systems. Unripe ackee fruit (*Blighia sapida*) contains a toxin (hypoglycin) which causes vomiting sickness and hypoglycaemia. Rhubarb contains oxalate, and although the amounts in stems are harmless, poisoning has resulted from eating the leaves which contain a much higher concentration.

Potatoes contain small, usually harmless, amounts of solanine, but this is increased to toxic levels by exposure to light and subsequent 'greening.'

Some plant foods have been the cause of occasional outbreaks of poisoning in special circumstances. For example Jimson weed, *Datura stramonium*, contains alkaloids including scopolamine which produce hallucination; and the hemp plant, *Cannabis indica*, and the peyote, containing mescaline, have been consumed deliberately for their psychic effects.

Contamination

Some vegetables accumulate environmental toxins such as lead and radioactive fallout but the main cause for concern is from residues of agricultural chemicals, pesticides, and weedkillers. The danger of these chemicals is mainly to those handling them in production and manufacture, but there is concern over the small amounts remaining in the crops since these may be consumed over a long period, and the toxins may possibly be cumulative. Among these are the organochlorine insecticides including DDT and dieldrin. Both the substances and their degradation products persist in crops and so find their way into the human food chain. The fact that they are fat soluble and accumulate in the adipose tissue has given rise to concern but there is no evidence that these quantities merit alarm. Nevertheless, they are restricted in use and are under continuous observation.

Role in Diet

There is considerable epidemiological evidence that a high intake of fruits and vegetables is protective against certain forms of cancer. Vitamins C, E, and β-carotene, when individually subjected to trials by dietary supplementation, have not been shown to be protective and it has been suggested either that there may be a synergistic effect between the various anti-oxidants found in plant foods or that some of the numerous other substances in the food may be the protective agent(s). These include lycopene, lutein, indoles, and phenols. Overall there are many hundreds of nonnutrients in plant foods whose functions in the diet, if any, have been little investigated.

Health Effects

Cardiovascular Disease

Observational studies during the past decade have shown consistent associations between consumption of fresh fruits and vegetables (FF&V) and reduction in cardiovascular disease (CVD) outcomes. An inverse correlation between fruits and vegetables intake and CVD mortality has also been shown in two studies. Similarly, a few, well-controlled randomized clinical trials have shown positive effects on CVD events or biomarkers of risk. The DASH trial (Dietary Approaches to Stop Hypertension), in which a high intake of fresh fruits and vegetables was a major component of the intervention, showed a significant decrease in systolic blood pressure after only 8 weeks. It is less clear whether these beneficial effects are primarily related to antioxidant effects of substances present in fresh fruits and vegetables or to the increased potassium intake associated with high consumption of fruits and vegetables.

Table 1 Sources of vitamin C in the average British diet

Food	*% Average intake*
Potatoes	16
Other vegetables	19
Fruit juices	18
Fruit	17
Salad vegetables	8
Milk products	5
Meat	4
Soft drinks	4
Enriched cereal products	3

From Gregory *et al.* (1990).

Diabetes

Cohort and cross-sectional studies indicate a protective effect of FF&V for type 2 diabetes. The European EPIC study reported an inverse association between FF&V intake and Hgb A1C levels. Other studies have shown a similar inverse correlation with fasting glucose and with glucose levels during an oral glucose tolerance test.

Cancer

The evidence of a protective effect of FF&V on several forms of cancer has been summarized by the International Agency for Research in Cancer (IARC), the World Health Organization, and the World Cancer Research Fund. The IARC report concluded that there is strong evidence of the protective effects of FF&V for cancers of the I gastrointestina tract (including colon) and lungs, but not for others. Some studies have suggested that the intake level necessary for cancer protection may be higher than the traditional recommendation five of servings per day.

See also: **Antioxidants**: Observational Studies; Intervention Studies. **Bioavailability**. **Dietary Fiber**: Physiological Effects and Effects on Absorption. **Food Safety**: Pesticides. **Legumes**. **Nutrition Policies In Developing and Developed Countries**. **Nutritional Surveillance**: Developing Countries. **Nuts and Seeds**. **Phytochemicals**: Epidemiological Factors. **Vegetarian Diets**.

Further Reading

Appel LJ *et al.* (1997) DASH: A clinical trial of the effects of dietary patterns on blood pressure. New England. *Journal of Medicine* **336**: 1117–1124.

FAO (1998) *Traditional food plants*. Food and Nutrition, Paper 42, Rome.

Ford ES and Mokhdad AH (2001) Fruit and vegetable consumption and diabetes mellitus incidence among US adults. *Preventive Medicine* **32**: 33–39.

Gregory J, Foster K, Tyler H, and Wiseman M (1990) *The Dietary and Nutritional Survey of British Adults*. London: HMSO.

Holland B, Welch AA, Unwin ID, Bass DH, Paul AA, and Southgate DAT (1991) *The Composition of Foods*, 5th edn. Cambridge: Royal Society of Chemistry.

Liener IE (1980) *Toxic Constituents of Plant Foodstuffs*, 2nd edn. New York: Academic Press.

Liu S, Manson JE, Lee IM *et al.* (2000) Fruit and vegetable intake and risk of cardiovascular disease: The Women's Health study. *American Journal of Clinical Nutrition* **72**: 922–928.

Oomen HAPC and Grubben GJH (1978) *Tropical leaf vegetables in human nutrition*. Communication 69 Amsterdam: Department of Agriculture Research.

Souci SW, Fachmann W, and Kraut H (1989) *Food Composition and Nutrition Tables*. Stuttgart: Wissen-schaftliche Verlagsgesellschaft mbH.

Watson DH (ed.) (1987) *Natural Toxicants in Food*. Chichester: Ellis Horwood.

World Cancer Research Fund and American Institute of Cancer Research (1997) *Nutrition and the Prevention of Cancer, a Global Perspective*. Washington, DC: AICR.

FUNCTIONAL FOODS

Contents

Health Effects and Clinical Applications

L Galland, Applied Nutrition Inc., New York, NY, USA

Introduction

Functional foods are foods with health benefits that exceed those attributable to the nutritional value of the food. The term is usually applied to foods that have been modified or combined in order to enhance the health benefits but may include any food that naturally possesses components with demonstrable pharmacologic activity. Functional foods are most often selected because they contain ingredients with immune-modulating, antioxidant, anti-inflammatory, antitoxic, or ergogenic effects. The most widely studied functional ingredients are plant-derived phenolic chemicals, probiotic bacteria, and fiber or other poorly digested carbohydrates, but colostrum, egg yolk, and other

nonplant foods may also serve as functional food sources. Although pharmacologic activity of most of these substances is well established *in vitro* or in small mammals, establishing clinical effects in humans poses a challenge for functional food research.

Concept and Definition

The concept of functional foods derives from the observation that certain foods and beverages exert beneficial effects on human health that are not explained by their nutritional content (i.e., macronutrients, vitamins, and minerals). The definition of functional foods varies among countries for reasons that are historical, cultural, and regulatory. In its broadest use, functional foods are food-derived products that, in addition to their nutritional value, enhance normal physiological or cognitive functions or prevent the abnormal function that underlies disease. A hierarchy of restrictions narrows the definition. In most countries, a functional food must take the form of a food or beverage, not a medication, and should be consumed the way a conventional food or beverage is consumed. If the ingredients are incorporated into pills, sachets, or other dosage forms they are considered dietary supplements or nutraceuticals, not functional foods. In Japan and Australia, the functional food appellation has been applied only to food that is modified for the purpose of enhancing its health benefits; in China, Europe, and North America, any natural or preserved food that enhances physiological function or prevents disease might be considered a functional food. If food is modified, there is lack of international consensus as to whether a vitamin or mineral-enriched food (e.g., folate-fortified flour or calcium-fortified orange juice) should be considered a functional food, or whether functional foods are described by the presence of their nonnutritive components (e.g., fiber or polyphenols). Future development of functional foods is likely to be driven by scientific research rather than government regulation, so it is likely that the concept (if not the definition) of functional foods will remain fluid and flexible.

History

If the broadest, least restrictive definition is employed, the use of functional foods for promoting health and relieving symptoms is as old as the practice of medicine. Specific dietary recommendations for treating or preventing various types of illness have been documented in Hippocratic and Vedic texts and the canons of traditional Chinese medicine. Traditional Chinese remedies frequently contain recipes for combining specific foods with culinary and nonculinary herbs to produce healing mixtures. Folk medicine, East and West, has always depended upon functional foods. Peppermint (*Mentha piperita)* tea has a long history of use for digestive complaints. Peppermint oil contains spasmolytic components that block calcium channels in smooth muscle. Cranberry (*Vaccinium macrocarpon*) juice contains proanthocyanidins that inhibit the attachment of *E.coli* to the epithelium of the urinary bladder, explaining its efficacy in prevention of bacterial cystitis and its traditional use for treatment of urinary infection.

Herbs and spices are added to food to enhance flavor and initially were used to inhibit spoilage. Many of these have documented medicinal uses that render them functional foods, broadly defined. Thyme (*Lamiaceae* spp.) was used to treat worms in ancient Egypt. Thyme oils possess potent antimicrobial properties. Ginger (*Zingiber officinale* root), cinnamon (*Cinnamomum* spp. bark), and licorice (*Glycyrrhiza glabra* root) are common ingredients in Chinese herbal tonics and have been widely used in Western folk medicine for treating digestive disorders. Ginger contains over four hundred biologically active constituents. Some have antimicrobial, anti-inflammatory, or anti-platelet effects; others enhance intestinal motility, protect the intestinal mucosa against ulceration and dilate or constrict blood vessels. Cinnamon oil contains cinnamaldehyde and various phenols and terpenes with antifungal, antidiarrheal, vasoactive, and analgesic effects. Recent research has identified phenolic polymers in cinnamon with actions that increase the sensitivity of cells to insulin, leading to the recognition that regular consumption of cinnamon may help to prevent type 2 diabetes. The most studied component of licorice, glycyrrhizin, inhibits the enzyme11 beta-hydroxysteroid dehydrogenase type 2, potentiating the biological activity of endogenous cortisol. Glycyrrhizin also inhibits the growth of *Helicobacter pylori*. Glycyrrhizin and its derivatives may account for the anti-inflammatory and anti-ulcerogenic effects of licorice.

Fermentation is a form of food modification initially developed for preservation. The health-enhancing effects of fermented foods have a place in folk medicine. Several fermented foods have health benefits that exceed those of their parent foods and can be considered functional foods, broadly defined. These include red wine, yogurt, and tempeh. Red wine is a whole fruit alcohol extract that concentrates polyphenols found primarily in the seed and skin of the grape. Its consumption is associated with protection against heart disease, perhaps because red wine polyphenols inhibit the production of free radicals and lipid peroxides that result from the simultaneous ingestion of

cooked meat. Fresh yogurt contains live cultures of lactic acid-producing bacteria that can prevent the development of traveler's diarrhea, antibiotic-induced diarrhea, rotavirus infection, and vaginal yeast infection, decrease the incidence of postoperative wound infection following abdominal surgery and restore the integrity of the intestinal mucosa of patients who have received radiation therapy. Tempeh is made from dehulled, cooked soy beans fermented by the fungus *Rhizopus oligosporus*. Not only is its protein content higher than the parent soy bean, but it also has antibiotic activity *in vitro* and the ability to shorten childhood diarrhea *in vivo*.

Modification of a food to make it less harmful by removing potential toxins or allergens may create a functional food. Using this criterion, infant formula, protein hydrolysates, low-sodium salt substitutes, low-fat dairy products, and low-erucic-acid rapeseed oil (canola oil) might be considered functional foods.

If the most restrictive definition of functional foods is employed, the functional food movement began in Japan during the 1980s, when the Japanese government launched three major research initiatives designed to identify health-enhancing foods to control the rising cost of medical care. In 1991, a regulatory framework, Foods for Special Health Uses (FOSHU), was implemented, identifying those ingredients expected to have specific health benefits when added to common foods, or identifying foods from which allergens had been removed. FOSHU products were to be in the form of ordinary food (not pills or sachets) and consumed regularly as part of the diet. Initially, 11 categories of ingredients were identified for which sufficient scientific evidence indicated beneficial health effects. The Japanese Ministry of Health recognized foods containing these ingredients as functional foods. They were intended to improve intestinal function, reduce blood lipids and blood pressure, enhance calcium or iron absorption, or serve as noncariogenic sweeteners (see **Table 1**). In addition, low-phosphorus milk was approved for people with renal insufficiency and protein-modified rice for people with rice allergy.

Interest in the development of functional foods quickly spread to North America and Europe, where the concept was expanded to include any food or food component providing health benefits in addition to its nutritive value. In Europe, functional food proponents distinguished functional foods from dietetic foods, which are defined by law. European dietetic foods are intended to satisfy special nutritional requirements of specific groups rather than to enhance physiologic function or prevent disease through nonnutritive influences. They include infant formula, processed baby foods (weanling foods), low-calorie foods for weight reduction, high-calorie foods for weight gain, ergogenic foods for athletes, and foods for special medical purposes like the treatment of diabetes or hypertension. In the US, functional food proponents have distinguished functional foods from medical foods, defined by law as special foods designed to be used under medical supervision to meet nutritional requirements in specific medical conditions. In both domains, functional foods have been viewed as whole foods or food components with the potential for preventing cancer, osteoporosis, or cardiovascular disease; improving immunity, detoxification, physical performance, weight loss, cognitive function, and the ability to cope with stress; inhibiting inflammation, free-radical pathology and the ravages of aging; and modulating the effects of hormones. Researchers have sought to validate biomarkers that demonstrate functional improvement in response to dietary intervention, identify the chemical components of functional foods responsible for those effects, and elucidate the mechanism of action of those components. The scientific substantiation of claims is a major objective.

In China, functional foods (referred to as health foods) have been viewed as part of an unbroken

Table 1 Some ingredients conferring FOSHU status on Japanese functional foods

Ingredient	*Physiological function*
Dietary fiber	Improve gastrointestinal function
Psyllium seed husk	
Wheat bran	
Hydrolyzed guar gum	
Oligosaccharides	Improve gastrointestinal function and mineral absorption
Xylo-, fructo-, isomalto-	
Soy-derived	
Polydextrose	
Bacterial cultures	Improve gastrointestinal function
Lactobacilli	
Bifidobacteria	
Soy protein isolates	Reduce cholesterol levels
Diacylglycerols	Reduce triglyceride levels
Sugar alcohols	Prevent dental caries
Maltitol	
Palatinose	
Erythritol	
Green tea polyphenols	Prevent dental caries
Absorbable calcium	Improve bone health
Calcium citrate malate	
Casein phosphopeptide	
Heme iron	Correct iron deficiency
Eucommiacea (tochu) leaf glycosides	Reduce blood pressure
Lactosucrose, lactulose, indigestible dextrin	Improve gastrointestinal function

medical tradition that does not separate medicinal herbs from foods. Over 3000 varieties of health foods are available to Chinese consumers, most derived from compound herbal formulas for which the active ingredients and their mechanism of action are unknown, all claiming multiple effects on various body systems, with little experimental evidence for safety and efficacy but widespread acceptance due to their history of use.

Edible Plants and Phytochemicals

Because their consumption is known to enhance health, vegetables, fruits, cereal grains, nuts, and seeds are the most widely researched functional foods. The health benefits of a plant-based diet are usually attributed to the content of fiber and of a variety of plant-derived substances (phytonutrients and phytochemicals) with antioxidant, enzyme-inducing, and enzyme-inhibiting effects. Some phytochemicals may also exert their health effects by modifying gene expression. Carotenoids, for example, enhance expression of the gene responsible for production of Connexin 43, a protein that regulates intercellular communication. The protective effect of carotenoid consumption against the development of cancer is more strongly related to the ability of individual carotenoids to upregulate Connexin 43 expression than their antioxidant effects or conversion to retinol. Dietary supplementation with beta-carotene reduces the blood levels of other carotenoids, some of which are more potent inducers of Connexin 43 than is beta-carotene. The unexpected and highly publicized increase in incidence of lung cancer among smokers taking beta-carotene supplements may be explained by this mechanism.

Phytochemicals associated with health promotion and disease prevention are described in **Table 2**. The most studied food sources of these phytonutrients are soy beans (*Glycine max*) and tea (*Camellia sinensis* leaves), but tomatoes (*Lycopersicon esculentum*), broccoli (*Brassica oleracea*), garlic (*Allium sativum*), turmeric (*Curcuma longa*), tart cherries (*Prunus cerasus)*, and various types of berries are also receiving considerable attention as functional food candidates. An overview of the research on soy and tea illustrates some of the clinical issues encountered in the development of functional foods from edible plants.

Soy protein extracts have been found to lower cholesterol in humans, an effect that appears to be related to amino acid composition. Soy protein extracts frequently contain nonprotein isoflavones, which have received considerable attention because of their structural similarity to estrogen. Soy isoflavones are weak estrogen agonists and partial estrogen antagonists. Epidemiologic and experimental data indicate that isoflavone exposure during adolescence may diminish the incidence of adult breast

Table 2 Phytochemicals associated with health promotion and disease prevention

Group	*Typical components*	*Biological activities*	*Food sources*
Carotenoids	Alpha- and beta-carotene cryptoxanthin, lutein, lycopene, zeaxanthin	Quench singlet and triplet oxygen, increase cell–cell communication	Red, orange and yellow fruits and vegetables, egg yolk, butter fat, margarine
Glucosinolates, isothiocyanates	Indole-3-carbinol sulphoraphane	Increase xenobiotic metabolism, alter estrogen metabolism	Cruciferous vegetables, horseradish
Inositol phosphates	Inositol hexaphosphate (phytate)	Stimulate natural killer cell function, chelate divalent cations	Bran, soy foods
Isoflavones	Genistein, daidzein	Estrogen agonist and antagonist, induce apoptosis	Soy foods, kudzu
Lignans	Enterolactone, enterolactone	Estrogen agonists and antagonists, inhibit tyrosine kinase	Flax seed, rye
Phenolic acids	Gallic, ellagic, ferulic, chlorogenic, coumaric	Antioxidant, enhance xenobiotic metabolism	Diverse fruits, vegetables
Phytoallexins	Resveratrol	Antioxidant, platelet inhibition, induce apoptosis	Red wine, grape seed
Polyphenols	Flavonoids, chalcones, catechins, anthocyanins, proanthocyanidins	Antioxidant, enhance xenobiotic metabolism, inhibit numerous enzymes	Diverse fruits, vegetables, red wine, tea
Saponins	Glycyrrhizin, ginsenosides	Antimicrobial, immune boosting, cytotoxic to cancer cells	Legumes, nuts, herbs
Sterols	Beta-sistosterol, campestrol	Bind cholesterol, decrease colonic cell proliferation, stimulate T-helper-1 cells	Nuts, seeds, legumes, cereal grains
Sulfides	Diallyl sulfides	Antimicrobial, antioxidant	Garlic, onions

cancer. *In vitro* studies show conflicting effects. On the one hand, soy isoflavones induce apoptosis of many types of cancer cells; on the other hand, estrogen receptor-bearing human breast cancer cells proliferate in tissue culture when exposed to isoflavones. Although the widespread use of soy in Asia is cited in support of the safety of soy foods, the intake of isoflavones among Asian women consuming soy regularly is in the range of 15–40 mg day^{-1}, significantly less than the isoflavone content of a serving of soymilk as consumed in the US. In clinical trials, soy isoflavones have not been effective in relieving hot flashes of menopausal women but do diminish the increased bone resorption that causes postmenopausal bone loss. In premenopausal women, soy isoflavones may cause menstrual irregularities. The successful development of soy derivatives as functional foods will require that these complex and diverse effects of different soy components in different clinical settings be better understood.

Regular consumption of tea, green or black, is associated with a decreased risk of heart disease and several kinds of cancer. These benefits are attributed to tea's high content of catechin polymers, especially epigallocatechin gallate (ECGC), which has potent antioxidant and anti-inflammatory effects, that may lower cholesterol in hyperlipidemic individuals and alter the activity of several enzymes involved in carcinogenesis. Catechin content is highest in young leaves. Aging and the fermentation used to produce black tea oxidize tea catechins, which polymerize further to form the tannins, theaflavin and thearubigen. Although ECGC is a more potent antioxidant than theaflavin, theaflavin is far more potent an antioxidant than most of the commonly used antioxidants, like glutathione, vitamin E, vitamin C, and butylated hydroxytoluene (BHT). Both ECGC and theaflavin are partially absorbed after oral consumption, but a clear dose–response relationship has not been established. Tea-derived catechins and polymers are being intensively studied as components of functional foods, because the results of epidemiologic, *in vitro*, and animal research indicate little toxicity and great potential benefit in preventing cancer or treating inflammation-associated disorders. Clinical trials have shown a mild cholesterol-lowering effect and perhaps some benefit for enhancing weight loss.

Probiotics and Prebiotics

Probiotics are live microbes that exert health benefits when ingested in sufficient quantities. Species of lactobacilli and bifidobacteria, sometimes combined with *Streptococcus thermophilus*, are the main bacteria used as probiotics in fermented dairy products. Most probiotic research has been done with nutraceutical preparations, but yogurt has been shown to alleviate lactose intolerance, prevent vaginal candidosis in women with recurrent vaginitis, and reduce the incidence or severity of gastrointestinal infections.

Prebiotics are nondigestible food ingredients that stimulate the growth or modify the metabolic activity of intestinal bacterial species that have the potential to improve the health of their human host. Criteria associated with the notion that a food ingredient should be classified as a prebiotic are that it remains undigested and unabsorbed as it passes through the upper part of the gastrointestinal tract and is a selective substrate for the growth of specific strains of beneficial bacteria (usually lactobacilli or bifidobacteria), rather than for all colonic bacteria, inducing intestinal or systemic effects through bacterial fermentation products that are beneficial to host health. Prebiotic food ingredients include bran, psyllium husk, resistant (high amylose) starch, inulin (a polymer of fructofuranose), lactulose, and various natural or synthetic oligosaccharides, which consist of short-chain complexes of sucrose, fructose, galactose, glucose, maltose, or xylose. The best-known effect of prebiotics is to increase fecal water content, relieving constipation. Bacterial fermentation of prebiotics yields short-chain fatty acids (SCFAs) that nourish and encourage differentiation of colonic epithelial cells. Absorbed SCFAs decrease hepatic cholesterol synthesis. Fructooligosaccharides (FOSs) have been shown to alter fecal biomarkers (pH and the concentration of bacterial enzymes like nitroreductase and beta-glucuronidase) in a direction that may convey protection against the development of colon cancer.

Several prebiotics have documented effects that are probably independent of their effects on gastrointestinal flora. Whereas the high phytic acid content of bran inhibits the absorption of minerals, FOSs have been shown to increase absorption of calcium and magnesium. Short-chain FOSs are sweet enough to be used as sugar substitutes. Because they are not hydrolyzed in the mouth or upper gastrointestinal tract, they are noncariogenic and noninsulogenic. Bran contains immunostimulating polysaccharides, especially beta-glucans and inositol phosphates, which have been shown to stimulate macrophage and natural killer cell activity *in vitro* and in rodent experiments. The poor solubility and absorption of beta-glucans and inositol phosphates are significant barriers to clinical effects in humans.

Immune Modulators

Several substances produced by animals and fungi have been investigated for immune-modulating effects. Fish oils are the most studied. As a source of *n*-3 fatty acids, fish oil consumption by humans has been shown to influence the synthesis of inflammatory signaling molecules like prostaglandins, leukotrienes, and cytokines. In addition to direct effects on prostanoid synthesis, *n*-3 fats have also been shown to directly alter the intracellular availability of free calcium ions, the function of ion channels, and the activity of protein kinases. Generally administered as nutraceuticals rather than as functional foods, fish oil supplements have demonstrated anti-inflammatory and immune suppressive effects in human adults. A high intake of the *n*-3 fatty acids eicosapentaenoic (20:5*n*-3) and docosahexaenoic (22:6*n*-3) acid (DHA) from seafood or fish oil supplements has also been associated with prevention of several types of cancer, myocardial infarction, ventricular arrhythmias, migraine headaches, and premature births, and with improved control of type 2 diabetes mellitus, inflammatory bowel disease, rheumatoid arthritis, cystic fibrosis, multiple sclerosis, bipolar disorder, and schizophrenia. 20:5*n*-3 but not 22:6*n*-3 is effective for schizophrenia and depression; 22:6*n*-3 but not 20:5*n*-3 improves control of blood sugar in diabetics. The benefits of fish oil supplements have prompted efforts at increasing the *n*-3 fatty acid content of common foods by adding fish oil or flax oil extracts. Consumption of these has been associated with decreased levels of some inflammatory biomarkers, including thromboxane B2, prostaglandin E2, and interleukin 1-beta.

Feeding flax seed meal or fish meal to hens enriches the *n*-3 fatty acid content of the yolks of the eggs they lay. Consumption of these eggs increases the *n*-3 fatty acid content of plasma and cellular phospholipids and produces an improved blood lipid profile when compared with consumption of standard eggs. Egg yolk is not only a source of fatty acids, but also of carotenoids and immunoglobulins. The xanthophyll carotenoids zeaxanthin and its stereoisomer lutein are readily absorbed from egg yolk. Their consumption is associated with a decreased incidence of macular degeneration and cataract. Immunizing hens to specific pathogens and extracting the antibodies present in their egg yolks yields a functional food that has been shown to prevent enteric bacterial or viral infection in experimental animals.

Bovine colostrum, the milk produced by cows during the first few days postpartum, has a long history of use as a functional food. Compared to mature milk, colostrum contains higher amounts of immunoglobulins, growth factors, cytokines, and various antimicrobial and immune-regulating factors. Consumption of bovine colostrum has been shown to reduce the incidence of diarrheal disease in infants and the symptoms of respiratory infection in adults. Specific hyperimmune bovine colostrums, produced by immunizing cows to pathogenic organisms like *Cryptosporidium parvum*, *Helicobacter pylori*, rotavirus, and *Shigella* spp., may prevent or treat infection by these organisms.

Human studies have also shown that consumption of bovine colostrum can improve anaerobic athletic performance and prevent the enteropathy induced by use of nonsteroidal anti-inflammatory drugs.

Mushrooms play a major role in traditional Chinese medicine and as components of contemporary Chinese health foods. Many *Basidiomycetes* mushrooms contain biologically active polysaccharides in fruiting bodies, cultured mycelium, or culture broth. Most belong to the group of beta-glucans that have both beta-(1→3) and beta-(1→6) linkages. Although they stimulate macrophages and natural killer cells, the anticancer effect of mushroom polysaccharide extracts appears to be mediated by thymus-derived lymphocytes. In experimental animals, mushroom polysaccharides prevent oncogenesis, show direct antitumor activity against various cancers, and prevent tumor metastasis. Clinical trials in humans have shown improvement in clinical outcome when chemotherapy was combined with the use of commercial mushroom polysaccharides like lentinan (from *Lentinus edodes* or shiitake), krestin (from *Coriolus versicolor*), or schizophyllan (from *Schizophyllum commune*). Mushroom extracts may fulfill their potential more as medicines than as functional foods.

Designer Foods

An important direction in the development of functional foods is the combination of numerous ingredients to achieve a specific set of goals, rather than efforts to uncover the potential benefits of a single food source. Infant formula was probably the first area for designer foods of this type, because of the profound influence of nutrients on the developing brain and immune system. The addition of DHA to infant formula for enhancing brain and visual development, the alteration of allergenic components in food, and the possible use of probiotics and

nucleotides to enhance immune response are important developments in this area.

Sports nutrition is another established arena for designer foods. Specific nutritional measures and dietary interventions have been devised to support athletic performance and recuperation. Oral rehydration products for athletes were one of the first categories of functional foods for which scientific evidence of benefit was obtained. Oral rehydration solutions must permit rapid gastric emptying and enteral absorption, improved fluid retention, and thermal regulation, to enhance physical performance and delay fatigue. Carbohydrates with relatively high glycemic index combined with whey protein concentrates or other sources of branched chain amino acids have been shown to enhance recovery of athletes. Caffeine, creatine, ribose, citrulline, L-carnitine, and branched chain amino acids have each been shown to improve exercise performance or diminish postexercise fatigue. Whether combinations of these ingredients, blended into foods or beverages, will perform better than the individual ingredients will help to determine the design of future sports foods.

Optimal cardiovascular health involves prevention of excessive levels of oxidant stress, circulating homocysteine, cholesterol, triglycerides and fibrinogen, and protection of the vascular endothelium. A mix of ingredients supplying all of these effects could consist of soy protein powder, oat beta-glucan, plant sterols and stanols, folic acid, L-arginine, 22:6*n*-3, magnesium, and red wine or green tea polyphenols. Evidence suggests that addressing multiple nutritional influences on cardiovascular health will be more beneficial than addressing only one influence, but more definitive studies are needed. Genetic factors may need to be incorporated for designer foods to achieve their full potential. Polyunsaturated fatty acids, for example, raise the serum concentration of HDL-cholesterol among individuals who carry the Apo A1-75A gene polymorphism, but reduce HDL-cholesterol levels of individuals who carry the more common Apo A1-75G polymorphism.

See also: **Alcohol**: Absorption, Metabolism and Physiological Effects. **Carotenoids**: Chemistry, Sources and Physiology; Epidemiology of Health Effects. **Dietary Fiber**: Physiological Effects and Effects on Absorption. **Fatty Acids**: Omega-3 Polyunsaturated. **Functional Foods**: Regulatory Aspects. **Microbiota of the Intestine**: Prebiotics; Probiotics. **Phytochemicals**: Classification and Occurrence; Epidemiological Factors. **Protein**: Quality and Sources. **Sports Nutrition**. **Tea**.

Further Reading

Ashwell M (2001) Functional foods: a simple scheme for establishing the scientific basis for all claims. *Public Health Nutrition* **4**(3): 859–862.

Bellisle F, Diplock AT, Hornstra G *et al.* (eds.) (1998) Functional food science in Europe. *British Journal of Nutrition* **80**(supplement 1): S1–S193.

Clydesdale FM and Chan SH (eds.) (1995) First International Conference on East–West Perspectives on Functional Foods. *Nutrition Reviews* **54**(11, part II): S1–S202.

Constantinou AI and Singletary KW (eds.) (2002) Controversies in functional foods. *Pharmaceutical Biology* **40**(supplement): 5–74.

Diplock AT, Aggett PJ, Ashwell M *et al.* (eds.) (1999) Scientific Concepts of Functional Foods in Europe: Consensus Document. *British Journal of Nutrition* **81**(supplement): S1–S27.

Farnworth ER (2003) *Handbook of Fermented Functional Foods.* USA: CRC Press.

Goldberg I (ed.) (1994) *Functional Foods, Designer Food, Pharmafoods, Nutraceuticals*. New York: Chapman and Hall.

ILSI North America Technical Committee on Food Components for Health Promotion (1999). *Food Component Report.* Washington, DC: ILSI Press.

Knorr D (1999) Technology aspects related to microorganisms in functional food. *Trends in Food Science and Technology.* **9**(8–9, Special Issue): 295–306.

Langseth L (1995) *Oxidants, Antioxidants and Disease Prevention: ILSI Europe Concise Monograph Series*. Washington, DC: ILSI Press.

Langseth L (1996) *Nutritional Epidemiology: Possibilities and Limitations: ILSI Europe Concise Monograph Series.* Washington, DC: ILSI Press.

Langseth L (1999) *Nutrition and Immunity in Man: ILSI Europe Concise Monograph Series*. Washington, DC: ILSI Press.

Meskin MS, Biidlack BI, Davies AJ, and Omaye ST (eds.) (2002) *Phytochemicals in Nutrition and Health*. USA: CRC Press.

Roberfroid MB (2000) Defining functional foods. In: Gibson G and Williams C (eds.) *Functional Foods*. Cambridge: Woodhead Publishing Ltd.

Truswell AS (1995) *Dietary Fat: Some Aspects of Nutrition and Health and Product Development: ILSI Europe Concise Monograph Series*. Washington, DC: ILSI Press.

Regulatory Aspects

H H Butchko, Exponent, Inc., Wood Dale, IL, USA
B J Petersen, Exponent, Inc., Washington DC, USA

Although there is no universally accepted definition of functional food, the International Life Sciences Institute of North America (ILSI NA) defines such foods as those that provide a health benefit beyond basic nutrition through the presence of physiologically active food components. Health Canada considers functional food as "similar in appearance to a conventional food, consumed as part of the usual diet, with demonstrated physiological benefits, and/

or to reduce the risk of chronic disease beyond basic nutritional functions." The Institute of Medicine of the US National Academy of Sciences has a more limited definition of functional foods as those in which the concentrations of one or more ingredients have been manipulated or modified to enhance their contribution to a healthful diet.

Under a broad definition, functional foods may include conventional foods; fortified, enriched, or enhanced foods; and dietary supplements because they provide essential nutrients often beyond quantities necessary for normal maintenance, growth, and development and/or other biologically active components that impart health benefits or desirable physiological effects. Thus, fruits and vegetables, such as broccoli, carrots, and tomatoes, are the simplest forms of functional foods because they provide physiologically active components such as sulforaphane, β-carotene, and lycopene, respectively.

Although functional foods can play a key role in promoting a healthier population, the government regulation of such foods is important to ensure protection of consumers from fraud and to ensure that any claims provide accurate information, are not misleading, and are scientifically valid. Governments have developed or are developing regulatory frameworks of nutrition and health claims to assist consumers in choosing foods for health promotion. The regulation of functional foods, the types and wording of claims communicated to consumers, and their place in national regulatory frameworks is evolving globally.

Regulation of Functional Foods in Japan

In the 1980s, the Japanese government funded large-scale research programs for systemic analysis of food functions and the physiological regulation of the function of food and the molecular design of functional foods. In 1991, the government established the Japanese Foods for Specified Health Use (FOSHU) to define foods with potential health benefits to help stem the rising cost of health care in Japan.

Under the FOSHU system, health claims are approved for specific products. Companies make an application for FOSHU approval to the Ministry of Health and Welfare (MHW). FOSHU are those foods that have a specific health benefit due to the presence of certain constituents or foods. Allergens cannot be present. Scientific substantiation, including the scientific evidence of safety and efficacy of the food and the medical or nutritional basis for the claim, must be provided to the MHW for consideration. To be classified as FOSHU, it must be demonstrated that the final food product, not just individual components, is likely to have a beneficial health effect when consumed as part of the normal diet. FOSHU must be in the form of food and not pills or capsules.

The labeling of FOSHU foods must not be misleading and must include the approved health claim, the recommended daily intake, relevant nutrition information, guidance on healthful eating, and any necessary warnings regarding excessive intake. Domestic products have an 'approved' mark from the MHW, whereas imported products have a "permitted" mark.

Regulation of Functional Foods in the United States

Current US food regulations do not specifically address functional foods but, rather, include them in several categories within conventional foods, food additives, dietary supplements, medical foods, or foods for special dietary use. All of these fall under the amended Federal Food, Drug and Cosmetic Act (FDCA) of 1938 and are implemented under regulations from the Food and Drug Administration (FDA). Four types of claims can be used to communicate the usefulness of functional foods to consumers: health claims, qualified health claims, structure–function claims, and nutrient content claims.

Health Claims

The Nutrition Labeling and Education Act (NLEA) of 1990 authorizes the FDA to allow approved disease risk-reduction claims, known as health claims, to appear on food labeling. NLEA allows claims that "characterize the relationship of any substance to a disease or health-related condition." For example, "diets low in sodium may reduce the risk of high blood pressure." Health claims may not be false or misleading in any respect and must not suggest that a food will diagnose, treat, mitigate, cure, or prevent any disease, or they would be considered drug claims under the FDCA. If a manufacturer fails to comply with all of the requirements for an approved health claim, the FDA would consider that the food is either misbranded (mislabeled and therefore illegal) or an illegal drug because it would not comply with all applicable drug requirements.

The scientific standard for authorization of a health claim under NLEA mandates that there be significant scientific agreement among qualified

experts about the validity of the relationship described in the proposed claim. The FDA has approved 12 health claims that meet the significant scientific agreement standard (**Table 1**).

The FDA Modernization Act of 1997 (FDAMA) provides an additional expedited process for manufacturers to use health claims. FDAMA allows health claims if they are based on current, published, authoritative statements from certain federal government official scientific bodies, such as the National Institutes of Health, the Centers for Disease Control and Prevention, and the National Academy of Sciences. Under FDAMA, manufacturers are required to notify and provide specific wording of the claim to the FDA 120 days in advance of use of the claim. During this time period, the FDA is expected to review the claim and may prohibit or modify the claim. If the FDA fails to act within the 120-day period, the claim is authorized by statue; the FDA is not required to issue a regulation. Since July 6, 1999, when the first health claim under the FDAMA was authorized, only one additional health claim has been allowed (**Table 2**).

Qualified Health Claims

Qualified health claims allow disease risk-reduction statements but, unlike health claims, must be qualified to indicate that the level of scientific support is not conclusive. Qualified health claims for dietary supplements were first authorized under a 1999 court decision in the case of *Pearson versus Shalala* regarding health claims for dietary supplements. In December 2002, the FDA announced the institution of a new labeling scheme for qualified health claims. The FDA indicated that it will depart from its standard of significant scientific agreement for health claims in evaluating qualified health claims. Qualified health claims may be based on the weight of the scientific evidence. In July 2003, the FDA announced a ranking system and proposed language for qualified health claims. Under the new scheme, claims are ranked by strength of scientific evidence. A claim designated as 'A' is actually an unqualified health claim with the standard of significant scientific agreement. Claims 'B,' 'C,' and 'D' would have progressively less supportive scientific evidence, and the FDA has suggested appropriate qualifying language for these claims (**Table 3**). To date, the FDA has allowed qualified health claims regarding certain foods, food components, and dietary supplements and the risk of cancer, cardiovascular disease, cognitive function and dementia, and neural tube defects (**Table 4**).

Structure–Function Claims

Structure–function claims for conventional foods focus on effects derived from nutritive value, whereas such claims for dietary supplements may focus on nutritive as well as nonnutritive effects. Structure–function claims describe the role of a nutrient or dietary ingredient that affects normal structure or function of the body (e.g., "calcium builds strong bones") without linking it to a specific disease. Structure–function claims may also characterize the mechanism by which a nutrient or dietary ingredient acts to maintain such structure or function (e.g., "fiber maintains bowel regularity"). They may also relate general well-being to consumption of a nutrient or dietary ingredient or describe a benefit related to a nutritional deficiency disease (e.g., deficiency of vitamin C and the occurrence of scurvy), which must be accompanied by a statement that describes the prevalence of such a disease in the United States. Structure–function claims on conventional foods are not preapproved by the FDA; it is the manufacturer's responsibility to ensure the accuracy and truthfulness of its claims and that such claims are not misleading.

The Dietary Supplement Health and Education Act of 1994 established special regulatory procedures for structure–function claims for dietary supplements. When such claims are used with a dietary supplement, the label must include a disclaimer that the FDA has not evaluated the claim and also that the product is not intended to "diagnose, treat, cure, or prevent any disease." Manufacturers of dietary supplements that make structure–function claims on labels or in labeling are required to submit a notification to the FDA no later than 30 days after marketing the dietary supplement. Although this notification must include the text of the structure–function claim, there is no requirement that the manufacturer include the scientific evidence supporting the claim with this notification.

Nutrient Content Claims

A nutrient content claim either expressly or implicitly characterizes the level of a nutrient in a product (e.g., "high in vitamin C" or "low in sodium"). In general, nutrient content claims cannot be used in food labeling unless the claim is made in accordance with existing FDA regulations or an authoritative statement by a scientific body. The FDA has allowed nutrient content claims for certain substances for which it has established Daily Reference Values or Reference Daily Intakes (RDIs). In general, the FDA allows nutrient content labeling for "high in," "rich in," or an "excellent source of" a vitamin or mineral

Table 1 FDA-approved health claims that meet the significant scientific agreement standard

Food or dietary component	*Disease claim*	*Model claim*
Calcium 21 CFR 101.72	Osteoporosis	"Regular exercise and a healthy diet with enough calcium helps teens and young adult white and Asian women maintain good bone health and may reduce their high risk of osteoporosis later in life."
Dietary fat 21 CFR 101.73	Cancer	"Development of cancer depends on many factors. A diet low in total fat may reduce the risk of some cancers."
Dietary saturated fat and cholesterol 21 CFR 101.75	Coronary heart disease	"While many factors affect heart disease, diets low in saturated fat and cholesterol may reduce the risk of this disease."
Dietary noncariogenic carbohydrate sweeteners 21 CFR 101.80	Dental caries	**Full claim**: "Frequent between-meal consumption of foods high in sugars and starches promotes tooth decay. The sugar alcohols in [name of food] do not promote tooth decay." **Shortened claim** (on small packages only): "Does not promote tooth decay."
Fiber-containing grain products, fruits, and vegetables 21 CFR 101.76	Cancer	"Low-fat diets rich in fiber-containing grain products, fruits, and vegetables may reduce the risk of some types of cancer, a disease associated with many factors."
Folate 21 CFR 101.79	Neural tube birth defects	"Healthful diets with adequate folate may reduce a woman's risk of having a child with a brain or spinal cord defect."
Fruits and vegetables 21 CFR 101.78	Cancer	"Low-fat diets rich in fruits and vegetables (foods that are low in fat and may contain dietary fiber, vitamin A, or vitamin C) may reduce the risk of some types of cancer, a disease associated with many factors. Broccoli is high in vitamin A and C, and it is a good source of dietary fiber."
Fruits, vegetables, and grain products that contain fiber, particularly soluble fiber 21 CFR 101.77	Coronary heart disease	"Diets low in saturated fat and cholesterol and rich in fruits, vegetables, and grain products that contain some types of dietary fiber, particularly soluble fiber, may reduce the risk of heart disease, a disease associated with many factors."
Sodium 21 CFR 101.74	Hypertension	"Diets low in sodium may reduce the risk of high blood pressure, a disease associated with many factors."
Soluble fiber from certain foods 21 CFR 101.81	Coronary heart disease	"Soluble fiber from foods such as [*name of soluble fiber source, and if desired, name of food product*], as part of a diet low in saturated fat and cholesterol, may reduce the risk of heart disease. A serving of [*name of food product*] supplies __ grams of the [necessary daily dietary intake for the benefit] soluble fiber from [*name of soluble fiber source*] necessary per day to have this effect."
Soy protein 21 CFR 101.82	Coronary heart disease	(1) "25 grams of soy protein a day, as part of a diet low in saturated fat and cholesterol, may reduce the risk of heart disease. A serving of [*name of food*] supplies __ grams of soy protein." (2) "Diets low in saturated fat and cholesterol that include 25 grams of soy protein a day may reduce the risk of heart disease. One serving of [*name of food*] provides __ grams of soy protein."
Stanols/sterols 21 CFR 101.83	Coronary heart disease	(1) "Foods containing at least 0.65 gram per serving of vegetable oil sterol esters, eaten twice a day with meals for a daily total intake of at least 1.3 grams, as part of a diet low in saturated fat and cholesterol, may reduce the risk of heart disease. A serving of [*name of food*] supplies __ grams of vegetable oil sterol esters." (2) "Diets low in saturated fat and cholesterol that include two servings of foods that provide a daily total of at least 3.4 grams of plant stanol esters in two meals may reduce the risk of heart disease. A serving of [*name of food*] supplies__ grams of plant stanol esters."

From *Label Claims, Health Claims That Meet Significant Scientific Agreement*. Available at www.cfsan.fda.gov/~dms/lab-ssa.html.

Table 2 Health claims allowed under the FDAMA

Food or dietary component	*Disease*	*Basis*
Potassium	High blood pressure and stroke	National Academy of Sciences report *Diet and Health: Implications for Reducing Chronic Disease Risk*
Whole grain foods	Heart disease and certain cancers	National Academy of Sciences report *Diet and Health: Implications for Reducing Chronic Disease Risk*

From *Label Claims, FDA Modernization Act of 1997 (FDAMA) Claims*. Available at www.cfsan.fda.gov/~dms/labfdama.html.

for which the agency has established an RDI if the food provides 20% or more of the RDI per reference amount customarily consumed. The FDA has also published regulations authorizing and establishing detailed requirements for "good source," "more," and "light" (or "lite") claims, and certain claims about calorie content, sodium content, and fat, fatty acid, and cholesterol content.

If a manufacturer wants to make a claim about a food as a good source of a nutrient for which no FDA nutrient content claim regulation exists, such a claim would not be allowed, even if the claim is truthful and not misleading, unless and until the FDA issues a regulation approving the use of the claim.

Table 3 Standardized language for qualified health claims by category

FDA category	*Level of scientific evidence*	*Proposed qualifying language*
B	Second level: Moderate/good level of comfort	"Although there is scientific evidence supporting this claim, the evidence is not conclusive."
C	Third level: Low level of comfort	"Some scientific evidence suggests ... however, FDA has determined that this evidence is limited and not conclusive."
D	Fourth level: Extremely low level of comfort	"Very limited and preliminary scientific research suggests ... FDA concludes that there is little scientific evidence supporting this claim."

From *Guidance for Industry and FDA. Interim Procedures for Qualified Health Claims in the Labeling of Conventional Human Food and Dietary Supplements*. Available at www.cfsan.fda.gov/~dms/hclmgui3.html.

European Regulations for Functional Foods

The concept of functional foods was first evaluated in Europe in the 1990s when the International Life Sciences Institute in Europe (ILSI Europe) developed a project on functional foods that became a European Commission (EC) concerted action, Functional Food Science in Europe. Approximately 100 experts in nutrition and medicine in Europe reviewed the scientific literature about foods and food components and their effects on body functions, and they developed a global framework that included a framework for the identification and development of functional foods and for the scientific substantiation of their health-related effects. From this evaluation, two types of claims for functional foods were suggested: enhanced function claims and reduction of disease risk claims. From this evaluation, the "Concepts of Functional Foods" was produced by ILSI followed by publication of "Scientific Concepts of Functional Foods in Europe: Consensus Document." According to this concept document, "a food can be regarded as functional if it is satisfactorily demonstrated to beneficially affect one or more target functions in the body, beyond adequate nutritional effects, in a way which is relevant to either an improved state of health and well-being, or reduction of risk of disease." As in the United States, this definition of a functional food specifically excludes the treatment of disease. However, unlike the case in the United States, where functional foods can include dietary supplements in pill or capsule form, in Europe functional foods must be foods that have a positive health benefit in amounts normally consumed in the diet.

There are currently no final regulations or legislation at the European Union (EU) level that define permissible nutrition and health claims on foods. However, various member states have adopted local legislation to regulate their use, which has resulted in numerous differences throughout the EU in the definition of terms and the circumstances when claims are warranted. In 2003, the EU Commission issued a "Proposal for a Regulation of the European Parliament and of the Council on Nutrition and Health Claims Made in Foods" for harmonization of claims throughout the EU. Under this proposed regulation, although functional food is not defined, both nutrition and health claims would be allowed. Nutrition and health claims must be based on, and substantiated by, generally accepted scientific data and not be false, ambiguous, or misleading or imply doubts about the safety or nutritional adequacy of other foods. This proposal is being evaluated by the member states.

Table 4 Qualified health claims permitted by FDA

Food or food component	*Disease*	*Eligible food*	*Required claim statement*
Selenium	Cancer	Dietary supplements containing selenium	(1) "Selenium may reduce the risk of certain cancers. Some scientific evidence suggests that consumption of selenium may reduce the risk of certain forms of cancer. However, FDA has determined that this evidence is limited and not conclusive." *or*, (2) "Selenium may produce anticarcinogenic effects in the body. Some scientific evidence suggests that consumption of selenium may produce anticarcinogenic effects in the body. However, FDA has determined that this evidence is limited and not conclusive."
Antioxidant vitamins	Cancer	Dietary supplements containing vitamin E and/or vitamin C	(1) "Some scientific evidence suggests that consumption of antioxidant vitamins may reduce the risk of certain forms of cancer. However, FDA has determined that this evidence is limited and not conclusive." *or*, (2) "Some scientific evidence suggests that consumption of antioxidant vitamins may reduce the risk of certain forms of cancer. However, FDA does not endorse this claim because this evidence is limited and not conclusive." *or*, (3) "FDA has determined that although some scientific evidence suggests that consumption of antioxidant vitamins may reduce the risk of certain forms of cancer, this evidence is limited and not conclusive."
Nuts	Heart disease	Whole or chopped nuts or foods containing nuts with at least 11 g per reference amount customarily consumed; types of nuts — almonds, hazelnuts, peanuts, pecans, some pine nuts, pistachio nuts, walnuts	"Scientific evidence suggests but does not prove that eating 1.5 ounces per day of most nuts [such as *name of specific nut*] as part of a diet low in saturated fat and cholesterol may reduce the risk of heart disease."
Walnuts	Heart disease	Whole or chopped walnuts	"Supportive but not conclusive research shows that eating 1.5 ounces per day of walnuts, as part of a low saturated fat and low cholesterol diet and not resulting in increased caloric intake, may reduce the risk of coronary heart disease. See nutrition information for fat [and calorie] content."
Omega-3 fatty acids	Coronary heart disease	Dietary supplements containing the omega-3 long-chain polyunsaturated fatty acids eicosapentanoic acid (EPA) and/or docosahexanoic acid (DHA)	"Consumption of omega-3 fatty acids may reduce the risk of coronary heart disease. FDA evaluated the data and determined that, although there is scientific evidence supporting the claim, the evidence is not conclusive."

Continued

Table 4 Continued

Food or food component	*Disease*	*Eligible food*	*Required claim statement*
B vitamins	Vascular disease	Dietary supplements containing vitamin B_6, B_{12}, and/or folic acid	"As part of a well-balanced diet that is low in saturated fat and cholesterol, Folic Acid, Vitamin B_6 and Vitamin B_{12} may reduce the risk of vascular disease. FDA evaluated the above claim and found that, while it is known that diets low in saturated fat and cholesterol reduce the risk of heart disease and other vascular diseases, the evidence in support of the above claim is inconclusive."
Phosphatidylserine	Cognitive function and dementia	Dietary supplements containing soyderived phosphatidylserine	(1) "Consumption of phosphatidylserine may reduce the risk of dementia in the elderly. Very limited and preliminary scientific research suggests that phosphatidylserine may reduce the risk of dementia in the elderly. FDA concludes that there is little scientific evidence supporting this claim." *or*, (2) "Consumption of phosphatidylserine may reduce the risk of cognitive dysfunction in the elderly. Very limited and preliminary scientific research suggests that phosphatidylserine may reduce the risk of cognitive dysfunction in the elderly. FDA concludes that there is little scientific evidence supporting this claim."
Folic acid (0.8 mg)[a]	Neural tube birth defects	Dietary supplements containing folic acid	"0.8 mg folic acid in a dietary supplement is more effective in reducing the risk of neural tube defects than a lower amount in foods in common form. FDA does not endorse this claim. Public health authorities recommend that women consume 0.4 mg folic acid daily from fortified foods or dietary supplements or both to reduce the risk of neural tube defects."

[a]FDA has approved a health claim for folic acid in food and does not endorse that the 0.8 mg is more effective than lower amounts found in food.
From *Summary of Qualified Health Claims Permitted*. Available at www.cfsan.fda.gov/~dms/qhc-sum.html.

Nutrition Claims

A nutrition claim is defined as "any claim which states, suggests, or implies that a food has particular nutrition properties due to" energy or nutrients or other substances (i.e., provides, provides at a reduced or increased rate, or does not provide). The proposed regulation calls for establishment of a list of permitted claims and their specific conditions of use, which would be regularly updated. In the current proposed regulation, the list of nutrients includes fat, saturated fat, unsaturated fat, monounsaturated fat, polyunsaturated fat, omega-3 fatty acids, sugar, sodium, fiber, protein, vitamins/minerals, or other substances.

Health Claims

A health claim is defined in the proposed regulation as "any claim that states, suggests, or implies that a relationship exists between a food category, a food, or one of its constituents and health." A reduction in disease risk claim is defined as "any health claim that states, suggests, or implies that the consumption of a food category, a food, or one of its constituents significantly reduces a risk factor in the development of a human disease." As in the United States, medicinal products in Europe are those that treat, prevent, or diagnose disease or restore, correct, or modify physiological functions. In the European Directive 2000/13/EC on labeling, presentation, and advertising of foods, there is a specific prohibition on attributing prevention, treatment, or cure of a human disease or any reference to such properties to a food. However, there is a distinction made between prevention and significant reduction of a major disease risk factor, and the directive acknowledges that diet and certain foods are important for supporting and maintaining health and can affect certain disease risk factors.

Under the proposed regulation, health claims will only be allowed if the following information is included on the label: (i) a statement describing the importance of a varied and balanced diet and healthy lifestyle; (ii) the quantity of the food and pattern of consumption required to obtain the claimed beneficial heath effect: (iii) a statement addressed to people who should avoid the food, if appropriate; and (iv) a warning for products that may result in a health risk if consumed in excess. A reduction of risk claim must also include a statement that diseases may have many risk factors and that altering only one of these factors may or may not have a beneficial effect. Claims regarding a "slimming, slimness-producing, or weight-reducing" effect, those that refer to reducing hunger or increasing satiety, and those that claim a reduction in available energy from the diet would not be permitted.

The European Food Safety Authority (EFSA) is designated as the body that will be responsible for evaluating and authorizing health claims. EFSA would be required to make a decision within 6 months of receipt of an application. Once EFSA makes a decision, it would be forwarded to the Commission, which would be required to draft a decision within 3 months, followed by publication in the *Official Journal of the European Communities*. It is also anticipated that the Commission would establish and maintain a "community register of nutrition and health claims on food."

Nutrition and Health Claims at Codex

Many countries use decisions of the Codex Alimentarius Commission (Codex) as a basis for national regulations. Codex decisions are also used by the World Trade Organization (WTO) as the basis for resolution of trade disputes between nations. In May 2004, the Codex Committee on Food Labeling adopted draft guidelines for the use of nutrition and health claims; the commission officially adopted these draft guidelines in June 2004. Nutrition and health claims would not be allowed for foods for infants and young children unless specifically provided for in relevant Codex standards or national legislation.

Nutrition Claims

According to Codex, a nutrition claim "states, suggests, or implies that a food has particular nutritional properties, including but not limited to the energy and to the content of protein, fat, and carbohydrates, as well as the content of vitamins and minerals." The only nutrition claims permitted are those relating to energy, protein, carbohydrate, and fat and components thereof, fiber, sodium, and vitamins and minerals for which Nutrient Reference Values (NRVs) have been laid down in the Codex Guidelines for Nutrition Labelling. Two types of nutrition claims were defined. A nutrient content claim is defined as a nutrition claim that describes the level of a nutrient contained in a food—for example, "source of calcium," "high in fiber," and "low in fat." A nutrient comparative claim is one that compares the nutrient levels and/or energy value of two or more foods—for example, "reduced," "less than," "fewer," "increased," and "more than" (**Table 5**).

Table 5 Nutrient content claims at Codex

Food component	*Content claim*	*Requirements Not more than*
Energy	Low	40 kcal (170 kJ) per 100 g (solids) or 20 kcal (80 kJ) per 100 ml (liquids)
	Free	4 kcal per 100 ml (liquids)
Fat	Low	3 g per 100 g (solids), 1.5 g per 100 ml (liquids)
	Free	0.5 g per 100 g (solids) or 100 ml (liquids)
Saturated fat	Low[a]	1.5 g per 100 g (solids), 0.75 g per 100 ml (liquids) and 10% of energy
	Free	0.5 g per 100 g (solid) or 100 ml (liquids)
Cholesterol	Low[a]	0.02 g per 100 g (solids), 0.01 g per 100 ml (liquids)
	Free[a]	0.005 g per 100 g (solids), 0.005 g per 100 ml (liquids) and, for both claims, less than 1.5 g saturated fat per 100 g (solids), 0.75 g saturated fat per 100 ml (liquids) and 10% of energy of saturated fat
Sugars	Free	0.5 g per 100 g (solids), 0.5 g per 100 ml (liquids)
Sodium	Low	0.12 g per 100 g
	Very low	0.04 g per 100 g
	Free	0.005 g per 100 g
Protein	Source	10% of NRV per 100 g (solids), 5% of NRV per 100 ml (liquids) or 5% of NRV per 100 kcal (12% of NRV per 1 MJ) or 10% of NRV per serving
	High	2 times the values for 'source'
Vitamins and minerals	Source	15% of NRV per 100 g (solids), 7.5% of NRV per 100 ml (liquids) or 5% of NRV per 100 kcal (12% of NRV per 1 MJ) or 15% of NRV per serving
	High	2 times the values for 'source'

[a]*Trans* fatty acids should be taken into account where applicable.
NRV, nutrient reference value.
From Codex Alimentarius Commission (2004) *Joint FAO/WHO Food Standards Programme. Twenty-Seventh Session Rome, 28 June–3 July 2004. ALINORM 04/27/22 Report of the Thirty-Second Session of the Codex Committee on Food Labeling. Montréal, Canada, 10–14 May 2004.* Available at www.codexalimentarius.net.

Any comparisons are to be based on a relative difference of at least 25% in the energy value or macronutrient nutrient content and a 10% difference in the NRV for micronutrients between the compared foods. The use of the word "light" follows the same criteria as for "reduced" and should include an indication of the characteristics that make the food "light."

Health Claims

Health claims are defined by Codex as "any representation that states, suggests, or implies that a relationship exists between a food or a constituent of that food and health." Health claims must be based on the current relevant scientific data and be sufficient to substantiate the claimed effect and the relationship to health; Codex also envisions a re-review of claims as additional data may become available. Health claims should include information on the physiological role of the nutrient or the diet–health relationship as well as information on the composition of the product relevant to the physiological role of the nutrient or the accepted diet–health relationship. Codex defined three types of health claims: nutrient function claims, other function claims, and reduction of disease risk claims.

A nutrient function claim describes the physiological role of a nutrient in growth, development, and normal functions of the body. For example, "Nutrient A (naming a physiological role of nutrient A in the body in the maintenance of health and promotion of normal growth and development). Food X is a source of/high in nutrient A."

Other function claims convey specific beneficial effects of the consumption of foods or their constituents in the context of the total diet on normal functions or biological activities of the body. Such claims relate to a positive contribution to health, to the improvement of a function, or to modifying or preserving health. For example, "Substance A (naming the effect of substance A on improving or modifying a physiological function or biological activity associated with health). Food Y contains x grams of substance A."

Reduction of disease risk claims convey that consumption of a food or food constituent, in the context of the total diet, is related to a reduced risk of developing a disease or health-related condition. Risk reduction is defined as "significantly altering a major risk factor(s) for a disease or health-related condition." Because diseases have multiple risk factors and altering one of these risk factors may or may not have a beneficial effect, the risk reduction claims must have appropriate language and reference to other risk factors to ensure that consumers do not interpret them as prevention claims. For example, reduction of risk claims may include statements such as "A healthful diet low in nutrient or substance A may reduce the risk of disease D. Food X is low in nutrient or substance A" or "A healthful diet rich in nutrient or substance A may reduce the risk of disease D. Food X is high in nutrient or substance A."

Claims Related to Dietary Guidelines or Healthy Diets

Claims describing a food as part of a healthy diet should be permitted only if they are related to the pattern of eating contained in dietary guidelines officially recognized by the appropriate national authority, contain a statement relating the food to the pattern of eating to that in the guidelines, and are consistent with such guidelines. Such claims should not describe the food itself as healthy or imply that the food in and of itself will impart health. Foods that are allowed such claims should meet certain minimum criteria for major nutrients in the dietary guidelines. Codex stated that there should be some flexibility in the wording of such claims as long as they "remain faithful to the pattern of eating outlined in the dietary guidelines."

Conclusion

With the growing recognition of the connection between diet and health along with soaring health care costs, both consumers and governments have had great interest in capitalizing on the benefits of functional foods for health promotion. Although there is no standard accepted definition of functional food, most regulations and guidelines incorporate the concept that such foods, food components, and supplements provide a benefit to health beyond basic nutrition. There has been considerable progress by governments and Codex to develop systems to allow such products and to communicate their benefits to consumers through labeling claims. Some national regulatory agencies have laid down scientific standards for demonstrating the safety and efficacy of functional foods for different types of nutrition and health claims. Regulatory agencies are especially concerned that claims on foods or supplements are truthful and not misleading and aid consumers in making informed choices for health promotion. In the future, it is expected that functional foods will gain in importance, and regulatory agencies will need to continue to ensure that any claims on such products meet consumer needs.

See also: **Bioavailability**. **Food Composition Data**. **Food Fortification**: Developed Countries; Developing Countries. **Fruits and Vegetables**. **Functional Foods**: Health Effects and Clinical Applications. **Supplementation**: Dietary Supplements; Role of Micronutrient Supplementation; Developing Countries; Developed Countries.

Further Reading

American Dietetic Association (2004) Position of the American Dietetic Association: Functional foods. *Journal of the American Dietetic Association*, **104**,: 814–826.

Ashwell M (2002) *Concepts of Functional Foods. ILSI Europe Concise Monograph Series*. Brussels: International Life Sciences Institute. Available at www.ilsi.org.

Codex Alimentarius Commission (2004) *Joint FAO/WHO Food Standards Programme. Twenty-Seventh Session Rome, 28 June–3 July 2004. ALINORM 04/27/22 Report of the Thirty-Second Session of the Codex Committee on Food Labelling. Montréal, Canada, 10–14 May 2004*. Available at www.codexalimentarius.net.

European Commission (2004) *Proposal for a Regulation of the European Parliament and of the Council on Nutrition and Health Claims Made in Foods*, 11028/04. Brussels: European Commission.

Food and Drug Administration, Center for Food Safety and Applied Nutrition (2004) *Label Claims*. Available at www.cfsan.fda.gov/~dms/lab-hlth.html.

Hasler CM (2002) Functional foods: Benefits, concerns and challenges—A position paper from the American Council on Science and Health. *Journal of Nutrition* **132**: 3772–3781.

Health Canada (2001) Product-specific authorization of health claims for foods. A proposed regulatory framework. Bureau of Nutritional Sciences. Food Directorate. Health Products and Food Branch. October, 2001.

ILSI North America Technical Committee on Food Components for Health Promotion (1999) Safety assessment and potential health benefits of food components based on selected scientific criteria. *Critical Reviews in Food Science and Nutrition* **39**(3): 203–316.

ILSI North America Technical Committee on Food Components for Health Promotion (2002) Scientific criteria for evaluating health effects of food components. *Critical Reviews in Food Science and Nutrition* **42**(supplement): 651–676.

Institute of Food Technologists (2004) *Expert Report on Functional Foods* (Draft). Chicago Institute of Food Technologists.

International Food Information Council (2002) *International Food Information Council: Functional Foods Attitudinal Research*. Available at www.ific.org/research/funcfoodres00.cfm.

Sloan AE (2004) The top ten functional food trends 2004. *Food Technology* **58**(4): 28–51.

G

GALACTOSE

A Abi-Hanna and J M Saavedra,
Johns Hopkins School of Medicine, Baltimore, MD, USA

This article is reproduced from the previous edition, pp. 915–922, © 1999, Elsevier Ltd.

Lactose, a disaccharide composed of glucose and galactose, is the principal sugar of mammalian milk and the principal carbohydrate energy source for infants and children; thus galactose plays a central metabolic role in human nutrition. Lactose is hydrolyzed in the intestine into glucose and galactose, which together with other sources of these monosaccharides are absorbed and metabolized and used as energy. Galactose additionally is an important constituent of complex polysaccharides, galactolipids, and other glycoconjugates of structural and functional importance. Both absorptive as well as metabolic defects affecting galactose have been described.

Dietary Sources of Galactose

Lactose is by far the most abundant source of galactose in the diet of most humans. However, lactose can also be found in a considerable number of sources. These include drugs and medications, which use lactose as an excipient, in part because of its excellent tablet-forming capacity. Additionally, small amounts of galactose can be present in many fruits and vegetables, and considerable amounts can also be found in legumes (beans and peas) and in other food plants. Galactose polysaccharides with various glycolytic linkages such as $\alpha(1–6)$, $\beta(1–3)$, and $\beta(1–4)$ are ubiquitous in animals and plants. The bioavailability of galactose in these linkages found in foods is not well known. Some galactosidasis in plants can liberate galactose, and foods fermented by microorganisms for preparation or preservation may also contain free galactose. The role of free and bound galactose in cereals, fruits, legumes, nuts, and other vegetables may contribute to sources of galactose that are not readily obvious. Bound galactose is also present in raffinose oligosaccharides and other sugars.

Galactose Absorption

Lactose is hydrolyzed in the intestine by the enzyme lactase-phlorizin hydrolase to glucose and galactose. In humans, D-glucose and D-galactose are the only nutritionally significant monosaccharides that are actively absorbed. Although glucose and galactose can cross the intestinal mucosa down a diffusion gradient, the slowness of this method is such that water would diffuse in the opposite direction leading to a lessening in the concentration gradient. Thus, a rapid transport mechanism exists for glucose and galactose, particularly in infants. The 'coupled carrier' hypothesis is generally accepted as the main mechanism. In the small intestine and proximal tubule of the kidney, D-glucose is absorbed by epithelial cells via a sodium-dependent cotransport system existing at the luminal membrane level and a sodium-independent transport system at the basolateral membrane level. It is suggested that the potential difference across the brush border membrane of the cell also plays an important role in the mechanism which concentrates sugar in the cell.

The genetic functional defects of this cotransport system are expressed in two main clinical entities; selective congenital glucose and galactose malabsorption by the intestine discussed below, and familial renal glycosuria. Once galactose is absorbed, it must be converted to glucose for utilization. This occurs primarily by the pathways explained below. Three distinct enzymatic defects are responsible for the conditions generally described as galactosemia.

Glucose–Galactose Malabsorption

Pathophysiology and Clinical Manifestation

Glucose and galactose malabsorption is a rare congenital disease resulting from a selective defect in the

intestinal transport of glucose and galactose. It is characterized by the neonatal onset of severe, watery, acidic diarrhea. The diarrhea is profuse and contains sugar. In children given lactose, fecal sugar mainly consists of glucose and galactose with only small amounts of lactose, since lactase activity is usually adequate. Hyperosmotic dehydration and metabolic acidosis are the rule. Related gastrointestinal signs and symptoms include increase of abdominal gas, distension, and vomiting. Intermittent or permanent glycosuria after fasting or after a glucose load is frequent. Thus the combination of reducing sugar in the stool and slight glycosuria despite low blood glucose levels is highly suggestive of glucose–galactose malabsorption.

The major characteristic in glucose–galactose malabsorption is the lack of intracellular glucose or galactose accumulation against a concentration gradient. The transport of other molecules such as alanine or leucine via a sodium cotransporter is typically intact.

The abnormality of carbohydrate metabolism is confined to glucose transport in the small intestine and the proximal renal tube. The main defect appears to be the absence of a functional sodium-dependent glucose contransporter. Electrolytes can be secreted in the jejunal mucosa together with fluids, suggesting that the combined glucose–sodium water absorption process is effective. Sucrose can undergo normal hydrolysis and fructose can be absorbed typically without problems. Glucose entry into the erythrocytes is normal and so are fasting blood glucose levels. Oral glucose tolerance tests usually yield a flat glucose curve while breath hydrogen tests done separately for glucose and galactose are consistent with malabsorption.

The functionality of the cotransporter at the brush border membrane is either absent or reduced. Additionally, the participation of a mutarotase in sugar transport has recently been suggested in the absence of this enzyme and has been demonstrated in glucose–galactose malabsorption. However, full understanding of this condition requires additional information on the liquid composition of the membrane and on other characteristics and genetic control of this transport system.

Diagnosis

Children affected with glucose–galactose malabsorption are of diverse origin. There is high consanguinity rate and no clear-cut vertical transmission, suggesting an autosomal recessive mode of inheritance.

The diagnosis can be established by a clinical history of watery diarrhea with glucose–water solution or milk and rapid cessation of the problem when these are discontinued. Oral glucose or galactose tolerance tests and breath hydrogen analysis can aid in the diagnosis. The differential diagnosis includes congenital lactase deficiency, sucrose–isomaltose deficiency, and congenital chloride-secreting diarrhea. Most other monosaccharide malabsorption and intolerance is secondary to mucosal injury and responds to adequate nutritional management with complete resolution.

Management

Treatment consists of immediate rehydration, adequate maintenance of hydration, and initiation of a glucose- and galactose-free diet. Since fructose is tolerated, most of the carbohydrate initially can be given as fructose, using other dietary modular products of protein and fat as well as micronutrients.

Biochemistry and Physiology of Galactose

The main pathway of galactose metabolism in humans is the conversion of galactose to glucose, without disruption of the carbon skeleton. The name 'galactosemia' has been associated with a syndrome of toxicity associated with the administration of galactose to patients with an inherited disorder of galactose utilization, leading to multiple clinical manifestations, including malnutrition, mental retardation, liver disease, and cataracts. The clinical manifestations are linked to specific enzymatic defects. Thus the term 'galactosemia' should be qualified by the specific defect. Three enzymatic steps are required to metabolize galactose to UDP-glucose. Two alternate pathways, oxidation and reduction, are used in the absence of enzymes of the main route.

Step 1: Galactokinase

Galactose is phosphorylated by galactokinase with ATP to form galactose 1-phosphate. The equilibrium is far in the direction of sugar phosphorylation, but the reaction is reversible. Galactokinase has been studied in detail in human red cells, leucocytes, fibroblasts, placenta, liver, and various human fetal tissues. It is detectable in fetal liver from 10 weeks of gestation onwards and the activity of the enzyme in liver and red cells is higher in the second and third trimester. Its activity is higher in red blood cells from human infants than in cells from adults, and in reticulocytes than with aged

red cells. Cultured human fibroblasts show enhanced galactokinase activity when grown in the presence of galactose, whereas in the liver the activity does not appear to be regulated by dietary galactose. The red cell enzyme, like that of the liver, undergoes substrate and product inhibition.

The assignment of the gene for galactokinase has been made to human chromosome 17, and its regional localization of the chromosome has been assigned to band q21–22.

Step 2: Transferase

Galactose 1-phosphate reacts with UDP-glucose to produce UDP-galactose and glucose 1-phosphate. This step is catalyzed by galactose-1-phosphate uridyltransferase, an enzyme present in bacteria and most mammalian tissues. Like galactokinase, galactose-1-phosphate uridyltransferase is detectable in fetal liver from 10 weeks of gestation, with the liver enzyme-specific activity being highest at 28 weeks of gestation. The rate of reaction may be regulated by substrate concentration and limited by UDP-glucose substrate inhibition of transferase. Glucose 1-phosphate is a potent inhibitor of the enzyme. Uridine nucleotides such as uridine di- and triphosphate are powerful competitive inhibitors of substrate UDP-glucose.

Galactose-1-phosphate uridyltransferase deficiency is the most commonly reported defect in galactosemic patients. In the young infant galactose is a major energy source and its metabolism to glucose 1-phosphate is essential, but this is not the case in the fetus in whom glucose is the main energy source. However, the metabolism of galactose in the fetus is important to prevent accumulation of toxic galactose metabolites. Thus in galactose-1-phosphate uridyl-transferase deficiency the fetus could be at a disadvantage as early as the 10th week of gestation. Dietary and hormonal influences on the liver enzyme have not been reported. In the rat a galactose-rich diet increases transferase activity.

Galactose-1-phosphate uridyltransferase is localized on chromosome 9p13. At least 32 variants in the nucleotide sequence of the galactose-1-phosphate uridyltransferase gene have been identified, with the most frequent being change in amino acid codon position 188 in which an arginine is substituted for a glutamine, the Q188R mutation. This Q188R mutation is associated with 'classical' galactosemia with virtually no galactose-1-phosphate uridyl-transferase activity detectable. However, there are other variant forms of the enzyme which have diminished but detectable activity, known as Duarte, Indiana, Rennes, Los Angeles, Münster, and Chicago. Heterozygotes for normal and Duarte alleles are presumed to have 75% of normal galactose-1-phosphate uridyltransferase activity. Homozygotes for the Duarte allele could have 50% activity, and compound heterozygotes for the Duarte allele and the classical galactosemia allele have 25% activity in peripheral erythrocytes.

Step 3: Epimerase

The UDP-galactose is converted to UDP-glucose by UDP-galactose 4′-epimerase. The UDP-glucose thus formed can then enter the reaction again in a cyclical fashion until all the free galactose coming into the pathway is converted to glucose 1-phosphate. This enzyme is responsible for the inversion of the hydroxyl group at the C-4 carbon of the hexose chain to form glucose from galactose; it is also important for the conversion of UDP-glucose to UDP-galactose when only glucose is available and galactose is required as a constituent of complex polysaccharides. The epimerase maintains a cellular equilibrium of UDP-glucose to UDP-galactose in a ratio of about 3:1.

The purified enzyme is a dimer of identical subunits that consists of a mixture of catalytically active subunits (epimerase-NAD^+) and inactive subunits (epimerase-NADH-uridine nucleotide). The NAD binds to the enzyme and induces a conformational change resulting in enzymatic activity. For liver enzyme activity, exogenous NAD is required and NADH is a potent inhibitor of the enzyme. Any process disturbing the NAD/NADH ratio, such as ethanol metabolism which generates NADH, will impair galactose utilization. Cellular levels of UDP-glucose and other uridine nucleotides may also exert rate-regulating effects. Cells not exposed to free galactose form the sugar from glucose in adequate amounts to satisfy normal growth and development. Epimerase activity of the intestinal mucosa increases with age, whereas human red cells have a higher activity in newborns than adults. The intestinal enzyme activity can be enhanced by feeding diets high in glucose or galactose content. Less information is available on fetal levels of UDP-galactose 4′-epimerase, but one fetus of 16 weeks' gestation had liver enzyme activity comparable with that of children and adults. In epimerase deficiency, when the amount of entering galactose is low, an elevated level of galactose 1-phosphate in red blood cells may be reduced to normal but the UDP-galactose level stays elevated. The gene for epimerase has been assigned to human chromosome 1.

Alternative Pathway: Reduction

The polyol pathway was first identified in placenta and seminal vesicles and is responsible for the fructose content of seminal fluid. Two enzymatic reactions involving aldose reductase and sorbitol dehydrogenase catalyze the conversion of glucose to fructose with sorbitol as the intermediate. In certain cells, such as renal collecting duct cells, retinal pigment epithelial cells, and renal glomerular endothelium, and under certain conditions, aldose reductase functions to produce sorbitol which acts as an intracellular osmolyte. The acyclic polyols such as sorbitol, galactitol, and mannitol are the end product of metabolism and have osmotic properties. The presence of galactitol in the urine and plasma of patients with transferase-, galactokinase-, and epimerase-deficiency galactosemia is suggestive of the importance of the reduction of galactose as an alternative pathway. However, the high K_m of this enzyme indicates that reduction will occur only when galactose levels in tissues are very high.

Patients with classical galactosemia have markedly elevated levels of galactitol in plasma and urine, which remain above age-matched control levels after treatment with galactose-free diet, whereas high urinary galactose levels return to normal in all patients. Aldose reductase has been localized to the Schwann cells of peripheral nerves and to renal paptillae cells. Kinetic studies suggest that neither glucose nor galactose are preferred substrates. Only when tissue levels of galactose are much elevated would reduction be important. Aldose reductase activity of lens and other tissue is stimulated by sulfate ions and ATP and is inhibited by various keto acids, fatty acids, and ADP. Increased production of galactitol is felt to play an important role in the pathogenesis of cataracts in the infant with galactose-1-phosphate uridyltransferase, galactokinase, and UDP-galactose epimerase deficiency. The toxicity of polyols in the ocular lens is probably related to their ability to act as osmotically active particles within the lens cells, which leads to accumulation of water and eventually cell dysfunction.

Cataracts are the primary manifestation of disease in untreated patients with galactokinase deficiency, who manifest accumulation of galactitol but not galactose 1-phosphate in tissues. Thus the galactose 1-phosphate and not galactitol toxicity is probably a necessary mediator in both transferase and epimerase deficiencies for expression of hepatic disease, renal tubular dysfunction, and increased red blood cell turnover.

Alternative Pathway: Oxidation

In the absence of galactose-1-phosphate uridyltransferase activity, galactose 1-phosphate and galactose accumulate behind the block. The second alternate pathway, besides reduction of galactose to sugar alcohol, galactitol, is the oxidation of galactose to sugar acid, galactonate. Galactonate, for example, appears in the urine of transferase-deficient individuals. Galactonate can be further metabolized to xylulose, a sugar capable of further metabolism. This pathway accounts for about 50% of oxidation of galactose by galactosaemic patients. Patients with transferase-deficient galactosemia excrete galactonate in urine after galactose is administered, and galactonate has been found in the liver of a transferase-deficient subject.

Disorders of Galactose Metabolism

Clinical Manifestations

Galactose is an important constituent of the complex polysaccharides which are part of cell glycoconjugates, key elements of immunologic determinants, hormones, cell membranes structures, endogenous animal lectins, and numerous other glycoproteins. In addition galactose is incorporated in galactolipids, important structure elements of the central nervous system. It is not difficult to assume that the abnormal galactose metabolism in galactosemic patients could have profound and widespread effects on glycoconjugate structures and their biological function.

Classically, the term 'galactosemia' was associated with an inherited disorder of galactose utilization characterized by malnutrition, liver disease, cataracts, and mental retardation, resulting from the specific deficiency of galactose-1-phosphate uridyltransferase. However, other enzymatic defects with variations of clinical presentation can also lead to galactosemia (**Table 1**). Thus it is preferably better to refer to these abnormalities of metabolism by the specific enzymatic deficiencies which are described below.

Transferase deficiency Failure to thrive is the most common initial clinical sign of galactose-1-phosphate uridyltransferase deficiency, and it is present in all cases. Vomiting or diarrhea is present in almost all patients, usually starting within a few days of milk ingestion. Jaundice, hepatomegaly, or both are present almost as frequently after the first week of life. The jaundice of intrinsic liver disease may be accentuated by severe hemolysis in some

Table 1 Disorders of galactose metabolism

Enzyme deficiency	*Primary clinical manifestations*
Galactose-1-phosphate uridyltransferase	Failure to thrive Emesis/diarrhea Jaundice, hepatomegaly Cataracts Galactosuria Gonadal dysfunction Developmental delay, neurologic symptoms Cataracts
Galactokinase	Similar manifestations as transferase deficiency, but with no liver, kidney, or gonadal dysfunction
UDP-galactose 4′-epimerase	Mostly asymptomatic Rarely same manifestations as transferase deficiency but with no gonadal dysfunction

patients. Abnormal liver function tests and ascites may develop. The reason for liver toxicity remains obscure. The liver of affected patients has a characteristic acinar formation, and liver biopsy on occasion has been helpful in establishing the diagnosis. There is high frequency of neonatal death due to *Escherichia coli* sepsis, possibly caused by the inhibition of leucocyte bactericidal activity.

Galactose 1-phosphate and galactitol have been detected in the kidneys of patients with galactosemia. Renal toxicity may manifest as renal tubular dysfunction and a defect in urine acidification mechanisms. Galactosuria, hyperchloremic acidosis, albuminuria, and aminoaciduria may also occur. Hyperchloremic acidosis could be also secondary to the gastrointestinal disturbance and poor food intake. Galactosuria may be intermittent, depending on oral intake, and can disappear within 3–4 days with the use of intravenous glucose. The finding of urinary reducing substances which do not react in a glucose oxidase test should raise the suspicion of galactosemia. This finding, however, does not establish the diagnosis, since galactosuria can also occur in intestinal lactase deficiency and in severe liver disease due to other causes.

Ovarian atrophy appears to be an important manifestation of galactose toxicity, with clinical and biochemical evidence of ovarian dysfunction present in nearly all affected females. The basis of the toxicity has not been defined. The consequences of the gonadal dysfunction range from failure of pubertal development, through primary amenorrhea to secondary amenorrhea or premature menopause (75–76% of affected females). Although gonadal function has been described as early as infancy based on elevations of follicle stimulating hormones (FSH) and abnormal stimulation testing, no predisposing factor for gonadal dysfunction can be found. Previous recommendations that dietary lactose restriction from birth may be beneficial have in fact not prevented gonadal dysfunction. In the galactosemic male, a complete understanding of gonadal dysfunction has not yet been described. The majority—but not all—of male galactosemic patients had normal pubertal development, and a few individuals have been found to have normal semen.

Cataracts have been observed within a few days of birth. These may be found only on slit-lamp examination and can be missed with an ophthalmoscope, since they consist of punctate lesions in the fetal lens nucleus. Several hypotheses have been postulated to account for their formation and are mentioned above. It seems conclusive that the initiator of the process in rats is galactitol and not galactose 1-phosphate. Galactose 1-phosphate accumulates only late in the process and is absent in patients with galactokinase deficiency who present with cataracts.

Development of mental retardation may be apparent after the first months of life. Signs of increased intracranial pressure and cerebral oedema have been observed as a presenting feature.

Many of the toxicity symptoms can rapidly resolve with institution of dietary lactose restriction. However, a substantial percentage of children have subnormal IQs and speech and language deficits, but rarely devastating neurological sequelae. Most galactosemic patients with lactose restriction are deficient of cognitive functioning in one or more areas. The deficits are variable and do not appear to be related to the age, diagnosis, or the severity of illness at presentation. The pathophysiology of these impairments in galactosemia remains unknown. Several hypotheses are suggested, including toxic oedema due to increased brain galactitol concentrations, changes in the second messenger pathway, and changes of the energy status of the brain.

Galactokinase deficiency Galactokinase deficiency is characterized by the occurrence of cataracts without liver, kidney, or ovarian dysfunction and no increased risk of infections. A number of infants are reported to have pseudotumor cerebri, with very rare neurological involvement, suggesting that retardation is not a feature. The absence of liver and kidney damage in galactokinase deficiency and the presence of damage to these organs in transferase deficiency make it likely that toxicity in the later condition is in some way associated with galactose 1-phosphate formation.

Epimerase deficiency Elevated red cell levels of galactose 1-phosphate with absence of UDP-galactose 4′-epimerase have been described in a patient with normal growth, development, and normal ability to metabolize ingested galactose. Several cases of biochemical deficiency have been described but symptomatic cases are extremely rare. A few had cataracts, sepsis, liver, kidney, and brain abnormalities, including a few with neurosensory deafness. There appears to be no ovarian dysfunction. The absence of ovarian dysfunction suggests that elevated UDP-galactose levels may protect the ovary from damage observed in transferase deficiency. Screening programs have been established in Japan, where the incidence is reported to be 1 in 23 000. In epimerase deficiency, when dietary galactose is low, galactose 1-phosphate concentrations in red blood cells may be reduced to normal, but UDP-galactose concentrations remain elevated. Despite the many phenotypic similarities between transferase and epimerase deficiency, the latter is characterized by elevated red cell levels of UDP-galactose even with modest galactose intake.

Diagnosis

The presence of reducing substance in urine which does not react with glucose oxidase reagents is consistent with galactosuria; however, occasionally some infants (particularly premature babies) also develop galactosuria. It is important to note that the presence of lactose, fructose, and pentose in the urine may give the same results. The presence of cataracts in infants without other systemic symptoms suggests the possibility of galactokinase deficiency. The presence of cataracts in older patients with the absence of gastrointestinal dysfunction or failure to thrive in galactosemic patients helps to differentiate between galactokinase deficiency and transferase deficiency.

The diagnosis of transferase deficiency is suggested by abnormally high amounts of red cell galactose 1-phosphate and confirmed by direct assay of red cell transferase activity. The red cell UDP-glucose consumption test may help to differentiate homozygous patients with a complete absence of transferase in red cells from heterozygous patients who have intermediate levels. Normal red cell values are of 6 mmol UDP-glucose consumed per hour per millilitre of red blood cells. In galactokinase deficiency the diagnosis can be made by the presence of normal amounts of galactose-1-phosphate uridyltransferase and the absence of galactokinase in the red blood cells.

Galactose-1-phosphate uridyltransferase deficiency can be diagnosed prenatally, by assay of galactose-1-phosphate uridyltransferase activity in cultured amniotic fluid cells or chorionic villi, and by galactitol measurement in amniotic fluid supernatant. The perinatal diagnosis is undertaken rarely, because the transferase deficiency is seen as a treatable condition.

Methods for mass screening of newborns for galactosemia are available, although galactosemia is rare. The incidence in Norway is 1 in 96 000, in Sweden 1 in 81 000, in the USA 1 in 62 000, in Switzerland 1 in 58 000, in Germany 1 in 40 000, and the worldwide incidence is about 1 in 70 000. Newborn screening has not been introduced in Great Britain, the Netherlands, or in some states of the USA. Most newborn screening programs designed for the detection of anomalies of galactose metabolism use tests to measure either blood galactose or the activity of galactose-1-phosphate uridyltransferase. Beutler and Baluda developed a fluorescence test in which the activity of uridyltransferase in the dried blood spot is required for the reduction of NADP, yielding fluorescence under long-wave ultraviolet light; the intensity of the fluorescence corresponds to the activity of uridyltransferase. The main advantage of this test is that it can be completed in short time, although false positive results do occur. The disadvantage of this test is that patients with galactokinase deficiency are not detected by this method. Guthrie and Paigen described a more efficient test using the principle of metabolite inhibition; galactose inhibited the growth of an *E. coli* mutant strain lacking uridyltransferase. Later, Paigen used an *E. coli* mutant strain which lacks UDP-galactose 4′-cpimerase activity. Using the Paigen test it is possible to detect galactokinase and uridyltransferase deficiencies. Epimerase deficiency can also be detected by the Paigen test if alkaline phosphatase is added to hydrolyze galactose phosphatase. In many screening laboratories the Beutler test is combined with the microbiological Paigen test.

Management

A galactose-free diet is the current treatment for galactosemia. It is important to know that galactose is present not only in milk but in other sources of food. A strict galactose-free diet in galactosemic patients with transferase deficiency is not harmful. The quality of the galactose-free diet and patient compliance are usually monitored by measuring free galactose in plasma and galactose 1-phosphate in erythrocytes.

Growth retardation, cognitive impairment, speech impediment, tremor, ataxia, and ovarian failure are frequent complications in spite of a strict galactose-free diet. Elevated galactose phosphate levels may occur in erythrocytes of even well-treated galactosemic patients. This elevation is attributed to endogenous production of the metabolite. A galactose-free diet is recommended from birth. It is recommended to restrict galactose in the diet of pregnant mothers diagnosed perinatally with transferase deficiency; a galactose-free diet should be started as soon as the diagnosis is made in the infant regardless of any preexisting manifestation of toxicity. The strict galactose-free diet will cause regression of symptoms and findings. It is important for the families to be aware of the high incidence of verbal dyspraxia even on a very strict diet. The speech intervention program and language stimulation are recommended as early as the first year of life. Many patients with normal IQ values who were treated from birth have learning disabilities, speech and language deficit, and psychological problems. Neurological sequelae have been described also in patients on strict galactose-free diets. These sequelae include cerebellar ataxia, tremor, choreoathetosis, and encephalopathy. Gonadal dysfunction in female galactosemic patients is an almost universal finding, even with a strict galactose-free diet. There is no current therapy for ovarian dysfunction except palliative replacement of oestrogen and progesterone. This is suggested in galactosemic females to develop secondary sexual characters and establish regular menses. There is no universal recommendation for the management of newborns screened positive nor for galactosemic heterozygotic patients.

In patients with epimerase deficiency, UDP-glucose cannot be converted to UDP-galactose. Thus a complete absence of galactose from the diet and the lack of formation of UDP-galactose via transferase would have serious consequences. There would be an inability to form complex polysaccharides and an inability to provide an adequate galactose component for brain cerebrosides. The treatment of epimerase deficiency relies on providing a small amount of dietary galactose.

See also: **Early Origins of Disease**: Fetal. **Glucose**: Chemistry and Dietary Sources; Metabolism and Maintenance of Blood Glucose Level; Glucose Tolerance. **Glycemic Index**. **Inborn Errors of Metabolism**: Classification and Biochemical Aspects. **Liver Disorders**.

Further Reading

Acosta PB and Gross KC (1995) Hidden sources of galactose in the environment. *European Journal of Paediatrics* **154**(supplement 2): S87–92.

Berry GT (1995) The role of polyols in the pathophysiology of hypergalactosemia. *European Journal of Paediatrics* **154**(supplement 2): S53–64.

Beutler E and Baluda M (1996) Biochemical properties of the human red cell galactose-1-phosphate uridyl transferase (UDP Glucose: alpha-D Galactose-phosphate uridyl transferase) from normal and mutant subjects. *Laboratory and Clinical Medicine* **67**: 947.

Gitzelmann R and Boshard NU (1995) Partial deficiency of galactose-1-phosphate uridyltransferase. *European Journal of Paediatrics* **154**(supplement 2): S40–44.

Jakobs C, Vleijer W, Allen Y, and Holton JB (1995) Prenatal diagnosis of galactosemia. *European Journal of Paediatrics* **154**(supplement 2): S33–36.

Jakobs C, Schweitzer S, and Dorland B (1995) Galactitol in galactosemia. *European Journal of Paediatrics* **154**(supplement 2): S50–52.

Kaufman FR, McBride-Chang C, Morris F, Wolf J, and Nelson M (1995) Cognitive functioning, neurologic starters and the brain imaging in galactosemia. *European Journal of Paediatrics* **154**(supplement 2): 2–5.

Liu Y, Vanhooke JL, and Perry PA (1996) UDP-galactose 4-epimerase: NAD^+ content and a charge-transfer band associated with the substrate-induced conformational transition. *Biochemistry* **35**(23): 7615–7620.

Ng WG, Xu YK, Kauffman DR, and Donnell GN (1989) Deficit of uridine diphosphate galactose in galactosemia. *Journal of Inherited and Metabolic Disease* **12**: 257–266.

Sagal S (1995) Defective galactosylation in galactosemia; low cell UDP galactose an explanation? *European Journal of Paediatrics* **154**(supplement 2): S65–71.

Segal S (1989) Disorders of galactose metabolism. In: Scriver CR, Beaudet AL, Sly WS, and Vale D (eds.) *The Metabolic Basis of Inherited Disease*, 6th edn, pp. 453–480. New York: McGraw-Hill.

Segal S (1995) Galactosemia unsolved. *European Journal of Paediatrics* **154**(supplement 2): S97–102.

Schweitzer S (1995) Newborn mass screening of galactosemia. *European Journal of Paediatrics* **154**(supplement 2): S37–39.

GALL BLADDER DISORDERS

B Nejadnik and L Cheskin, Johns Hopkins University, Baltimore, MD, USA

The incidence of gall bladder disease in the United States exceeds 20 million cases annually. The total cost of gall bladder disease in the United States is more than that of any other gastrointestinal illness, including colorectal cancer and peptic ulcer disease.

Gall bladder disease has an intimate relationship with diet and nutrition. On the one hand, there has been a great deal of interest in the role of dietary constituents and nutritional habits in the etiology of gall bladder disease. On the other hand, an individual's nutritional status has a direct impact on the risk of acquiring gall stones. For instance, obesity as well as rapid weight loss and total parenteral nutrition (TPN) predispose the patient to a higher risk of gall bladder disease. Of course, there are multiple other risk factors that play an important role in the formation and manifestations of gall bladder disease, including female gender, family history of first-degree relative with gall stone disease, pregnancy, and drug use. Approximately three-fourths of gall stones detected in the general population are cholesterol gall stones.

According to a large study of the Danish population published in 1991, the 5-year incidence of gall stones in men and women aged 30 years was 0.3% and 1.4%, respectively. At age 60, the incidence had increased to 3.3% and 3.7% for men and women, respectively. It seems that the difference in incidence between men and women disappears with age. This may be related to the difference in estrogenic hormones between the two genders, which follows the same pattern.

Normal Biliary Physiology

Bile, which is formed in the hepatic lobules, is secreted into the canaliculi, small bile ductules, and larger bile ducts that drain it into portal tracts. Interlobular bile ducts join to form larger septal bile ducts that coalesce to form the right and left hepatic ducts, which in turn join to form the common hepatic duct. The common hepatic duct is joined by the cystic duct of the gall bladder to form the common bile duct, which enters the duodenum (often after joining the main pancreatic duct) through the ampulla of Vater.

The largest bile components are water (82%), bile acids (12%), lecithin and other phospholipids (4%), and unesterified cholesterol (0.7%). The total daily basal secretion of hepatic bile is approximately 500–600 ml. The primary bile acids, cholic acid and chenodeoxycholic acid, are synthesized from cholesterol in the liver, conjugated with glycine or taurine, and excreted into the bile. Secondary bile acids, including deoxycholate and lithocholate, are formed in the colon as bacterial metabolites of the primary bile acids. The normal bile acid pool size is approximately 2–4 g. Bile salts play an important role not only in facilitating the biliary excretion of cholesterol but also in intestinal absorption of dietary fats. They are absorbed passively in the entire gut and actively absorbed by the terminal ileum. The bile acid pool circulates approximately 5–10 times daily.

Cholesterol is poorly soluble in water, and its solubility in bile depends on its lipid concentration and the quantity of bile acids and lecithin. Usually, cholesterol is solubilized and forms mixed micelles. Supersaturation of cholesterol provokes the precipitation of cholesterol crystals in bile.

Cholecystokinin (CCK) is the most powerful stimulator of gall bladder contraction. It is released from the duodenal mucosa in response to the ingestion of fats and amino acids. CCK plays its role through contraction of the gall bladder, reducing resistance of the sphincter of Oddi, increasing hepatic secretion of bile, and therefore enhancing flow of biliary contents into the duodenum.

Pathophysiology of Stone Formation

Morphology and Composition

There are three kinds of gall stone: cholesterol, black pigment, or brown pigment stones. Cholesterol stones constitute 75–90% of all gall stones. They are composed purely of cholesterol or have cholesterol as the major chemical constituent. Most cholesterol gall stones are of mixed composition. Pigmented stones get their color and their name from precipitated bilirubin. Increased production of unconjugated bilirubin causes black pigmentation. Formation of black pigment stones is typically associated with chronic hemolysis, cirrhosis, and pancreatitis. Brown pigment stones are usually associated with infection. Cytoskeletons of bacteria can be seen microscopically in brown pigment stones,

and bacterial infection seems to be a prerequisite for brown stone formation.

Pathogenesis

Three factors have been recognized in gall stone formation: cholesterol supersaturation, accelerated nucleation, and gall bladder hypomotility. Among these, the degree of cholesterol saturation in gall bladder bile is the most important factor in crystal formation.

Cholesterol supersaturation Cholesterol is hydrophobic and not easily soluble in water. Its solubility is dependent on the presence of bile salts and lecithin. It is easy to imagine that as the ratio of cholesterol to bile salts and lecithin increases, cholesterol precipitation, crystal formation, and therefore stone formation ensue.

Nucleating and antinucleating factors Pronucleators include mucin glycoproteins, immunoglobulin G (IgG) and IgM, aminopeptidase N, haptoglobin, and α_1 acid glycoprotein; the most prominent of these is mucin glycoproteins. The hydrophobic centers of these proteins can bind to cholesterol, phospholipids, and bilirubin.

Gall bladder hypomotility The gall bladder concentrates and acidifies the bile. The most powerful stimulant of gall bladder contraction is CCK. CCK release is stimulated by (in order of decreasing potency) long-chain fatty acids, amino acids, and carbohydrates.

Risk Factors Associated with Cholesterol Gall Stone Formation

Major risk factors predisposing to gall stones are age, sex, genetic profile, nutritional status (including the route of nutrition), hormones, drugs, and some other diseases such as diseases of the terminal ileum. A summary of these elements is provided in **Table 1**.

Table 1 Risk factors associated with cholesterol gall stone formation

Age
Female gender
Genetics
Prima Indians
Chileans
Family history of gall stone
Pregnancy
Small bowel diseases
Crohn's disease
Terminal ileum resection
Drugs
Estrogens
Ceftriaxone
Lipid-lowering agents (Clofibrate)
Octreotide
Nutritional status
Obesity
Rapid weight loss
Total parenteral nutrition
Diabetes
Other conditions
Immobility
Cirrhosis
Spinal cord injury

Age, sex, and genetic profile As mentioned previously, women are affected more than men, and the incidence of gall stone increases with age. A positive family history of gall bladder disease increases risk to more than twice that of the general population. Native Americans and Scandinavians are more predisposed to this disease than other ethnic groups.

Obesity, weight loss, and total parenteral nutrition Obesity is a well-known risk factor for cholelithiasis. A large prospective study of obese women found a strong linear association between body mass index and the reported incidence of cholelithiasis. In this study, those with the highest body mass index ($>45\,kg/m^2$) had a 7-fold increased risk of development of gall stones compared to non-obese controls. This relationship is somewhat weaker in men than in women. The association between obesity and gall stone formation may result from increased secretion of cholesterol into the bile as a result of higher 3-hydroxy-3-methylglutaryl-coenzyme A (HMG-CoA).

In studies of gall bladder motility in obese patients, no impairment in gall bladder contraction has been documented. Abnormal processing of the cholecystokinin receptor gene has been reported in one obese patient who had gall stones. Such an abnormality could lead to gall bladder stasis and ultimately to cholelithiasis.

Rapid weight loss is a recognized risk factor for cholesterol gall stone formation. As many as 30% of obese patients on restricted calorie intake may develop (usually asymptomatic) gall stones. This rate is higher, up to 50%, for obese patients who undergo gastric bypass surgery. It has been shown that hepatic cholesterol secretion increases in patients with low calorie intake. Other predisposing factors for the same patients are increased mucin secretion and decreased gall bladder motility. Gall stone formation may be prevented in this high-risk population possibly through prophylactic administration of a bile salt, ursodeoxycholic acid.

Low-fat diet by itself seems to be a predisposing factor. Cholecchia *et al*. studied 32 gall-stone-free obese patients and concluded that during a significant weight loss period, 54% of subjects following the low-fat diet, but none in the high-fat intake group, formed asymptomatic gall stones.

Total parenteral nutrition (TPN) is associated with the development of acalculous cholecystitis as well as cholelithiasis, cholecystitis, and gall bladder sludge. The latter can occur as early as 3 weeks after initiation of TPN. After 3 or 4 months of TPN, approximately 45% of patients will develop gall stones. Prolonged fasting resulting in gall bladder hypomotility seems to be the major cause of the bile stasis. Cholecystokinin–octapeptide 50 ng/kg intravenous infusion for 10 minutes once daily has been shown to prevent gall bladder sludge and gall stone formation in patients on TPN.

Hormones and drugs Hormones such as estrogen and progesterone have a significant effect on the risk of gall stone formation. One interesting illustration of these effects is seen in pregnant woman. Increased estrogen levels during pregnancy cause increased cholesterol secretion and supersaturation of bile, which results in more lithogenic bile. Progesterone, on the other hand, reduces gall bladder motility, resulting in stasis and sludge formation in 30% of cases.

Among lipid-lowering drugs, Clofibrate seems to have the greatest association with increased gall stone formation. The role of statins in gall bladder disease remains to be elucidated. Approximately one-third of patients treated with octreotide, a somatostatin analog, develop new gall stones. Ceftriaxone (Rocephin) has been shown to cause sludge formation in children. A large fraction of ceftriaxone is secreted in bile (40%) and forms complexes with calcium, resulting in an insoluble salt. The sludge disappears when ceftriaxone is discontinued.

Diet and lipid profile The ingestion of refined sugars has been shown to be associated with gall stone disease. However, no such association has been shown for alcohol or tobacco. It is not clear if high serum cholesterol predisposes to gall stone formation. In fact, the contrary has been shown in some studies. This is also the case for dietary cholesterol ingestion, which was shown to be a protective factor for gall stone formation in one study. Hypertriglyceridemia, on the other hand, is positively associated with an increased incidence of gall stones.

Dietary antioxidant deficiency, particularly of α-tocopherol, as well as low intakes of linoleic acid and essential amino acids may increase the incidence of gall stone disease. One study showed that there is an inverse correlation between the incidence of gall stone disease and the amount of certain foods, particularly fish and fruits, consumed per day. The gall stone subjects ate fewer meals per day but ate more cereals, oils, sugars, and meats. They also had more fluctuation in their weight. They consumed less fiber, folate, magnesium, vitamins, and minerals.

Other conditions predisposing to gall bladder disease Insulin-resistant diabetes predisposes to cholelithiasis. A Swedish study showed that the prevalence of gall stones in Crohn's disease was twice that seen in the general population. Cirrhosis is another major risk factor for gall stones. The incidence of gall stone formation in cirrhosis is 10 times that seen in the general population. The incidence increases with the severity of cirrhosis, being worse in Child's class B and C disease and in patients with higher body mass index. High estrogen level and reduced hepatic synthesis and transport of bile salts are reasons for the increased risk in cirrhosis. The Physicians' Health Study showed that 30 minutes of endurance-type exercise five times per week prevents approximately one-third of cases of symptomatic gall stones in men. The Nurses' Health Study confirmed the same trend in women.

Clinical Manifestations and Diagnosis of Gall Stone Disease

Approximately 80% of people with gall stones are asymptomatic. The presentation of gall bladder disease can be episodic pain when a brief cystic duct obstruction occurs or acute cholecystitis when the obstruction lasts longer and results in local and relatively extensive inflammation and edema. The complications include infection of the biliary system (cholangitis) and pancreatitis.

Symptoms and Signs

Pain related to the gall bladder is usually felt in the right upper quadrant or in the epigastrium. It may radiate to the back, going around the right flank. In some cases, it may radiate to the shoulder area or be felt in the chest. In acute cholecystitis, the pain is steady, as opposed to cramping or colicky. It typically occurs after a meal and may be accompanied by nausea and vomiting. Continuous obstruction of the cystic duct causes gall bladder distention and inflammation. Extension of the inflammation into the common bile duct area may cause edema and obstruction of the duct, resulting in jaundice. The

physical signs of acute cholecystitis include right upper quadrant tenderness and Murphy's sign, which refers to severe right upper quadrant tenderness and inhibition of inspiration on deep palpation under the right subcostal margin.

Laboratory Findings

In acute cholecystitis, liver enzymes are normal or mildly elevated. Marked increase in liver enzymes should raise the possibility of bile duct obstruction concomitant with, or instead of, acute cholecystitis. In the case of bile duct obstruction, alanine aminotransferase and aspartate aminotransferase increase rapidly to levels 10 times normal and then decrease quickly toward normal, even if the obstruction persists. Alkaline phosphatase, on the other hand, will continue to increase unless the obstruction resolves. Mild to moderate leukocytosis is common in acute cholecystitis. Bile levels may increase if the obstruction lasts long enough.

Imaging

Ultrasonography, with a sensitivity of 96%, is a major diagnostic tool in gall bladder disease. The sonographic evidence of acute cholecystitis includes gall bladder size, its wall thickness, and pericholecystic fluid conformation. Among these signs, the latter is the most sensitive. Computed tomography is less sensitive and more expensive than ultrasonography. Its main role is to rule out other intra-abdominal processes. Magnetic resonance imaging has become an important means of detecting bile duct stones. Its sensitivity is approximately 85% for bile duct stones. Oral cholecystography has almost completely been replaced by ultrasonography. Endoscopic retrograde cholangiopancreatography (ERCP) is mostly used for its diagnostic and especially its therapeutic capability for removing bile duct stones. Hepatobiliary scintigraphy consists of the uptake by the gall bladder of an intravenously administered, ^{99m}Tc-labeled iminodiacetic acid derivative. The liver excretes the isotope into the bile ducts. A normal hepatobiliary scan using diisopropyliminodiacetic acid effectively rules out acute cholecystitis. If, on the other hand, the isotope does not appear in the bile ducts within 4 h, the likelihood of acute cholecystitis is very high.

Management

The decision about the treatment of gall stones will depend strongly on the presentation of the patient. Asymptomatic gall stones should not be treated surgically. This recommendation is based on multiple studies, including several prospective studies, that showed that patients with asymptomatic gall stones, observed over many years, develop symptoms or biliary complications only on rare occasions. In one study of 123 people with asymptomatic gall stones followed for 11–24 years, biliary pain developed in 2% during each of the first 5 years, followed by a decreasing incidence thereafter. Complications were seen in only 3 people and were preceded by warning signs of pain in all cases.

Prophylactic cholecystectomy has been performed in diabetic patients because of concern of higher rates of complications from acute cholecystitis and also in patients with sickle cell disease. In the latter group, the main reason for prophylactic cholecystectomy is that the pain of sickle crises is not easily distinguished from acute cholecystitis. Cholecystectomy should be undertaken in patients with 'porcelain' gall bladder and Native Americans with gall stones because these groups have a higher incidence of gall bladder carcinoma. However, the risk of gall bladder carcinoma is not high enough to justify cholecystectomy in other asymptomatic patients. Symptomatic gall stones should be treated surgically unless contraindicated. More than 60% of patients with symptomatic gall stones will have recurrent episodes within 1 or 2 years. Moreover, approximately 3% will develop biliary complications annually.

Nutritional Considerations

Prevention of gall stone formation The importance of nutritional factors in the formation of gall stones was discussed previously. Thus, modification of those risk factors (obesity, rapid weight loss, TPN, etc.) or application of a remedy to correct the underlying offensive mechanism will perhaps reduce the chance of gall stone formation.

Some studies have shown beneficial effects of dietary fiber in the prevention of gall bladder disease during weight loss in obese patients. However, this effect seems to be limited. During rapid weight loss, patients are occasionally given a small amount of dietary fat in order to reduce the risk of gall stone formation. Some studies have confirmed this technique, especially when a lower amount of fat has been used (2 vs. 10 g fat/day); others have shown that this strategy is not effective when the comparison was made between higher amounts (16 vs. 30 g fat/day). However, some studies support the concept that factors other than gall bladder motility are involved in gall stone formation in patients undergoing rapid weight loss. The weight loss by itself will eventually be beneficial in the prevention of future gall stones.

In one study, fish oil (n-3-polyunsaturated fatty acids) was shown to prevent cholesterol gall stone formation in obese women during rapid weight loss.

A high-calcium diet has also been prescribed in order to change the output of deoxycholic acid. This diet has a modest beneficial effect on gall stone formation.

Multiple small meals as opposed to one or two larger meals per day will empty the gall bladder on a regular basis and prevent stasis. It also prevents long interruptions of the enterohepatic circulation of bile acids.

Cholecystokinin increases bile acid synthesis and, more important, promotes gall bladder contraction. Therefore, CCK is theoretically an effective agent in the prevention of gall stones in patients with TPN. In practice, though, this effect seems to be quite variable.

The Third National Health and Nutrition Examination Survey found that a history of clinical or asymptomatic gall stones was inversely correlated with serum ascorbic acid levels in women but not in men. Coffee consumption has been found to be associated with a reduced risk of symptomatic gall stones. Consumption of two or three cups of regular coffee per day reduces the risk of symptomatic gall stones by approximately 40% over 10 years. **Table 2** summarizes the different preventive measures available to reduce the risk of gall stone formation.

Complementary and alternative medicine Factors reducing hyperinsulinemia, such as dietary unrefined carbohydrates and aerobic exercise, have been suggested to reduce the risk of cholelithiasis. Holistic health providers have been prescribing herbal medicines, such as turmeric, Oregon grape, bupleurum, and coin grass, with the belief that they may reduce gall bladder inflammation and relieve liver congestion.

Medical Treatment

Ursodeoxycholic acid is approved for gall stone dissolution in appropriate patients. This drug is a bile acid that forms soluble vesicles and prolongs the nucleation time of bile. Moreover, it inhibits HMG-CoA reductase and secondarily reduces cholesterol saturation. A year of treatment with ursodeoxycholic acid will result in complete dissolution of 50% of gall stones. This agent works on the surface of the stones; thus, the larger the stone, the less the efficacy of ursodeoxycholic acid. This drug is used in patients who refuse surgery or have a high risk for cholecystectomy. Shiffman *et al.* studied the efficacy of ursodeoxycholic acid in patients undergoing dietary-induced weight reduction and found it to be highly effective in preventing gall stone formation. The cost of this drug ($4 per day) is an important impediment for its long-term use.

Nonsteroidal antiinflammatory drugs may have some beneficial effect in the prevention of gall stone formation, but studies have not been conclusive. Extracorporeal shock wave lithotripsy has been used for fragmentation of gall stones by high-amplitude

Table 2 Prevention of gall stone formation

Dietary fiber	Limited efficacy
During rapid weight less	
Small amount of dietary fats	Depends on the amount of fat
Fish oil	Prevents cholesterol gall stone formation in obese women
Nonsteroidal antiinflammatory drugs	Studies have not been conclusive
Ursodeoxycholic acid	Effective in secondary prevention
Reduce insulin resistance	
Exercise	
Gradual weight loss	
During total parenteral nutrition	
Cholecystokinin	Increases bile acid synthesis and promotes gall bladder contraction; efficacy variable
Nutritional factors	
Ascorbic acid	Inverse correlation with gall stone formation in women but not in men
Reduce ingestion of refined sugars	Prevents insulin resistance and secondarily prevents gall bladder immobility
Dietary antioxidants, fiber, folate, magnesium, vitamins, minerals, linoleic acid, and essential amino acids	Unknown mechanism
Multiple small meals as opposed to few large meals	Prevents gall bladder stasis
High-calcium diet	Preventive
Coffee consumption	May reduce risk of symptomatic gall stone

shock waves generated by external piezoelectric devices. The waves are guided toward the stones by ultrasound imaging. This procedure was first used for kidney stones. Its efficacy in gall bladder stone treatment has been much less impressive. Its complications are the consequence of migration of stone fragments and include postprocedure biliary colic and pancreatitis. The availability of laparoscopic cholecystectomy has limited the need for lithotripsy.

Surgical Treatment

Patients admitted with suspected acute cholecystitis should initially be made NPO (nothing by mouth) and intravenously hydrated. Administration of broad-spectrum antibiotics early in the course is recommended because secondary infection often supervenes in what is initially a noninfectious process. If the diagnosis of acute cholecystitis is made within 24–48 h of onset of symptoms, early surgery leads to reduced morbidity and mortality rates.

Laparoscopic cholecystectomy The first successful laparoscopic cholecystectomy in Europe in 1987 transformed gall bladder surgery very rapidly. The minimal injury to the abdominal wall tissues resulted in faster discharge and rapid return to normal activities. In approximately 1 in 10 patients, laparoscopic cholecystectomy must be converted to open surgery. The rates of complications in both procedures are approximately 5%.

Laparotomy (open cholecystectomy) Today, open cholecystectomy is almost always a consequence of conversion of a laparoscopic procedure. On rare occasions, such as when there is suspicion of gall bladder cancer, a history of prior extensive abdominal surgery resulting in adhesions, or in patients with common bile duct stones that cannot be removed by ERCP, surgeons will perform open cholecystectomy at the outset of the operation.

Conclusions

Gall bladder disease is intimately tied to nutrition with respect to its etiology, treatment, and prevention. Attention to nutritional considerations can have a meaningful impact in particular on the prevalence and incidence of asymptomatic and symptomatic gall stones. Certain nutriments can change the likelihood of gall stone formation or change the course for preexisting stones. Nutritional status, including obesity and rapid weight loss, is an important predisposing factor. TPN increases the risk of gall stones considerably. Diagnosis of gall stone disease is based on history and physical examination as well as the laboratory and ultrasonographic evidence. Therapeutic decisions depend on the patient's presentation. Asymptomatic patients should not be treated surgically, except those with sickle cell disease or if there is a higher risk of gall bladder carcinoma (as in Native Americans and in patients with porcelain gall bladder). Laparoscopic surgery is the treatment of choice in symptomatic patients.

Nutritional interventions, such as fish oil, ascorbic acid, and coffee, seem to be protective against gall stone formation. Dietary fiber and the administration of small amounts of dietary fat have shown limited preventative efficacy for patients at higher risk of gall stone formation. CCK may be useful in patients receiving TPN. Ursodeoxycholic acid is effective in preventing gall stone formation in patients undergoing diet-induced weight reduction. Some authors have recommended the use of nonsteroidal antiinflammatory drugs in the prevention of gall stones, but studies have not been conclusive.

See also: **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Diabetes Mellitus**: Etiology and Epidemiology. **Obesity**: Complications.

Further Reading

Andersen T (1992) Liver and gall bladder disease before and after very-low-calorie diets. *American Journal of Clinical Nutrition* **56**(1 supplement): 235S–239S.

Dam H (1969) Nutritional aspects of gall stone formation with particular reference to alimentary production of gall stones in laboratory animals. *World Review of Nutrition and Dietetics* **11**: 199–239.

Festi D, Colecchia A, Larocca A *et al.* (2000) Review: Low caloric intake and gall bladder motor function. *Alimentary Pharmacology and Therapeutics* **14**(supplement 2): 51–53.

Hayes KC, Livingston A, and Trautwein EA (1992) Dietary impact on biliary lipids and gall stones. *Annual Review of Nutrition* **12**: 299–326.

Hofmann AF (1988) Pathogenesis of cholesterol gall stones. *Journal of Clinical Gastroenterology* **10**(supplement 2): S1–S11.

Lee SP and Ko CW (2003) gall stones. In *Yamada Textbook of Gastroenterology*.

Rescorla FJ (1997) Cholelithiasis, cholecystitis, and common bile duct stones. *Current Opinion in Pediatrics* **9**(3): 276–282.

Rosen GH (1992) Somatostatin and its analogs in the short bowel syndrome. *Nutrition in Clinical Practice* **7**(2): 81–85.

Trotman BW (1991) Pigment gall stone disease. *Gastroenterology Clinics of North America* **20**(1): 111–126.

Geriatric Nutrition *see* **Older People**: Physiological Changes; Nutritional Requirements; Nutrition-Related Problems; Nutritional Management of Geriatric Patients

GLUCOSE

Contents

Chemistry and Dietary Sources

D J A Jenkins, R de Souza, L S A Augustin and C W C Kendall, University of Toronto, Toronto, ON, Canada

Glucose and its polymers are important energy sources for living organisms and structural components of plants. Because of the diversity of compounds in which glucose occurs, it may be helpful to first discuss nomenclature.

Nomenclature and Chemical Structure

Glucose

The compound D-glucose (Greek *gleucos*, 'sweet wine') or dextrose is 2,3,4,5,6-pentahydroxyhexaldehyde, more conventionally expressed as $C_6H_{12}O_6$, with a molecular weight of 180.16 kDa. Glucose is readily soluble in water in a powder form. Below 50 °C, α-D-glucose hydrate is the stable form; at 50 °C the anhydrous form is obtained; and at higher temperatures, α-D-glucose is obtained. Glucose is also present in the diet as part of the disaccharides sucrose (glucose and fructose), lactose (glucose and galactose), and maltose (glucose).

Glucose Oligosaccharides

Oligosaccharides (Greek *oligo*, 'few') are sugar polymers; the term usually refers to compounds containing 2–9 units but may include polymers containing up to 19 units. The dimer, trimer, and tetramer forms in which glucose molecules are joined by (1–4) linkages are referred to as maltose, maltotriose, and maltotetrose, respectively, since these substances are the products of starch digestion in the malting process. Sucrose, maltose, and lactose are common dietary disaccharides.

Starches

Starches are large-molecular-weight, α-linked polymers of glucose $(C_6H_{10}O_5)n$. Most starches show a mixture of α(1–4) and α(1–6) linkages. The α(1–4)-linked polymer forms a linear structure that allows for hydrogen bonding between polymer chains and a more compact starch structure. Introduction of (1–6) linkages results in branch points and a more open structure that allows the (1–4)-linked backbone with the hemiacetal bond in the alpha configuration to coil like a spring into a helical form. Branched starches with the (1–6) linkage are more readily hydrated and digested compared to the (1–4)-linked linear starch. The (1–4)-linked starches are referred to as amylose starch, and (1–6)-linked starches are amylopectin starches.

Resistant Starch

Resistant starches are defined by their resistance to digestion in the human upper gastrointestinal tract. As with the term 'dietary fiber,' the definition is largely physiological. One proposed classification divides resistant starches into three classes: RS_1, RS_2, and RS_3. The first class, RS_1, is starch that escapes small intestinal digestion owing to the food form and incomplete enzymatic attack (e.g., large particle size or compact nature of food, or starch entrapment by dietary fiber). The second, RS_2, includes the more crystalline starches that resist digestion (e.g., high-amylose starches that resist gelatinization). The RS_3 starches are retrograded starches (e.g., high-amylose starches that upon

cooling after cooking form a compact, hydrogen-bonded crystalline structure that excludes water).

Cellulose

Like starch, cellulose is a (1–4)-linked glucose polymer $(C_6H_{10}O_5)n$, but in this instance the glucose molecules are β-linked, allowing the development of a linear polymer with strong intrachain hydrogen bonding. Cellulose polymers may consist of as many as 10 000 glucose monomer units. Cellulose is both resistant to small intestinal digestion and insoluble in cold or hot water and most dilute acids and alkali. It is partially degraded by colonic bacteria; the proportion degraded is dependent on the source, with cellulose from vegetables broken down to a greater extent than cellulose from cereals such as wheat.

β-Glucans

In many ways, these predominantly (1–4)-linked glucose polymers are the cellulose equivalent of the starch amylopectin. Here, it is the (1–3) linkages interspersed throughout the polymer that prevent the compact structure achieved with the cellulose polymer where only the (1–4) linkages exist. As a result of the more open molecular structure of the β-glucan, unlike cellulose, it is readily hydrated and soluble in water, forming a solution of high viscosity. The viscosity, in turn, is dependent on the molecular weight and the presence of the (1–3) linkages. The greater the molecular weight, the greater the viscosity. Thus, reduction of molecular weight by acid or enzymatic hydrolysis, which may also occur during food processing, may greatly reduce viscosity. The common feature shared by cellulose and the β-glucans is that both are resistant to digestion by small intestinal enzymes. However, whereas cellulose is only partially fermented by the colonic bacteria, β-glucans are completely fermented.

Hemicellulose

The term 'hemicellulose' should not be taken to imply a class of (1–4)-linked glucose polymers. The similarity with cellulose lies not in the chemical structure but in the fact that hemicellulose is also insoluble in hot or cold water or hot dilute acid. It is, however, soluble in dilute alkali. The polymeric structure is heterosaccharitic with two or more sugars (e.g., arabinoxylans found in cereals), with a relatively small molecular size (50–200 saccharide units).

Occurrence

Glucose is the primary carbohydrate energy source of vertebrates. In healthy humans, fasting blood glucose levels are approximately 3.5–5.5 mmol/l (depending on the laboratory) and increase postprandially to values considerably less than 10 mmol/l (the renal threshold for complete reabsorption, above which glucose 'spills' over into the urine). Blood levels higher than 7.8 mmol/l 2 h after a 75-g glucose load are one of the diagnostic criteria for diabetes. Glucose is stored as glycogen, an α-linked polymer, predominantly in the liver and muscles ('animal starch'). On average, a 70-kg man may store 500 g of glycogen. Glucose can also be synthesized de novo by gluconeogenesis from the gluconeogenic amino acids lactate, glycerol, and pyruvate.

Erythrocytes, renal tissue, and nervous tissue require glucose as an energy source. In erythrocytes and renal tissue, the glucose is not oxidized but is returned to the liver as part of the Cori cycle for glucose synthesis. The brain oxidizes glucose and requires 140 g per day. From this figure the carbohydrate requirement was derived in the recent DRI assessment (**Table 1**). Despite this requirement, carbohydrate is still recommended to comprise between 45 and 65% of dietary calories.

Glucose is present in fruit and vegetables and, although less sweet on a per gram basis than fructose or sucrose, it is responsible together with fructose and sucrose for the sweet taste of vegetables and fruit. With the exception of fruit such as green banana, seeds (grain and dried legumes), and tubers, in which starch is the major carbohydrate form, foods containing glucose, fructose, and sucrose in various ratios comprise the major available (i.e., absorbable in the small intestine) carbohydrate sources. The relative proportions of the sugars have not been generally determined, and data are not available for many foods.

The main sources of dietary starch are cereal grains, dried legumes, and tubers. The major part of the available carbohydrate in these foods is

Table 1 Estimation of brain glucose requirements of adult humans

Glucose consumption (μmol/100 g brain/min)	*Estimated brain weight (g)*	*Brain glucose consumption*	
		mg/min	*g/day*
31–38	1450	81–99	117–142

Based on data from Sokoloff *et al.* (1977), Gottstein and Held (1979), Scheinberg and Stead (1949), and Reinmuth *et al.* (1965).

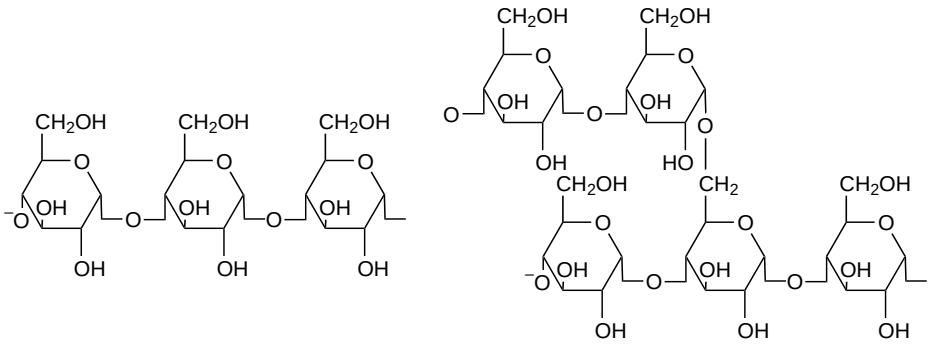


Figure 1 Partial structures of amylose (linear) and amylopectin (branched) starches.

starch. Starches contain both α(1–4) and α(1–6) linkages (i.e., amylose and amylopectin) (**Figure 1**). In most studies amylose predominates, with a ratio of amylose to amylopectin of 2–3:1. In general, legumes contain higher amylose levels than do cereals. Cultivars of corn have been bred with high amylose levels.

Resistant starches comprise a small proportion of most industrialized Western diets. Increased starch malabsorption may be induced by coarse milling or large particle size of cereal grains (e.g., whole-grain pumpernickel or bulgur wheat). Such foods may be said to contain resistant starch (RS_1). Resistant starches that are crystalline in nature and resist hydration (RS_2) are found in green banana, high-amylose corn, and relatively high-amylose legumes (peas, beans, and lentils). Starches, especially high-amylose starches that are cooked and then allowed to cool, undergo retrogradation with more crystalline realignment. These starches (RS_3) are produced in common foods such as potato, rice, and bread. Resistant starches in this category are produced commercially from high-amylose cornstarch by enzymatically debranching the remaining (1–6) linkages and allowing the resulting (1–4)-linked starch to 'retrograde' into a highly crystalline, digestion-resistant starch.

Cellulose is an important structural component of plant cell walls. In human nutrition, it forms an important part of the 'insoluble' dietary fiber component reported in food composition tables. However, values for the actual proportion of the total dietary fiber that is composed of cellulose are only available in special food composition tables for a relatively small number of foods.

From the standpoint of human nutrition, β-glucans are found predominantly in cereals, notably oats and barley, with trace amounts in wheat. In oats, the β-glucan is concentrated in the outer bran layer and may comprise 50% of the dietary fiber value and possibly 8 or 9% of the so-called oat bran derived from standard milling practices. In barley, the β-glucan is more dispersed through the endosperm, and thus a bran concentrate is less easy to achieve. In both cases, high β-glucan cultivars may greatly increase the yield of β-glucan. In addition, 'wet' processing techniques may yield a high-concentration β-glucan bran and purified β-glucan oat gum.

Analysis

Analysis of glucose may involve chemical, enzymatic, electrochemical, and high-performance liquid chromatography (HPLC) systems. Prior to the introduction of enzyme-based analyses, chemical techniques were based on the reducing ability of glucose, and techniques employing copper sulfate were popular. Such techniques were influenced by other reducing sugars and reducing substances, including uric acid and vitamin C. With the introduction of the more specific glucose oxidase-based tests, the chemical tests were abandoned, although there was debate over the potential carcinogenicity of the early chromogens, *O*-dianizadine and *O*-toluidine. Later, more specific, hexokinase-based enzyme assays were introduced. Current methods for rapid determination of blood glucose, which no longer require prior precipitation of plasma proteins, involve electrochemical detection. These methods rely on silver electrodes to detect electrons generated by the oxidation of glucose by glucose oxidase contained in membranes on the surface of the electrodes. For determination of glucose and α-limit dextrins resulting from starch digestion, HPLC techniques have proved useful.

Much attention has been given to the analysis of the glucose polymers—starches, resistant starches,

cellulose, and β-glucans—in the context of dietary fiber. The ultimate assessment depends on the use of specific enzymes or enzyme systems to break the macromolecules down to their component glucose and other sugars when mixtures containing other polymers (dietary fibers) are being analyzed. These are then assessed by gas chromatography or HPLC and the ratios of the sugars determined. More routine assessment may involve a variety of chemical techniques combined with enzymatic digestion and, in the case of a popular Association of Official Analytical Chemists (AOAC)-approved technique for dietary fiber analysis, with a gravimetric determination. However, there is debate as to whether the resistant starch, which in the gravimetric AOAC technique is analyzed as dietary fiber, should be included as fiber or whether it is physiologically distinct. It is also debated whether a determination of β-glucan is sufficient without knowledge of its viscosity and molecular weight—factors that determine its physiological effect.

Physiology

The physiology of the gastrointestinal absorption of (and the energy retrieved from) the glucose molecule along the length of the gastrointestinal tract in its various forms is discussed in the following sections, together with the influence of other dietary factors (**Figure 2**).

Absorption

In its simplest form, glucose ingested by mouth is rendered isotonic in the stomach by the gastric juices and expelled through the pylorus into the duodenum, where active transport takes place at the brush border by way of a sodium-linked glucose transporter. The absorbed glucose that is taken up by way of the portal vein suppresses hepatic glucose output but does not markedly alter the glucose balance across the liver. The major part of the absorbed glucose is taken up by muscle and also adipose tissue under the action of insulin. Similarly, sucrose, maltose, and lactose are both split and absorbed at the brush border by the brush border enzymes sucrase–isomaltase, maltase, and lactase. Although sucrose deficiency is exceedingly rare, hypolactasia is common in adult life in most of the world's populations, with the exception of those of northern European origin. Thus, unlike sucrose malabsorption, small-intestinal lactose malabsorption is common, with significant amounts of lactose entering

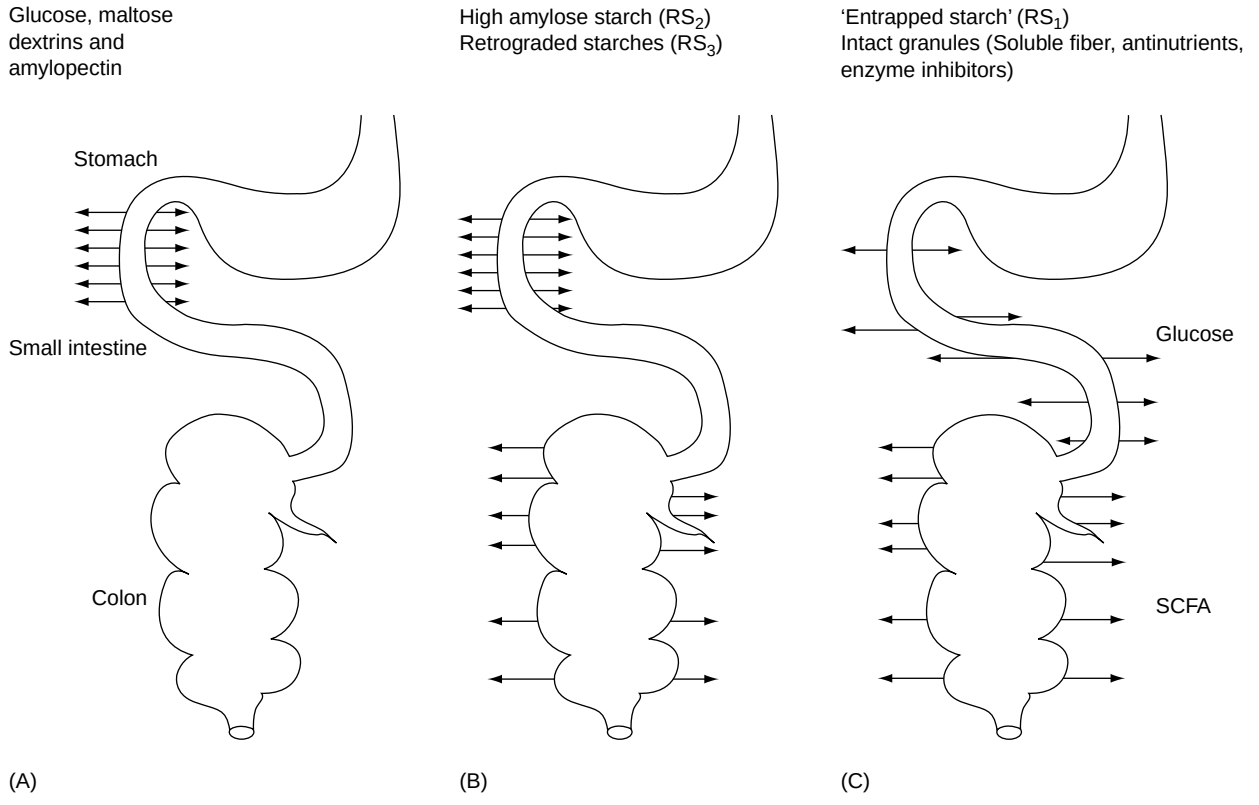


Figure 2 Effect of different forms of glucose on glucose absorption and short-chain fatty acid (SCFA) production and uptake from the gut.

the colon, resulting in gas production, short-chain fatty acid (SCFA) synthesis, and, in some instances, diarrhea.

On the other hand, purified, fully hydrated, cooked amylopectin starch commences digestion in the mouth under the action of salivary amylase. Enzyme activity ceases under the acidic conditions of the stomach and resumes in the duodenum under the action of pancreatic amylase. Amylolytic digestion in both mouth and stomach results predominantly in the production of free glucose, maltose, maltotriose, and the α-limit dextrins of greater polymeric length. The free glucose is taken up by the brush border glucose transporter, and the uptake of maltose and maltotriose is effected by brush border enzymes, notably the sucrase–isomaltase complex. In both situations, absorption in the small intestine is considered to be complete.

However, foods as eaten do not usually comprise pure glucose and pure amylopectin starch as their carbohydrate components. Many factors influence small-intestinal absorption in terms of both rate and amount (**Table 2**). Some of these factors were previously discussed in connection with amylose and resistant starch.

Food Components

Insoluble fiber may form a coat around starchy foods, limiting the penetration of enzymes and thus reducing the rate and amount digested. Viscous soluble fibers may also reduce the rate of

Table 2 Factors influencing glycemia and gastrointestinal events

Factors influencing the availability of carbohydrate	*Physiological effect*					
	Glycemia	*Stomach, gastric emptying*	*Small intestine, absorption rate*	*Motility*	*Colon*	
					Baterial fermentation	*Fecal bulk*
Food components						
Fiber						
Soluble (viscous)	−	−−−	−−	−?	+++	+
Insoluble	0 ?	+	+	+	+	+++
Macronutrients						
Protein–starch interaction	−	−?	−−	?	+	?
Fat–starch interaction	−	−	−−	?	+	?
Starches						
Amylopectin	+	?	++	?	0	0
Amylose	−	−	−−−	?	++	+
Sugars and glucose polymers						
Glucose	++	−?	+	0	0	0
Maltose	++	−?	+	0	0	0
Maltodextrins	++	0	+	0	0	0
Antinutrients						
Phytates	−	?	−	?	+	?
Tanins	−	?	−	?	+	?
Saponins	?	?	−	?	+	?
Lectins	−	−	−	?	+	?
Amylase inhibitors	−	0	−	?	+	?
Alpha-glucosidase inhibitors	−	0	−	?	+	?
Food processing						
Cooking						
Starch gelatinisation	+	0	++	?	0	0
Starch regtrogradation	−	?	−−	?	+	+
Parboiling (e.g., rice)	−	?	−	?	?	0
Particle size						
Milling						
Crushing	+	+	+	?	0	0
Flaking	+	?	+	?	0	0
Extruding	+	?	+	?	0	0
	−	?	−	?	0	0

+, increase, promote; −, inhibit, reduce; 0, no effect; ?, uncertain.

absorption through prolonging gastric emptying and by acting as a barrier to diffusion in the small intestine. Starch–protein interactions (as seen with gluten in wheat products) and starch–fat interactions have been shown to reduce the rate of digestion, and fat is known to slow gastric emptying. A number of the so-called 'antinutrients' present in foods, notably lectins, phytates, and tannins, have been shown to reduce the digestibility of foods. For example, it is considered that phytate, by binding calcium ions that catalyse starch digestion by amylase, reduces the rate of small-intestinal starch digestion.

Food processing may influence the rate of digestion by removing or reducing the level and activity of inhibitory food components. It may also modify the structure of the food or its components to make the food more available to digestive enzyme attack. Examples are cooking, resulting in starch gelatinization, and reducing the particle size (and hence increasing the surface area available to digestive enzymes) by milling, crushing, or flaking. On the other hand, processing may also reduce digestibility by parboiling, cooking with retrogradation of the starch, and extrusion, as in the production of pasta, producing a more compact physical structure.

Increasing the frequency of meals and reducing their size spreads the nutrient load over time and hence prolongs the time spent in the absorptive state. It is perhaps the 'clearest' model of slowing the rate of absorption and is referred to again to explain the metabolic consequences of reducing the absorption rate.

Finally, enzyme inhibitors of carbohydrate absorption have been developed for pharmacological use in the treatment of diabetes, and these work by reducing the rate of carbohydrate uptake from the small intestine. One example of this class of substances is acarbose, an α-glycoside hydrolase inhibitor that has antiamylase and anti-sucrase–isomaltase activity and thus inhibits both intraluminal and brush border carbohydrate digestion and absorption of starch, sucrose, and maltose.

Possible Effects of Prolonging Absorption Time of Carbohydrate

The question remains as to what physiological effects are produced when carbohydrate is absorbed more slowly (**Table 3**). Studies have demonstrated the effectiveness of carbohydrate-absorption enzyme inhibitors in treating diabetes but also in preventing the development of diabetes in high-risk subjects treated with acarbose over a 3-year period. A further way to reduce the rate of absorption of carbohydrate without altering its composition is to change the rate of ingestion of carbohydrate substrates.

Table 3 Possible effects of prolonging absorption time of carbohydrates

Flatter postprandial glucose profile
Lower mean insulin levels postprandially and throughout the day
Reduced gastric inhibitory polypeptide response
Reduced 24-h urinary C peptide output
Prolonged suppression of plasma free fatty acids
Reduced urinary catecholamine output
Lower fasting and postprandial serum total and LDL cholesterol levels
Reduced hepatic cholesterol synthesis
Lower serum apolipoprotein B levels
Lower serum uric acid levels
Increased urinary uric acid excretion

A number of effects appear to be beneficial when glucose is sipped slowly rather than drunk as a bolus or when starchy meals are eaten more frequently but in smaller amounts. Studies by Ellis in the 1930s first demonstrated a reduction in insulin requirements in patients with diabetes when glucose and insulin were administered in small, frequent doses. Since then, a range of metabolic benefits have been ascribed to increased meal frequency (the 'nibbling versus gorging' phenomenon). Early studies reported lower total cholesterol levels with increased meal frequency. Subsequent studies showed low-density lipoprotein (LDL) cholesterol reduction in subjects eating 3 meals a day compared to those eating from 6 to as many as 17 meals daily for periods of 2–8 weeks. An extreme model of slowing absorption, in which 17 meals daily were fed, demonstrated lower levels of apolipoprotein B in addition to total and LDL cholesterol. Population studies also indicated that total cholesterol levels were lower in those who ate more meals daily. Studies using stable isotopes showed that cholesterol synthesis was reduced at greater meal frequencies. Furthermore, mevalonic acid excretion (a water-soluble marker of cholesterol synthesis) suggested that the change in cholesterol levels was also related to the change in urinary mevalonic acid output. Since insulin is known to stimulate HMG-CoA reductase activity, a rate-limiting enzyme in cholesterol synthesis, the depressed cholesterol synthesis was attributed to the lower insulin levels observed. In addition, the reduction in serum cholesterol levels on 'nibbling' may have resulted from increased bile acid losses due to more frequent bile acid cycling through the gut following increased meal frequency.

Studies of non-insulin-dependent diabetes have shown depressed glucose and insulin levels during the day with increased meal frequency. In non-diabetic subjects, the major effect of reducing the absorption rate (by sipping glucose over 3 h instead of taking the same amount of glucose as a bolus within 5 minutes) was to reduce insulin secretion. In addition, insulin suppression of free fatty acids and branched-chain amino acid levels was prolonged, and following glucose challenge no counter-regulatory response was observed.

Finally, serum uric acid, an independent risk factor for coronary heart disease, was reduced and increased urinary uric acid excretion was seen with increased food frequency. As with the reduction in serum cholesterol levels, the effects of lower insulin levels were used to explain these differences. It was suggested that insulin promoted renal reabsorption of uric acid, as demonstrated in the context of sodium reabsorption and hypertension in hyperinsulinemic states.

Further effects of food frequency on diabetes have been assessed. It has been suggested that increased food frequency may limit obesity by reducing adipose tissue enzyme levels. Acute studies in humans failed to show an increased thermogenic response with increased meal frequency. Nevertheless, when satiety was assessed in acute studies, fluctuations in satiety were less over the whole day; long-term studies have yet to be undertaken. Concern still remains that 'snacking' may increase body weight in susceptible individuals. Despite these concerns, the demonstration that increased meal frequency can improve certain aspects of lipid and carbohydrate metabolism makes it a valuable model for other methods of 'spreading the nutrient load' (e.g., reducing the rate of glucose absorption).

Colonic Function

A portion of the starch, together with dietary fiber including cellulose and β-glucan, enters the colon and is fermented by the colonic microflora with the growth of the fecal biomass and the production of SCFA, hydrogen, and methane. The extent to which this occurs varies from individual to individual and is based on the nature of the resistant starch and the source of the cellulose (e.g., vegetable cellulose is more readily fermented than cereal cellulose). Although some individuals may have starch in their feces, the majority of subjects show little or no fecal starch. Furthermore, all the β-glucan is broken down by bacterial action in the colon. A large proportion of the cellulose escapes colonic bacterial fermentation and contributes directly to fecal bulk. Thus, a significant proportion of glucose molecules are not absorbed in the small intestine but enter the colon and are salvaged after conversion to SCFAs. The SCFAs are rapidly absorbed and contribute to the host's energy metabolism. They are usually produced in the ratio of 60% acetate, 20% propionate, and 20% butyrate, but the relative ratios of these three fatty acids vary depending on the substrate and the rate of fermentation. Of the three SCFAs, only acetate appears in the peripheral circulation to any significant extent. Propionate is of interest since it is gluconeogenic and has been suggested to inhibit hepatic cholesterol synthesis. It is largely extracted by the liver at first pass. Butyrate, on the other hand, is taken up and used by colonocytes. The slower the fermentation, the higher the butyrate levels. Starches have been claimed to increase colonic butyrate and in some instances propionate production, and butyrate is said to have antineoplastic properties.

The Glycemic Index

The differing effects of different carbohydrate foods in raising the blood glucose concentration postprandially have long been recognized. The glycemic index classification was proposed to indicate the rates at which different starchy foods were digested. It was hoped that selection of foods with lower glycemic indices would contribute to prolonging the absorption of nutrients and thus improve the glycemic profile and reduce levels of fasting blood lipids.

However, a number of acute (up to 1 day) mixed meal studies during the mid- and late 1980s suggested that a glycemic index classification of foods had no clinical utility. Nevertheless, a number of subsequent reports have documented improved glycemic control in both type 1 and 2 diabetes as judged by serum fructosamine and glycosylated hemoglobin levels in studies lasting from 2 weeks to 2 months. Furthermore, some studies also noted reductions in serum lipids. Many high-fiber foods that lower LDL cholesterol levels also have low glycemic indices (barley, beans, etc.). Extensive glycemic index tables have been published that will help in food selection for therapeutic and study purposes.

Many of the traditional starchy foods from different cultures have a low glycemic index (**Table 4**). Finally, results of cohort studies suggest that consumption of foods with a low glycemic index, especially in the context of a high-fiber

Table 4 Glycemic foods of staples from different cultures

Food	*Average GI*[a]	*Culture*
White bread rolls	100	North American, European
Pumpernickel	70–90	North European
Pasta	50–70	Mediterranean
Cracked wheat (tabouli)	60–70	Mediterranean, Middle Eastern
Beans, lentils, dried peas	40–70	Southern United States, Latin American, Middle Eastern, Indian, Oriental
Parboiled long-grain rice	70	Asian, North African

[a]Glycemic index (GI) is rounded to the nearest 10%.

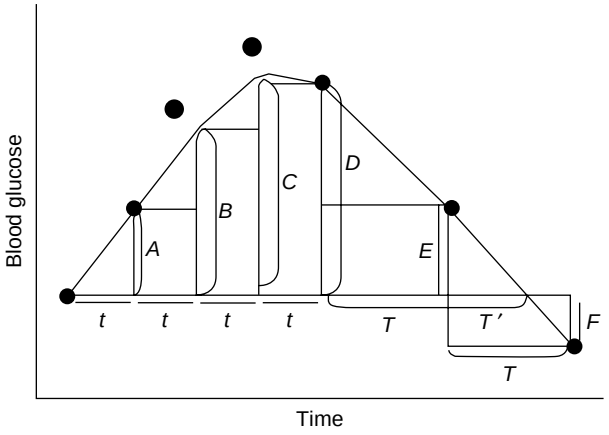


Figure 3 Schematic representation of postprandial blood glucose response (From Wolever TMS and Jenkins DJA (1986) The use of the glycemic index in predicting the blood glucose response to mixed meals. *American Journal of Clinical Nutrition* **43**:167–172.)

diet, protects from the development of type 2 diabetes. Therefore, the question is whether the rapid increase in diabetes in cultures in transition from traditional to Western lifestyle patterns is in part due to the high glycemic index of the diets eaten, in addition to the excess consumption of energy and reduced physical activity.

Calculation of the Glycemic Index

The glycemic index (GI) has been defined as the area under the blood glucose response curve for 50 g carbohydrate from the test food divided by the area under the blood glucose response curve for 50 g carbohydrate from the standard source, multiplied by 100. The standard carbohydrate source for modern assessments is white bread. In early studies, however, 50 g glucose was used rather than bread. On the 'bread scale' the glucose GI is approximately 130%. Other food GI values can be adjusted accordingly to allow direct comparison of the two scales.

The area under the blood glucose curve includes the area above the fasting level only. Any area beneath the fasting level is ignored. The incremental area under the blood glucose response curve is the sum of the areas of the triangles and rectangles. In **Figure 3**, *A*, *B*, *C*, *D*, *E*, and *F* represent the blood glucose increments above the baseline value (fasting level) at sequential time points, where *t* and *T* represent different time intervals between blood samples.

When the blood glucose concentration at *F* falls below the fasting concentration (**Figure 3**), only the area above the fasting level is included in the total area represented by the triangle *ET*, where *T′* represents the portion of the time interval *T* when the blood glucose level between *E* and *F* is above the fasting level.

The overall equation simplifies to

$$\text{Area} = \left(A + B + C + \frac{D}{2}\right)t + \frac{(D + E)T}{2} + \frac{E^2T}{2(E + F)}$$

If the last blood glucose concentration *F* is above the fasting level, then the term $(E+F)T/2$ is substituted for the last term in the equation, namely $E2\,T/2(E+F)$. An example of the incremental area calculation is shown in **Table 5**.

Calculation of mixed meal or total day's GI Each carbohydrate component is expressed as a percentage of the total carbohydrate in the meal or day and multiplied by the relevant GI. The sum of these values represents the meal's or the day's GI.

Table 5 Example of calculation of incremental area under the blood glucose response curve for glycemic response when the last glucose value falls below baseline

Time (min)	*Corresponding letter on* ***Figure 3***	*Blood glucose (mg dl^{-1})*	*Blood glucose increment (mg dl^{-1})*
0	—	100	—
15	*A*	120	20
30	*B*	140	40
45	*C*	160	60
60	*D*	150	50
90	*E*	120	20
120	*F*	90	−10

Calculation: Area $= (20 + 40 + 60 + 25) \times 15 + (25 + 10) \times 30 + (20^2 \times 30/2 \times (20 + 10)) = 3425$ mg min dl From Wolever TMS and Jenkins DJA (1986) The use of the glycemic index in predicting the blood glucose response to mixed meals. *American Journal of Clinical Nutrition* **43**: 167–172.

Calculation of glycemic load Glycemic load is the diet GI multiplied by daily dietary carbohydrate intake in grams per day.

See also: **Carbohydrates**: Resistant Starch and Oligosaccharides. **Diabetes Mellitus**: Classification and Chemical Pathology. **Glucose**: Metabolism and Maintenance of Blood Glucose Level; Glucose Tolerance.

Further Reading

Bertelsen J, Christiansen C, Thomsen C *et al.* (1993) Effect of meal frequency on blood glucose, insulin, and free fatty acids in NIDDM subjects. *Diabetes Care* **16**: 3–7.

Brand JC, Calagiuri S, Crossman S *et al.* (1991) Low-glycemic index foods improve long-term glycemic control in NIDDM. *Diabetes Care* **14**: 95–101.

Chiasson JL, Josse RG, Gomis R *et al.* STOP-NIDDM Trial Research Group (2003). Acarbose treatment and the risk of cardiovascular disease and hypertension in patients with impaired glucose tolerance: The STOP-NIDDM trial. *Journal of the American Medical Association* **23**: 486–494.

Englyst HN, Wiggins HS, and Cummings JH (1982) Determination of the non-starch polysaccharides in plant foods by gas–liquid chromatography of constituent sugars as alditol acetates. *Analyst* **107**: 307–318.

Englyst KN, Englyst HN, Hudson GJ *et al.* (1999) Rapidly available glucose in foods: An in vitro measurement that reflects the glycemic response. *American Journal of Clinical Nutrition* **69**: 448–454.

Fontvieille AM, Acosta M, Rizkalla SW *et al.* (1988) A moderate switch from high to low glycemic index foods for 3 weeks improves the metabolic control of type 1 (IDDM) diabetic subjects. *Diabetes Nutrition and Metabolism* **1**: 139–143.

Foster-Powell K and Brand Miller J (1995) International tables of glycemic index. *American Journal of Clinical Nutrition* **62**: 871S–893S.

Foster-Powell K, Holt SH, and Brand-Miller JC (2002) International table of glycemic index and glycemic load values: 2002. *American Journal of Clinical Nutrition* **76**: 5–56.

Jenkins DJA, Wolever TMS, Ocana AM *et al.* (1990) Metabolic effects of reducing rate of glucose ingestion by single bolus versus continuous sipping. *Diabetes* **39**: 775–781.

Jenkins DJA, Wolever TMS, Taylor RH *et al.* (1980) Rate of digestion of foods and post-prandial glycaemia in normal and diabetic subjects. *British Medical Journal* **2**: 14–17.

Jenkins DJA, Wolever TMS, Taylor RH *et al.* (1981) Glycemic index of foods: A physiological basis for carbohydrate exchange. *American Journal of Clinical Nutrition* **34**: 362–366.

Jones PJH, Leitch CA, and Pederson RA (1993) Meal frequency effects of plasma hormone concentrations and cholesterol synthesis in humans. *American Journal of Clinical Nutrition* **57**: 868–874.

Salmeron J, Manson JA, Stampfer M *et al.* (1997) Dietary fiber, glycemic load, and risk of non-insulin-dependent diabetes mellitus in women. *Journal of the American Medical Association* **277**: 472–477.

Schauberger G, Brinck UC, Sulder G *et al.* (1977) Exchange of carbohydrates according to their effect on blood glucose. *Diabetes* **26**: 415.

Wolever TMS and Jenkins DJA (1986) The use of the glycemic index in predicting the blood glucose response to mixed meals. *American Journal of Clinical Nutrition* **43**: 167–172.

Metabolism and Maintenance of Blood Glucose Level

V Marks, University of Surrey, Guildford, UK

This article is a revision of "Glucose: Maintenance of Blood Glucose Level" in the Encyclopedia of Food Science and Nutrition (Second Edition. pp. 2911–2916 © 2003 Elsevier Ltd).

Glucose is the only simple sugar found in most body fluids in anything more than trace amounts and for all practical purposes is confined to extracellular water. Lactose and fructose are the major sugars in milk and semen, respectively. This article reviews the major factors determining the concentrations of glucose in blood under everyday physiological and pathological conditions.

The Body Glucose Pool

The body of an adult subject seldom contains less than 8 g, or more than 28 g, of glucose at any one time (corresponding to blood glucose concentrations of 3.5–10 mmol l^{-1}), despite enormous fluctuations in demand and supply. This quantity of glucose can be considered as constituting a hypothetical body pool (**Figure 1**) confined within a glucose space equal in volume to the combined water in blood and the interstitial fluid (i.e., approximately 35% of total body water).

Glucose enters the cells by facilitated transport utilizing one or more of the genetically determined glucose transporter proteins that have been identified, depending on the tissue and which proteins are inducible. Upon entering a cell, glucose is immediately phosphorylated and consequently removed from the glucose pool.

Although its subsequent conversion into carbon dioxide and water or other metabolites (most notably glycogen, glycerol, fatty acids, and the glyco-moieties of mucopolysaccharides and glycoproteins) is the only way that glucose ordinarily leaves the glucose pool, its loss in the urine may become a major factor in diabetes mellitus. Glucose enters the glucose pool from food in the intestine after a meal via the portal vein or, in the postabsorptive subject, by release of glucose from preformed glycogen or molecules newly synthesized by liver cells into the hepatic veins.

Glucose Space

The glucose space (i.e., the extracellular water volume) is constant in any individual and

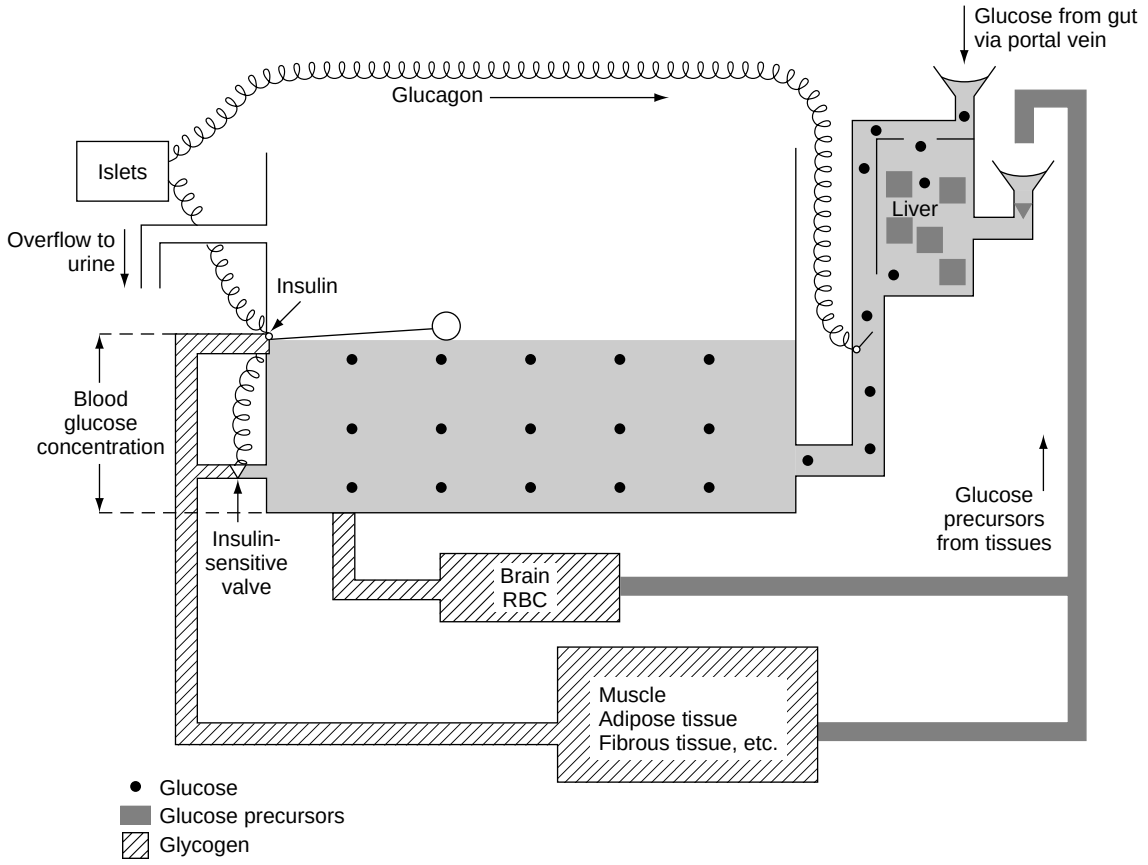


Figure 1 Schematic representation of blood glucose concentration and its relationship to the body glucose pool. The central system represents the hypothetical glucose pool, the actual size of which is represented by the horizontal axis (i.e., volume of distribution multiplied by blood (and extracellular fluid) glucose concentration). The postulated homeostatic switch is the cells of the endocrine pancreas, which respond to blood glucose concentration modulated by intestinal hormonal (incretin) and neural factors, which are themselves controlled by messages received from the gut (enteroinsular axis) and the autonomic nervous system. RBC, red blood cells.

consequently the amount of glucose in the body (the glucose pool) is directly proportional to its concentration in the blood. This is controlled through a series of complicated control mechanisms, the most important of which involve individual pancreatic islets of Langerhans. These function semiautonomously and release either insulin or glucagon according to need.

When pool size increases above a threshold, corresponding to a concentration in blood of approximately 10 mmol l^{-1}, glucose filtered at the glomeruli exceeds tubular capacity to reabsorb it and consequently spills over into the urine, producing glycosuria. Temporary increases in glucose pool size (hyperglycemia) are not immediately harmful but in the long term give rise to the so-called complications of diabetes. Decreases in glucose pool size (hypoglycemia), on the other hand, are immediately harmful and potentially so dangerous that many defence mechanisms have evolved to prevent or overcome them.

Blood Glucose

The brain, which can remove glucose from the extra-cellular fluid (ECF) in the absence of insulin, is the only important drain on the glucose pool in the fasting subject when plasma insulin levels are minimal. It consumes glucose at the rate of approximately 78 mg per gram of tissue per day. This amounts to approximately 110 g per day in an adult man or 75 g per day in a 1-year-old child. Estimates of glucose turnover suggest that approximately 9 g of glucose enters and leaves the glucose pool every hour in the average overnight fasting healthy subject.

The concentrations of glucose in venous and arterial blood are similar in the fasting subject because peripheral tissues, such as muscle, skin, and connective tissue, do not extract significant amounts of glucose from the blood in these circumstances. In the recently fed subject, however, glucose uptake by peripheral tissues increases markedly under the

influence of insulin released in response to the ingestion of a meal. This can produce a difference in arterial and venous blood glucose concentrations of $2\,mmol\,l^{-1}$ or more. This fact, known for more than 80 years, is still often forgotten or ignored by both experimentalists and clinicians. It not only has implications with regard to our understanding of the physiology of glucose homeostasis but also sometimes has unfortunate consequences for patients who may, if only venous blood is sampled, be misdiagnosed as suffering from hypoglycemia (i.e., blood glucose $<3.0\,mmol\,l^{-1}$).

It is, after all, arterial and not venous blood glucose that is homeostatically controlled and relevant to brain physiology, but because venous blood is more easily obtained, it is often used in studies of glucose homeostasis and clinical practice. Arterialized venous blood, collected from heat-distended veins on the back of the hands, is the best substitute for arterial blood in studies of glucose homeostasis. Finger-prick or earlobe capillary blood can also be used, although it is difficult to obtain in more than small amounts.

Blood glucose concentrations are generally $3.5–6.0\,mol\,l^{-1}$ in healthy fasting adult subjects and seldom rise above $11\,mmol\,l^{-1}$ in arterial or $10\,mmol\,l^{-1}$ in venous blood, even after a large carbohydrate-rich meal. Glucose and other simple sugars given in solution produce more rapid and greater increases in blood glucose than equal or larger amounts of glucose-yielding carbohydrate taken as part of a solid mixed meal. Conversely, prolonged starvation for as long as several weeks rarely causes the blood glucose concentration to fall below $3\,mmol\,l^{-1}$, except in children and adults with metabolic defects associated with impaired gluconeogenesis.

The remarkable ability of the body to regulate the size of the glucose pool under such widely diverse conditions depends mainly on two organs, the liver and the pancreas, although during prolonged starvation the kidneys become important generators of new glucose molecules.

The Effects of Feeding on Blood Glucose

Glucose

Glucose and the two lesser dietary monosaccharides, fructose and galactose, enter the circulation through the intestinal mucosa. The speed with which glucose can be absorbed is limited by the rate of transfer from the intestine but rarely exceeds 50 g (0.28 mol) per hour. This comparatively massive influx of glucose into a pool of less than 20 g ordinarily produces a remarkably small perturbation in blood glucose because the rate of removal from the glucose pool increases to match glucose input.

In healthy people, arterial blood glucose concentrations generally return to fasting levels within 2 h of eating a carbohydrate-rich meal and before all of it has been absorbed. This remarkable feat of homeostasis is achieved through the prompt and appropriate release of insulin into the circulation. This is a consequence of stimulation of pancreatic B cells (the source of insulin) by a rising arterial blood glucose concentration augmented by the insulinotropic hormones, GIP and GLP-1, released by endocrine cells in the intestinal mucosa. Nervous impulses originating in the brain in response to anticipation of eating (the cephalic phase) and from the mouth, gut wall, and portal vein may also play a role.

Under the influence of the rise in plasma insulin so produced and a simultaneous reduction in glucagon secretion, the liver reduces its rate of glucose input into the pool and increases its rate of extraction. Peripheral insulin-sensitive tissues, such as connective tissue, skin, fat, and especially striatal muscle, also start removing glucose. As a result of these duel actions, arterial blood glucose concentration decreases and the stimulus to insulin secretion declines until all of the food has been absorbed.

Ordinarily, the rates of change of glucose inflow from the gut into the glucose pool and the outflow of glucose into the tissues are so well aligned that arterial blood glucose levels rarely fall below fasting levels after a meal, and then only temporarily. Venous blood glucose levels do so more often. The somewhat unnatural conditions resulting from ingestion of large amounts of a glucose solution on an empty stomach may produce a 'reactive hypoglycemia' due to persistence of insulin action after plasma insulin has fallen to basal levels and all of the glucose has been absorbed from the gut. Such a reactive hypoglycemia may be, but rarely is, sufficiently severe to produce (neuroglycopenic) symptoms even in perfectly healthy individuals.

Disposal of an Oral Glucose Load

The exact disposition of glucose absorbed from the gut after a carbohydrate-rich meal by healthy subjects varies widely from individual to individual, and it depends on the size, composition, and physical nature of the meal. Within 4 h of ingestion, approximately 70% of a 70-g oral glucose load given in solution is removed by peripheral tissues, where most of it is used to generate energy by oxidation to carbon dioxide and water or turned into metabolites. The remaining 30% is removed by the liver

during its passage from the gut to the periphery and converted into glycogen, triglycerides, and other metabolites.

Volunteers given a meal consisting of glucose (1 g per kilogram bodyweight) as a 45% solution on an empty stomach reduced their normal basal release of glucose from preformed glycogen in the liver from $9\,g\,h^{-1}$ to approximately $2.2\,g\,h^{-1}$ (i.e., approximately 75%). This persists for the period (3 or 4 h) during which glucose is absorbed from the gut. In other words, although there is a small net uptake of glucose by the liver following a carbohydrate-rich meal, the liberation of glucose from preformed glycogen does not cease completely. Glycogenolysis and gluconeogenesis take place simultaneously but at different rates, depending on whether glucose is being absorbed as well as on the amount and nature of the hormones released by the pancreas and intestine in response to the presence of food.

Fructose and Galactose

Before being absorbed, sucrose is cleaved into glucose and fructose, and lactose is cleaved into glucose and galactose. Galactose shares a transporter mechanism with glucose, whereas fructose uses a less efficient one of its own. Fructose and galactose, and a percentage of absorbed glucose, are removed on their first pass through the liver and converted into glycogen. This provides a store of carbohydrate that is released as glucose into the body pool when absorption from the gut is no longer occurring and gluconeogenesis has not yet become fully reestablished.

Starches and their partial hydrolytic products are converted enzymatically into glucose in the gut lumen and mucosal brush border at a rate that depends on their composition and physical form. Some starches are absorbed as rapidly as preformed glucose, whereas others are absorbed much more slowly. This is reflected in the rate and magnitude of the increase in blood glucose concentration that follows their ingestion and is referred to as the glycemic index.

The Postabsorptive Stage

The exact duration of the absorptive phase following ingestion of a meal depends on many factors, including the size of the meal, its composition, physical nature, and energy density as well as the rate of gastric emptying. Studies based on measurement of intestinal hormones that are released only in response to the absorption of food indicate that the average adult who eats three meals per day is rarely truly postabsorptive except during the night. During this comparatively brief fasting period, glucose lost from the glucose pool by its constant drain into the brain is replaced by glucose derived from the breakdown of liver glycogen. This can come either from reserves built up during the absorptive phase of a meal or by synthesis from glucose precursors, such as lactate, pyruvate, glycerol, and alanine, brought to it in the blood from peripheral tissues during fasting. Glucose synthesis, or gluconeogenesis, is increased by rising levels of glucagon and fatty acids in the blood, which are a consequence of decreasing plasma insulin levels.

The amount of glycogen in the liver varies with the nature of the diet and the size, composition, and timing of the last meal. The average amount of glycogen in the liver after an overnight fast is approximately 44 g (range, 15–80 g) and surprisingly does not increase very much after a meal. Nevertheless, after 36 h without food, liver glycogen stores may decline to as low as 4–8 g. Paradoxically, more prolonged fasting has little additional effect: Indeed, hepatic glycogen stores may actually be replenished as the brain shifts from using glucose to β-hydroxybutyrate as its main source of energy.

Glycogen probably never disappears from the liver completely, except *in extremis*, and there is evidence that it may be an intermediary in the production of glucose by the gluconeogenic pathway. Striatal muscles, which lack glucose-6-phosphatase, cannot convert the glycogen they contain into glucose and release it into the blood. Instead, they release its main breakdown products, lactate and pyruvate (and the latter's transamination partner, alanine), into the blood for conversion into glucose in the liver.

Gluconeogenesis

Gluconeogenesis is the process wherein the liver and, to a smaller but often significant extent, the kidneys make new glucose molecules from chemically simpler compounds. In humans, lactate is probably the most important glucose precursor, especially during exercise. Others, in order of importance, are alanine, pyruvate, glycerol, and some glucogenic amino acids, including glutamate. Glutamate is especially important in gluconeogenesis in the kidney. Fatty acids, apart from propionate formed in the colon by bacterial fermentation of nonabsorbable carbohydrates, do not serve as glucose precursors to any significant degree but do provide the conditions under which it can take

place. So too do specific hormones, such as glucagon and cortisol.

The contribution by alanine to gluconeogenesis has probably been exaggerated. Although formed along with other amino acids by proteolysis of non-structural muscle proteins during periods of prolonged fasting and starvation, its main role under normal conditions is to transport, after transamination, three-carbon skeletons (e.g., pyruvate) derived from muscle glycogen to the liver, where it is converted into glucose during fasting.

Eating inhibits gluconeogenesis mainly through an increase in insulin and decrease in glucagon action. Fasting produces the opposite effect. Alcohol specifically inhibits gluconeogenesis from lactate but not other substrates, such as alanine. It does so by adversely changing the redox potential within the hepatocytes and reducing the availability of nicotinamide adenine dinucleotide, which is an essential component in the formation of glucose from lactate. The inhibition of gluconeogenesis by quite modest amounts of alcohol can sometimes be so profound that people, especially children, with reduced liver glycogen stores may develop hypoglycemia of a severity that can be fatal.

Hormones and Glucose Homeostasis

Insulin is the only major hormone capable of lowering blood glucose levels (**Table 1**). It does so by inhibiting glycogen breakdown in the liver and inhibiting gluconeogenesis and by encouraging glucose uptake by peripheral tissues. It achieves this mainly by activating the glucose transporter protein, GLUT-4, an action that is enhanced by exercise and hyperglycemia. Consequently, insulin lowers

Table 1 Hormones that affect blood glucose concentrations

Hormones concerned with glucose homeostasis

Hormone	*Source*	*Stimuli*	*Inhibitors*	*Main effect on glucose homeostasis*
Insulin	B-cells of islets	Hyperglycemia, incretins (i.e., GIP & GLP-1), glucagon, some amino-acids e.g., arginine, leucine: parasympathetic nervous system i.e., vagus	Hypoglycemia, sympathetic nervous system and adrenaline, somatostatin	Reduced blood glucose concentration by Inhibition of hepatic glycogenolysis and gluconeogenesis, permitting peripheral glucose uptake
GIP	K-cells of duodenum, jejunum, ileum	Actively absorbed sugars e.g., glucose and galactose, actively absorbed fats, especially polyunsaturated	Glucagon; insulin	Incretin: stimulates insulin secretion only in presence of hyperglycemia
GLP-1	L-cells of the ileum and colon	Ingested food; whether absorbed or not	Glucagon; insulin	Incretin: stimulates insulin secretion only in presence of hyperglycemia: inhibits glucagon secretion. Inhibits gastric emptying: reduces appetite
Glucagon	A-cells of islets	Hypoglycemia, adrenaline, some amino-acids e.g., arginine	Insulin, hyperglycemia.	Increases glycogenolysis in liver (not peripheral tissues), enhances gluconeogenesis
Adrenaline	Adrenal medulla	Hypoglycemia, through sympathetic nervous stimulation, physical and mental stress.		Increases glycogenolysis in liver and peripheral tissues, inhibits insulin secretion, stimulates glucagon secretion, impairs peripheral glucose utilization and increases lipolysis (i.e., raises plasma NEFA levels)
Cortisol	Adrenal cortex	Hypoglycemia through hypothalamic release of ACTH		Decreases peripheral glucose uptake, induces insulin resistance, permits hepatic glycogenesis
Growth Hormone	Anterior pituitary	Hypoglycemia, through hypothalamus, ghrelin released from stomach following ingestion of food	Hyperglycemia, Somatostatin, alcohol	Decreases peripheral glucose uptake; increases adipocytes lipolysis
Vasopressin	Hypothalamus and posterior pituitary	Hypoglycemic stress; dehydration	Hypo-osmolality: alcohol	Stimulates hepatic glycogenolysis

blood glucose by two independent mechanisms. Which of the two actions predominates depends on the circumstances, including the actual concentration of insulin in the blood. Another is whether it is of exogenous or endogenous (pancreatic) origin. Exogenous insulin reaches peripheral tissues at a concentration in blood equal to or greater than that in blood perfusing the liver and is unaccompanied by C-peptide. Endogenous insulin, on the other hand, reaches the liver at a higher concentration than peripheral tissues and is accompanied by C-peptide, for which there is increasing evidence of synergism with insulin action.

Insulin released into the portal circulation is partially or, when its concentration in portal blood is low, almost completely removed by the liver. Not all tissues on which insulin acts are equally sensitive to its actions.

At the concentration at which insulin normally circulates in peripheral blood of fasting subjects ($<30\,\text{pmol}\,\text{l}^{-1}$) it depresses, but does not completely suppress, the release of fatty acids from adipocytes and completely inhibits glucose uptake by striatal muscle. At insulin concentrations seen in peripheral blood in the absorptive phase of a meal ($\sim$150–600 $\text{pmol}\,\text{l}^{-1}$) it enhances peripheral glucose uptake and is responsible for the marked arteriovenous glucose difference observed at this time.

The release of insulin from the B cells of the pancreatic islets depends on the concentration of glucose in the blood perfusing them. At blood glucose levels less than approximately 3.5–4.0 $\text{mmol}\,\text{l}^{-1}$, insulin secretion is minimal (constitutive). This means that as the arterial blood glucose declines toward its basal level in the postabsorptive state, plasma insulin levels also decline. However, they never decline low enough in the nondiabetic subject to permit uncontrolled liberation of glucose by the liver or fatty acids by adipocytes. This does, of course, happen when the B cells are destroyed, as in C-peptide negative type 1 diabetes, and is the cause of the hyperglycemia and ketosis that are the hallmark of this illness.

During prolonged starvation (20–50 days without food), small amounts of insulin still reach the liver. However, the amount reaching the adipocytes is sufficiently small to permit lipolysis by adipocytes, and their release of fatty acids into the circulation is sufficient to produce hyperketonemia (10–20 $\text{mmol}\,\text{l}^{-1}$) comparable to that seen in diabetic ketoacidosis. The situation differs, however, from that in diabetes, in that the restraining effect of insulin on hepatic gluconeogenesis and glycogenolysis remains. Consequently, blood glucose levels remain normal rather than grossly elevated and few ill effects develop except weight loss.

An important consequence of reduced insulin secretion as blood glucose levels decline after absorption of a meal is emancipation of the A cells, which are 'down stream' of B cells in the islet, from its suppressive effect on their own release of glucagon. Glucagon reaching its target organ, the liver, promotes both glycogenolysis and gluconeogenesis, thus reversing the effect exerted by insulin during the absorptive phase. In other words, each of the approximately 1 million pancreatic islets functions independently as a miniature glucostat.

Counterregulatory Hormones

Although it is possible to explain the control of blood glucose largely by means of the servoregulatory control of insulin and glucagon secretion described previously, the body has many other neural and hormonal mechanisms to correct or overcome a decline in blood glucose below the critical level (approximately 3.5 $\text{mmol}\,\text{l}^{-1}$) necessary to maintain normal brain function. The sensors for this regulatory function are located in at least two anatomically distinct sites—within the brain and in the portal vein.

The following are the most important mechanisms involved:

Stimulation of the sympathetic and parasympathetic nervous systems, which in turn leads, respectively, to release of adrenaline from the adrenal medulla and noradrenaline from nerve terminals in the liver, and glucagon from the pancreas

Secretion of growth hormone, prolactin, and adrenocorticotrophic hormone by the anterior pituitary gland, cortisol by the adrenal cortex, and vasopressin by the posterior pituitary gland

None of these hormones, apart from cortisol, appears to be absolutely essential for the maintenance of normal glucose homeostasis, but all are brought into play under adverse conditions. They produce their hyperglycemic effects in a variety of ways that can be summarized as follows:

1. Increasing the liberation of glucose by the breakdown of preformed glycogen in the liver (e.g., glucagon, adrenaline, noradrenaline, and vasopressin)
2. Increasing gluconeogenesis in the liver (e.g., glucagon and cortisol)
3. Decreasing peripheral glucose utilization by peripheral tissues (e.g., growth hormone, cortisol, and prolactin)

Hyperglycemia and the Glycemic Index

In contrast to the numerous processes that protect against blood glucose falling too low, there is only one that protects the body from hyperglycemia—the release of insulin into the blood in response to the ingestion of food. Plasma insulin concentration, although neither its rate of increase nor its effectiveness (which depend on intrinsic physiology of the B cells and peripheral insulin sensitivity, respectively), is in large part determined by the increase in arterial blood glucose concentration that follows ingestion of a carbohydrate-containing meal. People who develop type 2 diabetes often have a delayed B cell response to intravenous glucose before showing overt evidence of impaired glucose tolerance in response to a meal. This is because their B cells, although relatively insensitive to hyperglycemia alone, remain sensitive to the hormones GIP and GLP-1, collectively known as incretins and that are released into the circulation from endocrine cells in the intestine in response to actively absorbed carbohydrates and fats. The incretins are able to stimulate insulin secretion only in the presence of hyperglycemia and do not do so when the meal contains little or no carbohydrate, thereby avoiding the risk of hypoglycemia following ingestion of a high-fat meal that does release GIP but virtually no insulin.

Other hormones released from the intestine in response to dietary constituents indirectly affect disposal of absorbed nutrients into the tissues and appetite as well as regulate gastrointestinal functions, such as the rate of gastric emptying and secretion of digestive enzymes.

There is a school of thought that maintains that the rate and magnitude of increase in blood glucose following ingestion of a carbohydrate-containing food compared to those of a comparable amount of glucose in solution or white bread eaten alone can usefully be expressed as its glycemic index. It further avers that low glycemic index foods are nutritionally preferable to high glycemic index foods, the ingestion of which various chronic illnesses are attributed. This has been disputed, and there is little evidence that the concept has much relevance to people with healthy pancreatic endocrine function eating combinations of diverse foods in their everyday life. There is indirect evidence that it may have value in determining dietary choices by people with diabetes whose pancreatic endocrine function is faulty or absent.

Glycosuria

Approximately 100 g of glucose is normally filtered from the blood at the glomeruli of the kidneys each day, more than 99% of which is reabsorbed by the kidney tubules. As a result, healthy people lose less than 150 mg of glucose in their urine each day, an amount too small to be detected by most simple screening procedures for glycosuria. When, for any reason—the most common cause of which is a blood glucose concentration of 10 mmol l^{-1} or higher—the amount of glucose filtered at a glomerulus is more than can be reabsorbed by its tubule, glucose appears at high concentration in the urine. The osmotic diuresis so produced is associated with an increased excretion of water, sodium, chloride, and potassium and is often the first clue to the existence of hyperglycemia, the characteristic hallmark of diabetes. Moreover, it is their loss and not that of glucose that leads to the fatal outcome of diabetic ketoacidosis in patients with untreated type 1 diabetes.

Because their concentration in blood is ordinarily extremely low, except in certain rare diseases, neither fructose nor galactose occur in the urine of healthy people.

See also: **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Fructose**. **Galactose**. **Glucose**: Chemistry and Dietary Sources; Metabolism and Maintenance of Blood Glucose Level; Glucose Tolerance.

Further Reading

Bolli GB and Fanelli CG (1999) Physiology of glucose counterregulation to hypoglycemia. *Endocrinology and Metabolism Clinics of North America* **28**: 467–493.

Foster-Powell K, Holt SHA, and Brand-Miller JC (2002) International table of glycaemic index and glycaemic load values: 2002. *American Journal of Clinical Nutrition* **76**: 5–56.

Jackson RA, Blix PM, Mathews JA *et al.* (1983) Comparison of peripheral glucose uptake after oral glucose loading and a mixed meal. *Metabolism* **32**: 706–710.

Ludwig DS (2002) The glycaemic index: Physiological mechanisms relating to obesity, diabetes and cardiovascular disease. *Journal of the American Medical Association* **287**: 2414–2423.

Marks V (1989) Glycaemic responses to sugars and starches. In: J Dobbing (ed.) *Dietary Starches and Sugars: A Comparison*, pp. 150–167. London: Springer-Verlag.

Marks V, Samols E, and Stagner J (1992) Intra-islet interrelationship. In: Flatt PR (ed.) *Nutrient Regulation of Insulin Secretion*, pp. 41–57. London: Portland Press.

Olson AL and Pessin JE (1996) Structure, function, and regulation of the mammalian facilitative glucose transporter gene family. *Annual Review of Nutrition* **16**: 235–256.

Wahren J and Johansson B-L (1998) New aspects of C-peptide physiology. *Hormone & Metabolic Research* **30**: A2–A5.

Glucose Tolerance

B Ahrén, Lund University, Lund, Sweden

Definition and Impact of Glucose Tolerance

The prevalence of type 2 diabetes is steadily increasing and it has been estimated that the prevalence will increase during the next 25 years to reach epidemic levels. It is assumed that within 25 years, if the trend is not altered, more than 25% of the global adult population older than 65 years of age will be affected by diabetes. The prevalence also shows ethnic differences, with prevalence ranging by a factor of 10 between different populations. Also, the prevalence of diabetic complications is high, resulting in high and significant morbidity. Diabetes is also a major risk factor for cardiovascular diseases, and the majority of deaths in those with diabetes are due to cardiovascular or cerebrovascular disorders. Altogether, this makes diabetes a major burden for global health and health economy.

A most important factor underlying the morbidity in diabetes, its complications and concurrent cardiovascular diseases, is hyperglycemia. Importantly, even at such a low degree as not to reach the limit criteria for diabetes, hyperglycemia is related to morbidity. Lifestyle changes and pharmacological interventions to reduce or even normalize the hyperglycemia exist, and consistent adherence to such regimen will reduce the morbidity. However, hyperglycemia is initially without symptoms and therefore usually remains undetected for a long period of time. Therefore, it is important to have reliable methods for the detection of hyperglycemia in its initial stages for proper actions to be taken. Such detection relies on analysis of the circulating glucose in the fasting state or after a challenge. Thus, it is important to recognize that hyperglycemia is subdivided into two different entities. The first entity is fasting hyperglycemia. This is mainly due to inappropriately high release of glucose from the liver, which is in turn caused by excessive glucagon levels in combination with low insulin levels and/or deficient action of insulin to restrain glucose release from the liver. Standardization of the sample is usually defined as 8- or 12-h fast. The second entity of hyperglycemia is postchallenge hyperglycemia, which occurs after meal or glucose ingestion. This is called 'glucose intolerance' and is equivalent to an impairment to dispose glucose after a challenge. Several modes to diagnose glucose intolerance exist. However, the gold standard for its diagnosis is the oral glucose tolerance test (OGTT). This article describes this test, its advantages and limitations, and its potential role for early detection of patients with increased risk for developing type 2 diabetes and cardiovascular diseases. The article also summarizes the basic mechanisms determining glucose tolerance as well as epidemiological and clinical aspects of glucose intolerance, including the potential of treating the condition for prevention of diabetes and cardiovascular diseases.

Glucose Tolerance Tests

History and Definition of Oral Glucose Tolerance Tests

A major breakthrough in the understanding of glucose intolerance as a basis and risk factor for development of type 2 diabetes and cardiovascular diseases was the introduction of a worldwide standardization of the OGTT in the 1970s. By this introduction, glucose tolerance became a standardized entity, which enabled studies in metabolism, physiology, and clinical medicine with detection of risk factors as well as progressive follow-up studies using a standard recognized worldwide. At the same time, and also of significant importance for the generation of present-day knowledge within the field, was the introduction of the clinical entity impaired glucose tolerance (IGT), which replaced the term "borderline" diabetes. A problem with the term borderline diabetes was that its definition was not uniform, which was partly due to inconsistencies in the procedure of performing a glucose tolerance test, with the amount of glucose ingested varying from 50 to 100 g or given on a kilogram basis. IGT as an entity was thus introduced simultaneously with the suggestion that glucose tolerance in a clinical test should be determined following ingestion of 75 g glucose, with a blood sample for the measure of glucose to be taken 2 h later.

The evaluation of the standardized OGTT in the clinical setting relies on the 2-h glucose value. This value during the 75 g OGTT usually displays a normal distribution slightly skewed to the right. **Figure 1** shows the distribution pattern of 2-h glucose levels obtained from 802 Caucasian subjects in the Malmö Prevention Study, an epidemiological study from Sweden. From this distribution, normal values may be defined statistically from mean and variance values for statistical definition of the distribution. The mean value, as in most studies, is $\approx$7 mmol/l and standard deviation is $\approx$1 mmol/l. By defining reference values as 95% confidence intervals, the cutoff value for normality would be

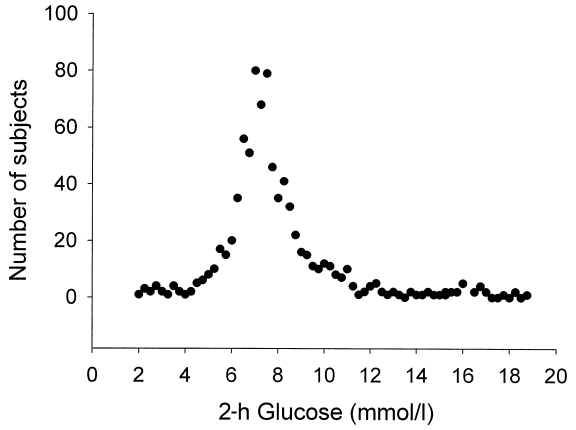


Figure 1 Distribution of the 2-h glucose value in an OGTT performed in 802 Caucasian subjects from the Malmö Prevention Study (unpublished data).

approximately 9 mmol/l and, hence, values higher than 9 mmol/l would indicate diabetes. By using such a definition of diabetes, a large number of subjects would have the disease, the clinical relevance of which is doubtful. Therefore, the definition of diabetes has instead been based on prospective studies evaluating the risk for microvascular disease and the cutoff-levels have been defined as levels increasing this risk. Therefore, a cutoff value of 11.1 mmol/l glucose has been used for the definition of type 2 diabetes.

OGTT was frequently used during the 1980s and 1990s for the clinical diagnosis of type 2 diabetes and in epidemiological studies, which markedly increased our knowledge of these conditions. By the end of the 1990s, however, definitions of IGT and clinical tests to be performed were again discussed. This resulted in revised cutoff levels and the introduction of a new entity called impaired fasting glycemia (IFG), which is defined as high fasting glucose. **Table 1** shows the new cutoff values for the three diagnostic entities—IFG, IGT, and diabetes. It was also suggested that fasting glucose was sufficient for the diagnosis of glucose intolerance and type 2 diabetes.

Table 1 Cutoff values for fasting and 2-h glucose values (mmol/l) of impaired glucose tolerance (IGT), impaired fasting glucose (IFG), and type 2 diabetes (T2D) in an oral glucose tolerance test according to guidelines by the American Diabetes Association

	Plasma	*Whole blood*	
	Venous	*Venous*	*Capillary*
T2D			
Fasting glucose	≥7.0	≥6.1	≥6.1
2-h glucose	≥11.1	≥10.0	≥11.1
IGT			
2-h glucose	7.8–11.0	6.7–9.9	7.8–11.0
IFG			
Fasting glucose	6.1–6.9	5.6–6.0	5.6–6.0

The suggestion that a fasting sample is sufficient for the diagnosis of abnormal glycemia has been questioned, however, mainly because studies have shown that such a strategy will reduce the numbers at risk who are diagnosed and detected. This is because a large proportion of subjects with IGT have a normal fasting glucose but an elevated 2-h glucose value. In fact, there are populations with IFG alone, IGT alone, and IFG and IGT together, and these populations may represent different risks for diabetes and cardiovascular diseases. Consequently, those having a high 2-h glucose value but a normal fasting glucose, who also have increased risk for cardiovascular diseases, will be missed by the suggested strategy. A study by Larsson and collaborators from Sweden identified this dilemma since it was demonstrated that out of 414 subjects with abnormal fasting or 2-h glucose values during an OGTT, only 140 (34%) had elevation of both values. The largest group comprised subjects with high 2-h glucose values but normal fasting glucose values (i.e., IGT but not IFG), which were seen in 235 subjects (57%), whereas only 39 subjects (9%) had high fasting but normal 2-h glucose values (i.e., true IFG). The individual subgroups were shown to have similar risk factor patterns in terms of degree of obesity, blood pressure, and lipid levels. Therefore, it is now obvious that for a proper strategy to detect early cases at risk for diabetes and cardiovascular diseases, an OGTT needs to be performed since this test includes both fasting and postchallenge glucose determination.

Procedures and Evaluation of the Oral Glucose Tolerance Test

Glucose tolerance is defined as the ability to dispose a glucose load, and therefore glucose intolerance is defined as an impaired ability for glucose disposal. The gold standard technique is to challenge with an oral glucose load, with measurement of circulating glucose before and after the challenge—the OGTT. As routinely performed, this test determines the ability to dispose glucose after oral administration of 75 g glucose. The test is standardized such that it is performed in the morning after a 12-h overnight fast and blood samples are taken before the glucose load and after 2 h. Furthermore, the diet during the 3 days preceding the test should contain at least 250 g carbohydrates per day and the subjects should rest during the test in a semirecumbant position

without smoking. The glucose given should be dissolved in 250–300 ml fluid, and sometimes fruit-flavored water is used to improve the taste. There has been much debate about how to take the blood sample. The original diagnostic criteria used values obtained from plasma derived from blood taken venously in tubes containing additives for prevention of coagulation. However, valid results are also obtained when glucose is measured in whole blood and when capillary samples are taken, although cutoff levels need to be adjusted for the different glucose concentrations in these samples. Arterial samples are also theoretically possible but rarely, if ever, used. Sometimes, mainly for research purposes, more frequent samples are taken and the test may last 3 h; however, for clinical purposes, the routine OGTT lasts 2 h, with a sample taken at that time point.

As shown in **Figure 2**, in a normal person, circulating levels of glucose increase within the first 15 min after the oral ingestion of glucose to reach a peak after 30 min. Thereafter, a progressive decline occurs, with the 2-h value usually approximately 25% higher than the fasting value. Usually, it takes 3 h for a return to baseline glucose levels. In subjects with IGT, there is usually also a peak at 30 min, albeit at a higher level than in normal subjects, but the main difference versus normal subjects is that the glucose disposal is impaired, which results in a higher 2-h glucose value. In diabetics, there is usually not a peak at 30 min but a continuous rise throughout the 2-h study period. The currently used cutoff values are shown in **Table 1**. Note that the mode of measuring glucose is important with regard to the cutoff values used.

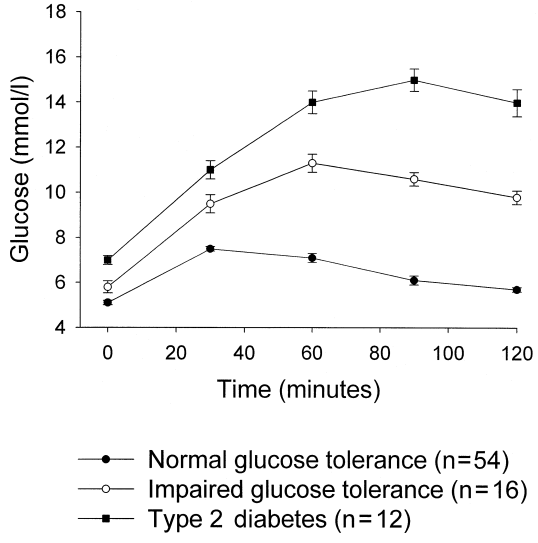


Figure 2 Venous plasma glucose levels during OGTT in subjects with normal impaired glucose tolerance, impaired glucose tolerance, and diabetes. From Ahrén B, unpublished data. Means ± SEM are shown.

Limitations of the Oral Glucose Tolerance Test

An important limitation of the OGTT is the variability in results when the test is repeated. Actually, the coefficiency of variance (CV) is usually 15% and in some studies 20%, which is higher than that for most other clinical tests. It is not clear why OGTT has such a high CV. The variance is not, however, dependent on CV in the measurement of glucose, which is a procedure with very small error and CVs usually below 3%. Therefore, biological variation probably explains the high CV of OGTT. Factors explaining this variation may be preceding diet, exercise, emotions, stress, drugs taken for various diseases, and gender, which are all factors influencing gastric emptying, carbohydrate absorption, islet hormone secretion, hepatic glucose production, and peripheral glucose uptake (i.e., all processes contributing to the 2-h glucose value). Because of the high variability in the 2-h glucose value, a diagnosis of IGT or diabetes, particularly if intervention is planned, should not be based on a single OGTT. Instead, a clinical recommendation is to perform two OGTTs and use the mean of the two 2-h glucose values as the diagnostic value. The time interval between the two OGTTs should not exceed 3 months.

Metabolic Basis for Oral Glucose Tolerance

Oral ingestion of glucose initiates a series of metabolic perturbations, which comprise the 2-h glucose value. These metabolic perturbations are complex and involve glucose entering the bloodstream, changes in neural activity and islet hormone secretion, suppression of hepatic glucose production, and stimulation of peripheral glucose uptake. From a quantitative standpoint, of most importance with regard to the 2-h value are the changes in islet hormone secretion, which include stimulation of insulin secretion and inhibition of glucagon secretion, and the suppression of hepatic glucose production. In fact, there is an inverse linear relation between the inhibition of hepatic glucose production and the 2-h glucose value and, similarly, a linear inverse relation between stimulation of the early (first 30 min) insulin secretion and 2-h glucose. This section briefly summarizes these processes.

A first series of events in the OGTT is initiated during the anticipation of the oral glucose ingestion, through olfactory stimuli and through receptors located in the oral cavity. This response is called

the cephalic phase and activates sensory nerves, which give input to the central nervous system. This information is integrated in the hypothalamus for initiation and adjustment of a vagal nerve response to release insulin from the pancreatic islets. Therefore, when analyzed in detail, there is an increase in circulating insulin after glucose or meal ingestion already before glucose levels become elevated. After passage of glucose through the oral cavity, glucose passes to the stomach and through a regulated mechanism delivered into the gut. Since glucose is a monosugar, it is readily absorbed in the small intestine and reaches the splanchnic venous drainage. Glucose then passes to the portal vein and the liver. In the portal vein, glucose activates glucosensitive receptors, which through afferent sensory nerves send signals centrally to the brain for further integration with the previous signals in the hypothalamus for adjustment of efferent nerve activity. Furthermore, glucose in the liver inhibits hepatic glucose production, which is high after the overnight fast. Then, glucose passes to the general circulation to reach the pancreatic islets and the peripheral cells. The glucose load to the gut also stimulates the release of intestinal hormones, such as glucose-dependent insulinotropic polypeptide and glucagon-like peptide-1 (GLP-1). These hormones then pass through the circulation to reach the pancreas, where they stimulate insulin secretion and, as for GLP-1, inhibit glucagon secretion. In the pancreatic islets, vagal activation, intestinal hormones, and glucose stimulate insulin secretion, and glucose, GLP-1, and insulin inhibit glucagon secretion. These islet responses are of major importance for a normal glucose tolerance, and defects in these islet responses are major determinants of IGT and type 2 diabetes. Following passage of insulin into the venous drainage of the pancreas, the islet hormones reach the portal vein and the liver, and a main function of insulin is to potently suppress hepatic glucose production. This is a major process with regard to the degree of hyperglycemia during the test; in subjects with inappropriately high hepatic glucose production, the glucose level after oral glucose is high. This suppression of hepatic glucose production is augmented by the reduction in circulating levels of glucagon, which is initiated by the direct action of glucose and GLP-1 on the glucagon-producing cells and also by the action of insulin to inhibit glucagon secretion. After the liver, glucose and insulin reach the peripheral circulation and peripheral cells, where glucose is transported across the cell membranes and therefore leaves the circulation. In most cells, a major proportion of glucose uptake is sensitive to insulin; therefore, the amount of insulin, in relation to the insulin sensitivity of the cell, is of major importance for the delivery rate of glucose. However, insulin-independent mechanisms also exist, even in tissues, which are also insulin sensitive, and glucose uptake is thus also dependent on glucose. Of most importance for glucose disposal after oral glucose is the muscle cells, which have a high capacity for glucose uptake. From all these processes, the glucose level at 2 h can be determined.

It is important to realize that the metabolic processes underlying glucose tolerance are different from those underlying the fasting glucose value. Fasting glucose is mainly determined by hepatic glucose delivery during the night, which in turn is governed by the ability to maintain normal basal insulin and glucagon levels. Therefore, mechanisms underlying IFG include defective insulin secretion, defective suppression of glucagon secretion, defective sensitivity in the liver for the action of insulin, and defective peripheral glucose disposal at low glucose levels, which is a sign of insulin resistance. Although mechanisms underlying fasting and 2-h glucose values differ, there is a high correlation between fasting and 2-h glucose values in normal subjects, as shown in **Figure 3**. Nevertheless, there is a limited overlap between IGT and IFG in a population; in fact, most subjects with IGT have normal fasting glucose, and most subjects with IFG have a normal 2-h glucose value. This suggests that different pathophysiological processes underlie IGT and IFG, which in turn suggests that OGTT should be undertaken more frequently than performed today.

Differential Tests for Glucose Tolerance

Diagnoses of type 2 diabetes or stages preceding its occurrence can be undertaken by other means

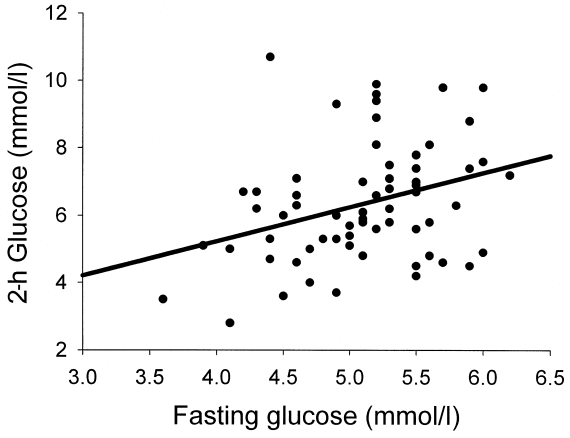


Figure 3 Correlation between fasting glucose and 2-h glucose during an OGTT in nondiabetic subjects. The regression is significant ($r = 0.32$, $p = 0.008$). From Ahrén B, unpublished data.

besides OGTT. As previously stated, the use of fasting glucose has been suggested as the gold standard for diagnosis during recent years. Although it has a lower CV than the 2-h glucose value after OGTT and is simpler and more convenient for both the subject and the staff, the problem with this method is that a large number of subjects with IGT, namely those with a normal fasting glucose, will be missed.

As an alternative to OGTT, glucose tolerance may also be determined by administering glucose intravenously. In the intravenous glucose tolerance test (IVGTT), glucose is injected intravenously, usually at a dose of 0.3, 0.5, or 1 g/kg, and circulating glucose is determined before and 8, 10, 15, 20, 30, 40, 50, 60, and 80 min after injection. Glucose tolerance is estimated from the elimination rate, where a glucose elimination constant (k_g) is calculated. The theory behind this is that the glucose elimination after intravenous glucose displays an exponential function (i.e., after logarithmic transformation of the data, the elimination is linear. k_g is thus calculated as the slope for the glucose curve following logarithmic transformation of the individual glucose values and is calculated from the formula $k_g = (0.693 \times 100)/t_{1/2}$, where $t_{1/2}$ is the half-time of glucose elimination (in minutes). The unit for k_g is percentage of glucose decay per minute. **Figure 4** shows this condition. Before OGTT was routinely used, IVGTT was undertaken more frequently. Unless very specific questions are asked, it is currently not used in clinical practice because it is more cumbersome to perform and it identifies only some of the metabolic processes underlying glucose tolerance, mainly insulin secretion, insulin sensitivity, and glucose uptake. Thus, other important aspects, such as glucagon secretion, release of incretin hormones, and hepatic glucose output, which are involved in the overall glucose tolerance and included in the 2-h glucose value after OGTT, contribute only marginally to the k_g after IVGTT.

It has been suggested that measurement of HbA_{1c} (i.e., the fraction of hemoglobin being glycosylated) may be used for the diagnosis of IGT and type 2 diabetes. The rationale is that hemoglobin is irreversibly glycosylated in proportion to the glucose level, and therefore the proportion of HbA_{1c} should reflect the mean of the glucose levels during the preceding 2 or 3 months. However, although this theoretical assumption is true, measurements of HbA_{1c} are not precise, have not been standardized at levels near the normal levels, and, consequently, slight elevations of HbA_{1c} will not distinguish normal from impaired glucose tolerance with fairly high precision. In addition, most subjects with IGT have HbA_{1c} levels within the normal distribution.

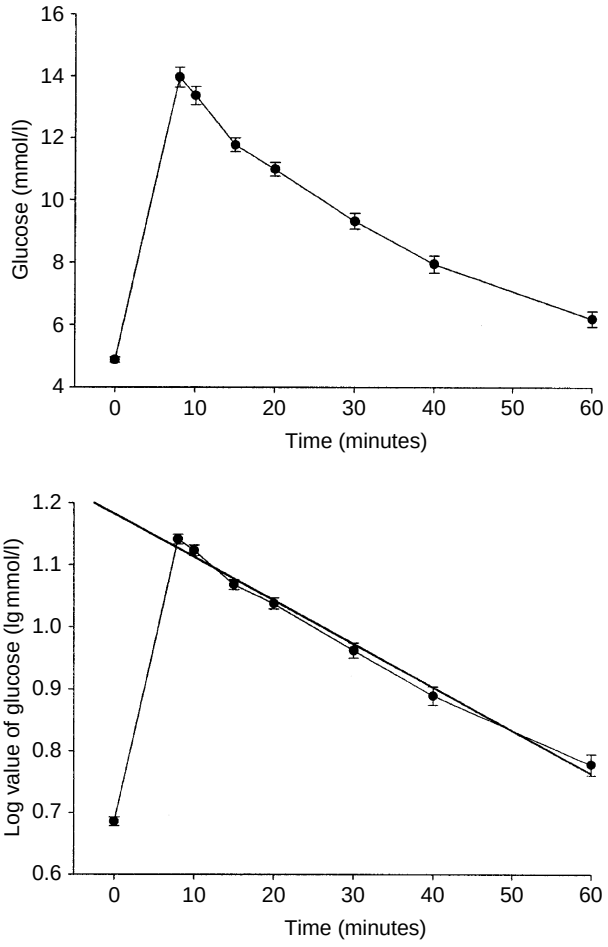


Figure 4 Glucose levels during an IVGTT in 41 healthy subjects with normal glucose tolerance. Glucose (0.3 g/kg) was injected intravenously at time 0. Linear regression curve for the logarithmic values from minutes 8 to 60. k_g value = 1.61 ± 0.08%/min. From Ahrén B, unpublished data. Means ± SEM are shown.

Clinical Aspects of IGT

Epidemiology of IGT

During the 1980s, studies on the prevalence of IGT and type 2 diabetes were performed in several different populations. It became apparent that the prevalence of these conditions, although high in many populations, varied markedly between different populations. Thus, for some populations, mainly in Africa, data from only a few percent were published, whereas an exceedingly high prevalence (60%) was reported in some populations, such as Pima Indians. **Figure 5** shows a collaborative study from 1993 in which data from several populations throughout the world are summarized. Studies during the past 10 years have further increased our knowledge since they have included additional populations and demonstrated that the

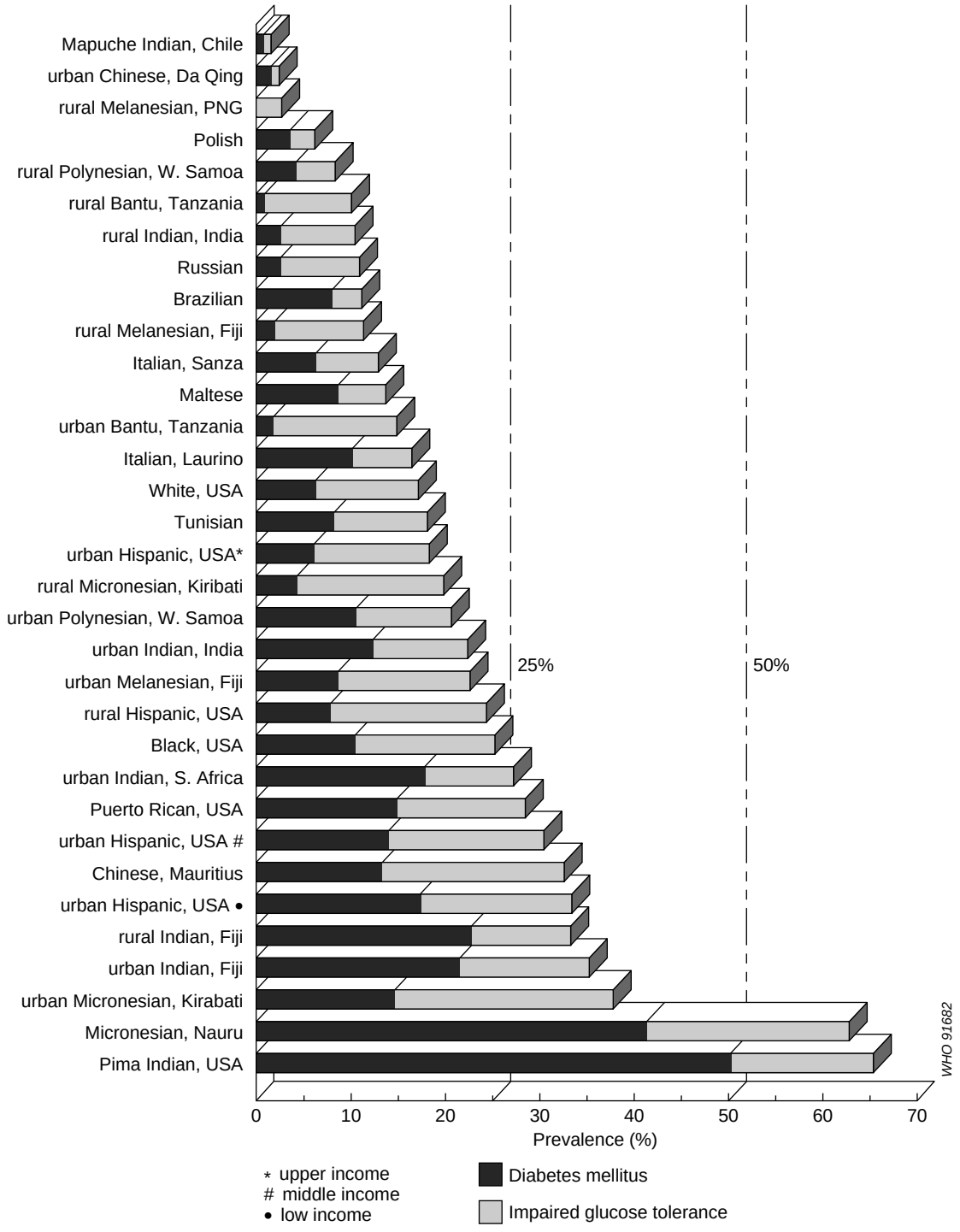


Figure 5 Prevalence (%) of diabetes and IGT in selected populations in the age range of 30–64 years; genders combined. Copyright © 1993 American Diabetes Association. From *Diabetes Care* volume 16, page 170. Reprinted with permission from the American Diabetes Association.

prevalence of IGT and type 2 diabetes is steadily increasing. Hence, the prevalence reported in 1993 is an underestimation of the prevalence today. It has to be emphasized, however, that the difference in reported prevalence rates between different populations may be partially explained by methodological differences. For example, the prevalence of IGT and type 2 diabetes is increased by age,

and in many populations there is also a higher prevalence in women than in men, at least in younger age groups. Different studies have not controlled for these confounders. Furthermore, due to migration patterns in some populations, generalization of study results is questionable, and there may also be differences in the likelihood of subjects attending a study between different populations. Nevertheless, a true ethnic difference seems to exist, with extremely high values in some Pacific island and North American Indian populations and a low prevalence in South American Indian and Bantu populations. An interesting observation is that the increase in prevalence of IGT and type 2 diabetes seems to be higher in populations with low prevalence rates and vice versa, which probably will result in diminished differences in prevalence rates between different populations in the future.

Clinical Consequences of IGT

IGT is an important risk factor for development of type 2 diabetes. However, prospective and long-term studies report different predictive values for the development of type 2 diabetes in different populations. In general, the risk of transition of IGT into type 2 diabetes ranges from 1–2% to 5% and as high as 15–20% per year. The risk is higher for those older than 50 years of age. There is also evidence that hyperglycemia, even at levels not reaching the threshold for type 2 diabetes, is associated with a substantial risk for the development of cardiovascular diseases. One explanation for this is that glucose initiates metabolic perturbations of importance for developing angiopathy, such as tissue peroxidation, production of plasminogen activation inhibition-1, and impairment of endothelial function, such as nitric oxide production. Another explanation is that hyperglycemia is associated with a number of risk factors for cardiovascular diseases, such as high blood pressure, hyperinsulinemia, dyslipidemia, and microalbuminuria, which all are included in the metabolic syndrome. In fact, if hyperglycemia is present, the risk for developing cardiovascular diseases for each of the other risk factors is augmented. Attempts to define cutoff values of glucose for cardiovascular risks have been problematic, however, probably due to the fact that the risk is continuously increased across the glucose ranges. Hence, the use of defined cutoff values is more a convenient practical issue, which is important in a clinical setting, but offers limitations from a theoretical standpoint.

Since IGT is a risk factor for type 2 diabetes and cardiovascular diseases, it is also a risk factor for overall mortality. Alberti and coworkers attempted to quantify this by performing a meta-analysis on 13 prospective studies, and they identified a hazard ratio of 1.34 (95% confidence interval, 1.14–1.57) by comparing subjects with IGT to those with normal glucose tolerance. The hazard ratio is higher for subjects with IGT than for subjects with IFG, suggesting that the 2-h glucose value is more predictive of mortality than the fasting glucose value. This shows that an individual with IGT has an increased risk not only for type 2 diabetes but also for cardiovascular diseases and hence mortality. This indicates that attempts should be made to prevent IGT from progressing to cardiovascular diseases and type 2 diabetes.

Treatment of IGT

During recent years, the issue of whether IGT may be treated to prevent progression to type 2 diabetes or cardiovascular diseases has gained considerable interest. On the one hand, it has been argued that it is important to prevent progression of IGT. On the other hand, it has been argued that treating such a large population group as those with IGT would be risky. **Table 2** lists criteria that need to be fulfilled to justify prevention of a condition. In fact, most of these criteria are met for IGT; therefore, it may be argued that treating IGT is now justified. The optimal preventive intervention for IGT is not known, however. The intervention may include lifestyle changes, notably increased physical activity and dietary regulations. Such interventions have been shown to be efficient in highly motivated populations and study centers. However, whether generalization of these results to the general population is possible is not known. A clinical experience is that the outcome of advice on lifestyle changes is often disappointing in the long term. Another mode of

Table 2 Criteria for recommending population-based intervention for preventing a disease[a]

Criterion 1:	The disease (IGT and type 2 diabetes) should pose a major health problem.
Criterion 2:	Early development and natural history of the disease (IGT and type 2 diabetes) should be understood to identify parameters that measure its progression.
Criterion 3:	Tests should exist for diagnosing the presumptive population (OGTT).
Criterion 4:	Preventive methods should be safe, efficient, and reliable.
Criterion 5:	Effort to find subjects and cost of intervention should not be burdensome and should be cost-effective.

[a]Based on recommendations from the American Diabetes Association (2004).
IGT, impaired glucose tolerance; OGTT, oral glucose tolerance test.

intervention is pharmacological treatment using compounds to stimulate insulin secretion, suppress hepatic glucose production, and/or enhance insulin sensitivity. These may be efficient, perhaps more efficient than advice on lifestyle changes, but may in turn pose other questions concerning long-term efficiency and potential adverse events. These two strategies are not mutually exclusive, however, and introducing pharmacological intervention without giving lifestyle advice is not appropriate in a clinical setting.

Recently, interesting data from large population studies on the prevention of progression of IGT have been obtained. Two studies, the Finnish diabetes prevention study and the Diabetes Prevention Program, have shown that lifestyle changes (i.e., individualized diet and exercise counseling) in subjects with IGT reduced the incidence of diabetes by more than 50%. In addition, in the Diabetes Prevention Program, it was shown that metformin (which reduces glucose output from the liver) reduces the risk by approximately 30%. This suggests that pharmacological treatment of IGT prevents development of type 2 diabetes. Several large studies are ongoing and results are expected within a few years.

Whether interventional programs on IGT are valid also for the prevention of cardiovascular diseases is not clearly established, mainly because long-term studies have not been performed. The STOP-NIDDM study, however, showed that acarbose, which reduces glucose absorption from the gut, reduced cardiovascular events by more than 30% during a 3-year study period. This suggests that cardiovascular diseases may be prevented by treating IGT. It should be noted, however, that for prevention of cardiovascular diseases and mortality, more studies and longer follow-up periods are required.

Conclusion

Because of the risk of developing type 2 diabetes and cardiovascular diseases in subjects with IGT, and also in those with IFG albeit at a lower level, it is important to diagnose and treat these conditions. This means that OGTT should be undertaken more frequently, at least in subjects found to have high fasting glucose. These subjects should be regarded similarly as other subjects with risk factors for cardiovascular diseases (i.e., those with hypertension and dyslipidemia). This implies that subjects with IGT should be given lifestyle advice and be checked regularly. Ongoing large and long-term prevention trials will also provide information on whether pharmacological treatment should be added.

See also: **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Glucose**: Metabolism and Maintenance of Blood Glucose Level.

Further Reading

American Diabetes Association (2004) Clinical practice and recommendations 2004. *Diabetes Care* **27**(supplement 1).

Chiasson JL, Josse RG, Gomis R *et al.* (2002) Acarbose for prevention of type 2 diabetes mellitus: The STOP-NIDDM randomised trial. *Lancet* **359**: 2072–2077.

DECODE Study Group (1999) Glucose tolerance and mortality: Comparison of WHO and American Diabetes Association diagnostic criteria. *Lancet* **354**: 617–621.

De Vegt F, Dekker JM, Ruhe HG *et al.* (1999) Hyperglycaemia is associated with all-cause and cardiovascular mortality in the Hoorn population: The Hoorn Study. *Diabetologia* **42**: 926–931.

Expert Committee on the Diagnosis and Classification of Diabetes Mellitus (1997) Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care* **20**: 1183–1197.

Isomaa B (2003) A major health hazard. The metabolic syndrome. *Life Sciences* **73**: 2395–2411.

King H, Rewers M and WHO Ad Hoc Diabetes Reporting Group (1993) Global estimates for prevalence of diabetes mellitus and impaired glucose tolerance in adults. *Diabetes Care* **16**: 157–177.

Knowler WC, Barrett-Connor E, Fowler SE *et al.* (2002) Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *New England Journal of Medicine* **346**: 393–403.

Larsson H, Ahrén B, Lindgärde F, and Berglund G (1995) Fasting blood glucose in determining prevalence of diabetes in a large, homogenous population of Caucasian middle-aged women. *Journal of International Medicine* **237**: 537–541.

Larsson H, Berglund G, Lindgärde F, and Ahrén B (1998) Comparison of ADA and WHO criteria for diagnosis of diabetes and glucose intolerance. *Diabetologia* **41**: 1124–1125.

Larsson H, Lindgärde F, Berglund G, and Ahrén B (2000) Prediction of diabetes using ADA or WHO criteria in postmenopausal women: A 10-year follow-up study. *Diabetologia* **43**: 1224–1228.

Tuomilehto J, Lindström J, Eriksson JG *et al.* (2001) Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *New England Journal of Medicine* **344**: 1343–1350.

Unwin N, Shaw J, Zimmet P, and Alberti KGMM (2002) Impaired glucose tolerance and impaired fasting glycaemia: The current status on definition and intervention. *Diabetic Medicine* **19**: 708–723.

Weyer C, Bogardus C, and Pratley RE (1999) Metabolic characteristics of individuals with impaired fasting glucose and/or impaired glucose tolerance. *Diabetes* **48**: 2197–2203.

World Health Organization (1999) *Definition, diagnosis and classification of diabetes mellitus and its complications. Report of a WHO consultation. Part 1: Diagnosis and classification of diabetes mellitus.* Geneva: World Health Organization.

Zimmet P, Alberti KG, and Shaw J (2001) Global and societal implications of the diabetes epidemics. *Nature* **414**: 782–787.

Zimmet P, Shaw J, and Alberti KGMM (2003) Preventing type 2 diabetes and the dysmetabolic syndrome in the real world: A realistic view. *Diabetic Medicine* **20**: 693–702.

GLYCEMIC INDEX

G Frost and A Dornhorst, Imperial College at Hammersmith Hospital, London, UK

In the past 10 years, a number of important epidemiological and experimental studies have linked glycemic index to postprandial glucose metabolism, insulin resistance, and cardiovascular risk factors. The World Health Organization (WHO) and the Food and Agriculture Organization (FAO) recommended that the physiological effects of dietary carbohydrates be classified according to their glycemic index. This review examines the historical and scientific background of the glycemic index.

Background and Definition

In 1939, Conn and Newburgh noted how different carbohydrate-containing foods could have the same macronutrient composition but different glycemic responses. Insulin responses elicited by different carbohydrates also vary. These observations led to the first classification of carbohydrate foods according to their glycemic response, which then allowed different dietary carbohydrates to be exchanged within a meal without altering postprandial glucose levels. The 'glycemic index' was introduced as a means of quantifying the glycaemic response of different dietary carbohydrates.

Glycemic indices of several foods are published in international nutritional tables, the most recent of which was published in 2002. Methodology on their derivation is available from previous reviews. Glycemic index of a food is a measure of postprandial glucose response after a 50-g load of available carbohydrate from the food (**Figure 1**) and provides a standardized comparison of a carbohydrate's 2-h postprandial glucose response with that of glucose (**Table 1**). Low glycemic index carbohydrates have lower 2-h incremental areas under the glucose curve than glucose, whereas high glycemic index foods have higher areas. Although the insulin response is not used to define glycemic index, the lower the glycemic index of a food, the more attenuated is the insulin response to a standard test meal. It has been argued that it is the insulin response to foods and not the glycemic response that is important in the pathogenesis of insulin resistance and related metabolic disturbances and disease risk. Although still an area of debate in general, glycemic index is a surrogate marker of the insulin response to different carbohydrates, with the possible exception of diary products. Indeed, the insulin response in nondiabetic subjects to a wide range of foods (glycemic indices between 32 and 100) are highly correlated. The exception to this is possibly diary products which have an insulin response high than predicted but the glycemic index. This remains unexplained at present. Dietary carbohydrates stimulate insulin secretion both directly by stimulating the pancreatic β cell and indirectly through their secretion effect. The pattern of insulin secretion caused by different

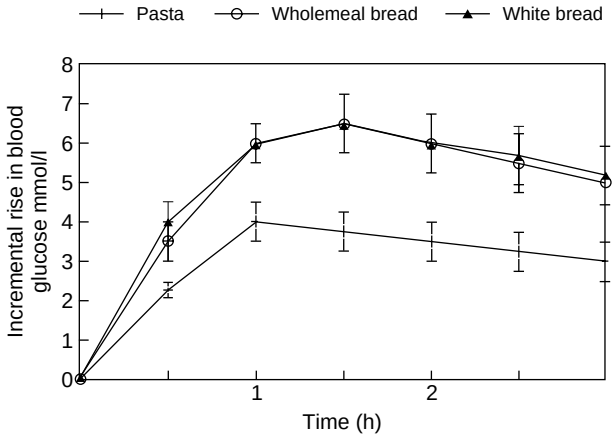


Figure 1 Mean blood glucose increment after equi-available carbohydrate meals. (Data from Jenkins DJ, Wolever TM and Jenkins AL (1988) Starchy foods and glycemic index. *Diabetes Care* **11:** 149–159.)

Table 1 The glycemic index model

$$\frac{\text{Incremental area under blood glucose response curve for the test food containing 50 g of carbohydrate}}{\text{Corresponding area after equicarbohydrate portion of glucose}} \times 100$$

Calculation of the glycemic index of a mixed meal containing three separate carbohydrate-containing foods

$$\text{Glycemic index/mixed meal} \quad (GI_1)(PCF_1) + (GI_2)(PCF_2) + (GI_3)(PCF_3)$$

Where
The three carbohydrate-containing foods are 1, 2, and 3
The glycemic index for each carbohydrate-containing food is GI_1, GI_2, and GI_3
The carbohydrate content is **C** 1, **C** 2, and **C** 3 g
The total meal carbohydrate (TMC) is [**C** 1 + **C** 2 + **C** 3] g
The proportion of carbohydrate from each food (PCF) is PCF1 **C** 1/TMCg, PCF2 **C** 2/TMCg, and PCF3 **C** 3/TMCg

carbohydrates reflects their different intestinal transit times.

Type of Dietary Carbohydrate and the Glycemic Index

The glycemic index of a carbohydrate is influenced by its rate of intestinal absorption, which in turn is influenced by its composition, tertiary structure, type of starch, and susceptibility to enzymic digestion.

Chain Length and Composition

Complex carbohydrates are polymeric chains of repeating monosaccharide units. Starches comprise repeating glucose units. The glycemic indexes of different starches are determined by their susceptibility to enzymic digestion, not chain length. White bread and pasta have similar chain lengths, but bread has a higher glycemic index due to its tertiary structure and solubility that ensures greater exposure to salivary and pancreatic amylases.

Short-chain carbohydrates are rapidly absorbed; however, when they contain nonglucose sugars, the glycemic index is lowered proportionally. The disaccharides sucrose and lactose consist of 50% glucose and 50% fructose or galactose, respectively, and both have a lower glycemic index than maltose, the disaccharide formed from two molecules of glucose.

Amylose and Amylopectin

The starches in cereal grains, rice, potatoes, and all green plants are composed of repeating glucose units arranged in straight (amylose) and branched-chained (amylopectin) polysaccharides. The absorption rate, and hence the glycemic index, of these starches is influenced by the ratio of amylose to amylopectin. The more compact structure of amylose than amylopectin results in a smaller surface area being available for amylase digestion. Amylose-enriched starches therefore have lower glycemic indexes than those enriched in amylopectin.

Relationship of Insoluble and Soluble Nonstarch Polysaccharides (Fiber) to Glycemic Index

Nondigestable complex carbohydrates are commonly known as dietary fiber; the more correct terminology is nonstarch polysaccharides (NSPs). NSPs are either soluble or insoluble. Clinical studies have shown that diets rich in soluble fiber/NSPs, such as guar gum, pectin, and sugar beet fibers, lower postprandial blood glucose and insulin levels. Guar gum, a β-glactomannan from the Indian locust bean, also reduces postprandial lipemia. Nonsoluble NSPs have no effect on dietary glycemic index.

Soluble NSPs, such as pulse vegetables, whole fruits, oats, and barley, form gelatinous gels within the stomach that delay gastric emptying and enzymic digestion, the latter by forming a physical barrier around the carbohydrate. Insoluble NSPs have little effect on gastric emptying and no effect on glucose absorption. High-fiber/high-NSP diets are therefore not necessarily synonymous with low glycemic foods. Cellulose is the most widely used NSP in household cereals, whole meal bread, and brown rice, and since it is insoluble, these foods have the same glycemic index whether replete or deplete of their dietary fiber/NSPs. For unknown reasons, Albran is an exception, and despite its high insoluble fiber content, it has a low glycemic index.

The solubility of dietary fiber/NSPs have benefits on postprandial glycemia and hyperinsulinemia. The reason for this are multifactorial including slowing of gastric emptying, a physical barrier to amylase, possible thickening of the unstirred lear and possative effects on gut incretin hormones such as GLP-1 and GIP. The lack of effect on increasing nonsoluble fiber NSPs on glucose and insulin should not detract from important effects on bowel function and bowel pathology.

Cell Structure, Food Preparation, and Processing

Cooking and food preparation can modify the glycemic index. Highly processed convenience foods tend to have high glycemic indexes. When cooking and processing disrupt the cell wall, the starch granules are broken open, optimizing amylase digestion and increasing the glycemic index. Cooked pulse vegetables have low glycemic indexes because their cell walls are resistant to cooking. The intact cereal grains of pumpernickel rye bread, granary bread, and bulgur wheat all have low glycemic indexes. However, when granary bread is processed to wholemeal bread, these grains are disrupted and the glycemic index rises. Cooling can paradoxically lower the glycemic index of certain starches, such as potatoes, due to the formation of retrograde starches that are resistant to amylase digestion.

Effects of the Upper Gastrointestinal Tract

For many foods, their glycemic index is determined by the process of chewing and swallowing. Chewing can reduce food particle size, which increases absorption rates. This explains why boiled and mashed potatoes have different glycemic indexes. Chewing can also change the constituency of the food such that with bread the particle size is reduced

to such an extent that it behaves more as a fluid on swallowing and is therefore very rapidly absorbed. In contrast, pastes retain their structure on swallowing and are more slowly absorbed. The rate of gastric emptying also influences the glycemic index, with lower glycemic index foods being retained in the stomach for longer periods than high glycemic index foods.

Concerns Related to the Glycemic Index

Whereas the 1998 WHO/FAO dietary carbohydrate guidelines and the 1998 European dietetic guidelines advocate greater use of the glycemic index, the American Diabetes Association's evidence-based guidelines are more cautious, suggesting a "B"-level evidence grade. This is basically due to the lack of long-term studies, there is only one randomized control trial which had a study period of longer than 6 months. Also there remains concerns regarding the effects of mixed meals are difficult to predict. Against this is the observation in well conducted randomized control trials, blood glucose during the low glycemic index diet is lower than the high glycemic index diet. Studies of the long-term efficacy of low glycemic index diets in diabetes, obesity, and coronary risk groups using randomized control methodology are under way and will report during the next 5 years. The issue regarding the predictability of the glycemic index of mixed meals remains a matter of debate, but evidence suggests that the glycemic index of a mixed meal is reasonable when the fat content is low and deteriorates as the fat content of the meal rises. However, this academic debate should not detract from the fact the evidence from randomized control trial suggests positive benefits on glucose, insulin and lipids from low glycemic index diets.

Reproducibility

Within-subject variation The variability of the glycemic response for a given food for any individual is similar to that seen for the oral glucose tolerance test. In one study, a 25% coefficient of variation (CV) within individuals was seen when 11 healthy subjects had their glycemic response to different carbohydrates tested on eight separate occasions. In another study, the CV of the glycemic response in 22 healthy subjects given 50 g of white bread was 22%. This variability is reduced when the glycemic response is expressed in terms of the 'glycemic index.'

Between-individual variation The variability of the glycemic responses between individuals is larger than that within individual subjects. In a study that included 11 nondiabetic individuals, 10 non-insulin-treated type 2 diabetic subjects, 12 insulin-treated type 2 diabetic subjects, and 14 type 1 diabetic subjects, the CV between individuals within each group was 26, 34, 23, and 34%, respectively. From this it can be seen that comparing the absolute glycemic responses both within and between subjects is unreliable. However, this problem is considerably lessened when the glycemic response to any given food is expressed as a percentage of that individual's glycemic response to a standardized food substance, which in the case of the glycemic index is usually 50 g of glucose. By expressing the glycemic response of a test food against an equal amount of a standard carbohydrate in an individual, variations that occur with age, sex, body mass index, and ethnicity as well as medical conditions such as diabetes are minimized. By using the glycemic index, the between-individual CV of the glycemic response is reduced from approximately 40 to 10%.

Reproducibility of the Glycemic Index

The glycemic index measurement of certain foods can vary between individuals. For example, one study reported that the glycemic index of lentils ranged between 23 and 70 for different subjects. However, this large variability can be minimized to approximately 10% when both the food to be tested and the standard, usually white bread, are each measured in triplicate.

Problems arising from Different Methodologies Used to Calculate the Glycemic Index

Prior to the 1998 WHO nutritional report that standardized the methodology of assessing the glycemic index, different groups used different techniques to calculate the area under the glucose curve and to assess the postprandial glycemic response. The biggest change has been the standardization of the standard used from white bread to glucose. To allow comparison to historic data published glycemic index tables provide conversion factors or present tables using different methods.

Mixed Meals and Other Nutrients

Carbohydrate foods are frequently taken as part of a mixed meal, and the addition of fat and protein to a carbohydrate-containing meal tends to lower the glycemic response. Although the addition of protein or fat to carbohydrate foods reduces the glycemic response, the relative response of one carbohydrate to another remains, such that lentils

will always have a lower response than white bread when part of a mixed meal.

The glycemic index of a mixed meal can be calculated from the different proportions of each of the carbohydrate-containing foods and their individual glycemic index values. For example, when bread and beans are mixed in equal portions, the resulting glycemic response is midway between that of bread alone and that of beans alone. A formula for calculating the glycemic index of mixed meals has been derived by Wolever and Jenkins (**Table 1**). For accuracy, this method requires all individual carbohydrate components of a mixed meal to be pretested. Other methods of calculating the glycemic responses of mixed meals relying on a single measurement of the area under the glycemic curve for the mixed meal or an estimation that does not account for all the carbohydrate-containing foods will be less accurate. To be fair, this remains an area of debate; a recent study suggested that the ability to predict the glycemic index of a mixed meal is poor, particularly those with a high fat content.

The Second Meal Effect

Dietary carbohydrates can influence the glycemic response of a second meal consumed during the postprandial period. The blood glucose response to a lunchtime meal is lower when taken after a low glycemic index breakfast than after a high glycemic index breakfast. Similarly, the glycemic response of a second meal taken during the postprandial period after lunch or dinner is influenced by the glycemic index of the preceding meal.

Wolever attributed the differences in the glycemic response to a second meal during the postprandial period to differences in intermediary metabolism and insulin action associated with rapidly and slowly absorbed carbohydrates. Rapidly absorbed carbohydrates produce large increases in blood insulin levels that result in blood glucose levels decreasing sufficiently quickly to stimulate several counterregulatory hormones that inhibit insulin action and glucose disposal. Both carbohydrate drinks and meals consumed rapidly rather than sipped or eaten slowly are associated with significantly higher serum concentrations of glucagon, catecholamines, growth hormone, and nonesterified fatty acid (NEFA) levels postprandially. The addition of guar to a meal, which slows glucose absorption and lowers the glycemic response, reduces postprandial NEFA and β-hydroxybutyrate levels and improves insulin action. In contrast, nibbling high glycemic index foods between meals increases the glycemic response of a subsequent meal.

Clinical Significance of Postprandial and Fasting Hyperglycemia in Diabetic and Nondiabetic Populations

As with fasting blood glucose levels, postprandial hyperglycemia in nondiabetic populations is a predictor of insulin resistance and cardiovascular disease (CVD). The combined 20-year mortality data on men from the Whitehall, Paris prospective, and Helsinki policemen studies showed that the highest quintile compared with the lowest for the 2-h postplasma glucose load was associated with a 2.7 increased risk of CVD mortality. The fasting glucose values were less predictive for CVD, with only the top 2.5% conferring a 1.8-fold increased mortality risk. During a 7-year period, elderly women with isolated postprandial hyperglycemia and a 2-h value more than 11.1 mmol/l and fasting value less than 7.0 mmol/l on a 75-g oral glucose tolerance test had an approximately threefold increased risk of heart disease compared with women whose 2-h values were less than 11.1 mmol/l.

In established diabetes, postprandial glycemia appears to have a stronger relationship with microvascular and macrovascular disease than fasting blood glucose. Similarly, in gestational diabetes adverse pregnancy outcome is more closely related to postprandial glycemia than fasting and premeal glycemic values.

Benefits of Low Glycemic Index Carbohydrates on Diabetic Control

This is the area in which there is most evidence of clinical efficacy. Two independent systematic reviews of the world evidence demonstrated the efficacy of low glycemic index diets on glycemic control in both type 1 and type 2 diabetes. Clinical studies have shown that after 3 months of a diet containing low glycemic index carbohydrates, glycemic control is improved in both type 1 and type 2 diabetes. With low glycemic diets, postprandial glucose and insulin concentrations decrease in type 2 diabetic subjects, whereas both postprandial glucose values and insulin requirements decrease in type 1 diabetic subjects. Good glycemic control and favorable lipid and fibrinolytic profiles have also been reported in individuals with either type 1 or 2 diabetes who habitually consume low glycemic index dietary carbohydrates. It remains to be shown whether these diets bestow

long-term benefits on micro- or macrovascular complications.

Benefits of Low Glycemic Index Carbohydrates on Cardiovascular Disease Risk Factors

High glycemic index foods induce postprandial hyperinsulinemia, which is a powerful predictor for metabolic risk factors and CVD in epidemiological studies. Both cross-sectional and prospective population studies have shown favorable lipid profiles in association with high carbohydrate diets. Initially, these benefits were attributed to a high fiber content. However, when the glycemic index is controlled for, it is the low glycemic index diets rather than high fiber content that have the greatest influence on high-density lipoprotein (HDL) cholesterol, insulin sensitivity, and fibrinolytic parameters. In a cross-sectional study on more than 2000 middle-aged subjects, the glycemic index was a stronger determinant of HDL cholesterol than any other dietary factor, be it carbohydrate or fat. In this study, the HDL cholesterol of the women whose habitual diet was within the lowest quintile for glycemic index was 0.25 mmol/l higher than that for women whose dietary carbohydrate was within the highest quintile. Extrapolating from the Framingham data that showed a 3% decrease in female and a 2% decrease in male cardiovascular morbidity to be associated with a 0.026 mmol/l increase in HDL cholesterol, one would predict a 29% difference in CHD morbidity between women in the lowest and highest quintile for dietary glycemic index. A similar calculation for men with dietary carbohydrates in the lowest and highest quintile for glycemic index found a 7% decrease in CHD morbidity associated with the 0.09 mmol/l difference in HDL cholesterol concentrations. Low glycemic index diets have also been shown to lower serum cholesterol and triglyceride levels in hyperlipidemic subjects.

Glycemic Index and the Prevention of Type 2 Diabetes

Changes in diet and physical activity levels, both alone and in combination, reduce the progression of impaired glucose tolerance to diabetes. Two large US prospective population studies have demonstrated a doubling of the relative risk of developing type 2 diabetes for both men and women when the habitual diet is characterized by a high glycaemic index and high fat content. A similar protective effect against diabetes has been reported in populations consuming high-fiber foods and high quantities of fruit, and one would predict that these diets would also have a low glycemic index.

Obesity and Glycemic Index

Obesity contributes to the pathogenesis and morbidity of type 2 diabetes. Obesity is associated with changes in carbohydrate and fat metabolism that are central to the development of insulin resistance. Although low glycemic index diets enhance insulin sensitivity and improve metabolic cardiovascular risk factors, they will not reduce weight unless part of an energy-deficient diet. However, in obese subjects, when low glycemic carbohydrates are incorporated into a hypocaloric diet, there is a greater decrease in insulin resistance than can be accounted for by weight loss alone. Evidence from both animal and human studies demonstrates a change in body composition (decrease in fat but no change on overall weight) when exposed to a low glycemic index diet.

Pregnancy and Glycemic Index

Throughout pregnancy in well-nourished urbanized women consuming typical Western diets, glucose tolerance deteriorates. During pregnancy, African women living in traditional rural populations and consuming high-carbohydrate/low glycemic index diets do not invariably experience deterioration in their glucose tolerance. Clinical studies in the West show that women consuming similar high-carbohydrate/low glycemic index diets throughout pregnancy also have no deterioration of glucose tolerance despite the physiological increase in insulin resistance that occurs secondary to maternal and placental hormones. When the proportion of dietary carbohydrate increases above 50% in women with gestational diabetes, if no emphasis on low glycemic index carbohydrates is given, glucose tolerance will deteriorate.

Proposed Mechanism by which Dietary Carbohydrates/Glycemic Index Influence Insulin Resistance

Adipocyte metabolism is central to the pathogenesis of insulin resistance and dietary carbohydrates influence adipocyte function. The previous simplistic view that insulin resistance resulted from the down-regulation of the insulin receptors in response to hyperinsulinemia is being replaced by the hypothesis that high circulating NEFA levels both impair insulin action and reduce pancreatic β cell secretion. It is plausible that low glycemic index carbohydrates

reduce insulin resistance by their ability to reduce adipocyte NEFA release. There is evidence of a loss of suppression of hormone-sensitive lipase (HSL), an enzyme that breaks down triglyceride to free fatty acids and glycerol, to small physiological amounts of insulin and, to a lesser extent, insulin insensitivity of lipoprotein lipase. HSL is normally very sensitive to small increases in insulin levels and is totally suppressed at much lower concentrations than those required for glucose uptake. In insulin-resistant subjects, HSL is less sensitive to small changes in insulin levels and adipocyte NEFA release is increased. A relationship between increased adipocyte NEFA release and insulin resistance has been shown in subjects with coronary heart disease. The metabolic consequences of increased circulating NEFA are multiple and are beyond the scope of this review, but they include adverse lipoprotein and coagulation changes and have been reported to affect insulin secretion and have a lipotoxic effect on the β cell. Accumulation of triglyceride within the β cell also impairs insulin secretion.

Many of the metabolic benefits associated with low glycemic index carbohydrates can be attributed to their ability to reduce adipocyte NEFA release. Low glycemic index foods have been consistently shown to reduce insulin resistance, and animal studies have shown that improvements in fat and muscle insulin sensitivity are accompanied by decreases in fatty acid synthatase activity, adipocyte size, and lipid storage. Although human studies have shown that low glycemic index diets consumed for 3 weeks increase adipocyte insulin sensitivity, no direct effect on adipocyte metabolism has been identified.

Low glycemic index diets attenuate the insulin response for approximately 4 h postprandially. This slightly high postprandial insulin is insufficient to affect glucose transport but does suppress the insulin-sensitive enzyme, HSL, and thus ensures prolonged suppression of postprandial NEFA output. The ability of low glycemic carbohydrates to do this is in stark contrast with high glycemic diets that can cause an elevation of NEFA release postprandially by stimulating the counterregulatory hormones, as discussed previously. Low glycemic meals taken in the evening can effectively suppress circulating NEFA concentrations and hepatic glucose output throughout the night. These metabolic effects are predicted to promote insulin sensitivity.

Our own work has shown that insulin-resistant adults with a history or who are at risk of CHD improve their adipocyte insulin sensitivity after consuming a low glycemic index diet for 3 weeks and their circulating NEFA levels decline. These human studies complement animal work showing that low glycemic index diets improve insulin sensitivity by modulating adipocyte metabolism.

Conclusion

The glycemic index of a diet is an indicator of postprandial metabolism, which is important in contributing to cardiovascular risk. Dietary carbohydrates are absorbed and metabolized differently and therefore influence postprandial glucose, insulin, and NEFA concentrations differently. In Western society, the proportion of the day that we spend in the postprandial state is increasing as the tendency to snack throughout the day replaces sit-down meals. The known detrimental consequences of high glycemic foods and snacks on postprandial metabolism should encourage us to advocate low glycemic diets to counter the current epidemic of insulin resistance-related diseases, notably CVD and diabetes. The relevance of the glycemic index to these two major preventable diseases of the Western world argues strongly for its greater acceptance in current nutritional guidelines.

See also: **Carbohydrates**: Chemistry and Classification; Regulation of Metabolism; Requirements and Dietary Importance; Resistant Starch and Oligosaccharides. **Diabetes Mellitus**: Dietary Management. **Dietary Fiber**: Physiological Effects and Effects on Absorption. **Fructose**. **Galactose**. **Glucose**: Chemistry and Dietary Sources. **Obesity**: Complications. **Pregnancy**: Nutrient Requirements; Safe Diet for Pregnancy. **Sucrose**: Nutritional Role, Absorption and Metabolism. **World Health Organization**.

Further Reading

American Diabetic Association (2004) Nutrition principles and recommendations in diabetes. *Diabetes Care* **27**: S36.

Connor H, Annan F, Bunn E *et al.* Nutrition Subcommittee of the Diabetes Care Advisory Committee of Diabetes UK (2003) The implementation of nutritional advice for people with diabetes. *Diabetic Medicine* **20**(10): 786–807.

FAO/WHO (1998) *Carbohydrates in Human Nutrition. Report of a Joint FAO/WHO Committee, Rome 14–18 April, 1997*, Paper No. 66. Rome: FAO.

Foster-Powell K, Holt SH, and Brand-Miller JC (2002) International table of glycemic index and glycemic load values: 2002. *American Journal of Clinical Nutrition* **76**(1): 5–56.

Ha KK and Lean MEJ (1998) Recommendations for the nutritional management of patients with diabetes mellitus. *European Journal of Clinical Nutrition* **52**: 467–481.

Jenkins DJ, Wolever TM, and Jenkins AL (1988) Starchy foods and glycemic index. *Diabetes Care* **11**: 149–159.

Lean MEJ, Brenchley S, Connor H *et al.* (1992) Dietary recommendations for people with diabetes: An update for the 1990s. Nutrition Subcommittee of the British Diabetic Association's Professional Advisory Committee. *Diabetic Medicine* **9**(2): 189–202.

Goitre *see* **Iodine**: Deficiency Disorders

GOUT

L A Coleman, Marshfield Clinic Research Foundation, Marshfield, WI, USA
R Roubenoff, Millennium Pharmaceuticals, Inc., Cambridge, MA, USA and Tufts University, Boston, MA, USA

A diagnosis of gout refers to a group of metabolic conditions resulting from the deposition of monosodium urate crystals around and in the tissues of joints. The precise mechanism by which uric acid leads to gouty arthritis remains somewhat unclear; various contributing factors are discussed. Clinically, gout typically involves an episodic monoarthritis; if untreated, acute gout can segue into a deforming, chronic polyarthritis that may be difficult to distinguish from rheumatoid arthritis. Improved prevention and treatment of gout have occurred during the latter half of the twentieth century; however, recent research has focused on the link between serum urate, coronary artery disease, and insulin resistance syndrome. Dietary management of gout no longer seems to be focused on restriction of foods with a high purine content but, rather, on the treatment of metabolic disorders commonly associated with gout: obesity, insulin resistance syndrome, and dyslipidemia.

Definition and Etiology

Gout, from the Latin *gutta* or drop (of evil humor), is an ancient disease that was included in Hippocrates' Aphorisms. In the first edition of his textbook, *Principles and Practice of Medicine* (1892), Osler defined gout as "a nutritional disorder associated with an excess formation of uric acid." Today, we recognize that this definition is partly true, but that most cases of gout are not due to excess formation of uric acid but, rather, to insufficient clearance of the substance. Hyperuricemia occurs when there is too much uric acid in the blood, a condition that is generally agreed to exist when the serum or plasma uric acid exceeds the saturation point at 37 °C, which is approximately 7.0 mg dl^{-1}. Hyperuricemia is a requirement for gout, but it is not always present when a patient presents with a first episode of gout, presumably because the acute deposition of uric acid in a joint reduces blood levels transiently. However, hyperuricemia is present at some point in virtually all gout patients. It is important to distinguish hyperuricemia, an asymptomatic condition, from gout, a painful disease that afflicts only a minority of people with elevated uric acid levels. Hyperuricemia can result from overproduction of uric acid in 10–15% of cases (generally because of enzyme deficiency or overactivity) or from underexcretion of uric acid, which accounts for 85–90% of cases of gout (due to decreased renal clearance of uric acid, even in the setting of a normal glomerular filtration rate).

Chemical Pathology

Uric acid is a by-product of purine metabolism in humans and certain apes who lack uricase, the enzyme that breaks down uric acid (**Figure 1**).

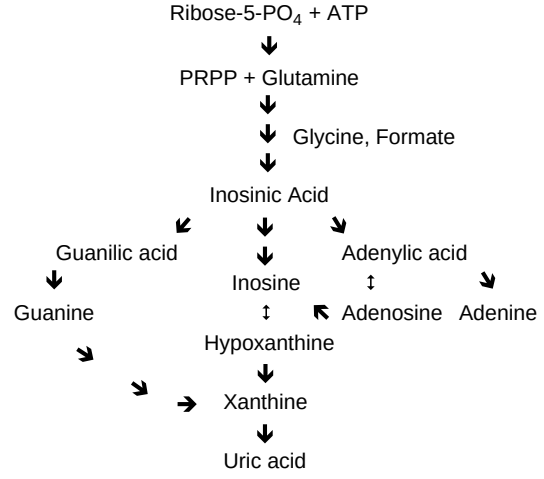


Figure 1 Simplified pathway of uric acid metabolism. PRPP, 5-phosphoribosyl-*l*-pyrophosphate. (Modified from Seegmiller JE, Rosenbloom FM and Kelly WN (1967) Enzyme defect associated with a sex-linked human neurological disorder and excessive purine synthesis. *Science* **155:** 1682–1684.)

When uric acid production is normal, and its clearance by the kidneys is normal, this metabolic quirk has no ill effects. However, this minor metabolic inconvenience becomes of pathological importance because uric acid is so poorly soluble in aqueous solutions that it can crystallize and cause the various conditions recognized as gout. Uric acid can be ingested directly in the diet (especially in organ meats such as liver, kidney, and sweetbreads), or it can be produced in the body by two pathways involved in purine metabolism (**Figure 1**). The de novo synthesis of uric acid proceeds directly from ribose-5-phosphate, whereas the salvage pathway consists of production of the uric acid precursors inosine from adenosine and xanthine from guanine. The medication allopurinol, which blocks the conversion of xanthine to uric acid by xanthine oxidase, is effective because xanthine is far more soluble in aqueous solutions than is uric acid.

The precise mechanism by which uric acid leads to gouty arthritis remains somewhat unclear. However, uric acid is known to be proinflammatory in that it can initiate an immune response with recruitment of white blood cells after uric acid crystals are phagocytosed by polymorphonuclear leukocytes or macrophages. These white blood cells also release tumor necrosis factor and interleukin-1, recruiting more white cells, which release lysosomal enzymes that lead to cartilage destruction and joint erosions with repeated attacks. In addition, ingestion of uric acid leads to death of the phagocytosing cells, leading to release of the uric acid and additional proteolytic enzymes, thus reinforcing the inflammatory condition. However, the crystals become progressively less phlogistic after several cycles of ingestion and release, and the inflammation relents over a period of 10–14 days. The natural history of untreated gout progresses through four stages from (i) asymptomatic hyperuricemia to (ii) acute gouty arthritis, (iii) intercritical gout, and (iv) chronic tophaceous gout. In addition, renal manifestations of gout develop in up to 50% of patients, depending on the amount of uric acid they excrete.

Prevalence and Risk Factors

Gout is the most common inflammatory arthritis in men; more than 2 million men and women in the United States are afflicted. The prevalence of gout in the United States tripled between 1969 and 1981 but recently seems to have stabilized. This increase is thought to be due to a combination of factors, including aging of the US population, increased prevalence of diuretic treatment of hypertension, better access to health care, and better diagnosis and reporting of gout. The incidence of gout (i.e., the development of new cases) is linked to serum uric acid levels, increasing from 0.9 cases per 1000 person-years for uric acid levels less than $7.0\,\mathrm{mg\,dl^{-1}}$ to 70 cases per 1000 person-years for levels higher than $10.0\,\mathrm{mg\,dl^{-1}}$. However, even in the highest category, only 30% of men developed gout during the 5 years after their uric acid level was determined, confirming that only a minority of hyperuricemic men develop acute gout.

Risk factors for acute gout other than hyperuricemia have been identified. All risk factors act either by increasing serum uric acid levels or by reducing the solubility of uric acid in the joints. For example, male sex, alcohol ingestion, obesity, and weight gain are associated with increased uric acid production, whereas diuretics (thiazides and loop diuretics), low-dose salicylates, and renal insufficiency lead to reduced clearance of uric acid. Hypertension has been associated with increased risk of gout, but this effect probably operates through renal insufficiency, which occurs as a result of hypertension and diuretic therapy. Lead, on the other hand, has been shown to directly reduce the solubility of uric acid in synovial fluid, whereas lead nephropathy also leads to reduced clearance of uric acid; the gout associated with lead toxicity is known as saturnine gout. Joint trauma and cooling of distal joints also reduce solubility of uric acid and increase the risk of an acute attack. Gout was known in the eighteenth century as 'pheasant hunter's toe' when aristocratic gentlemen developed *podagra* (acute inflammation of the first metatarsophalangeal (MTP) joint) after a day of hunting in the cold marshes and a night of drinking alcohol, especially sherry shipped in lead-lined casks. In more recent times, saturnine gout has been associated with drinking illegal 'moonshine' whiskey distilled through lead-lined stills. An independent association has been shown between kidney stone disease and gout, strongly suggesting that they share a common underlying pathophysiological mechanism.

Insulin resistance has been increasingly implicated in the pathogenesis of gout. The lipoprotein abnormalities described in subjects with hyperuricemia are similar to those found in individuals with insulin resistance, and insulin has an impact on renal urate excretion. Although the precise frequency of insulin resistance syndrome in patients with gout is not known, it is estimated to be as high as 76% for insulin resistance syndrome and 95% for hyperinsulinemia. It has been suggested that elevated serum urate may even serve as a surrogate marker for insulin resistance syndrome.

Clinical Features

Gout is typically an episodic monoarthritis, although polyarticular gout (involving three or more joints) occurs in approximately 10% of cases. The description of the pain of acute gout by Thomas Sydenham in the seventeenth century remains among the best:

> The victim goes to bed and sleeps in good health. About two o'clock in the morning he is awakened by a severe pain in the great toe; more rarely in the heel, ankle, or instep. This pain is like that of a dislocation ... then follow chills and a little fever. The pain ... becomes more intense. ... Now it is a violent stretching and tearing of the ligaments—now it is a gnawing pain and now a pressure and tightening. So exquisite and lively meanwhile is the feeling of the part affected, that it cannot bear the weight of the bed-clothes nor the jar of a person walking in the room. The night passes in torture.

More than half of patients present with podagra, and 75–90% of patients eventually develop podagra. This joint is thought to be most susceptible to gout because it is very prone to trauma and cooling, both of which reduce the solubility of uric acid. After the first MTP, acute gout most commonly involves the ankles, knees, instep, but it can also involve the wrists, elbows, and small joints of the hands and feet. Large axial joints, such as hips, shoulders, and vertebral joints, are rarely affected. Acute gout often involves a component of tenosynovitis (inflammation of tendon sheaths), and gouty cellulitis (sterile inflammation with urate crystals in the skin) and bursitis have also been described. As Sydenham stated, the onset is generally explosive, but many patients also describe a series of minor attacks leading up to the full-blown episode. Untreated gouty arthritis lasts from days to weeks, but minor bouts may resolve spontaneously in a few hours. At this stage, joint radiographs are normal except for soft tissue swelling. If untreated, acute gout can segue into a deforming, chronic polyarthritis that may be difficult to distinguish from rheumatoid arthritis.

Because a significant proportion of people, perhaps as many as one-third, who have a single acute gouty attack do not have another for 1 year or longer, no further therapy is indicated after the first attack has subsided. Once a patient has demonstrated recurrent attacks of acute gout, or if he or she has had a uric acid stone (or another type of stone in the setting of hyperuricosuria), treatment aimed at reducing serum uric acid below the point of solubility is indicated. In general, patients who develop a second attack and have a serum creatinine concentration less than 2.0 mg dl^{-1} should be evaluated further with a 24-h urine collection for creatinine clearance and uric acid output while consuming their regular diet. If the 24-h urinary uric acid totals more than 1000 mg, the patient is classified as an overproducer of uric acid and should be treated with allopurinol if he or she is not allergic. If the 24-h uric acid production is under 700 mg, then the patient is an underexcreter and may first be treated with a uricosuric agent, which is safer and less expensive than allopurinol. Renal insufficiency will reduce both creatinine clearance and urinary uric acid output, and allopurinol is the drug of choice in this situation, so the utility of a 24-h urine collection is reduced. Patients who produce between 700 and 1000 mg of uric acid are in a gray zone, and clinical judgment regarding optimal therapy is necessary, balancing issues of safety, cost, and convenience in the management of a chronic disease.

Chronic tophaceous gout occurs with an average of 10 years of untreated or inadequately treated gout. Over time, the acute attacks become less noticeable, and the patient develops a chronic, often deforming arthritis. This arthritis may mimic rheumatoid arthritis, although it should be less symmetric. At this time, the radiological hallmarks of gout, which include large, well-demarcated erosions in the absence of joint space narrowing ('rat-bite erosions'), are often visible. Tophi, which are subcutaneous deposits of uric acid, may be found in and around joints, bursae (especially the olecranon), tendons (Achilles and infrapatellar), and the extensor surfaces of the forearms. Less commonly, they may arise in the pinna of the ear, cardiac valves, cornea and sclera, and nasal cartilage. Needle aspiration or spontaneous rupture of tophi elicits a white, chalky material that is full of urate crystals under microscopy and is diagnostic of tophaceous gout. The presence of tophi is always an indication for allopurinol in nonallergic patients.

Dietary Management

There has been a substantial change in the predominant view regarding the relationship between diet and gout. It has even been said that "dietary considerations now play a minor role in the treatment of hyperuricemia, despite a fascinating history and abundant literature."

The relationship between gout and gluttony (overindulgence of food and alcohol) dates back to ancient times. In the fifth- century BC, Hippocrates attributed gout to dietary excesses of food and wine; he advised dietary restriction and reduction of alcohol consumption. Historically, the dietary management of gout has focused on two goals: (i) reducing the amount

of uric acid that may be deposited as crystals in joints or soft tissues, leading to the clinical syndrome of gout, and (ii) managing the disorders that occur with increased frequency among patients with gout, including diabetes mellitus, obesity, hyperlipidemia, hypertension, and atherosclerosis.

Although some practitioners may still advocate the traditional low-purine, low-protein, alcohol-restricted diet, there is increasing support for the more 'contemporary' view that dietary management should focus on weight reduction with a restricted intake of calories and carbohydrates along with proportional increases in both protein and unsaturated fats and no restriction of purine content.

Traditional Low-Purine Diet

The primary dietary modification that has traditionally been recommended to reduce uric acid production is a low-purine diet (<75 mg/24 h; **Table 1**). Uric acid is the end product of purine metabolism in humans, formed by oxidation of its precursors, the oxypurines, hypoxanthine and xanthine. With the advent of more powerful and effective urate-lowering drugs, however, dietary restriction of purine-rich foods is of decreasing importance. Although patients may be advised to avoid large quantities of food and alcoholic beverages that they know may precipitate a gouty attack (i.e., large amounts of organ meats or beer), a rigid purine-restricted diet is no longer viewed as a mainstay of dietary management.

Many patients with gout are overweight, and a combination of caloric reduction and exercise can have a beneficial impact on any associated hypertension, hyperlipidemia, and insulin resistance syndrome via enhanced renal excretion of urate and reduced serum urate levels. However, although weight reduction, purine restriction, and reduced alcohol consumption may transiently reduce serum urate, there are no long-term studies demonstrating the efficacy of such an approach. Any benefit that does occur is likely to be small, and any limited reduction in serum urate levels is likely to be offset by the difficulty of maintaining such an improvement over the long term.

Table 1 Foods to avoid on a purine-restricted diet

Meats, organ meats (sweetbreads, liver, kidney), fish, eggs, sausages, meat extracts and gravies
Beans, peas, spinach, asparagus, cauliflower, mushrooms
Oatmeal
Legumes
Chocolate
Yeast and yeast extracts
Tea, coffee, cola beverages, alcoholic beverages

Touger-Decker R (1996) Nutritional Care in Rheumatic Diseases. In: Mahan LK and Escott-Stumps S (eds.) *Krause's Food, Nutrition, & Diet Therapy*, 9th edn. pp. 889–898. Philadelphia, PA: W.B. Saunders Company.

Contemporary Low-Calorie, Carbohydrate-Restricted Diet

In view of the well-recognized link between insulin resistance syndrome, hyperuricemia, and gout, a diet emphasizing reduced calorie intake with moderate restriction of carbohydrates and liberalization of protein and unsaturated fat consumption has been espoused for patients with gout. Low-purine foods are often high in both carbohydrate and saturated fats; these foods tend to further decrease insulin sensitivity, thereby contributing to even higher levels of insulin, glucose, triglycerides, and low-density lipoprotein cholesterol and lower high-density lipoprotein cholesterol levels, all of which result in increased risk of coronary heart disease among these patients. Conversely, a calorie-restricted, weight-reduction diet that is low in carbohydrates (40% of total calories) and relatively high in protein (approximately 120 g per day compared to 80–90 g in the typical Western diet) and unsaturated fat content, with no limitation of purine content, has been studied and found to result in weight loss and reductions in serum urate, lipids, and gouty attacks. These benefits seem to relate to the coexistence of hyperlipidemia and glucose intolerance in patients with gout.

In summary, both the traditional and the contemporary approaches to dietary modification need to be studied over the long term, but it appears that the most benefit can be gained by focusing efforts on reducing calorie intake and carbohydrate consumption rather than on limiting the purine content of the diet. It appears that restriction of alcoholic beverages is advisable in the management of gout.

Prognosis

Gout is unusual among the rheumatic diseases in that its etiology, treatment, and prevention are well understood. Thus, the long-term sequelae of gout should be completely avoidable with adequate treatment, making the overall prognosis excellent. Noncompliance with medication, lack of access to adequate medical care, and inability to tolerate one or more of the medications used to treat gout can lead to a worse outcome. A number of dietary and lifestyle factors may contribute to the increased uric acid production among patients with gout. If these factors can be identified and appropriate changes made, the serum uric acid concentration may decline substantially. However, many patients require medication to control the hyperuricemia. The

predominant dietary approach to gout is that dietary advice, other than the restriction of overly excessive alcohol intake, is likely to be limited to weight reduction.

See also: **Alcohol**: Disease Risk and Beneficial Effects. **Arthritis**. **Diabetes Mellitus**: Etiology and Epidemiology. **Hypertension**: Dietary Factors.

Further Reading

Brand FN, McGee DL, Kannel WB, Stokes J, and Castelli WP (1985) Hyperuricemia as a risk factor of coronary heart disease: The Framingham Study. *American Journal of Epidemiology* **121**: 11–18.

Dessein PH, Shipton EA, Stanwix AE, Joffe BI, and Ramokgadi J (2000) Beneficial effects of weight loss associated with moderate calorie/carbohydrate restriction, and increased proportional intake of protein and unsaturated fat on serum urate and lipoprotein levels in gout: A pilot study. *Annals of Rheumatic Disease* **59**: 539–543.

Emmerson BT (1996) The management of gout. *New England Journal of Medicine* **334**: 445–451.

Fam AG (2002) Gout, diet, and the insulin resistance syndrome. *Journal of Rheumatology* **29**: 1350–1355.

Lawrence RC, Helmick CG, Arnett FC *et al.* (1998) Estimates of the prevalence of arthritis and selected musculoskeletal disorders in the United States. *Arthritis & Rheumatism* **41**: 778–799.

Roubenoff R (1990) The epidemiology of gout and hyperuricemia. *Rheumatic Disease Clinics of North America* **16**: 539–550.

Roubenoff R (1996) Gout and other crystal diseases. In: Stobe JD, Ledenson PW, Traill TA, Petty BG, and Helliman DB (eds.) *Principles and Practice of Medicine*, 23rd edn., pp. 233–239. Hartford, CT: Appleton.

Roubenoff R, Klag MJ, Mead LA *et al.* (1991) Incidence and risk factors for gout in white men. *Journal of the American Medical Association* **266**: 3004–3007.

Snaith ML (2001) Gout: Diet and uric acid revisited. *Lancet* **358**: 525.

Sydenham T (1850) *The Works of Thomas Sydenham* (translated from Latin by RG Lathan), vol. 2, pp. 124. London: New Sydenham Society.

Terkeltaub RA (2001) Gout: Epidemiology, pathology, and pathogenesis. In: Klippel JH (ed.) *Primer on the Rheumatic Diseases*, 12th edn., pp. 307–312. Atlanta: Arthritis Foundation.

Touger-Decker R (1996) Nutritional Care in Rheumatic Diseases. In: Mahan LK and Escott-Stump S (eds.) *Krause's Food, Nutrition, & Diet Therapy*, 9th edn. pp. 889–898. Philadelphia, PA: W.B. Saunders Company.

Wortmann RL (2002) Gout and hyperuricemia. *Current Opinions in Rheumatology* **14**: 281–286.

Grains *see* **Cereal Grains**

GROWTH AND DEVELOPMENT, PHYSIOLOGICAL ASPECTS

W W Hay Jr, University of Colorado Health Sciences Center, Aurora, CO, USA

Introduction

Growth and development refers to the growth of the individual in size as determined by anthropometric measurements of body weight, length, circumference, and weight/length ratio, as well as changes in body composition. This article will focus on growth and development of the fetus, as most of the relevant concepts about growth and development apply to the fetal period of development and this period encompasses the greatest changes in body proportion and composition during the life of the individual. Fetal growth occurs by increases in cell number and size. In the first third of gestation, during the embryonic period, growth occurs primarily by increased cell number (hyperplasia); in the middle third of gestation, cell size also increases (hypertrophy), while the rate of cell division becomes stable. In the last third of gestation, the rate of cell division declines, while cell size continues to increase.

Many terms are used to describe variations in growth. For example, human newborns are classified as having normal birth weight (greater than 2500 g), low birth weight (less than 2500 g), very low birth weight (less than 1500 g), or extremely low birth weight (less than 1000 g). Obviously,

classification by weight alone says little about growth rate, as most infants with less than normal birth weights are the result of a shorter than normal gestation, i.e., they are preterm. Similarly, classifying newborns according to duration of gestation (e.g., preterm, term, or post-term) on the basis of birth weight also is erroneous, because infants with intrauterine growth restriction (IUGR) are smaller and macrosomic infants of diabetic mothers are larger than normal at any gestational age. Furthermore, it is inappropriate to label newborns as abnormally grown when their birth weight is less than some arbitrarily determined 'normal' birth weight, but their mother was quite small to begin with; such newborns are considered constitutionally small but not abnormal.

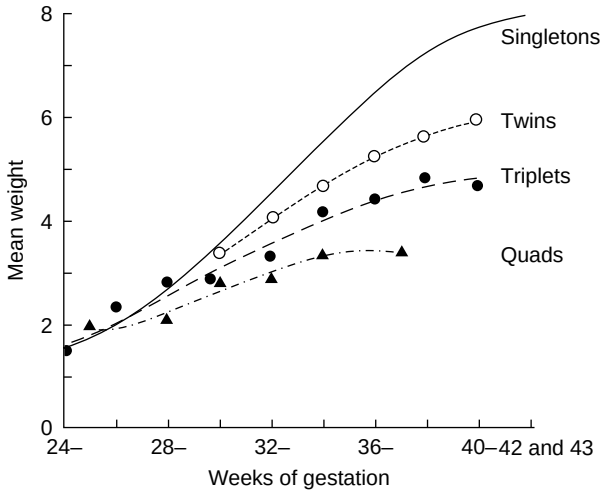


Figure 1 Mean birth weight of single and multiple fetuses related to duration of gestation. (Reproduced with permission from Ounsted M and Ounsted C (1973) *On Fetal Growth Rate*. Spastics International Medical Publications (Clinics in Developmental Medicine No. 46), p. 17. London: William Heinemann Medical Books Ltd.)

Growth of Fetal Size

What should be considered more important for growth assessment than birth weight (at any gestational age) is the genetic growth potential of the infant, which may or may not be limited by maternal size. Under usual conditions, the fetus grows at its genetic potential. Small fetuses of small parents or large fetuses of large parents do not reflect fetal growth restriction or fetal overgrowth, respectively; in fact, their rates of growth are normal for their genome. If the mother is unusually small, however, she might limit fetal growth by 'maternal constraint,' which represents a limited uterine size (primarily endometrial surface area) and thus the capacity to support placental growth and nutrient supply to the fetus. A clear example of maternal constraint is shown in **Figure 1**, showing the reduced rate of fetal growth of multiple fetuses in a species, i.e., human, that optimally supports only one fetus.

Fetal weight tends to increase exponentially in the middle part of gestation, producing the typical S-shaped curve of fetal weight versus gestational age that is derived from cross-sectional measurements of newborn weights at different known gestational ages (**Figure 2**). The length of gestation is more strongly related to the growth of neural tissue (range 0.015–0.033 $g^{1/3}$/day – a 2.2-fold range) than to the growth of the fetal body (range 0.033 to 0.25 $g^{1/3}$/day – a 7.6-fold range). The physiological significance of this relationship is not known, but intrauterine development of a large brain/body mass ratio in humans is favored in a single fetus and is made possible by a slow rate of somatic growth.

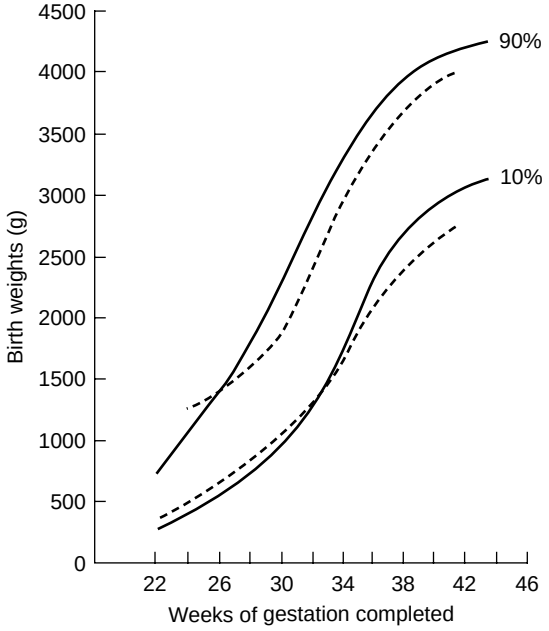


Figure 2 Birth-weight percentiles for gestational age. Solid lines represent California total singleton live births, 1970–1976: dotted lines represent Colorado General Hospital (Denver, Colorado) live births, 1948–1960. (Reproduced with permission from Creasy R and Resnik R (1989) Intrauterine growth retardation. In: Creasy R and Resnik R (eds.) *Maternal-Fetal Medicine*, 2nd edn, pp. 549–564. Philadelphia: W.B. Saunders.)

Developmental Change of Fetal Body Composition

Fetal growth during the last third of gestation requires large increases in nutrient supplies and

appropriate utilization of these nutrients. Nutrient substrate supply is coupled with increased development of anabolic hormones and growth factors in fetal tissues and fetal plasma to produce increased nitrogen and carbon deposition in protein, carbohydrate deposition in glycogen, and fatty acid, glycerol and triglyceride deposition in adipose tissue. Growth of these tissues gradually replaces water in the fetal extracellular space.

Chemical composition studies of normal human infants are limited. Based on data from 15 studies that included 207 infants, nonfat dry weight and nitrogen content (predictors of protein content) show a linear relationship with fetal weight and an exponential relationship with gestational age (**Figure 3**). As gestation proceeds, larger fetuses grow faster than smaller fetuses, and protein accretion follows accordingly.

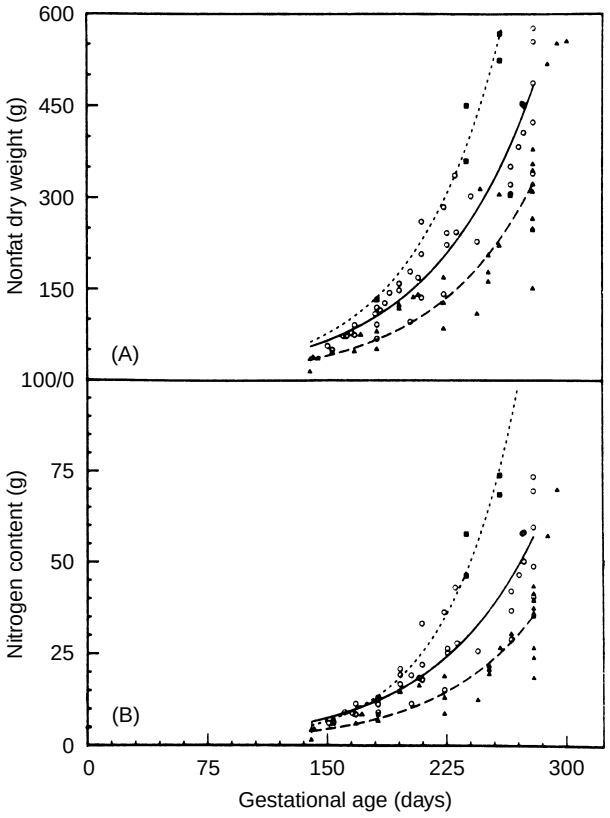


Figure 3 Nonfat dry weight (A) and nitrogen content (B) are plotted against gestational age for LGA (■, ····), AGA (O, —), and SGA (◆, – – –) infants. (Reproduced with permission from Sparks JW (1992) Intrauterine growth and nutrition. In: Polin RA and Fox WW (eds.) *Fetal and Neonatal Physiology*, p. 184. Philadelphia: W.B. Saunders.)

Water

Fetal water content increases directly with body weight, but not proportionally to body weight, as fetal body water, expressed as a fraction of body weight, decreases with advancing gestation. The relatively large growth of adipose tissue in the human fetus further dilutes the body concentration of water. Extracellular water, as a fraction of fetal body weight, also decreases more than intracellular water as gestation advances; this is mainly due to increasing cell number and increasing cell size, rather than the intracellular concentration of osmotic substances.

Nonfat dry weight

Comparative aspects of fetal chemical and physical growth in six species are summarized in **Table 1**. Despite growth rate variations up to 20-fold and weight-specific fat content variances at term up to 16-fold among these species, nonfat dry weight and protein weight-specific contents (as percentages of total weight at term) are constant. Protein concentration is about 12% in all species at term and fetal protein content is linearly related to fetal weight; thus, protein accretion in the fetal rat occurs about 23 times as fast as it does in the human. These species-related differences in growth rate are remarkable and require marked differences in the placental capacity to supply nutrients to the fetus.

Table 1 Growth characteristics and chemical composition at term of selected mammals and a representative human fetus

	Human	*Monkey*	*Sheep*	*Pig*	*Rabbit*	*Rat*
Gestation (days)	280	163	47	67	30	21.5
Number of fetuses	1	1	1	3–5	4–6	10–12
Growth rate ($g\,kg^{-1}\,day^{-1}$)	15	44	60	70	300	350
Fetal weight (g)	3500	500	4000	100	60	5
Dry weight (g/% body wt)	1050/30	125/25	760/19	25/25	9/15	0.2/4
Nonfat dry weight (g/% body wt)	490/14	–	640/16	14/14	–	–
Protein (g/% body wt)	420/12	–	480/12	12/12	7.2/12	0.6/12

From McCance RA and Widdowson EM (1985) In: Falkner F and Tanner JM (eds.) *Human Growth*, 2nd edn, vol. 1 p. 139. New York: Plenum Press.

Nitrogen Balance, Protein Turnover, and Protein Synthesis

According to animal data, only about 80% of the nitrogen content of the fetus is found in protein; the rest is found in urea, ammonia, and free amino acids. Additional nitrogen requirements for urea excretion and for other possible nitrogen excretion products are not known for human fetuses.

Radioactive and stable isotopic tracers of selected amino acids, especially essential amino acids such as leucine and lysine, have been used to measure fetal protein synthesis, breakdown, and accretion. Limited human data is consistent with data in the fetal sheep, the only species studied in significant detail. **Figure 4** shows results of experiments in fetal sheep over the second half of gestation, comparing fractional protein synthesis rates derived from tracer data and fractional body growth rates derived from body composition data. Whole body weight-specific protein turnover rate is higher in the early-gestation fetus primarily from increased rates of amino acid uptake from the placenta (exogenous entry of amino acids into the fetal circulation) and protein synthesis. These processes produce a 50% higher rate of net protein accretion in the mid-gestation fetus.

Mechanisms underlying the decrease in protein synthesis rate over gestation are not well understood, but they appear to be intrinsic to the fetus and not to a limitation of nutrient supply by the placenta. At least a partial explanation can be offered according to the changing proportion of body mass contributed by the major organs (**Table 2**). Based on the increased mass of skeletal muscle with advancing gestation, fetal whole body fractional synthesis rate should be lower, as skeletal muscle has a relatively lower fractional protein synthetic rate in late gestation than in earlier gestation. A direct relationship between anabolic growth-promoting substances acting as principal regulators of fetal protein synthesis rate, and thus fetal growth rate, cannot be made, however, as plasma concentrations or secretion rates of these substances increase in the fetus as gestation proceeds, while protein synthetic rates decline.

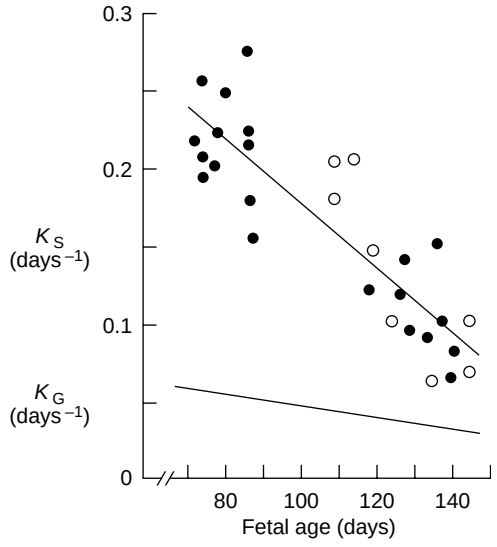


Figure 4 Fractional rate of protein synthesis (K_s) over gestation in fetal sheep studied with leucine (•) and lysine (O) radioactive tracers compared with the fractional rate of growth (K_G) (lower portion of the figure, —). (Reproduced with permission from Hay WW Jr (1992) Fetal requirements and placental transfer of nitrogenous compounds. In: Polin RA and Fox WW (eds.) *Fetal and Neonatal Physiology*, p. 439. Philadelphia: W.B. Saunders.)

Table 2 Fetal organ weight as per cent of body weight

	50% Gestation	*67% Gestation*	*90% Gestation*
Liver	6.5	5.1	3.1
Kidneys	1.6	1.2	0.7
Heart	0.9	0.8	0.8
Brain	3.4	2.9	1.7
Hindquarters	14.5	15.1	22.0

Reproduced with permission from Bell AW *et al.* (1987) Relation between metabolic rate and body size in the ovine fetus. *Journal of Nutrition* **117**: 1181–1186. Used with permission.

Glycogen

Many tissues in the fetus, including brain, liver, lung, heart, and skeletal muscle, produce glycogen over the second half of gestation. Liver glycogen content, which increases over the gestation period, is the most important store of carbohydrate for systemic glucose needs, because only the liver contains sufficient glucose-6-phosphatase for release of glucose into the circulation. Skeletal muscle glycogen content increases during late gestation and forms a ready source of glucose for glycolysis within the myocytes. Lung glycogen content decreases in late gestation with change in cell type, leading to loss of glycogen-containing alveolar epithelium, development of type II pneumocytes, and onset of surfactant production. Cardiac glycogen concentration decreases with gestation owing to cellular hypertrophy, but cardiac glycogen appears essential for postnatal cardiac energy metabolism and function. At term, fetal liver glycogen concentration in most species (80–120 $mg\,g^{-1}$) is at least twice the adult concentration, but in the relatively slow-growing human fetus, glycogen synthesis rates are low (about 2 $mg\,day^{-1}\,g^{-1}$), representing less than 2% of estimated whole body glucose utilization rate.

The principal source of fetal glycogen is glucose derived from placental transport of glucose from the mother. Smaller fractions come from lactate and amino acids such as glutamine. Glycogen content of the fetus and selected fetal organs is directly related to the maternal and thus fetal plasma glucose concentrations. Thus, macrosomic fetuses of diabetic mothers have very high body and organ contents of glycogen, while IUGR fetuses that result from sustained maternal hypoglycemia or placental insufficiency and decreased placental glucose supply to the fetus have markedly decreased glycogen contents.

Cortisol and glucose provide developmental regulation, while adrenaline (epinephrine) and glucagon provide acute and more variable regulation. Experimentally, cortisol infusion decreases glycogen content of the liver while deficiencies in hypothalamic-pituitary regulation of the adrenal gland leads to cortisol deficiency and glycogen deficiency. Insulin acts synergistically with glucose to increase hepatic glycogen stores. Glucose also acts independently to activate glycogen phosphorylase and glycogenolysis to keep hepatic glycogen content constant at higher glucose concentrations.

Fat

Fetal fat content as a fraction of fetal weight varies several fold among species (**Figure 5**). The fat content of newborns at term of almost all land mammals is 1–3% and is considerably less than that of the human (15–20%). Differences in body fat content among species are due primarily to the capacity of the placenta to transfer fat to the fetus and to the capacity of the fetus to synthesize triglycerides and fat. Even in those species that take up fat from the placenta and deposit fat in fetal tissues, the rate of fetal fatty acid oxidation is presumed low, because plasma concentrations of fatty acids (and keto acid products such as β-hydroxybutyrate and acetoacetate) are low, and because the carnitine palmitoyl transferase enzyme system is not sufficiently developed to deliver long-chain fatty acids to the respiration pathway inside the mitochondria.

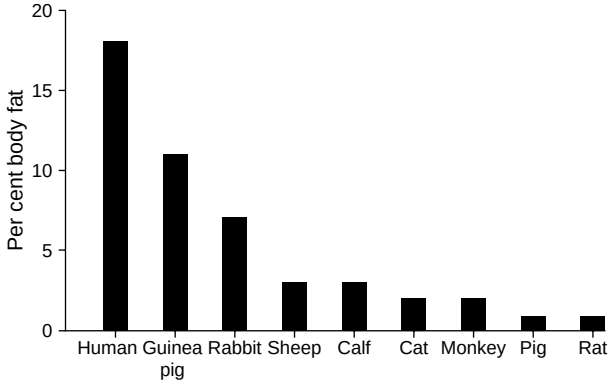


Figure 5 Fetal fat content at term as a per cent of fetal body weight among species. (Reproduced with permission from Hay WW Jr (1996) Nutrition and development of the fetus: carbohydrate and lipid metabolism. In: Walker WA and Watkins JB (eds.) *Nutrition in Pediatrics*, 2nd edn, p. 376. Hamilton: B. C. Decker.)

In the human fetus, calories produced by the complete oxidation of glucose and lactate can fully meet energy required for maintenance metabolism and for conversion of glucose and lactate to fatty acids. The portion of glucose converted into fat has been estimated to be $23\,\mathrm{kcal\,kg^{-1}\,day^{-1}}$. This would permit accumulation of $2.4\,\mathrm{g\,kg^{-1}\,day^{-1}}$ of fat. In the human fetus between 26 and 30 weeks' gestation, nonfat and fat components contribute equally to the carbon content of the fetal body. After that period, fat accumulation considerably exceeds that of the nonfat components. At 36 weeks' gestation, 1.9 g of fat accumulates for each gram of nonfat daily weight gain, and by term, the deposition of fat accounts for over 90% of the carbon accumulated by the fetus. The rate of fat accretion is approximately linear between 36 and 40 weeks' gestation, and by the end of gestation, fat accretion ranges from 1.6 to $3.4\,\mathrm{g\,kg^{-1}\,day^{-1}}$. At 28 weeks' gestation, it is slightly less and ranges between 1.0 and $1.8\,\mathrm{g\,kg^{-1}\,day^{-1}}$. By term, fat content of the human fetus is 15–20% of body weight, ranging from less than 10% in IUGR fetuses to 25% or more in macrosomic infants of diabetic mothers.

Energy Accretion in the Fetus

Fat has a high energy content ($9.5\,\mathrm{kcal\,g^{-1}}$) and a very high carbon content (approximately 78%). Thus, differences in fetal fat concentration among species lead to large differences in calculated energy accretion rates and carbon requirements of the fetal tissues for growth. The energy concentration of nonfat dry weight is fairly consistent across species and also within species at different developmental stages, indicating that the ratio of protein to nonprotein substrates in the tissues is relatively constant. Thus, energy accretion rate of any fetus can be estimated from the growth curve of the fetus in question and the changing fat and water concentrations.

Data for energy accretion and distribution in the human fetus are shown in **Table 3**. Because growth of fat and nonfat (protein plus other) tissues are metabolically linked through energy supply that is used for protein synthesis and the production of

Table 3 Calculation of the energy distribution in the term human infant

	Wet weight	*Fat*	*Nonfat wet weight*	*Nonfat dry weight*
Weight (g)	3450	386	3064	511
Total calories (kcal)	5950	3650	2300	2300
Energy concentration (kcal g^{-1})	1.73	9.45	0.75	4.5

From Ziegler EE *et al.* (1979) Body composition of the reference fetus. *Growth* **40**: 329–341.

anabolic hormones that promote positive protein, fat, and carbohydrate growth, restriction of nutrient supply is likely to produce growth deficits of all tissues, not just fat (i.e., growth retardation involves limitation of muscle growth as well as fat and glycogen). Indeed, chronic experimental selective energy (glucose) restriction in the fetal sheep leads to increased protein breakdown as well as to lower rates of fetal growth and lipid content. In contrast, as shown by the growth curves in **Figure 6** from human infants born prematurely at different times

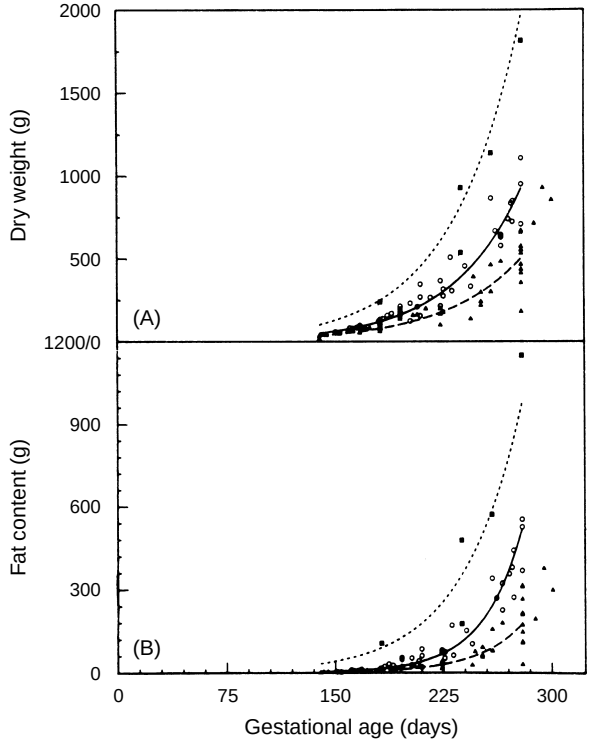


Figure 6 Dry weight (A) and fat content (B) plotted against gestational age in the same newborn human infants shown in **Figure 3** for LGA (■, ····), AGA (O, —), and SGA (◆, – – –) infants. (Reproduced with permission from Sparks JW (1992) Intrauterine growth and nutrition. In: Polin RA and Fox WW (eds.) *Fetal and Neonatal Physiology*, p. 184. Philadelphia: W. B. Saunders.)

over the last third of gestation, there is a bias towards thinner, SGA infants with less fat relative to nonfat weight and nitrogen content, raising the possibility that in a species that does lay down considerable fetal fat during late gestation, differences in intrauterine growth rate may reflect fat deposition more than the growth of nonfat, protein-containing tissues.

Mineral Accretion in the Fetus

Fetal calcium content is best correlated with fetal body length; this is true for both AGA and SGA infants. Using this index, fetal calcium content increases exponentially with a linear increase in length. Using this estimate, the human fetal rate of calcium accretion is about 85 mg kg^{-1} day^{-1}. Accretion of other minerals varies more directly with body weight, and according to the distribution of the minerals into extracellular (e.g., sodium) or intracellular (e.g., potassium) spaces.

Regulation of Fetal Growth

Fetal growth is the result of interaction among maternal, placental, and fetal factors, representing a mix of genetic mechanisms and environmental influences through which the genetic factors are expressed and modulated. The single most important environmental influence that affects fetal growth is the nutrition of the fetus. Nutrient supply to the fetus and the resulting increases in fetal tissue and plasma concentrations of anabolic hormones and growth factors are regulated by maternal health, maternal nutrition, uterine growth (including uterine blood flow and endometrial surface area), and placental growth and function.

Genetic Factors

Many genes contribute to fetal growth and birth weight of the normal term fetus. Maternal genotype is more important than fetal genotype in the overall regulation of fetal growth. **Table 4** presents estimates of the quantitative contribution of fetal and parental factors to fetal growth and birth weight at term. The more modest regulation by the paternal genotype, acting through the fetal genotype, is essential for trophoblast development. In fact, overexpression of the paternal genotype can produce trophoblast tumors. More specific gene targeting studies have shown the importance of genomic imprinting on fetal growth. For example, in mice normal fetal and placental growth require that the IGF_2 gene be paternal and the IGF_2 receptor gene be maternal, and paternal disomy producing IGF_2 gene overexpression results in fetal overgrowth while

Table 4 Factors determining variance in birth weight

	Per cent of total variance
Fetal	
Genotype	16
Sex	2
	18
Maternal	
Genotype	20
Maternal environment	24
Maternal age	1
Parity	7
	52
Unknown	30

From Penrose LS (1954) *Proceedings of the 9th International Congress on Genetics*, Part 1; and Milner RDG and Gluckman PD (1996) Regulation of intrauterine growth. In: Gluckman PD and Heymann MA (eds.) *Pediatrics and Perinatology*, 2nd edn, p. 285. London: Arnold.

maternal disomy producing IGF_2 underexpression results in fetal dwarfism. In humans, isopaternal inheritance of IGF_2 alleles is associated with the Beckwith-Wiedemann syndrome, which includes hyperinsulinism and fetal macrosomia.

Nongenetic Maternal Factors

There is a high correlation between birth weights of siblings that extends to cousins. The nongenetic, maternal nature of this effect is demonstrated by embryo transfer and cross-breeding experiments. For example, a small-breed embryo transplanted into a large-breed uterus will grow larger than a small-breed embryo remaining in a small-breed uterus. Furthermore, partial reduction in fetal number in a polytocous species such as the rat produces greater than normal birth weights in the remaining offspring. Conversely, embryo-transfer of a large-breed into a small-breed uterus will result in a newborn that is smaller than in its natural large-breed environment. Such evidence demonstrates that fetal growth is normally constrained, and that this constraint comes from the maternal environment. This is a physiological process and includes the maternal-specific capacity of uterine size, placental implantation surface area of the uterus, and uterine circulation, which together support the growth of the placenta and its function.

Maternal Nutrition

Normal variations in maternal nutrition have relatively little effect on fetal growth, because they do not markedly alter maternal plasma concentrations of nutrient substrates or the rate of uterine blood flow, the principal determinants of nutrient substrate delivery to and transport by the placenta. Human epidemiological data from conditions of prolonged starvation, as well as nutritional deprivation in experimental animals, indicate that even severe limitations in maternal nutrition only limit fetal growth by 10–20%. Restriction of caloric and protein intakes to less than 50% of normal for a considerable portion of gestation are needed before marked reductions in fetal growth are observed; such severe conditions often result in fetal loss before the impact on late gestation fetal growth rate and fetal size at birth are manifested. Similarly, fetal macrosomia is only common in pregnancies complicated by gestational diabetes mellitus in which maternal plasma hyperglycemia and hypertriglyceridemia, plus fetal hyperinsulinemia, combine to produce excessive fetal adiposity.

The Placenta

The placenta exerts strong control over fetal growth by providing nutrients directly or in metabolically altered form and amount. Naturally and experimentally, placental growth precedes fetal growth, and failure of placental growth is directly associated with decreased fetal growth. There is considerable variation in this control. For example, experiments in sheep that limited placental growth did not result in proportionately reduced fetal weight, indicating that either the capacity of the smaller placenta to transport nutrients to the fetus increased adaptively, or that the fetus developed increased capacity to extract nutrients from the placenta and direct those nutrients to growth. More characteristically, though, limitation in placental function to transfer nutrients to the fetus directly limits fetal growth. In fact, fetal growth retardation is seen as a natural and reproductively successful (though not perfect) adaptation to nutrient limitation. There is a direct relationship between fetal weight and placental weight in humans, indicating that placental size and fetal size are directly interrelated, although functional interrelationships between placenta and fetus also are important to fetal growth and development.

Maternal Endocrine Influences on Fetal Growth

Changes in maternal circulating growth hormone and growth hormone-like peptides such as placental lactogen, which increase during pregnancy, have combined effects that induce maternal insulin resistance and lead to higher circulating concentrations of glucose and lipids. These in turn are transported in increased amounts to the fetus where, combined with their stimulatory effects on fetal insulin and IGF_1 and IGF_2, promote fetal adiposity (or macrosomia, as in the infant of the diabetic mother) and

limit fetal protein breakdown, both of which promote fetal growth.

Influence of Fetal Endocrine and Autocrine/Paracrine-Acting Growth Factors on Fetal Growth

Growth hormone, which classically acts as the major regulator of postnatal growth, has no demonstrable influence on fetal growth. Fetal insulin does regulate fetal growth, although the complete absence of insulin does not abolish fetal growth. In sheep, for example, fetal pancreatectomy in late gestation limits fetal growth rate only by 20–30%, and pancreatic agenesis in humans produces IUGR fetuses who are 30–50% less than normal weight near term. Insulin infusions into the fetus and excessive fetal insulin secretion enhance fetal glucose utilization and produce increased adiposity, but only a 10–15% increase in fetal nonfat growth. Such hyperinsulinemic conditions also limit protein breakdown, which leads to increased protein accretion, although by limiting protein breakdown, as well as by enhancing amino acid synthesis into proteins, insulin actually decreases plasma concentrations of amino acids, thereby limiting protein synthesis at the same time. Therefore, it is not clear how increased fetal insulin enhances, or its deficiency limits, protein accretion. The primary action of fetal insulin may be to promote glucose utilization and, in turn, enhance protein accretion by providing more energy substrate to fuel protein synthesis and to substitute glucose carbon for amino acids to fuel oxidative metabolism. For example, removal of insulin from the fetus increases fetal glucose concentration and the transfer of glucose from mother to fetus via the placenta, which reduces net fetal carbon accretion. Insulin also directly activates proteins in its signal transduction pathway, promoting the incorporation of amino acids into protein synthesis, and by activation of the MAP-kinase pathway; this occurs even when glucose utilization and oxidation rates are limited by reduction of glucose supply.

Interaction of amino acids with insulin also regulates amino acid synthesis into proteins. Recent studies in the fetal sheep have shown that amino acid infusion, independent of insulin, increases the skeletal muscle concentration of mTOR and elF4E, the key regulatory proteins for ribosomal synthesis of amino acids into protein. In contrast, increases in phosphorylated mTOR and 4EBP1 were only demonstrated when insulin concentrations were also elevated. These observations indicate that amino acids can independently upregulate particular signal transduction proteins during late gestation fetal growth, and emphasize, as does the data showing insulin activation of the MAP-kinase pathway, that nutrient–hormone interaction is central to regulation of growth.

Both IGF_1 and IGF_2 regulate fetal growth. Mice lacking the IGF_1 gene have markedly reduced rates of fetal growth in late gestation. IGF_2 knockouts also have delayed fetal growth that is more pronounced in early to mid-gestation. IGF_1 receptor knockout mice are more growth retarded than either IGF_1 or IGF_2 knockouts alone. These IGF_1 receptor knockouts are growth restricted to the same extent as mice in which both IGF_1 and IGF_2 genes are deleted, confirming that receptor activation is the principal growth-regulating step in IGF_1 and IGF_2 action. Infusions of IGF_1 into fetal sheep demonstrate limited insulin-like effects on fetal glucose metabolism, but they do limit fetal protein breakdown, particularly when sustained hypoglycemia is present in the presence of increased proteolysis. IGFs also regulate fetal growth by regulating placental growth. IGF_2 gene knockout mice have small placentas and, in turn, lower IGF_1 and IGF_2 binding proteins. IGF binding proteins modulate effects of IGF_1 and IGF_2 on fetal growth. Circulating IGF_2 receptors limit IGF_2 effects by binding most of it in the circulation. IGFBP-1 and -2 levels are relatively high in fetal plasma, perhaps limiting the effectiveness of IGF_1, while IGFBP-3 is low in the plasma of fetuses with IUGR, perhaps due to simultaneous insulin deficiency.

Interpretation of Growth Curves

Cross-sectional growth curves have been developed from anthropometric measurements in populations of infants born at different gestational ages. Such curves have been used to estimate whether growth of an individual fetus or preterm newborn is within or outside of the normal range of fetal growth, which is defined as between the 10th and 90th percentile, although what the curves actually show is simply how big a given fetus or newborn is relative to others at any given gestational age. Fetuses and newborn infants who are between the 10th and 90th percentiles for weight vs. gestational age are considered appropriate for gestational age (AGA), those who are less than the 10th percentile are considered small for gestational age (SGA), and those who are greater than the 90th percentile are considered large for gestational age (LGA). In general, SGA infants come from small parents (particularly the mother) and LGA infants come from large parents (again, particularly when the mother is big as well as the father).

Standard fetal and preterm neonatal growth curves represent the third trimester in humans. Each curve is based on local populations with variable composition of maternal age, parity, socioeconomic status, race, ethnic background, body size, degree of obesity or thinness, health, pregnancy-related problems, and nutrition, as well as the number of fetuses per mother, the number of infants included in the study, and how and how well measurements of body size and gestational age were made. Estimates of gestational age often are imprecise because of variable maternal postimplantation bleeding and irregular menses, onset and appearance of physical features of maturation in the infant, and interobserver assessments of an infant's developmental stage.

Mathematical analyses of various growth curves have been used to determine growth rates over relatively short gestational periods or at discrete gestational ages. The data used in the Lubchenco growth curves (**Figure 2**), for example, reflect a simple exponential function showing fetal weight increasing at about $15\,g\,kg^{-1}\,day^{-1}$ for average-sized infants; this rate will be lower for smaller infants and greater for the larger infants.

More recent growth curves have been developed from serial ultrasound measurements of fetal growth in normal pregnancies, providing continuous rather than cross-sectional growth patterns. The growth of a preterm infant is better correlated with serially determined fetal growth rates than with cross-sectional neonatal growth curves. Serial ultrasound measurements of fetal growth also more accurately determine how environmental factors can inhibit (for example, maternal undernutrition globally, or hypoglycemia specifically) or enhance (for example, maternal overnutrition globally or hyperglycemia specifically) growth.

Extremes of Growth and Development: Intrauterine Growth Restriction and Macrosomia

Newborn birth weights have been steadily increasing since the 1970s throughout much of the developed world, although in developed countries, this increase has been tempered by the increased number of preterm infants born as multiple births following *in vitro* fertilization procedures. However, the relative proportions of the two extremes of birth weight, very small infants with intrauterine growth restriction (IUGR) and those who are excessively large with macrosomia, remain constant, and within some populations are actually increasing.

Intrauterine Growth Restriction (IUGR)

In developed countries, 3–7% of newborns are classified as IUGR. These infants weigh less than two standard deviations below the mean of a population born at the same gestational age. Most of these infants experienced suboptimal nutrient supply, and consequently a restriction of fetal growth, as a result of some form of placental insufficiency. IUGR imposes increased risks of specific types of fetal and neonatal morbidity and mortality (**Table 5**).

Table 5 Risks of specific types of fetal and neonatal morbidity and mortality in IUGR infants

Problem	*Pathogenesis/pathophysiology*
Intrauterine death	Chronic hypoxia
	Placental insufficiency
	Growth failure
	Malformation
	Infection
	Infarction/abruption
	Pre-eclampsia
Asphyxia	Acute hypoxia/abruption
	Chronic hypoxia
	Placental insufficiency/pre-eclampsia
	Acidosis
	Glycogen depletion
Meconium aspiration	Hypoxia
Hypothermia	Cold stress
	Hypoxia
	Hypoglycemia
	Decreased fat stores
	Decreased subcutaneous insulation
	Increased surface area
	Catecholamine depletion
Persistent pulmonary hypertension	Chronic hypoxia
Hypoglycemia	Decreased hepatic/muscle glycogen
	Decreased alternative energy sources
	Heat loss
	Hypoxia
	Decreased gluconeogenesis
	Decreased counter-regulatory hormones
	Increased insulin sensitivity
Hyperglycemia	Low insulin secretion rate
	Excessive glucose delivery
	Increased catecholamine and glucagon effects
Polycythemia/ hyperviscosity	Chronic hypoxia
	Maternal–fetal transfusion
	Increased erythropoiesis
Gastrointestinal perforation	Focal ischemia
	Hypoperistalsis
Acute renal failure	Hypoxia/ischemia
Immunodeficiency	Malnutrition
	Congenital infection

From Anderson S, Hay WW, Jr. The small-for-gestational-age Infant. In: Avery GB, Fletcher MA, MacDonald MG (Eds), *Neonatology*: Pathophysiology and Management of the Newborn, 5th Edition. Lippincott-Raven, Philadelphia, pp. 411–444, 1999.

Possible adult disorders resulting from intrauterine growth restriction Interest in IUGR has been enhanced recently by retrospective epidemiological, clinical follow-up, and animal studies that indicate long-term consequences in adult life of IUGR offspring, including higher incidences of obesity, insulin resistance, impaired glucose tolerance, enhanced hepatic glucose production, pancreatic insulin secretion deficiency, type 2 diabetes mellitus, hypertriglyceridemia, and cardiovascular disease, particularly hypertension. These conditions, often called syndrome X or the metabolic syndrome, may represent an example of 'programing,' in which an insult, when applied at a critical or sensitive stage in development, produces lasting, even lifelong, effects on the structure or function of the organism. Mechanisms responsible for these later-life morbidities are not yet established. There is some evidence of diminished pancreatic growth and development, which might become manifest in later life as pancreatic insufficiency when the adult starts and then continues eating a diet rich in simple carbohydrates and lipids. Peripheral insulin resistance may develop in the same way, and hypertension in adulthood may be the result of restricted renal and adrenal development. A common theme among these observations is that excessive weight gain starting at any weight percentile is the strongest predictor of syndrome X or metabolic syndrome disorders.

Macrosomia

At the other end of the birth weight spectrum are macrosomic or large-for-gestational age (LGA) infants. These infants were exposed to excess nutrient supply *in utero*, principally carbohydrates and lipids. Macrosomic newborns have increased specific morbidities primarily associated with metabolic complications of maternal diabetes mellitus during pregnancy and associated birth complications and birth injuries as a result of excessive fetal size.

Macrosomia is defined in a newborn as a birth weight more than two standard deviations above the mean percentile for gestational age or a birth weight greater than 4000 g at term. Neonatal macrosomia has a strong ethnic predisposition affecting up to 50% of Latino and Native American pregnant women versus 19% of African-American pregnant women. Macrosomia is characteristic of infants of diabetic mothers (IDMs) who were hyperglycemic during pregnancy. The diabetes can be long standing, but the most common group producing macrosomic infants are women with gestational diabetes mellitus (GDM), which complicates 3–5% of all pregnancies. The risk of macrosomia is not consistent a cross all classes of diabetes; it primarily reflects the degree and duration of maternal hyperglycemia and hypertriglyceridemia and particularly high spikes of these conditions following meals that are more common in gestational diabetes. The hyperglycemia results in fetal hyperglycemia and hyperinsulinemia, while the hypertriglyceridemia contributes to the effect of the excess glucose and insulin to produce excess fat deposition.

Development of type 2 diabetes in later life in macrosomic offspring IDMs, particularly those with macrosomia, have increased risk of developing type 2 diabetes earlier in life. Mechanisms responsible for this sequence of events include insulin resistance and insufficient insulin secretion (β-cell dysfunction) in response to hyperglycemia. Typically, glucose intolerance from obesity and increased insulin resistance progresses to fasting hyperglycemia and the inability of β-cells to compensate by increasing insulin secretion. This form of β-cell failure appears to be reversible over short periods by improved glycemic control, but long-term exposure to hyperglycemia can lead to β-cell exhaustion and specific inhibition of insulin secretion, which are irreversible by glycemic normalization. The insulin resistance also extends to the liver where glucose production increases. This triad of insulin resistance, reduced β-cell insulin secretion, and increased hepatic glucose production results in type 2 diabetes.

See also: **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology. **Growth Monitoring**. **Infants**: Nutritional Requirements. **Low Birthweight and Preterm Infants**: Nutritional Management. **Pregnancy**: Role of Placenta in Nutrient Transfer; Nutrient Requirements; Energy Requirements and Metabolic Adaptations.

Further Reading

Barker DJ (1993) The fetal and infant origins of adult disease. *British Medical Journal* **301**: 1111.

Barker DJ (1995) Intrauterine programming of adult disease. *Molecular Medicine Today* **1**(9): 418–423.

Battaglia FC and Meschia G (1986) *An Introduction to Fetal Physiology*. Orlando: Academic Press.

Davis JA and Dobbing J (1974) *Scientific Foundations of Paediatrics*. Philadelphia: W.B. Saunders.

Hay WW Jr (1991) Glucose metabolism in the fetal-placental unit. In: Cowett RM (ed.) *Principles of Perinatal-Neonatal Metabolism*, pp. 337–367. New York: Springer-Verlag.

Hay WW Jr (1995) Current topic: Metabolic interrelationships of placenta and fetus. *Placenta* **16**: 19–30.

Hay WW Jr (2003) Nutrition and development of the fetus: carbohydrate and lipid metabolism. In: Walker WA, Watkins JB, and Duggan CP (eds.) *Nutrition in Pediatrics (Basic Science and*

Clinical Applications), 3rd edn, pp. 449–470. Hamilton, ON: BC Decker Inc Publisher.
Hay, WW Jr and Anderson MS. Fuel homeostasis in the fetus and neonate. In: DeGroot LJ and Jameson JL (eds.) *Endocrinology*, 5th edn. Philadelphia: W.B. Saunders (in press).
Hay WW Jr, Catz CS, Grave GD, and Yaffe SJ (1997) Workshop summary: fetal growth: its regulation and disorders. *Pediatrics* **99**: 585–591.
Hay WW Jr and Regnault TRH (2003) Fetal requirements and placental transfer of nitrogenous compounds. In: Polin RA, Fox WW, and Abman SH (eds.) *Fetal and Neonatal Physiology*, 3rd edn, pp. 509–527. Philadelphia: W.B. Saunders.
Milner RDG and Gluckman PD (1993) Regulation of intrauterine growth. In: Gluckman PD and Heymann MA (eds.) *Pediatrics and Perinatology, The Scientific Basis*, 2nd edn. pp. 284–289. London: Arnold.
Molteni RA, Stys SJ, and Battaglia FC (1978) Relationship of fetal and placental weight in human beings: Fetal/placental weight ratios at various gestational ages and birth weight distributions. *Journal of Reproductive Medicine* **21**: 327–334.
Nimrod CA (1992) The biology of normal and deviant fetal growth. In: Reece EA, Hobbin JC, Mahoney MJ, and Petrie RH (eds.) *Medicine of the Fetus & Mother*, pp. 285–290. Philadelphia: JB Lippincott Co.
Ounsted M and Ounsted C (1973) *On Fetal Growth Rate: Clinics in Developmental Medicine*, No. 46. Philadelphia: J. B. Lippincott.
Philipps AF (2003) Oxygen consumption and general carbohydrate metabolism in the fetus. In: Polin RA, Fox WW, and Abman SH (eds.) *Fetal and Neonatal Physiology*, 3rd edn, pp. 465–477. Philadelphia: W.B. Saunders.
Robinson JS, Owens JA, and Owens PC (1994) Fetal growth and fetal growth retardation. In: Thorburn GD and Harding R (eds.) *Textbook of Fetal Physiology*, pp. 83–94. Oxford: Oxford University Press.
Sharp F, Fraser RB, and Milner RDG (1989) *Fetal Growth*. London: Royal College of Obstetricians and Gynaecologists.
Smart J (1986) Undernutrition, learning and memory: review of experimental studies. In: Taylor TG and Jenkins NK (eds.) *Proceedings of XII International Congress of Nutrition*, p. 74. London: John Libbey.
Sparks JW (1984) Human intrauterine growth and nutrient accretion. *Seminars in Perinatology* 8(2): 74–93.
Sparks JW and Cetin I (1991) Intrauterine growth. In: Hay WW Jr (ed.) *Neonatal Nutrition and Metabolism*, pp. 3–41. St. Louis: Mosby Year Book.
Sparks JW, Girard JR, and Battaglia FC (1980) An estimate of the caloric requirements of the human fetus. *Biology of the Neonate* 38(3–4): 113–119.

GROWTH MONITORING

T J Cole, Institute of Child Health, London, UK

Growth is the single quality that most clearly distinguishes between children and adults—children grow, whereas adults do not. This in turn means that healthy children grow well and ill children often grow poorly. For this reason, the monitoring of growth is a logical and effective procedure for detecting child ill health—not just specific growth disorders but also more general conditions that affect growth indirectly.

A Cochrane Review on growth monitoring defined it as "the regular recording of a child's weight, coupled with some specified remedial actions if the weight is abnormal in some way." So the key elements are one or more measurements of weight, plus a protocol for recording, plotting, and interpreting the measurements, leading in suitable cases to some intervention.

The primary purpose of growth monitoring is to detect and treat illness in the individual child. The reasonable question asked by the Cochrane Review is "Does it work?" The Review found only two randomised clinical trials measuring the impact of growth monitoring, and they differed in their conclusions. One found that infants whose growth was monitored for 30 months were no healthier than age-matched controls, whereas the other showed that mothers trained to use a growth chart were more knowledgeable after 4 months. So there is little research on the subject, and even less evidence to justify its use, which is surprising given the enormous resources devoted each year to growth monitoring throughout the industrialized and developing worlds.

A possible reason for this lack of evidence is that growth monitoring is seen as intrinsically 'a good thing.' Parents are always interested to know how their children are growing, and the benefit to them of measuring the child regularly, although difficult to quantify, is assumed. So there is uncertainty as to exactly what growth monitoring is for and what outcome it might lead to.

Purpose and Outcome

The primary aim of growth monitoring is to improve child health by regular anthropometry (which literally means measurement of man and refers to body measures such as weight, height, and mid-upper arm circumference). However, the measurements are useful only if they are properly

recorded and interpreted and a suitable intervention is introduced when the child's growth is suboptimal.

It is important to target the purpose, measurement, and intervention to the environment in which they are used. In practice, this makes growth monitoring a very different proposition in the developing and industrialized worlds due to differences in the burden of disease, resources, and training.

The disease burden is much greater in the developing world, with diarrhea, malaria, respiratory infection, tuberculosis, and HIV all common. Here, effective growth monitoring can in principle save lives. In contrast, in the industrialized world such conditions are less common and milder, so the focus is more on growth disorders such as growth hormone deficiency or Turner's syndrome, where mortality is not the issue.

This different focus also affects the target age range when children are monitored. Mortality risk in the developing world is greatest during infancy, and it is increased by low birth weight. So poor early growth is a potent risk factor for infant mortality, and infancy is the period when growth monitoring is likely to be of greatest value. In contrast, the main concerns in the industrialized world are growth disorders that usually show themselves after infancy, although infant failure to thrive is also a concern. So in the developing world growth monitoring targets the preschool years, whereas in the industrialized world it covers all childhood up to and including puberty.

Throughout the world mothers are encouraged to take their infants to the clinic for regular anthropometry and immunizations. But in the developing world, where infants are much more likely to grow poorly, it makes economic sense to educate mothers, who have the greatest influence over their infant's environment, about the principles of growth monitoring. This education component is not stressed in the same way in the industrialized world.

Maternal education is thus a secondary aim of growth monitoring in the developing world. If growth monitoring makes mothers more aware of their child's state of health, then it should also have an impact on the child's health.

In addition to detecting disease and raising parental awareness at the individual level, growth monitoring in the sense of information gathering has potential benefits at the population level. It provides information about average child growth that is useful for comparison, policy, and planning. For example, a knowledge of mean height for age and weight for age in children from different regions is useful for identifying areas where the prevalence of malnutrition is highest, which in turn allows resources of emergency aid and support staff to be effectively targeted.

Growth monitoring also supports scientific research on the prevention and treatment of disorders affecting growth. Evidence-based child health relies on well-designed studies to test the impact of interventions on child health outcomes. Growth is a proxy for child health and is a common choice of outcome. So growth monitoring fits naturally into the framework of a randomized clinical trial, in which it is used to measure the impact of the intervention. This is different from the situation considered by the Cochrane Review, in which growth monitoring was the intervention. Strictly, the use of repeated anthropometry as an outcome should not be called growth monitoring because it omits the important final stage in which some intervention depends on it.

The Process of Growth Monitoring

The process of growth monitoring involves three stages: anthropometry, interpretation, and referral.

Anthropometry

In infancy, the most common routine measurement is weight. It is simple to do, the required equipment is reasonably cheap, and it provides a convenient global summary of the infant's size. Birth weight in particular is a useful proxy for fetal growth. An advantage of weight is that it relates closely to the mother's own perception of her child's size.

Infant length is more difficult to measure for several reasons. The optimal equipment is a length board with a sliding footboard, which is expensive and needs regular calibration. Simpler equipment such as a tape measure increases the measurement error dramatically. Most important, proper length measurement requires two trained observers—one to hold the infant's head against the headboard and the other to position the footboard and take the measurement. For these reasons, infant length is often measured either poorly or not at all.

Arm circumference (or mid-upper arm circumference (MUAC)) is a popular alternative to weight in the developing world but less so in the industrialized world. This is because arm circumference measurement can detect malnutrition using a simple cutoff. The equipment (a specially marked inextensible tape) is cheaper and easier to use than weighing scales, and arm circumference is highly correlated with weight.

In the industrialized world, head circumference measured in infancy can detect some rare conditions

such as hydrocephalus, which is indicated by a rapid increase in head circumference at approximately the time of birth.

Once past infancy, priorities change. Height becomes much more important, particularly in the industrialized world, where the emphasis is on detecting primary growth disorders such as growth hormone deficiency or Turner's syndrome. Children can be measured standing at approximately 2 years of age using a freestanding stadiometer with counterbalanced headboard. The child's head is positioned in the Frankfort plane (looking straight ahead with the line between the ear hole and eye horizontal), with the child's shoulders, buttocks, and heels touching the back plate, and the observer brings the headboard down gently and reads the height. Alternatively, a second observer can take the measurement while the first checks the child's position. Measurement technique is crucial for height, particularly when it is measured repeatedly. Height velocity is relatively low after infancy, so the height increment over a period of 6 months, for example, may be only 2 or 3 cm depending on age, which can result in excessive measurement error. A competent observer should be able to achieve a measurement error of less than 0.3 cm.

Weight and height are highly correlated, so their assessments are often similar: On average, a tall child is heavy and a short child is light. Once past infancy, weight can be more informative when expressed as an index of weight adjusted for height. There are many weight-for-height indices, but among the most common in the industrialized world is the body mass index, calculated as weight (measured in kilograms) divided by the square of height (in meters). It has been used in adults for decades (the index was originally proposed by Quetelet in the nineteenth century) but in children only relatively recently, and mainly in the area of child obesity. Note that it requires adjustment for age, in the same way as for weight and height. Other weight-for-height indices, used mainly in the developing world, adjust weight for height ignoring the child's age, which is an advantage when the age is not known. However, this leads to biases at certain ages, notably infancy and puberty, when a child's expected weight depends on their age as well as their height.

With the recent steep increase in the prevalence of child obesity, waist circumference has become useful as a measure of central body fatness. Body mass index is less useful because it does not distinguish between fat mass and muscle mass. A child may become fatter over time without becoming heavier simply by losing muscle mass (through inactivity) and gaining an equal mass of fat, as occurred in US adolescents during the 1980s. It is easier to measure waist circumference than skinfold thickness (e.g., triceps or subscapular skinfold), and the required equipment is also simpler—an inextensible tape as opposed to a skinfold caliper (the appearance of which often frightens parents and young children).

Table 1 The suitability of anthropometry for growth monitoring at different ages[a]

Anthropometry	*Infancy*	*Preschool*	*Childhood and adolescence*
Weight	***	**	*
Length/height	*	***	***
Body mass index	*	***	***
Arm circumference	**	*	*
Head circumference	***	*	*
Waist circumference	*	***	***

[a]More asterisks indicate better suitability.

Table 1 summarizes the value of anthropometry at various stages of childhood, as described previously. The process of anthropometry requires attention to detail: suitable equipment that is regularly maintained and calibrated; observers who are trained in correct measurement technique; and regular quality control sessions in which observers are checked, against both themselves and each other, for measurement precision and accuracy. Only in this way can accuracy and precision be maintained.

Plotting and Chart Interpretation

The second stage of growth monitoring involves interpreting the anthropometry. The growing child increases in size over time, so the way to assess the child's growth is by comparison with a set of age-specific norms, usually in the form of a growth chart. This involves plotting the child's measurement on the chart and then interpreting it in the context of any previous measurements. First, growth charts and how they are constructed are discussed.

Growth Chart Construction

Charts to Measure Size

A growth chart summarizes the distribution of the measurement (e.g., weight) as it changes with age in some prespecified reference population (e.g., British children measured in 1990). At its simplest, the chart consists of the median curve, a smooth curve connecting median weight for the population at different ages. Usually, however, there are curves for other distribution centiles as well,

extending typically from the 3rd to the 97th centile, to give an idea of the spread of measurements at each age. A centile corresponds to a given percentage (between 0 and 100) and is a measurement below which that percentage of children in the reference population will be found. For example, in the British 1990 boys weight reference, median (or 50th centile) weight at 1 year is 10.1 kg, whereas the 3rd and 97th centiles are respectively 8.3 and 12.3 kg. Therefore, 50% of British boys aged 1 year weigh less than 10.1 kg, whereas 3% weigh less than 8.3 kg and 97% weigh less (and 3% weigh more) than 12.3 kg.

The lowest centile curve on the chart is often used as a cutoff to detect poor growth. Children with measurements falling below this centile are viewed as 'at risk' and may be referred for more detailed examination. **Figure 1** shows the British weight reference for boys in infancy, which has nine centile curves ranging from the 0.4th at the bottom to the 99.6th centile at the top. The value of these more extreme centiles for screening is explained later.

Using a centile curve as the cutoff on the chart is just one approach to identifying at-risk children. In the developing world, where the concept of centiles can be difficult to explain, a simpler alternative is 'percent of the median.' Here, the cutoff is constructed as a curve that is, for example, 80% of the median weight curve, which is broadly similar in shape to the 3rd centile curve. A child whose weight falls below the cutoff is said to be below 80% weight for age.

Another approach to defining cutoffs involves standard deviation (SD) scores, also called *z* scores. These are linked to centiles through the underlying frequency distribution, in particular the SD of the measurement at each age. With a normal distribution, the median and mean coincide and the SD

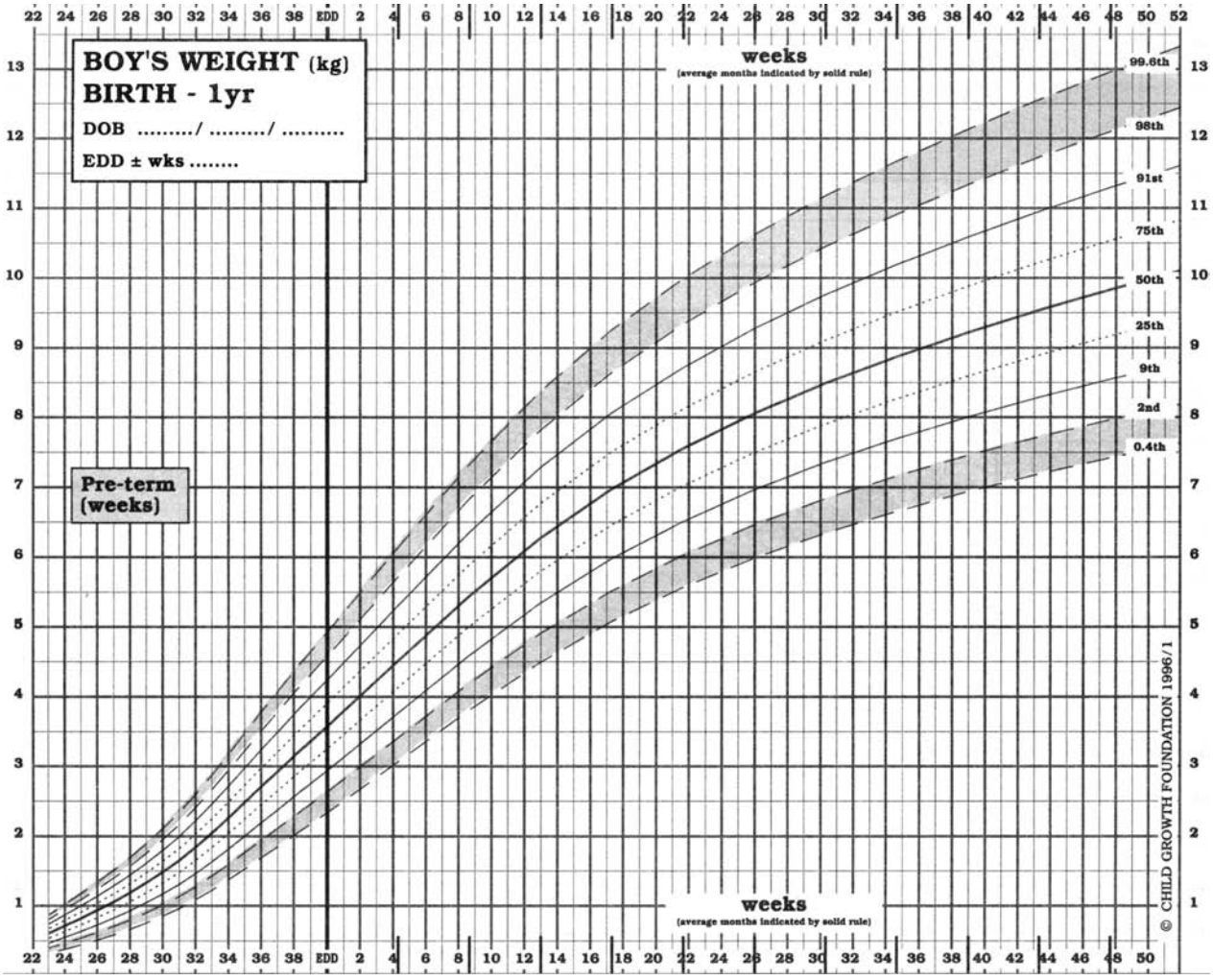


Figure 1 The British 1990 boys infant weight chart, covering 22 weeks of gestation to 12 months. The nine centiles range from the 0.4th to the 99.6th, and infants whose weights fall outside the extreme centiles are referred. (© Child Growth Foundation.)

score is 0. The 2nd centile is approximately 2 SDs below the median at all ages, whereas the 75th centile is 1.3 SDs above the median. Therefore, the 2nd centile corresponds to an SD score of −2 and the 75th centile to +1.3.

Height is normally distributed, but much anthropometry including weight is not—it has a skew distribution. The link between centiles and SD scores can be extended to measurements that are not normally distributed using a technique called the LMS method, which is the basis for many growth reference charts. The telltale sign of a skewness-adjusted chart is that the centile curves are asymmetrically spaced at each age; for example, the gap between the 25th and 50th centile curves is less than the 50th–75th centile gap.

The British growth reference centiles (**Figure 1**) are defined by their corresponding SD scores. The centiles are spaced two-thirds of an SD score apart, so the 0.4th centile corresponds to −2.67 SD scores. Similarly, the WHO chart spaces its centiles 1 SD score apart, from −3 to +3, which allows it to cover the wide range of anthropometry seen internationally.

Charts for use in the developing world tend to be simpler in design, often with only two or three centile curves. They highlight a region on the chart where children's individual growth curves should lie, and mothers are taught that children with curves within this region are healthy children. One common chart is the Road to Health chart (**Figure 2**). The advantage of the chart is that it can also include public health information about the timing of breast feeding, immunizations, etc.

The most difficult part of the measurement process is plotting the data accurately. For use in the developing world, the Direct Reading scale is a clever device that links the weighing scale to the chart so that as the child is weighed a pen records the scale deflection on the chart. This and the Road to Health chart can be obtained via Teaching Aids at Low Cost (TALC).

Another direct-reading chart is the Nabarro height-for-weight wall chart (also available through TALC). The child is first weighed, and then he or she stands against the chart at the point corresponding to his or her weight. The chart consists of

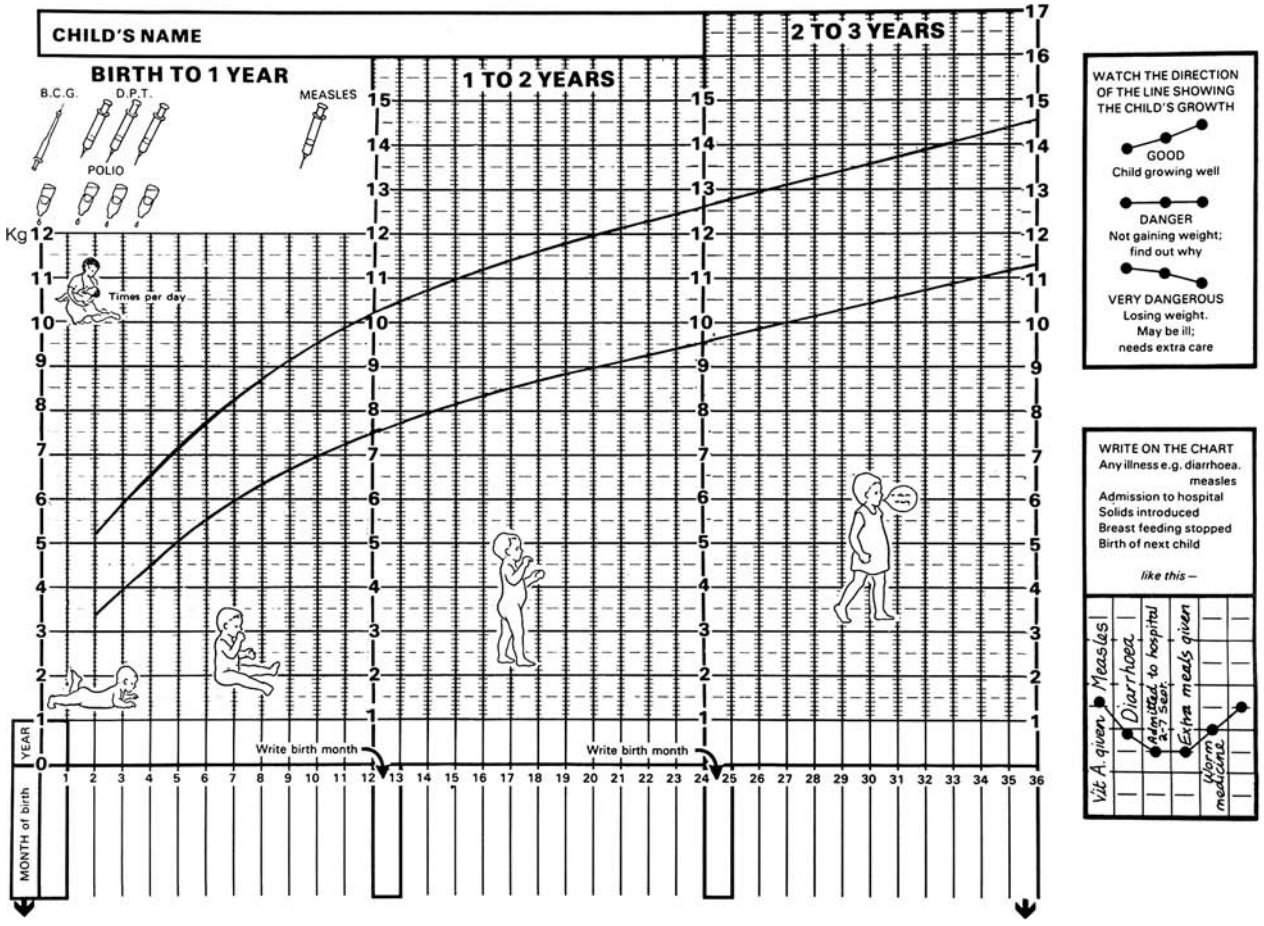


Figure 2 The TALC Road to Health chart. (Reproduced by kind permission of Teaching-Aids at Low Cost.)

vertical bars, and the height of the relevant bar is compared to the child's height. If the child is thin, he or she will be taller than other children of the same weight, and the upper section of the bar is color coded in red to flag excessive thinness, also known as wasting.

Familial Height Adjustment

Child height is strongly correlated with parental height, so many short children have short parents. Short children from tall families are easier to detect if their height is adjusted for midparent height. Tanner published a chart to do perform this adjustment in 1970, but it proved cumbersome to use, and currently the target height method is preferred. This method uses midparent height to estimate the child's likely height centile as an adult and compares it with the child's current centile.

However, the target height approach is also cumbersome and relies on the heights of both parents, one of which may not be available. A familial height chart has been described that compares the child's current height centile with that of his or her mother and/or father and/or sibling(s). The advantages of the chart are that it uses all the available familial height information (a sibling is as close genetically as a parent), it avoids all calculations (apart from reading the centiles off the height chart), and it adjusts for the secular trend in height by using an older height reference for the parents.

Ethnicity

A growth chart reflects the size and growth of its reference population. When the target population is materially different in size, as happens with ethnic minorities in the industrialized world (e.g., Hispanics in the United States or Asians in the United Kingdom), the chart's centiles can be misleading. More than 3% of such children may be found below the 3rd height centile, and the relevant cutoff for referral should take ethnicity into account. One simple way to do this is to estimate an offset, measured in SD score units, to apply to the chart for a given ethnic minority group. For example, Southeast Asian children in the United Kingdom are approximately 0.4 SDs shorter than ethnic Caucasians, corresponding to approximately half a channel width on the chart, the distance between adjacent centiles.

The differences in height between ethnic groups can be explained partly by genetics (i.e., differences in parental height) and partly by differences in the environment. However, the parental height differences reflect the environment of previous generations, so it is not feasible to ascribe the differences purely to genes or the environment—the two are inextricably linked. If the child's height is appropriate for the heights of his or her parents, this should provide reassurance.

Charts to Measure Growth

Growth charts are usually constructed using cross-sectional data—each reference child contributes a single measurement to the data set. This allows children in each age group to be ranked by size to identify the required centiles, which can then be plotted against age. James Tanner coined the term "distance to indicate size attained," meaning the distance the child has travelled on the journey from conceptus to adult.

Growth is the rate of change of size, or velocity in Tanner's notation. To measure growth in an individual child requires at least two measurements separated in time. However, the growth chart is based on cross-sectional data, which provide no information about growth. This is an irony of the conventional growth chart: It is designed to measure size, not growth. It flags poor growth when the child's growth curve rises more slowly than the centile curve, but it does not distinguish between mild and severe faltering. The chart effectively has only three growth categories: normal (i.e., tracking along centiles as recommended by the Road to Health chart), above average growth (i.e., crossing centile curves upwards), and below average growth or crossing centiles downwards. Within the latter two categories, the chart does not grade the rate of centile crossing.

Tanner introduced notation to distinguish between distance or size charts, on the one hand, and velocity or growth charts, on the other hand. Velocity charts display centiles of growth velocity by age, and probably the most useful such chart is for height velocity measured over 1 year. In theory, velocity is better than distance for detecting short-term growth faltering, but in practice it is more difficult to measure because it involves two measurement errors, not one. It is also more complicated to monitor because the process of charting velocity involves taking the two measurements, calculating the velocity in units of cm/year or kg/year, and then plotting it at the mid-age point on the velocity chart.

Other forms of charts have been described that provide more information about growth velocity. Thrive lines are extra lines superimposed on the centile chart to quantify the rate of centile crossing of weight in infancy. Infants with monthly weight measurements who are growing on the 5th velocity centile

track along the thrive lines. If they track in this way for 1 month it is a sign of moderate weight faltering, but if it continues then the faltering is progressively more severe. The thrive lines take into account the child's age and sex and adjust for regression to the mean. **Figure 3** illustrates the thrive lines superimposed on the weight chart of **Figure 1**. An infant measured twice 4 weeks apart whose weight curve tracks along the thrive lines (i.e., crosses centiles downwards) is growing on the 5th velocity centile, indicating moderate weight faltering.

Wright designed a weight monitoring chart in which the centile curves are spaced according to the infant's chance of crossing them in a given period of time. Large infants tend to cross centiles downwards more rapidly than small infants—this is a consequence of regression to the mean. Therefore, the centiles are relatively widely spaced at the top end and become progressively closer together at lower centiles. The chart is designed to simplify the assessment of infants recovering from failure to thrive.

Interpretation

The choice of charts for growth monitoring is bewildering—size charts, growth charts, and parental adjustment, each with many different cutoffs. In the industrialized world, the aim is to detect growth disorders as early as possible, and the key question is "Which form of growth monitoring is most effective at detecting disease?"

A common view is that growth monitoring requires measurements on two or more occasions, whereas growth screening involves a single measurement, and monitoring is therefore better than screening. Indeed, this is a fundamental tenet of growth monitoring as practiced in the developing world. Yet there is no direct evidence either way.

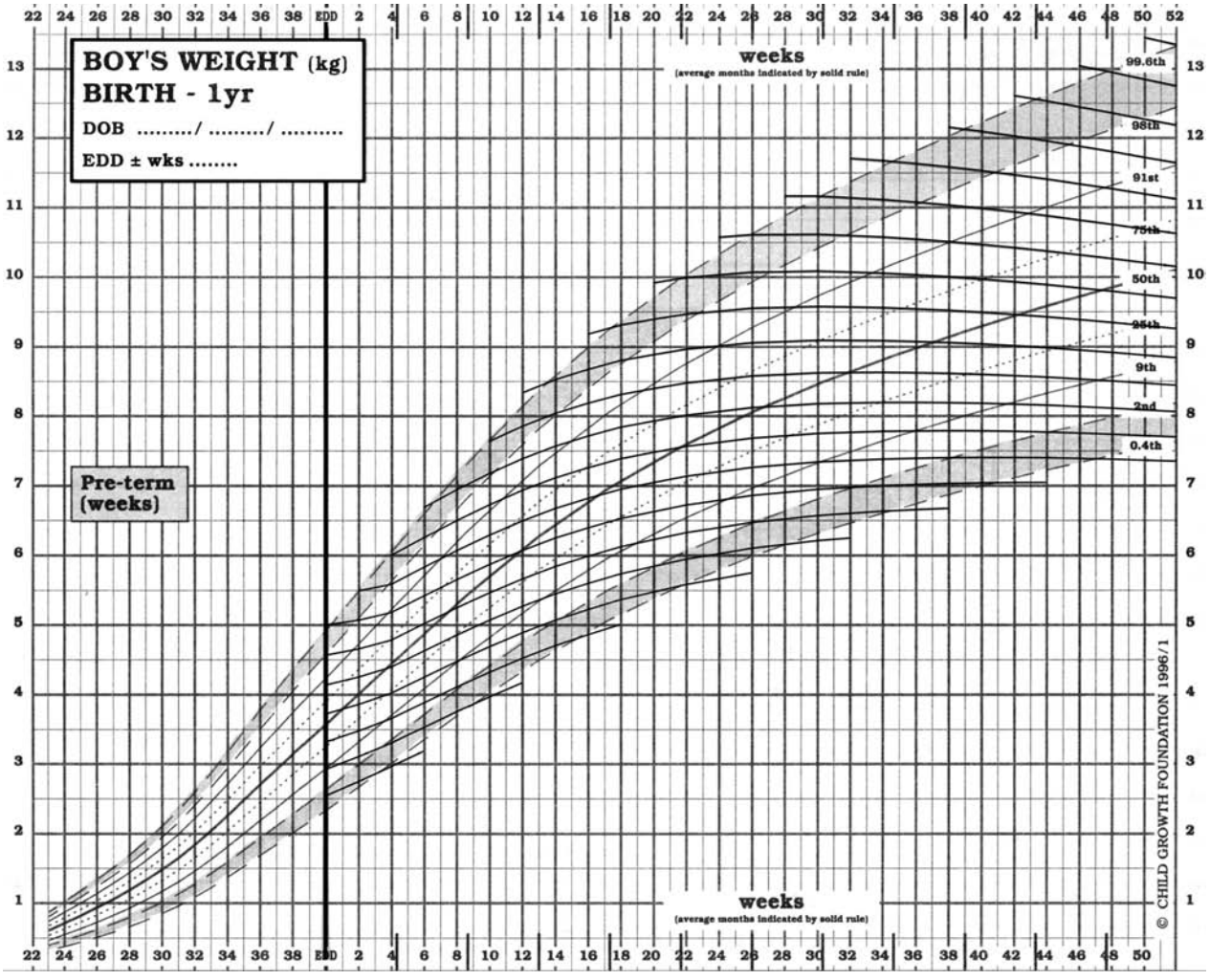


Figure 3 The British 1990 boys infant weight chart with thrive lines superimposed. The thrive lines represent downwards centile crossing corresponding to the 5th velocity centile, i.e., moderate weight faltering. (© Child Growth Foundation.)

A Dutch study by van Buuren and colleagues for the first time addressed this question, treating height monitoring as a diagnostic test with associated sensitivity and specificity. The study aimed to detect Turner's syndrome in girls using a regional growth survey as the corresponding normal population. Three measures of size and growth were used: height SD score, height SD score adjusted for parental height, and height velocity, each with a range of cutoffs. The study found that height alone was not very effective at identifying cases of Turner's syndrome, whereas height adjusted for parental height was very effective. Height velocity was useful for ruling out Turner's syndrome, but it was poor at ruling it in. Overall, the study emphasized the value of height adjusted for parental height and tended to discount the value of repeated measurements.

Similar studies need to be done for other outcomes, but until they are performed, the choice of charts and cutoffs needs to be based on simpler criteria. The British height reference uses the 0.4th centile to screen for short stature (**Figure 1**), which screens in 0.4% of the population and corresponds to a specificity of approximately 99.6%. It is very important for the false-positive rate to be as low as 0.4% to avoid overwhelming growth clinics with referred patients. In the United Kingdom, height velocity is viewed as too noisy, because of the two measurement errors involved, to justify its routine use.

Of course, these conclusions apply to growth monitoring in the industrialized world. They should not be extended uncritically to the developing world, where the purpose of growth monitoring is different—to reduce malnutrition. Here, underweight is judged by the Road to Health chart and is the universal indicator of malnutrition. Weight can fluctuate rapidly due to disease, and encouraging mothers to weigh their infants regularly is a logical way of encouraging the child to grow along the 'road to health.' This philosophy is just one component of UNICEF's GOBI program, which combines growth monitoring, oral rehydration, breast feeding, and immunization.

When height is also available, weight-for-age can be separated into height-for-age and weight-for-height, where low values are known as stunting and wasting. Stunting reflects long-term malnutrition and wasting short-term malnutrition. The implications of the two conditions are different, the latter indicating a need for medical intervention, possibly urgent, and the former is a proxy for more deeply seated socioeconomic problems that are less amenable to intervention. In practice, however, height is rarely measured in infancy and the main focus is on detecting underweight.

Nature of the Intervention

The intervention arising from growth monitoring may be quite specific (e.g., the identification and treatment of a particular growth disorder) or it may be more general (e.g., referral to a growth clinic, a dietician, or a feeding station). If the mother is involved, it may alter her view of her child's health and so modify her child care. At the population level, it may affect the allocation of resources (e.g., between regions for malnutrition relief).

If growth monitoring is evaluated in the spirit of the Cochrane Review, the outcome it leads to needs to be quantifiable and objective. Also, the Cochrane Review evidence, such as it is, suggests that growth monitoring in the developing world is ineffective. However, several potential outcomes are too diffuse to quantify (e.g., increased parental interest and education), and this needs to be recognized. The absence of an evidence base in favor of growth monitoring should not necessarily be interpreted as evidence that it lacks benefit. The benefits may simply be too subtle to detect using conventional trials.

See also: **Growth and Development, Physiological Aspects**. **Low Birthweight and Preterm Infants**: Causes, Prevalence and Prevention. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **Nutritional Assessment**: Anthropometry. **Nutritional Surveillance**: Developed Countries; Developing Countries. **Obesity**: Childhood Obesity.

Further Reading

British 1990 growth charts at http://www.healthforallchildren.co.uk.

Cole TJ (1993) The use and construction of anthropometric growth reference standards. *Nutrition Research Reviews* **6**: 19–50.

Cole TJ (1997) 3-in-1 weight monitoring chart [Research letter]. *Lancet* **349**: 102–103.

Cole TJ (2000) A simple chart to assess non-familial short stature. *Archives of Disease in Childhood* **82**: 173–176.

Hall DMB (2000) Growth monitoring. *Archives of Disease in Childhood* **82**: 10–15.

International Information Support Centre at http://www.asksource.info/

Jelliffe DB and Jelliffe EFP (1990) *Growth Monitoring and Promotion in Young Children*. Oxford: Oxford University Press.

Panpanich R and Garner P (2003) *Growth Monitoring in Children* [Cochrane review], The Cochrane Library, Issue 3. Oxford: Update Software.

Teaching-Aids at Low Cost at http://www.talcuk.org/

US CDC 2000 growth charts at http://www.cdc.gov/growthcharts/

van Buuren S, van Dommelen P, Zandwijken GRJ *et al.* (2004) Towards evidence based referral criteria for growth monitoring. *Archives of Disease in Childhood* **89**: 336–341.

WHO (1986) *The Growth Chart: A Tool for Use in Infant and Child Health Care*. Geneva: WHO.

WHO (1995) *Physical Status: The Use and Interpretation of Anthropometry*. Geneva: WHO.

Wright CM (2000) Identification and management of failure to thrive: A community perspective. *Archives of Disease in Childhood* **82**: 5–9.

Gut Flora *see* **Microbiota of the Intestine**: Probiotics; Prebiotics

H

HANDICAP

Contents

Down's Syndrome

M Collins and A Laverty, Muckamore Abbey Hospital, Antrim, UK

Down's syndrome, named after John Langdon Down, is the most widely recognized chromosomal disorder found in humans and falls into a category of chromosomal disruptions known as trisomies; hence the other term for the condition, trisomy 21. People with Down's syndrome vary widely in their abilities, but the syndrome is the most common genetic cause of learning disability. The syndrome is characterized by abnormalities of both structure and function, some of which may be amenable to nutritional intervention.

More than 90% of Down's syndrome individuals have a total of 47 chromosomes in cells instead of the usual 46. The remaining cases are mainly either translocations, where there is a rearrangement of fragments of chromosomes, or mosaics, in whom there are both normal and trisomic cells (i.e., mosaic trisomy 21). There is a relationship between the frequency of Down's syndrome births and age, with both very young mothers and older mothers having a higher incidence of the syndrome. It has been suggested that nutrition may be implicated in the nondisjunction of the chromosomes. The additional chromosomal material in Down's syndrome usually comes from the mother, but it can come from the father, and one study reported that the maternal age relationship had been found associated with paternal origin of the additional chromosome. This observation may be indicative of hormonal changes in the older mother that reduce the likelihood of spontaneous abortion in an abnormal pregnancy.

The incidence of Down's syndrome is approximately 1 in 600–1000 live births. Prevalence is rising as life expectancy has improved in recent years with advancing medical knowledge and higher standards of care, but concurrently there is a declining incidence of live births in industrialized countries due to prenatal diagnostic screening and abortion.

Physical defects common in Down's syndrome include congenital anomalies of the gastrointestinal tract, which occur in approximately 12% of infants with Down's syndrome. Most of these anomalies require the neonate to be operated on immediately to allow nutrition. Congenital heart disease occurs in approximately 40% of infants with Down's syndrome. Children with congenital heart disease may present with failure to thrive, but after surgical repair of heart defects these children usually improve. Immune dysfunction, increased susceptibility to leukemia, and premature aging with Alzheimer-like changes in the brain are major features of the syndrome.

Thyroid dysfunction is more common in people with Down's syndrome, with the incidence increasing as they get older. Hypothyroidism is most frequently reported, but hyperthyroidism can also occur. Correction of thyroid function is essential to allow normal learning processes to take place and to aid weight control.

There are many biochemical anomalies associated with the syndrome, mainly quantitative rather than qualitative. It is presumed that the overexpression of genes on chromosome 21 contributes to both the structural and the functional pathology. Overdose effects of the genes already mapped to chromosome

Table 1 Nutritional complications of Down's syndrome

Physical	Problems with muscle tone, oral health and dentition, chewing, and swallowing
Metabolic	Anomalies in carbohydrate protein and lipid metabolism
	Increased demands on antioxidant defence system and methylation pathways
	Increased incidence of diabetes, coeliac disease, obesity, and thyroid disorders
Behavioral	Food consumption and exercise choices

21 are thought to alter pathways controlling the production of monocarbons, purines, pyrimidines, tubulins, and myelin.

The nutritional complications associated with Down's syndrome are summarized in **Table 1**.

Nutritional Status

It is debatable how relevant reference data from normal groups are for people with Down's syndrome.

Dietary Assessment

In children with Down's syndrome, conflicting reports have shown energy intake to be less than, similar to, or more than that of age-matched comparison groups, with a small percentage of children exceeding the recommended daily intake by more than 50%. However, because children with this syndrome tend to be shorter than age-matched children, energy intake comparisons need to be calculated per unit of body height.

Lower than recommended intakes of nonstarch polysaccharide coupled with higher than recommended consumption of protein and fats have also been reported. Some, but not all, researchers have reported low intakes of calcium, particularly in preschool and school-age children who refuse or limit milk consumption. Iron intakes have been reported to be low, particularly non-hem iron. Vitamins A and C intakes are limited in those who have a poor intake of fruit and vegetables. Intake of B vitamins has also been reported as low.

Laboratory Assessment

Carbohydrate Metabolism

Fasting blood glucose levels are usually in the normal range, but the glucose tolerance curve has been reported to be flatter and often with a double-humped curve, suggestive of delayed absorption. There is an increased incidence of both type 1 (insulin dependent) and type 2 diabetes in Down's syndrome.

Protein Metabolism

Disturbances in protein metabolism are common in Down's syndrome. An increased level of immunoglobulin A and immunoglobulin G antibodies to food antigens has been reported, and several studies have reported an increased prevalence of coeliac disease. Abnormal levels of fasting plasma and urinary amino acids have been reported.

Lipid Metabolism

One study reported no significant differences between study and control groups, drawn from within the same families, in levels of total cholesterol, low-density lipoprotein, apolipoprotein B, and the apolipoprotein B-to-apolipoprotein A-I ratio. Triacylglycerol levels were significantly increased and serum high-density lipoprotein cholesterol-to-total cholesterol ratio was significantly decreased in Down's syndrome. This suggests increased risk for coronary heart disease. The results of this study and other studies reporting no difference between Down's syndrome and comparison groups with regard to atherosclerosis contrast with early reports that suggested a decreased incidence of coronary artery disease in Down's syndrome. It is not clear whether the differences reflect nutritional variables or population variable changes reflecting the increased survival rate in infancy.

There is evidence of increased lipid peroxidation in Down's syndrome.

Vitamins

Some, but not all, studies have reported biochemical evidence of deficiency of thiamin, nicotinic acid, pyridoxine, cobalamin, folate, ascorbic acid, retinol, β-carotene, and α-tocopherol. Vitamin D metabolites have been reported to be in the normal range in a Spanish study that demonstrated wide seasonal variation linked to intensity of solar radiation.

Minerals

Low iron, calcium, manganese, and zinc concentrations have been reported, and the iron-to-copper ratio has been reported to be decreased. Studies reported that intracellular zinc in blood mononuclear cells was approximately 47% lower than that of normal controls, and it is possible that this may play a role in thyroid dysfunction, immunodeficiency, retarded growth, and faulty DNA repair. Further

research is required to determine if zinc supplements are beneficial and at what level. Supplementation with selenium aimed at increasing levels of the selenium-dependent enzyme glutathione peroxidase is reported to have led to a decrease in initially high blood mononuclear cell levels of copper, but it did not affect iron or zinc.

Vitamin and mineral levels have been held to reflect not just nutrient intake but also abnormal metabolism. Assessments of antioxidants and oxidation by-products are useful indicators of nutritional status in people with Down's syndrome. The overexpression of the superoxide dismutase system, the purine synthesis pathway, and cystathionine β-synthase are thought to create extra demands for antioxidants and for folate, but despite gene dosage effects the many biochemical anomalies that have been reported in people with Down's syndrome show a great deal of individual variation.

Anthropometric Assessment

Growth delay is one of the main characteristics of Down's syndrome, but impaired growth velocity is particularly evident at certain stages of development.

Fetal growth has usually been reported to be relatively normal and the length of the neonate is often within normal limits, allowing for gestation. Some studies have reported prenatal growth delay, and a major Italian study comparing neonatal length, weight, head circumference, and weight/length squared reported all percentiles of growth variables lower in Down's syndrome infants except for weight/length squared percentiles.

At approximately the age of 6 months, when growth starts to become regulated by growth hormone, growth velocity usually begins to show a marked reduction from normal levels. Although for the Down's syndrome child the period between birth and 2 years and the period between 6 years and 10 years of age are times of accelerated growth, the deviation from normal levels remains significant. Slow growth velocity is also a particular feature of adolescence, although there is a pubertal growth spurt. The deviation of adult stature from the means of reference groups is greater than the deviations in early infancy.

The short stature in Down's syndrome seems to be mainly the result of impaired growth of the long bones of the leg, because sitting height measurements show that the growth of the vertebral column is closer to normal.

Why there is growth delay in Down's syndrome is not entirely clear, and several hypotheses have been advanced. Both human growth hormone therapy and zinc sulfate supplementation of the diet have been reported to accelerate growth.

Children with Down's syndrome tend to be not only shorter but also heavier than reference children. Charting the height and weight of a child with Down's syndrome using reference norms from the general population will show the abnormality of the growth pattern. However, it is more useful clinically to compare the height and weight of an individual against syndrome-specific norms because this will identify any deviation from the growth patterns of children with Down's syndrome.

Italian percentile charts have been drawn up for neonates with Down's syndrome based on a large sample of consecutively born infants. Specific growth charts for children with Down's syndrome have been constructed based on anthropometric assessments of US children, Sicilian children (thought to be representative of southern European children), and Dutch children (thought to be representative of northern European children). On average, the Dutch children were taller than the US children, and the US children were taller than the Sicilian children.

Nutritional Requirements

Children and adults with Down's syndrome need the same range of nutrients as the general population. Energy intake standards based on age groups are not appropriate for children with Down's syndrome. Energy intakes in both children and adults need to be tailored to height and weight and to physical activity.

Nutritional Therapy

In the 1970s and 1980s, hopes were raised that megadoses of vitamins and minerals would boost intelligence in children with Down's syndrome, but rigorous studies have shown these doses lead neither to higher intelligence nor to better health. In addition, there is anxiety about possible side effects, particularly from the fat-soluble vitamins.

As more has been learned about the genes on chromosome 21, interest has shifted to targeted nutritional intervention aimed at correcting the metabolic anomalies that are common in Down's syndrome due to genetic overexpression, with the emphasis on nutrients to maintain health and prevent disease. Targeted nutritional supplementation with vitamins, minerals, amino acids, digestive enzymes, and essential fatty acids is still

controversial. Clinicians have reported differences between children treated and not treated in health, growth, and cognitive and speech functions, and extensive double-blind studies are planned.

Dietary Management

Dietary Guidelines

Dietary recommendations are as for the general population until research proves otherwise. There are no specific dietary guidelines for the woman pregnant with a Down's syndrome child or for the pregnant Down's syndrome woman. There are indications that antioxidant and essential fatty acid intake may be particularly important, and folic acid supplements beneficial, but dietary advice is currently the same as for other pregnant women.

The situation is similar for infant feeding. Brain lipids in the human infant are known to change with changing intakes of fatty acids. The needs of a newborn with Down's syndrome for the long-chain polyunsaturated fatty acids docosahexenoic acid and arachidonic acid have not been determined. Since breast milk contains the preformed dietary very long-chain fatty acids that seem to be essential for the development of the brain and the retina, it seems prudent to encourage breast-feeding.

The antioxidant defence system has a particularly important role in Down's syndrome, and parents and caregivers can be advised on providing a diet rich in antioxidants. Dietary intakes need to be considered for the sulfur amino acids (which are needed for glutathione synthesis); fat-soluble vitamins A, C, and E; water-soluble vitamins B_6, B_{12}, and folic acid; and the minerals selenium and zinc. In latitudes where no vitamin D is synthesized in the winter months, it is particularly important to ensure exposure to sunlight during summer months to maintain adequate stores of the vitamin throughout the year because studies indicate an increase in the incidence of osteoporosis in Down's syndrome.

Feeding Behavior

Feeding skills tend to be delayed in the young child with Down's syndrome, but the sequence of the emergence of the skills is the same as that for other children if appropriate learning opportunities are provided.

Infants with Down's syndrome have a smaller oral cavity, which makes it easier for liquids to spill from the sides of the mouth. If a child is hypotonic, the tongue is likely to flatten out when the child sucks instead of forming a groove around the nipple, so the child will have a weak suck, may gag, and milk will leak from the mouth. Feeding will be exhausting, and particularly when the child has a cardiac defect, the child may have difficulty taking in enough milk to meet energy requirements. Tube feeding may be necessary until the child develops better tongue control. As infants with Down's syndrome are often placid, sleepiness may be overlooked and feeding will be easier if the infant is wide awake. Extra support for the infant during feeding, and in particular supporting the infant's chin to help steady the jaw, can help encourage intake. Because of the benefits of breast feeding, it is essential that nursing mothers are given help and advice when their infants have initial difficulties. Breathing during feeding may be helped if the mouth and nose are cleared of mucus with a syringe before feeding.

As with other children, it is important to introduce textured food when the child is developmentally ready, and information should be provided to parents and caregivers regarding both appropriate expectations and helpful feeding techniques as well as dietary advice. In children with Down's syndrome, poor neuromotor control of the tongue may result in the continued use of pureed food. There may be slow initiation of the swallow response, possibly because of hypotonic pharyngeal muscles, and oral sensitivity problems may also make the transition to textured foods difficult. Persistent feeding problems merit multidisciplinary assessment and therapy. Impaired swallow can result in food being aspirated and contribute to respiratory problems. The presence of the tongue protrusion reflex past the age of 12–18 months can result in delayed progression to solid food and can contribute to malocclusion of teeth. Also, dental abnormalities can exacerbate difficulties with chewing and can contribute to poor nutrition because children who have problems chewing may be offered soft, often high-energy food and be given little opportunity to accept meats, fresh fruits, and vegetables, which are lower in energy.

Fresh fruit and vegetables also provide the non-starch polysaccharide that can help prevent the constipation common in Down's syndrome. Prunes, fruit juices, and water between meals also help with constipation. Because the hypotonia in Down's syndrome also contributes to sluggish bowel habits, this is another reason for children and adults to be encouraged to take part in physical activity. If constipation does not respond to dietary management, there should be a medical

assessment to exclude gastrointestinal and thyroid problems.

Dental Problems

Dental anomalies in Down's syndrome include changes in tooth structure, reduced total number of teeth, and delayed or abnormal eruption. Together with the physical abnormalities of the facial appearance and oral cavity, these can all impact on feeding. Dental disease is common in Down's syndrome because teeth are more at risk of wear through bruxism and decay due to fragile enamel. In addition, gum disease (gingivitis) and oral infections due to mouth breathing can lead to teeth becoming loose and falling out. A healthy balanced diet, low in sugar-containing fluids and fizzy drinks (including 'diet' varieties), without frequent snacks and plenty of fruit and vegetables will help preserve teeth.

Obesity

Obesity is not inevitable in Down's syndrome, but it is common. Obesity in children with Down's syndrome has been reported from different cultures and different ethnic backgrounds. From Australian and North American studies, it has been reported that by 2 or 3 years of age more than 30% of children with trisomy 21 are overweight, and by 9 years of age the average child with Down's syndrome is obese; from the age of 10 years, the average weight of Dutch children with Down's syndrome is above the 90th percentile of weight-for-height curves of healthy children. Since the 1960s, obesity has been increasing rapidly in school-age children in Japan, and in 1994 a survey of children at special schools reported that more than 20% of school children with Down's syndrome were obese.

High rates of overweight and obesity have been reported in adults with Down's syndrome, both living in the community and at home, and more commonly in females than males. Overweight and obesity are particularly associated with living in the family home compared to supervised community units or hospitals, but they are not significantly associated with the degree of learning disability.

Because excessive weight gain in childhood often leads to adult obesity, it is important to encourage healthy choices in childhood. Why children with Down's syndrome have a tendency to become fat is not clear, but it is likely that several factors influence the weight gain. Retardation of growth resulting in short stature may be of prime importance. Obesity in people with Down's syndrome has also been linked to poor eating behaviour, excessive energy intake, depressed resting metabolic rate, hypotonia, reduced exercise, and endocrine abnormalities such as hypothyroidism. Abnormal substrate fat oxidation may also be implicated (**Table 2**).

Prepubescent children with Down's syndrome have a decreased resting metabolic rate compared to control children matched for body mass index. Children of approximately the same body composition, whether or not they have Down's syndrome, expend similar levels of energy in physical activity. Since obesity is negatively correlated with motor performance, it is likely to lead to a reduction in sporting and physical recreation activities, and thus obesity has social as well as health implications, in children with Down's syndrome as in other children. However, children and adolescents with Down's syndrome have been shown to have difficulty with sustained physical exercise in both laboratory and recreational situations, and this has been attributed to physiological impairments, notably cardiovascular, as well as to lack of motivation.

Children, adolescents, and adults with Down's syndrome have a deficit in isokinetic strength, and by the age of 14 years adolescents with testosterone levels in the normal range fail to show the pubertal muscle strength increase. Progressive resistance exercise programs can help to build muscle strength, and regular aerobic exercise will improve exercise tolerance. Often, individuals can attain high standards in competitive gymnastics and swimming. The overexpression of collagen genes on chromosome 21 affects both muscle and connective tissue, and it has been claimed that targeted nutritional treatment leads to rapid improvement in both muscle strength and joint stability.

In a cross-sectional study of men and women with Down's syndrome, body mass index (weight in kilograms divided by height in meters squared)

Table 2 Factors predisposing to obesity in Down's syndrome

Factor	*Increased*	*Decreased*
Poor eating behavior	↑	
Calorie intake	↑	
Resting metabolic rate		↓
Muscle tone		↓
Exercise		↓
Thyroid function		↓
Substrate fat oxidation		? ↓

declined with increasing age. Further research is needed to clarify whether individuals lose weight as they age or whether there is a shorter life expectancy for individuals with higher body mass indices.

Aging

The rapid aging that characterizes Down's syndrome is in line with the accumulating evidence that many degenerative diseases are associated with deleterious activated oxygen species reactions. Activated oxygen species can damage genetic material and inactivate membrane-bound enzymes as well as cause lipid peroxidation in cell membranes. Of particular relevance to Down's syndrome is evidence relating to cancer, inflammatory joint disease, diabetes, degenerative vascular disorders, degenerative eye disease, and senile dementia—all reported to have increased prevalence in Down's syndrome.

The gene for copper/zinc superoxide dismutase is on chromosome 21, and copper/zinc superoxide dismutase levels are elevated by 50% in a range of cells of people with Down's syndrome, including erythrocytes, blood platelets, leucocytes, and fibroblasts. The increase has also been reported in fetal cerebral cortical cells. Although copper/zinc superoxide dismutase usually functions as an antioxidant, it seems likely that in Down's syndrome the raised levels lead to oxidative stress. When the increased production of hydrogen peroxide through catalysis of superoxide free radicals is not matched by a sufficient increase in glutathione peroxidase to metabolize the additional hydrogen peroxide to water and oxygen, there is thought to be an increase in highly reactive hydroxyl radicals leading to increased lipid peroxidation.

Fibroblasts derived from people with Down's syndrome show elevated lipid peroxidation, and levels of thiobarbituric reaction products, which indicate the extent of lipid peroxidation, have been reported to be raised in erythrocytes from Down's syndrome subjects compared to controls.

A reported increase in the activity of the hexose monophosphate pathway in Down's syndrome is thought to be a compensatory mechanism to deal with increased hydrogen peroxide, allowing greater production of the reduced form of nicotinamide-adenine-dinucleotide phosphate, thus improving the ability of cells to reduce oxidized glutathione. However, it has been suggested that this shift of glucose utilization from energy production to reducing power may compromise cellular cation pumps.

Among the genes identified on chromosome 21 is that for the β-amyloid precursor protein. Amyloidosis is evident in the brain tissue of both patients with Alzheimer's disease and those with Down's syndrome. Studies are investigating the implications of the anomalies in the expression of the β-amyloid precursor protein and also the effect on cobalamin/folate metabolism of the gene for the enzyme cystathionine β-synthase, also on chromosome 21. The overexpression of both these genes is believed to contribute substantially to the development of dementia of the Alzheimer type. Although all people with Down's syndrome have evidence of brain pathology similar to Alzheimer's disease by their early thirties, not all show Alzheimer-like behavior changes as they age.

It may be that an increase in dietary antioxidants could delay the onset of Alzheimer-type symptoms, but more research is required. However, standard dietary recommendations for healthier lifestyles (i.e., eating more fruit and vegetables and including more oily fish in the diet) may have the added potential benefits of increasing antioxidant intake. Unfortunately, these are often the foods least favored by individuals with Down's syndrome.

Low vitamin E levels have been found to be associated with dementia, not only in the elderly but also in those with Down's syndrome. Vitamin E may have a potential therapeutic role in Alzheimer-like neurological changes by protecting the integrity of the muscarinc receptors. Continuing research into the etiology of the Down's syndrome phenotype is expected to lead to advances in the treatment of both Down's syndrome and Alzheimer's disease.

Care in the Community

Most people with Down's syndrome live in the community; some live with parents or caregivers, but adults often live independently or semi-independently. Many people with Down's syndrome can learn about healthy eating and manage their own diets. A dietitian's role in a community learning disability support team is likely to encompass not only individual assessment but also teaching and educating people with Down's syndrome as well as parents, caregivers, and other professionals.

See also: **Aging**. **Antioxidants**: Diet and Antioxidant Defense; Observational Studies; Intervention Studies. **Dental Disease**. **Fatty Acids**: Metabolism. **Growth and**

Development, Physiological Aspects. **Immunity**: Physiological Aspects. **Infants**: Nutritional Requirements. **Obesity**: Definition, Etiology and Assessment; Fat Distribution; Childhood Obesity; Complications; Prevention; Treatment. **Weight Management**: Approaches; Weight Maintenance; Weight Cycling. **Zinc**: Physiology.

Further Reading

Ani C, Grantham-McGregor S, and Muller D (2000) Nutritional supplementation in Down's syndrome: Theoretical considerations and current status. *Developmental Medicine and Child Neurology* **42**: 207–213.

Antila E, Westermarck T, Huovinen K, Lehto J, and Johansson E (1996) Indications for nutritional supplementation in Down's syndrome, 10th World Congress of the International Association for the Scientific Study of Intellectual Disabilities. *Trends in Biomedicine in Finland* **7**(supplement): 50–53.

Bell EJ, Haidonis J, and Townsend GC (2002) Tooth wear in children with Down's syndrome. *Australian Dental Journal* **47**(1): 30–35.

Chumlea WC and Cronk CE (1981) Overweight among children with trisomy 21. *Journal of Mental Deficiency Research* **25**: 275–280.

Clementi M, Calzolari E, Turolla L, Volpato S, and Tenconi R (1990) Neonatal growth patterns in a population of consecutively born Down syndrome children. *American Journal of Medical Genetics* **7**(supplement): 71–74.

Cremers MJG, van der Tweel I, Boersma B, Wit JM, and Zonderland M (1996) Growth curves of Dutch children with Down's syndrome. *Journal of Intellectual Disability Research* **40**(5): 412–420.

Cronk CE (1978) Growth of children with Down's syndrome: Birth to age three years. *Pediatrics* **61**: 564–568.

Cronk C, Crocker AC, Pueschel SM *et al.* (1988) Growth charts for children with Down's syndrome: 1 month to 18 years of age. *Pediatrics* **81**(1): 102–110.

Hennequin M, Faulks D, Veyrune JL, and Bourdiol P (1999) Significance of oral health in persons with Down syndrome: A literature review. *Developmental Medicine and Child Neurology* **41**: 275–283.

Hewitt P and Smith D (1983) Down's syndrome—Nutritional aspects. In: Macrea K, Robinson RK, and Sadler MJ (eds.) *Encyclopaedia of Food Science, Food Technology and Nutrition*, vol. 2, London: Academic Press.

Lejeune J (1990) Pathogenesis of mental deficiency in trisomy 21. *American Journal of Medical Genetics* **7**(supplement): 20–30.

Pipes PL (1988) Feeding management of children with Down syndrome. In: Dmitriev V and Oelwein P (eds.) *Advances in Down's Syndrome*. Seattle: Special Child Publications.

Piro E, Pennino C, Cammarata M *et al.* (1990) Growth charts of Down syndrome in Sicily: Evaluation of 382 children 0–14 years of age. *American Journal of Medical Genetics* **7**(supplement): 66–70.

Prasher VP (1995) Overweight and obesity amongst Down's syndrome adults. *Journal of Intellectual Disability Research* **39**(5): 437–441.

Pruess JB, Fewell RR, and Bennett FC (1989) Vitamin therapy and children with Down syndrome: A review of research. *Exceptional Children* **55**(4): 336–341.

Prader–Willi Syndrome

A O Scheimann, Johns Hopkins School of Medicine, Baltimore, MD, USA

Introduction

Prader–Willi syndrome (PWS) is a genetic disorder caused by deletion of the paternally derived genes in the proximal arm of chromosome 15. Patients with PWS manifest several common features, including hypotonia, decreased fetal movement, obesity, hyperphagia, short stature, growth hormone deficiency, hypogonadism, strabismus, and small hands and feet. Clinical features of PWS present ongoing medical and nutritional management issues.

Clinical History

Infants with PWS exhibit decreased fetal movement, weak cry, neonatal hypotonia, genital hypoplasia (cryptorchidism and clitoral hypoplasia), and failure to thrive (due to hypotonia and poor feeding). Toddlers with PWS acquire major motor milestones later than controls (walk at 24 months). Hyperphagia becomes evident between 18 months and 7 years of age. The majority of patients with PWS have growth hormone deficiency with short stature manifest during childhood and lack of a pubertal growth spurt. Individuals with PWS have an elevated pain threshold and vomiting threshold, with reports of delayed diagnoses of fractures, appendicitis, and gastroenteritis with significant morbidity. Obesity-related comorbidities, including sleep apnea, diabetes, and cor pulmonale, will shorten life expectancy without aggressive interventions. Behavioral problems, including obsessive–compulsive behavior (skin picking and rectal digging), stubbornness, and food foraging (including garbage and frozen food), are common; 5–10% of adults with PWS have features of psychosis.

Genetics of Prader–Willi Syndrome

Prader–Willi syndrome (PWS) is the first human disorder caused by altered imprinting. The incidence has been estimated between 1 in 8,000 to 1 in 25,000. During the process of imprinting, genes are differentially expressed based upon the parent of origin. Prader–Willi syndrome results from loss of imprinted genetic material localized to the paternal 15q11.2–13 region; loss of maternal genes in the same region results in Angelman's syndrome. The majority of cases are sporadic mutations. Nearly 70% of PWS patients have

deletions of the paternal 15q11.2–13 region; 28% of patients have maternal uniparental disomy. Approximately 1% of patients have mutations within the imprinting center, which has a higher recurrence risk unlike patients with deletions and uniparental disomy.

Fulfillment of diagnostic criteria and genetic testing confirm in individuals suspected with PWS. In 1993, age-stratified diagnostic criteria were published by Holm *et al*. PWS is very likely in children <3 years of age with 5 points (3 from major criteria) or in those >3 years of age with 8 points (4 from major criteria). Major diagnostic criteria for PWS (1 point for each) include infantile central hypotonia, feeding difficulties in infancy, accelerated weight gain in early childhood, hypgonadism, developmental delay and typical facial features (narrow bifrontal diameter, almond palpebral fissures, narrow nasal bridge, down-turned mouth). Current minor diagnostic criteria for PWS (1/2 point each) include decreased fetal movement, sleep apnea, short stature, hypopigmentation, small hands/feet, narrow hands with straight ulnar border, esotropia/myopia, thick saliva, skin picking and speech problems. Other commonly reported features of individuals with PWS include high pain threshold, decreased vomiting, temperature instability, premature adrenarche and osteoporosis.

In those suspected of having Prader–Willi syndrome, genetic testing should be pursued. Genetic testing for PWS includes chromosomal analysis and assessment for methylation patterns in the PWS region on chromosome 15. Flourescent in situ hybridization (FISH) is diagnostic in patients with deletions of the 15 q11.2–13 regions. Analysis for underlying uniparental disomy requires samples from both parents and the index case for DNA methylation patterns.

Nutritional Assessment

Nutritional monitoring in Prader–Willi syndrome requires attention to growth, body composition, and intake. Infants and children should be weighed on calibrated scales in minimal garments with heights obtained via a stadiometer. PWS-specific charts are available through the Prader–Willi Syndrome Association (www.pwsausa.org) to monitor linear growth.

Nutritional Management

Long-term management of patients with PWS presents unique evolving nutritional challenges. During infancy, muscle hypotonia impairs oral feeding and causes inadequate caloric intake and failure to thrive. The combination of altered body composition (with diminished metabolically active lean mass), growth hormone deficiency, and excessive intake results in obesity unless intake is restricted. Without aggressive nutritional interventions, the life expectancy is shortened for individuals with PWS due to obesity comorbidities.

Prior studies have reported metabolic differences among individuals with PWS. Adults with PWS have a low basal metabolic rate (BMR) dependent on the technique of body composition analysis. Despite differences in body composition, energy expenditure during physical activity is similar to that of controls but their overall activity level is less than that of controls. The combination of diminished BMR and activity level necessitates lower caloric intake to avoid significant obesity.

Dietary Interventions

Nutritional support for patients with PWS requires ongoing adaptation to meet age-specific needs. During infancy, hypotonia of the muscles associated with sucking limits the volume of caloric intake during feedings. Through the use of adaptive bottles/nipples, thickening agents (Thick-It and cereal), formula concentration, and short-term nasogastric tubes, infants with PWS can meet caloric requirements without placement of a gastrostomy. The feeding therapy utilized is determined by the adequacy of swallowing skills and nutritional status under the supervision of an oromotor therapist and nutritionist.

During infancy and early childhood, caloric intake should conform to the current guidelines from the Nutrition Committee of the American Academy of Pediatrics. During the first 6 months of life, breast milk or infant formulas are primary nutritional sources, followed by introduction of solids at 5 or 6 months of age. Solid textures are gradually advanced based on oromotor skills (jaw strength and tongue mobility). Due to the high likelihood for development of hyperphagia and obesity, the majority of parents avoid exposure of the PWS child to high-calorie solids, desserts, and juices. Via close nutritional follow-up during the first 2 years, oral intake can be appropriately adjusted to maintain weight for height between the 25th and 80th percentiles. Caloric restriction under the guidance of an experienced nutritionist is employed only if weight gain becomes excessive.

Nutritional strategies beyond the toddler years focus on avoidance of obesity. A number of studies have evaluated the caloric requirements for individuals with PWS. Weight maintenance has been reported with intakes of 8–11 kcal/cm/day (non-PWS children require 11–14 kcal/cm/day); weight loss has been documented with intakes of 7 kcal/cm/day. Proper implementation of caloric restrictions requires

attention to all potential sources of intake, including cafeterias, school buses, classroom activities ('life skills'), vending machines, neighbors, convenience stores, as well as home access (e.g., pantry, garbage cans, refrigerator, and tabletop).

Individuals with PWS and significant obesity-related comorbidities may require more aggressive weight loss interventions. To promote aggressive inpatient weight loss, a protein-sparing modified fast diet with micronutrient supplementation has been used over short time periods. Ongoing monitoring of food access is essential for long-term weight management.

Bariatric surgery causes weight loss through either a diminished capacity for intake or malabsorption. A long-term analysis of 10 non-PWS adolescents with a mean weight of 148 kg demonstrated a mean 5-year weight loss of >30 kg in 90% of patients; only 1 patient regained the weight. Bariatric surgery was initially attempted in PWS in the early 1970s. Gastroplasties were performed with the goal of decreasing PWS-related hyperphagic tendencies. More than half of PWS patients required subsequent revisions of the gastric pouch due to inadequate weight loss. The overall experiences with bariatric surgery in PWS are summarized in **Table 1**. The reported outcomes of bariatric

Table 1 Outcomes of bariatric surgery in Prader–Willi syndrome

Year	*Author*	*Type of surgery*	*No. of patients*	*Media age (years)*	*Median weight (kg)*	*Success rate*	*Complications*
1974	Soper	Gastroplasty	7	15	92.5	43%?	57% required revisions due to inadequate weight (wt) loss
1980	Anderson	91% gastric bypass; 9% gastroplasty	11	13	85		1 (9%) wound infection 54% required revision due to inadequate wt loss 1 dumping/diarrhea 1 death from uncontrolled wt gain
1981	Fonkalsrud	Vagotomy	1	17	120	?	29-kg initial wt loss followed by 20-kg gain
1983	Touquet	Jejunoileal bypass	1	24	181	62 kg (1 year)	Postoperative wound infection DVT/pulmonary embolus 4 or 5 stools/day
1991	Laurent-Jacard	Biliopancreatic diversion	3	27.6	84.5	Significant wt loss 1st year followed by wt gain ($2\frac{1}{2}$–6 years)	Diarrhea Vitamin D, vitamin B_{12}, folate, and iron deficiency
1992	Dousei	Vertical banded gastroplasty	1	21	57.4	Initial improved DM control	Short-term wt loss followed by break of staple line and wt gain
1997	Chelala	Laparoscopic adjustable gastric band	1	?	?		Death 45 days postoperatively from GI bleeding
2000	Grugni	Biliopancreatic diversion	1	24	80	Initial wt loss but wt gain without restriction	Diarrhea, severe osteopenia, anemia, hypoproteinemia
2001	Marinari	Biliopancreatic diversion	15	21	127	56–59% wt loss at 2–3 years; then regain 10–20% of wt lost	2 deaths from unrelated causes; no vitamin levels or bone density data provided

Adapted from Scheimann (2003) Management of nutrition issues in Prader–Willi syndrome. In *Management of Prader–Willi syndrome*, 3rd edn.

procedures have been less than satisfactory, with short-term weight loss followed by weight gain.

Micronutrients

Adequate vitamin and mineral supplementation is imperative for the patient with PWS. Hypocaloric diets required for patients with PWS preclude acquisition of adequate vitamin and minerals from traditional dietary sources. Commonly used meal plans for individuals with PWS are deficient in calcium, iron, vitamin D, vitamin E, biotin, pantothenic acid, magnesium, zinc, and copper.

See also: **Obesity**: Definition, Etiology and Assessment; Childhood Obesity; Treatment. **Weight Management**: Weight Maintenance.

Further Reading

Anderson AE, Soper RT, and Scott DH (1980) Gastric Bypass for Morbid Obesity in Children and Adolescents. *J Pediatr Surg* **15**: 876–881.

Bistrian BR, Blackburn GL, and Stanbury JB (1977) Metabolic Aspects of a Protein-sparing Modified Fast in the Dietary Management of Prader-willi Obesity. *NEJM* **296**: 774–9.

Chelala E, Cadiere GB, Favretti F, Himpens J, Vertruyen M, Bruyns J, Maroquin L, and Lise M (1977) Conversions and Complications in 185 Laparoscopic Adjustable Silicone Gastric Banding Cases. *Surg Endosc* **11**: 268–71.

Collier SB and Walker WA (1991) Parenteral Protien-sparing Modified Fast in an Obese Adolescent with Prader-willi Syndrome. *Nutr Rev* **49**: 235–8.

Dousei T, Miyata M, Izukura M, Harada T, Kitagawa T, and Matsuda H (1992) Long-term Follow-up to Gastroplasty in a Patient with Prader-willi Syndrome. *Obesity Surgery* **2**: 189–93.

Fonkalsrud EW and Gray G (1981) Vagotomy for Treatment of Obesity in Childhood Due to Prader-willi Syndrome. *J Pediatr Surg* **16**: 888–89.

Gavranich J and Selikowitz M (1989) A Survey of 22 Individuals with Prader-willi Syndrome in New South Wales. *Aust Paediatr J* **25**: 43–46.

Greenswag LR and Alexander RC (1995) *Management of Prader-willi Syndrome Second Edition.* Springer-Verlag.

Grugni G, Guzzaloni G, and Morabito F (2000) Failure of Biliopancreatic Diversion in Prader-willi Syndrome. *Obesity Surgery* **10**: 179–81.

Hill JO, Kaler M, Spetalnick B, Reed G, and Butler MC (1990) Reating Metabolic rate in Prader-willi Syndrome. *Dysmorphol Clin Genet* **4**: 27–32.

Pipes PL and Holm VA (1973) Weight Control of Children with Prader-willi Syndrome. *J Am Diet Assoc* **62**: 520–23.

Holm VA and Pipes PL (1976) Food and Children with Prader-willi Syndrome. *Am J Dis Child* **130**: 1063–7.

Laurent-Jacard A, Hofstetter J-R, Saegesser F, and Chapuis G (1991) Long-trem Result of Treatment of Prader-willi Syndrome by Scopinaro's Bilio-pancreatic Diversion. Study of Three Cases and the Effect of Dextrofenfluramine on the Postoperative Evolution. *Obesity Surgery* **1**: 83–87.

Marinari GM, Camerini G, Novelli GB, Papadia F, Murelli F, Marini P, Adami GF, and Scopinaro N (2001) Outcome of Biliopancreatic Diversion in Subject with Prader-willi Syndrome. *Obesity Surgery* **11**: 491–95.

Nardella MT, Sulzbacher SI, and Worthington-Roberts BS (1983) Activity Levels of Persons with Prader-willi Syndrome. *Am Jour Mental Deficiency* **87**: 498–505.

Schoeller DA, Levitsky LL, Bandini LG, Dietz WW, and Walczak A (1988) Energy Expenditure and Metabolism in Prader-willi Syndrome. *Metabolism* **37**: 115–20.

Soper RT, Mason EE, Printen KJ, and Zellweger H (1975) Gastric Bypass for Morbid Obesity in Children and Adolescents. *J Pediatr Surg* **10**: 51–58.

Stadler DD (1995) Nutritional Management in *Management of Prader–willi Syndrome: 2nd Edition.* Greenswag LR and Alexander RC (eds.) Springer.

Strauss RS, Bradley LJ, and Brolin RE (2001) Gastric Bypass Surgery in Adolescents with Morbid Obesity. *J Pediatr* **138**: 499–504.

Touquet VLR, Ward MWN, and Clark CG (1983) Obesity Surgery in a Patient with the Prader-willi Syndrome. *Br J Surg* **70**: 180–81.

Van Mil EA, Westerterp KR, Gerver WJ, Curfs LM, Schrander-Stumpel CT, Kester AD, and Saris WH (2000) Energy Expenditure at Rest and During Sleep in Children with Prader-willi Syndrome Is Explained by Body Composition. *Am J Clin Nutr* **71**: 752–6.

Cerebral Palsy

J Krick and P Miller, Kennedy–Krieger Institute, Baltimore, MD, USA

This article focuses on cerebral palsy (CP) and its nutritional implications. The first section defines CP and describes its causes, prevalence, and classification types. Associated deficits related to CP are also explored. The topic of nutritional assessment of children with CP includes discussions on growth, body composition, and energy, protein, fluid, and nutrient needs. Feeding and swallowing problems and the influence of muscle tone on the ability to eat safely are discussed in-depth, as are alternative feeding routes. The interdisciplinary approach is emphasized throughout as the ideal model to provide services to people with CP in order to ensure quality of life in the community.

Definition and Etiology

Cerebral palsy is a term that refers to a number of nonprogressive disorders of movement and posture that result from an injury to the central nervous system during early brain development (**Table 1**).

Table 1 Causes of cerebral palsy

Cause	*% of cases*
Perinatal	44
First trimester	
– Teratogens	
– Genetic syndromes	
– Chromosomal abnormalities	
– Brain malformations	
Second and third trimesters	
– Intrauterine infections	
– Problems in fetal/placental functioning	
Labor and delivery	19
Preeclampsia	
Complications of labor and delivery	
Perinatal	8
Sepsis/central nervous system infection	
Asphyxia	
Prematurity	
Childhood	5
Meningitis	
Traumatic brain injury	
Toxins	
Not obvious	24

Adapted from Hagberg B and Hagberg G (1984) Prenatal and perinatal risk factors in a survey of 681 Swedish cases. In: Stanley F and Alberman E (eds.) *The Epidemiology of the Cerebral Palsied*, pp. 116–134. Philadelphia: JB Lippincott.

Classification

There are several different classifications of CP. The three most predominant types are pyramidal, extrapyramidal, and mixed-type. The type of CP and the degree of involvement play an important part in nutritional assessment and treatment.

Pyramidal (spastic) cerebral palsy Children with spastic CP have increased muscle tone with a clasped-knife quality. In spastic quadriplegia (30% of cases of pyramidal CP), all four extremities are involved. In spastic diplegia (25%), both lower extremities are spastic with minimal upper extremity involvement. Hemiplegia (45%) implies involvement on only one side of the body, with the upper extremity usually more affected than the lower extremity.

Extrapyramidal cerebral palsy Choreoathetosis involves the presence of abrupt, involuntary movements of the upper and lower extremities. This condition can greatly increase energy expenditure and is further discussed in the energy needs section.

Mixed-type cerebral palsy Mixed-type CP includes characteristics of both the pyramidal and the extrapyramidal types. For example, a child may have rigidity in the upper extremities and spasticity in the lower extremities.

Associated Disabilities/Deficits

Associated deficits of CP are important to note since they impact on nutritional status. Cognitive impairments are quite common. Mental retardation occurs in 60% of CP cases, with the remainder at high risk for some type of learning disability. Sensory deficits are prevalent, including those in the visual and auditory modalities. Seizures occur in 20–30% of cases, with the highest proportion in the spastic type. Feeding, behavioral, or emotional problems are also frequently noted.

Nutritional Assessment

The goal for nutritional assessment and intervention is to have healthy, alert, interactive individuals who are able to take advantage of all that the environment has to offer. Each person must be able to participate to his or her capacity in the learning and therapeutic habilitative processes and in social, community, and leisure activities.

Growth

The literature describes children with CP who are shorter and lighter than the reference standard. This may be the result of several factors. Individuals with CP have alterations in muscle tone affecting their limbs and torso, depending on the level of severity and topography. They often exhibit muscle contractures, depending on the type of CP; muscle spasticity may retard bone growth. Limited physical activity may impede growth. Immobilization may be required after orthopedic surgery. Immobilization inhibits bone formation and longitudinal growth and results in suppression of certain growth-stimulating hormones. It has been suggested that dysregulation of growth hormone secretion may be another factor affecting growth.

A growth reference for children with spastic quadriplegia has been developed to facilitate uniformity in clinical appraisal as well as to simplify comparative interpretation of growth data. These growth curves can be seen in **Figures 1–6**. It is important to view the velocity of rate of growth from one measurement to another to aid clinical management. The rate of growth in children with CP is slower so that as they get older, the difference from the standard becomes greater.

Both nutritional and nonnutritional factors influence growth in children with CP. Nonnutritional

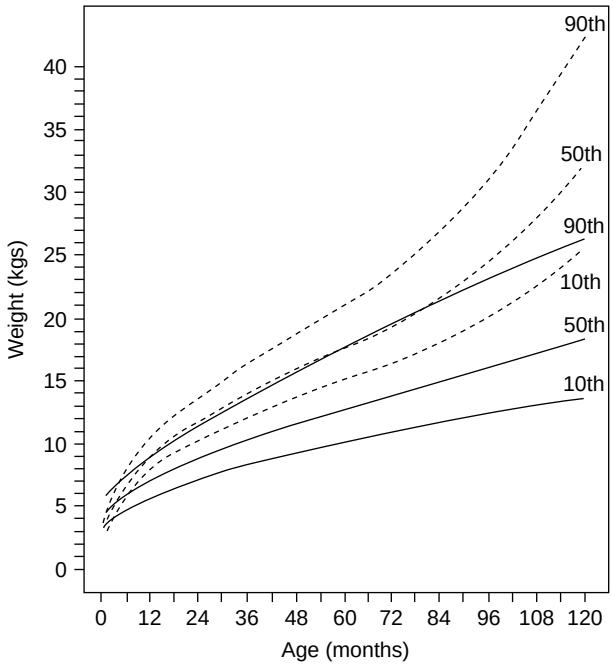


Figure 1 Weight for age for girls aged 0–120 months. The solid line represents girls with quadriplegic cerebral palsy, and the dotted line represents the National Center for Health Statistics standard curve for 10th, 50th, and 90th percentiles. (Reproduced with permission from Krick J, Murphy-Miller P, Zeger S, and Wright E (1996) Pattern of growth in children with cerebral palsy. *Journal of the American Dietetic Association* **96**: 680–685.)

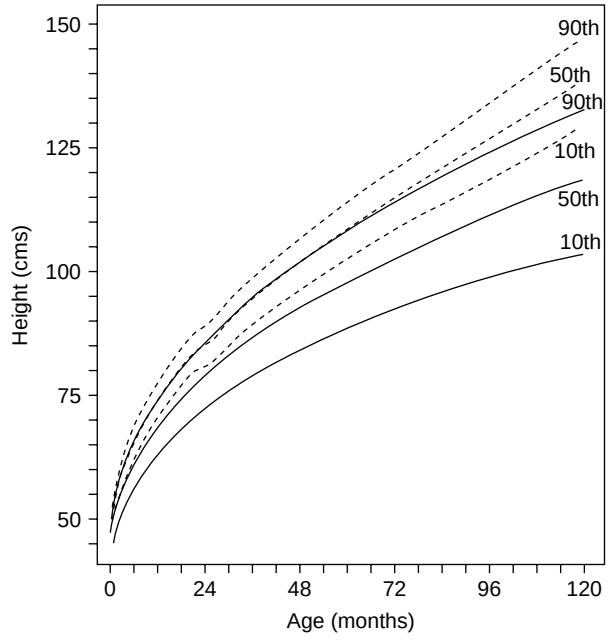


Figure 2 Length for age for girls aged 0 to 120 months. The solid line represents girls with quardriplegic cerebral palsy and the dotted line represents the National Center for Health Statistics standard curve for 10th, 50th and 90th percentiles.

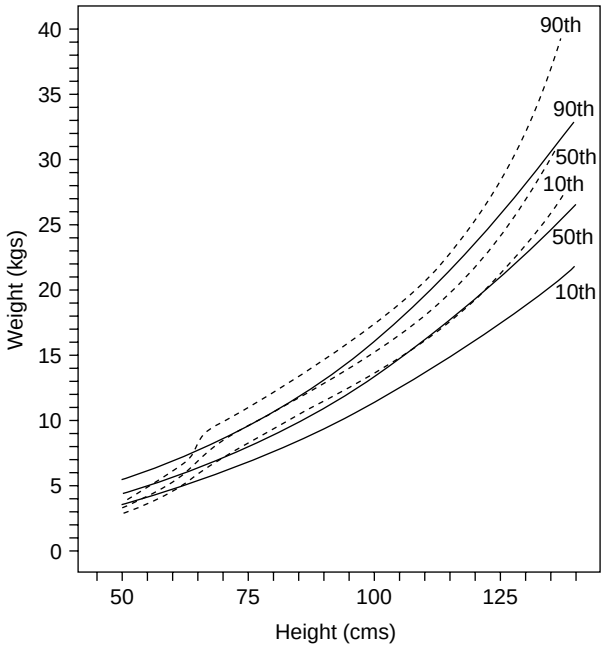


Figure 3 Weight for length for girls aged 0 to 120 months. The solid line represents girls with quadriplegic cerebral palsy and the dotted line represents the National Center for Health Statistics standard curve for 10th, 50th and 90th percentiles.

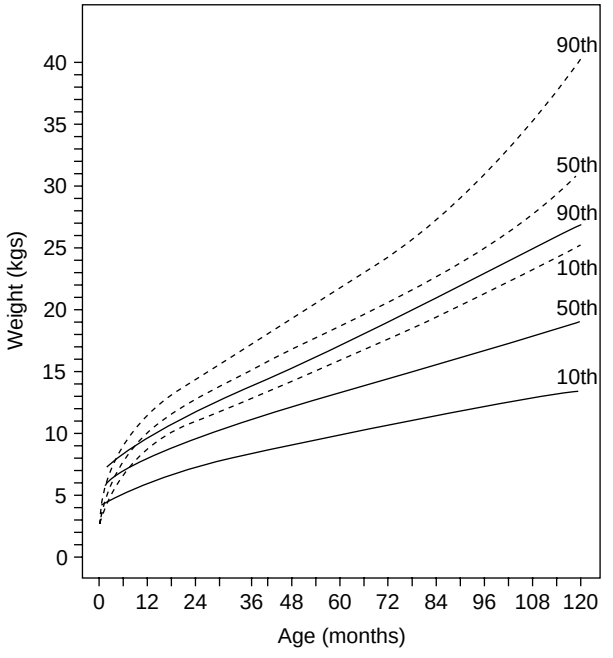


Figure 4 Weight for age for boys aged 0 to 120 months. The solid line represents boys with quadriplegic cerebral palsy and the dotted line represents the National Center for Health Statistics standard curve for 10th, 50th and 90th percentiles.

influences that have been suggested to impact growth include weight-bearing opportunities and, by extension, interventions using aggressive physical therapy, growth hormones, and electrical stimulation of muscle. In 1995, Stevenson reviewed growth in hemiplegics and noted that there is diminished

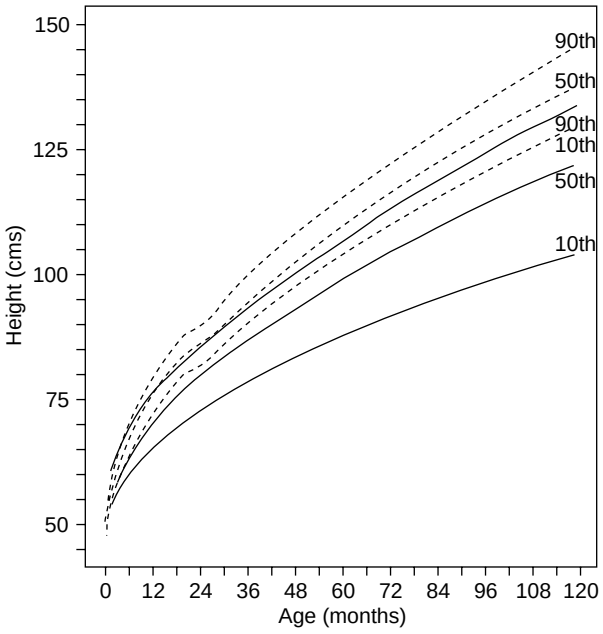


Figure 5 Length for age for boys aged 0 to 120 months. The solid line represents boys with quadriplegic cerebral palsy and the dotted line represents the National Center for Health Statistics standard curve for 10th, 50th and 90th percentiles.

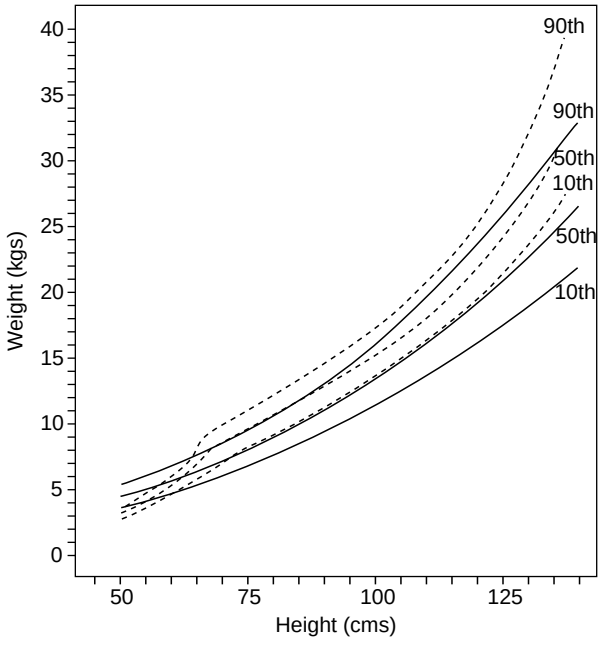


Figure 6 Weight for length for boys aged 0 to 120 months. The solid line represent boys with quadriplegic cerebral palsy and the dotted line represents the National Center for Health Statistics standard curve for 10th, 50th and 90th percentiles.

growth, decreased muscle mass, and decreased fat stores on the affected side, and that the magnitude of the differences increases with age and functional severity. Gender, age, cognitive impairment, and ambulatory status have also been noted to contribute to the slow growth seen in this population.

Measurement of length or height for individuals with CP may require techniques and standards using arm span, lower leg length, or segmental measurements because of the difficulties encountered with joint contractures and/or scoliosis. The use of height age, rather than chronological age, is a common technique and is defined as the age at which the child's height crosses the 50th percentile on the National Center for Health Statistics chart.

The use of z scores for length-for-age, weight-for-age, and weight-for-length promotes an accurate evaluation of discrete changes from one measurement date to another. Percentile tables describe ranges, and consequently detection of movement within the range is difficult to describe. The z score denotes standard deviation units from the median and allows the practicing clinician and investigator to pinpoint precisely any given measurement.

For screening purposes, conventional length/height and weight measures can be completed and compared to the Centers for Disease Control and Prevention growth charts. Reference standards for body mass index for children with CP do not exist; therefore, one must use body mass index data in conjunction with body composition data to determine adequacy of growth. Samson-Fang and Stevenson recommend using the TSF as a screening for identifying suboptimal fat stores in children with CP.

When trying to obtain growth measurements, joint contractures, muscle spasms, and poor cooperation will impact accuracy. Upper extremity (arm) length, tibial length, and knee height are often noted in the literature as valid proxies for length in children with CP up to the age of 18 years. (See **Table 2** for estimation of height using segmental measures.)

Researchers from the multicenter North American Growth in Cerebral Palsy Project suggest that a practical method to assess nutritional status in a child with CP is to measure body fat. This can be done in the form of either the triceps skin fold or

Table 2 Estimation of height from segmental measures

Age 0–12 years
$(4.35 \times \text{UAL}) + 21.8$
$(3.26 \times \text{TL}) + 30.8$
$(2.68 \times \text{KH}) + 24.2$
Age 6–18 years
White male $(2.22 \times \text{KH}) + 40.54$
Black male $(2.18 \times \text{KH}) + 39.60$
White female $(2.15 \times \text{KH}) + 43.21$
Black female $(2.02 \times \text{KH}) + 46.59$

UAL, upper arm length; TL, tibia length; KH, knee height.

both the TSF and subscapular skin folds. However, patient cooperation with the measuring techniques, required for accuracy and safety, may be difficult to obtain or maintain. For some individuals with CP, the process may be difficult, and training is needed to learn the technique for body fat measures and segmental measures mentioned previously.

Ideal Body Weight

The estimate of ideal body weight (IBW) is also in part determined by the severity of the CP. The IBW should be aimed at maintaining adequate fat and muscle stores to endure repeated surgeries or a common virus while facilitating daily physical care and management. Weight-for-length is an indicator of nutritional status, which obscures the issue of chronological age and addresses whether the individual is proportionate. IBW can be expressed as this ratio. Those with cerebral palsy should attain and maintain an IBW that takes into account their age, level of physical ability, and their independence. Measurement of arm anthropometry will provide a description of body composition and support clinical judgments related to IBW. For example, children with spastic quadriplegia are the most dependent and the 10th percentile weight-for-length would be designated as the IBW. However, this assignment is done in tandem with assessment and monitoring of body composition, and if either the arm fat or the arm muscle area were less than the 5th percentile, then the IBW would be adjusted upward.

Body Composition

Since the 1970s, researchers reviewing body composition have noted reduced lean body mass in children with CP. Recent work examining adults with CP and their age-matched controls found no difference in lean body mass or percentage of body fat.

Bone Mineral Density

Bone mineral density (BMD) is markedly reduced in nonambulatory children with CP, placing them at risk for nontraumatic fractures. Osteopenia defined as <2 standard deviations below the mean was found in the femur of most nonambulatory children by the age of 10 years. Decreased BMD results from a combination of factors, including immobilization, antiepileptic therapy, and nutritional deficiencies. Serum levels of calcium, phosphate, alkaline phosphatase, and osteocalcin were not found to be reliable indicators of low BMD when studied by Henderson. The same author noted that fracture rate is fourfold higher following spica casting and more than threefold higher following an initial fracture.

Many nonambulatory children require and are given less calories than recommended for their non-CP counterparts; therefore, the clinician is obliged to review the adequacy of the micronutrients, specifically calcium. Most likely, their diets will require supplementation to meet 100% of the DRI standards for age and gender.

Methods to increase BMD include weight-bearing activities, dietary adequacy, and the use of bisphosphonates. In several studies, bisphosphonate use has demonstrated increased bone density by 20–89% with no obvious adverse effects.

Energy Needs

Equations that are frequently used to predict energy requirements were developed using healthy children and adults in usual environmental and physical activity conditions and do not provide an accurate assessment of the needs of those with CP. From a nutrition perspective, wide-ranging studies demonstrate underreporting of energy needs on food records, which at best provide a qualitative measure of intake. Therefore, clinicians have turned to the use of more sophisticated technology, such as doubly labeled water and indirect calorimetry, to assess the energy needs of this population. Additionally, the energy cost of movement, whether it be wheelchair propulsion, crutch ambulation, or the involuntary movements of the individual with athetosis, must be considered. Those with CP may undergo repeated orthopedic surgery that insults nutritional status, resulting in increased nutrient and energy demands. It has also been hypothesized that whole body metabolic rate may be related to differences in skeletal muscle fiber proportions and/or differences in enzymatic activity. People with CP have abnormal variation in the size of muscle fibers and altered distribution of fiber types.

Altered energy needs are common among those with CP and differ widely from the norm. Clinicians use a variety of approaches to estimate energy needs, such as the Dietary Reference Intakes (DRIs) for chronological age, the RDAs for height age, and the World Health Organization equation. When estimating energy needs, information related to muscle tone, activity level, and needs for growth or catchup growth must be added to the estimate for resting energy expenditure (REE).

One equation designed specifically for this population is:

$$\text{REE} \times \text{muscle tone factor} \times \text{activity factor} + \text{growth factor(s)} = \text{kcal per day}$$

The REE can be determined using indirect calorimetry or can be derived from estimating body surface

area standard metabolic rate 24 h. Body surface area (m^2) is calculated from length and weight using the nomogram derived from the formula of DuBois and DuBois, and the standard metabolic rate (kcal/m^2/h) is identified using height age and sex applying Fleisch data. The modifying factors are applied as follows:

- Muscle tone factors: Multiply by 10% for high tone (hypertonicity) and decrease by 10% for low tone (hypotonicity); no adjustment for normal tone.
- Activity factors: Multiply by 15% for bedridden state, 20% for wheelchair, and 30% for ambulation.
- Growth factors: Add 5 kcal (20.92 kJ) per gram of desired growth, expected growth, and catchup.

Energy needs must be viewed on an individual basis assimilating the concepts noted previously. The use of any approach is regarded as a guidepost and requires careful monitoring of body weight. Modifications to the diet should be based on clinical observation and measurement. There is a subset of individuals with CP who require significantly less kilocalories than anticipated (as few as perhaps 15 kcal/kg). Care should be exercised to provide adequate nutrients, protein/kg, and fluid despite the very low calorie needs.

Nutrient and Fluid Needs

Nutrient and protein needs are based on DRIs similar to those of the population without CP. Height age is often used in these determinations.

Fluid needs are based on body size rather than calorie intake. **Table 3** demonstrates how to calculate fluid needs. Constipation is a chronic problem for most children with CP and is related to muscle tone, loss of sensation, limited physical activity, medication side effects, and inadequate dietary fiber and/or fluid intake. Oral motor dysfunction results in diminished intake as well as in food and fluid loss. Modified food and fluid textures result in less free water and fiber in the diet. Discomfort associated with constipation may decrease appetite and increase gastroesophageal reflux. Dietary intervention may therefore be limited and medical management may be necessary.

Table 3 Fluid needs based on body weight[a]

Body weight (kg)	*Fluid need ($cm^3\,kg^{-1}$)*
≤10	100
11–20	+50
≥21	+25

[a]Suggest monitoring urine-specific gravities when available and quantity, color, and odor of urine, and adjust for periods of stress and temperature. Example: 28-kg child
100 ml × 10 kg = 1000 ml
50 ml × 10 kg = 500 ml
25 ml × 8 kg = 200 ml
Total need = 1700 ml

Assessment of Feeding Skills and Safety

Eating skills are acquired in a sequential pattern so that a developmental history will be helpful in evaluating current function and planning treatment options. Factors affecting feeding performance are shown in **Table 4**.

Oral Motor Evaluation

Feeding and swallowing problems are common in the child with CP, depending on the type of muscle tone, the presence of primitive reflexes, movement patterns, and the integrity of the sensory system. Clinical indicators of feeding and swallowing dysfunction are shown in **Table 5**. Problems often include poor intake, inefficient and lengthy mealtimes, abnormal oral motor patterns, inappropriate progression of feeding skills, and/or physiological compromise with feeding. Sensory, cognitive, and language deficits may also complicate the feeding process. An interdisciplinary team evaluation is

Table 4 General factors affecting feeding performance

Neuromotor performance	Constipation
Perceptual deficits	Amount of physical and verbal assistance required
Cognition and communication skills	Physiological support
Vision and hearing	Oral motor skills and swallowing status
Behavior/interaction	Medications
Growth	Dental and gum disease
Dietary adequacy	Multiple orthopedic procedures
GER and other gastrointestinal-related issues	Family/psychosocial stressors

Table 5 Clinical indicators of feeding and swallowing dysfunction

Congestion	Difficulty managing secretions
Noisy 'wet' sounds	History of upper respiratory infections
Multiple swallows to clear bolus	Apnea during feeding
Unexplained fevers, unexplained irritability	Failure to thrive, failure to maintain weight
Coughing/choking/gagging before, during, or after swallow	
Food refusal	

essential for the assessment, development of appropriate goals, and facilitation of a treatment plan that respects the developmental progression. A clinical assessment of the feeding process should include observance of facial muscle tone, oral reflex activity, functional oral motor skills, structural abnormalities, sensory responses, behavior and interaction during feeding, respiratory and phonatory status, and posture and positioning.

Radiographic and ultrasound studies can provide more detailed information about the oral structures and the competency of the oral, pharyngeal, and esophageal phases, including the detection of aspiration. Cervical auscultation can also be helpful in evaluating the pharyngeal phase of swallowing. In addition, these techniques can assist in determining the suitable solid and liquid texture and appropriate head and neck positioning. Hypertonicity leads to abnormal movements of the tongue, lip, and jaw. These abnormal movements can be manifested as tongue retraction, tongue tip elevation, tongue thrust, tonic biting, jaw thrust, jaw instability, lip retraction, and lip/cheek instability. An abnormally strong gag reflex, tactile hypersensitivity in the oral area, and drooling can also complicate feeding. Individuals with CP are also at risk for dental problems due to poor oral hygiene, teeth grinding, hypersensitivity in the oral area, and hyperplasia of the gums from long-term use of phenytoin, a medicine commonly prescribed for seizure management.

Aspiration and Gastroesophageal Reflux

Clinical signs of aspiration may include coughing, choking, gagging, inability to handle oral secretions, wet upper airway sounds with poor vocal quality, apnea, food refusal, frequent upper respiratory infections, and aspiration pneumonia. Aspiration of food may occur without physical evidence if the protective cough or gag is not functioning, sensory deficits exist, and/or the swallowing mechanism is dysfunctional. This results in what is termed silent aspiration. Although aspiration from solid food can be detected, the possibility of aspiration from gastroesophageal reflux (GER) may also need to be considered. The regurgitation of gastric contents from the stomach into the esophagus can lead to irritability during or after feeding, arching, esophagitis, and ultimately food refusal. Other symptoms of GER include respiratory compromise, apnea, and drooling. Treatment for GER includes the use of antacids, H2 blockers, medications to increase gut motility, reduction in feeding rate, positioning, thickening of foods or liquids, or surgical intervention. Small, frequent feedings help to decrease the volume in the stomach at one time.

Fatigue may occur in the child who is not able to sustain the work involved with feeding and may be expressed by an increase in respiratory rate, diaphoresis, or increased work of breathing. The causes may be muscular, respiratory, or cardiac, and they may increase the risk of aspiration or hypoxia. The work required to eat a meal is accomplished at a higher physiological cost to the child, thereby increasing caloric needs.

Muscle Tone and Positioning

It is important to understand the influences of muscle tone and proper positioning on the ability to eat safely and efficiently in this population. Increased or decreased muscle tone contributes to difficulty preserving a patent airway, compromised self-feeding skills, poor rib cage expansion and esophageal motility, and difficulty in maintaining a stable supported base for seating. Fluctuating muscle tone leads to involuntary movements and limited postural stability. Despite the type of muscle tone, optimal positioning is crucial for feeding and swallowing. The proper feeding position includes neutral alignment of head and neck, midline orientation, symmetrical trunk position, 90° pelvic/femoral alignment, and symmetrical arm position with neutral shoulders. An example of proper positioning can be seen in **Figure 7**. Consultations with orthopedists and/or rehabilitation physicians to address current and potential musculoskeletal problems, physical and occupational therapists for functional assessment, orthotists for deformity management, and durable medical equipment specialists to customize standard wheelchair components are valuable.

Underweight and Overweight

Overweight Most children with CP who are overweight or obese have low muscle tone. Their nutritional status impacts sleeping and breathing patterns, mobility, physical care, and peer relationships. It is difficult to attain an ideal body weight because energy needs are significantly reduced and the options for exercise are limited.

Underweight Typically, children with athetosis struggle to maintain weight given their excessive involuntary movements, which significantly increase energy needs. As these children age, the problem becomes more apparent, and many of these children will require enteral supplementation. One evaluation of this population noted that the basal energy requirement was 40% higher than expected.

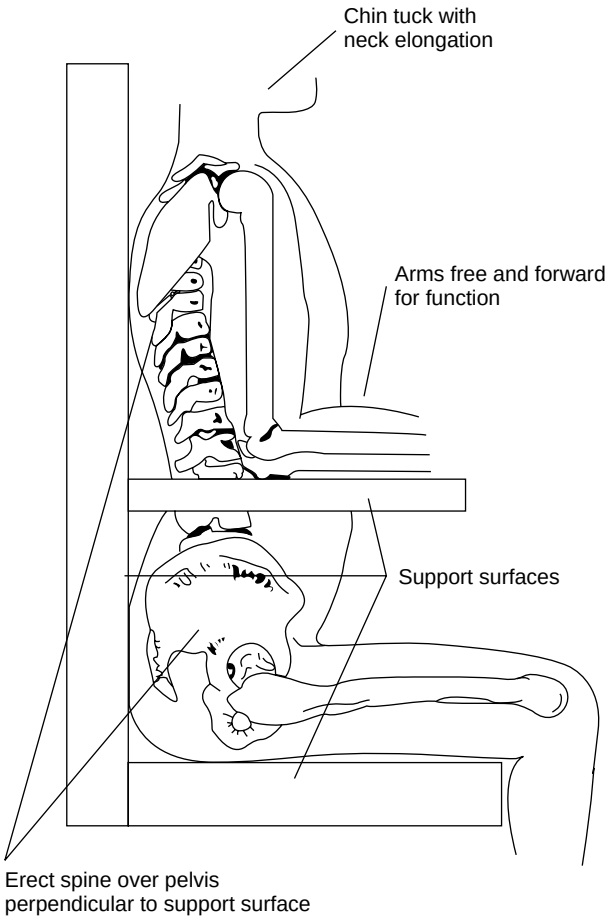


Figure 7 Proper seating position. (Reproduced with permission from: *The Handbook of Assistive Technology*. Singular Publishing Group, Inc.)

Superior Mesenteric Artery Syndrome

Superior mesenteric artery (SMA) syndrome is a condition in which the third portion of the duodenum is intermittently compressed by the overlying SMA, resulting in gastrointestinal obstruction. Symptoms include recurrent vomiting, abdominal distension, weight loss, and postprandial distress. People with CP are at high risk for several of the reported causes of SMA syndrome, including body cast compression, severe weight loss, prolonged supine positioning, and scoliosis surgery. Consequently, it is important to recognize the symptoms and know the appropriate treatments for this syndrome. Most people can be treated nonsurgically with gastric aspiration and nasojejunal or gastrojejunal feedings distal to the obstruction. One study also found that turning to the left from a supine position displaces the SMA from the right to the left side of the aorta in scoliosis cases. Thus, positioning can help alleviate symptoms and special considerations may be indicated in light of the limitations imposed by the CP.

Behaviors at Mealtimes

Parent–child interactions can also influence feedings. Ineffective communication, lack of bonding, the absence of social interaction or poor interactive skills, family dysfunction, and decreased environmental stimuli can exacerbate feeding difficulties or lead to frustration and anxiety with subsequent food refusal or parental withdrawal. Aversion to oral feeds can also be an outcome of medical complications, such as esophagitis and GER, or lack of feeding experience at critical milestones secondary to prolonged tube feedings. Behavioral treatment should only be undertaken after thorough medical, nutritional, and neurodevelopmental assessments are completed.

Feeding Issues

The feeding plan should be safe, promote growth or weight maintenance without excessive energy expenditure in order to obtain the required calories, and meet the needs of the family. It should reflect their resources in time and skill, and it should address their concerns and expectations. The goals for treatment once feeding and swallowing problems are identified are to prevent aspiration and thereby respiratory compromise; provide adequate calories, protein, vitamins, minerals, and fluid; and educate caregivers regarding nutritional requirements.

Oral Motor Considerations

Management strategies for daily mealtime feeding include positioning, modification of the sensory properties of the food, oral motor facilitation techniques, and equipment adaptations. For individuals with increased energy needs, the nutrient density of their meals may need to be maximized. **Table 6** lists

Table 6 Calorie boosters

Instant breakfast	Margarine, butter, oils, gravy
Powdered, evaporated milk	Sugar, honey, syrup
Whole milk cheeses	Cream cheese
Peanut butter	Sour cream
Wheat germ	Concentrate juices
Yogurt, pudding, custards	Breading or cracker meal
Milkshakes, eggnog	Fruit canned in heavy syrup
Supplements such as Polycose, Promod, Microlipid, Pediasure, and Ensure	

commonly used calorie boosters. It is important to acknowledge the inability to change the underlying feeding problem while providing a method of circumventing the problem to allow adequate nutrition and growth. For example, facilitative techniques to minimize excessive jaw movement may entail the feeder providing physical jaw control/support; a change in the food consistency, texture, temperature, or taste to improve the ability to propel a bolus through the oropharynx; the careful selection of adaptive feeding equipment to assist with self-feeding and/or increased intake; and an appropriate seating system. Proper positioning also allows the feeder use of both hands.

Alternative Feeding Routes

Many children with CP are not able to meet some or all of their calorie needs by mouth due to one or more of the following conditions: oral motor dysfunction, excessive energy needs, recurrent infections, illnesses, and orthopedic surgical interventions. Consequently, if the gastrointestinal tract is functioning, supplemental or total tube feedings may be indicated. Early intervention with enteral nutrition may prevent protein–energy malnutrition and its complications. Studies have shown improvements in weight gain (fat mass as opposed to fat-free mass) with supplemental tube feedings, which better enables individuals to endure short-term medical insults.

Enteral nutrition may be delivered by nasogastric, nasojejunal, gastrostomy, gastrostomy–jejunal, and jejunostomy tubes. The degree of GER and risk of aspiration determine where the tube is placed, whereas the length of time needed for tube feedings determines whether a nasoenteral or surgically placed tube is required. The decision regarding continuous, intermittent, or combination tube feeds is dependent on the individual needs of the patient.

Tube feedings should be considered a tool to improve nutritional status rather than failure of the child's ability to eat. Based on the medical diagnosis and developmental stage of the child, the prognosis for return to oral feeding varies, and the length of time to achieve this goal is extremely variable. For some children, the goal of returning to full or partial oral feeding is not realistic. In a study evaluating the health of children with CP, Liptak describes those who were tube fed as having the lowest mental age, requiring the most health care resources, using the most medications, and having the most respiratory problems. These children were characterized as especially frail and required numerous health-related resources and treatments. Oral motor therapy should focus on maintaining existing oral motor skills, encouraging pleasurable oral experiences, and tolerance of oral hygiene practices. Nonnutritive oral stimulation must be performed when tube feedings are employed as the route of nutrition. The benefits are listed in **Table 7**. Improvement in nutritional status can result in positive changes in oral feeding.

Table 7 Benefits of nonnutritive oral stimulation

Maintains oral sensation and tolerance
Facilitates saliva production, swallowing, and other oral motor patterns
Maintains or develops coordination of respiration and swallowing
Facilitates parent–child interactions

Parenteral nutrition should only be used when the gastrointestinal tract is dysfunctional. When initiating feedings in patients with major weight loss or failure to thrive, whether enteral or parenteral nutrition is used, it is important to be aware of the 'refeeding syndrome.' This syndrome refers to phosphorus depletion and alterations in potassium, magnesium, and glucose metabolism, resulting in severe metabolic and physiological complications. It is imperative to increase calorie delivery slowly with close laboratory monitoring.

Medications

Drug–nutrient interactions should be considered for all children receiving long-term medications for seizure disorders, alterations in muscle tone, attentional deficits, gastrointestinal disorders, and/or other chronic conditions. One drug or the combination of multiple drugs may affect nutrition in many ways, such as causing decreased appetite, interference with absorption of specific nutrients, nausea, and vomiting.

Medication treatment options offer challenges to nutrition. For instance, diazepam, often used to decrease spasticity, increases the potential for drooling. This raises concerns of fluid loss/balance as well as loss of the protective effect of saliva on esophageal mucosa. Additionally, attention must be paid to tone reduction in the trunk and oral structures that would compromise safety of feeding skills.

Tone-lowering drugs potentially reduce energy expenditure and, as a result, require increased vigilance to avert excessive weight gain. Anecdotally, as tone is significantly reduced in children for whom intrathecal Baclofen pumps are used, so is the energy requirement. Most of these children seen have been on tube feedings with a constant intake over time.

With the use of the Baclofen pump, weight gain is seen and adjustments in the kilocaloric level may be necessary.

Repeated Orthopedic Surgeries

These are common in children with CP, and each surgery must be preceded by an evaluation of nutritional status and assessment of the child's ability to physically heal and recover quickly from the trauma. Many children who are marginal oral feeders will decompensate, lose weight, and have a difficult time healing because of a cascade of events including pain, poor positioning for safe feeding, worsening constipation, minimal intake, lethargy, and increased medications for pain that may have a sedative effect. They may require supplemental feedings prior to surgery or during the postoperative period.

Coordinated Services

The provision of nutrition services and prevention of further disabling conditions can be done in a variety of health care, school, vocational, home, and community settings. It is the responsibility of the family in concert with the health care team to promote nutrition care planning in these settings. More than 90% of children with CP live to adulthood; however, their life expectancy is less than that of the general population. The chronicity of nutrition problems for individuals with CP is recognized and has in part created a need for care coordination and integrated service planning to provide meaningful and cost-effective services.

See also: **Energy Expenditure**: Indirect Calorimetry; Doubly Labeled Water. **Nutritional Support**: Adults, Enteral.

Further Reading

Capute A and Acardo PJ (eds.) (1996) *Developmental Disabilities in Infancy and Childhood*, 2nd edn., vol. 2. Baltimore: Paul H Brookes.

Case-Smith J (ed.) (1993) *Pediatric Occupational Therapy and Early Intervention*. Stoneham, UK: Butterworth-Heinemann.

Cherney L (1994) *Clinical Management of Dysphagia in Adults and Children*. Gaithersburg, MD: Aspen.

Eicher PS and Batshaw ML (1993) Cerebral palsy. *Pediatric Clinics of North America* **40**: 537–551.

Ekvall SW (ed.) (1993) *Pediatric Nutrition in Chronic Diseases and Developmental Disorders: Prevention, Assessment and Treatment*. New York: Oxford University Press.

Henderson CR, Lark KR, Gurka JM *et al.* (2002) Bone density and metabolism in children and adolescents with moderate to severe cerebral palsy. *Pediatrics* **110**(1).

Hogan SE (1999) Knee height as a predictor of recumbent length for individuals with mobility-impaired cerebral palsy. *Journal of the American College of Nutrition* **18**(2): 201–205.

Klein M and Delaney T (1994) *Feeding and Nutrition for the Child with Special Needs*. Tucson, AZ: Therapy Skill Builders.

Krick J, Murphy-Miller P, Zeger S, and Wright E (1996) Pattern of growth in children with cerebral palsy. *Journal of the American Dietetic Association* **96**: 680–685.

Liptak GS *et al.* (2001) Health status of children with moderate to severe CP. *Developmental Medicine and Child Neurology* **43**: 364–370.

Samson-Fang LJ and Stevenson RD (1998) Linear growth velocity in children with cerebral palsy. *Developmental Medicine and Child Neurology* **40**(10): 689–692.

Samson-Fang LJ and Stevenson RD (2000) Identification of malnutrition in children with cerebral palsy: Poor performance of weight-for-height centiles. *Developmental Medicine and Child Neurology* **42**(3): 162–168.

Stallings VA, Charney EB, Davies JC, and Cronk CE (1993) Nutrition related growth failure in children with quadriplegic cerebral palsy. *Developmental Medicine and Child Neurology* **35**: 126–138.

Stevenson RD (1995) Use of segmental measures to estimate stature in children with cerebral palsy. *Archives of Pediatrics & Adolescent Medicine* **149**(6): 658–662.

Heart Disease *see* **Coronary Heart Disease**: Hemostatic Factors; Lipid Theory; Prevention

Height *see* **Nutritional Assessment**: Anthropometry

HOMOCYSTEINE

J W Miller, UC Davis Medical Center, Sacramento, CA, USA

Introduction

Homocysteine is a sulfur amino acid and an intermediate in the biochemical conversion of methionine to cysteine, a process called trans-sulfuration. Vincent Du Vigneaud and others elucidated the biochemistry of homocysteine over the period from the 1930s to the 1950s. In the early 1960s, the description and characterization of the inborn error of metabolism, homocystinuria, initiated a 40-year (and continuing) period of investigation that has revealed homocysteine as an independent risk factor for vascular disease. The association between elevated blood levels of homocysteine (hyperhomocysteinemia) and vascular disease may be similar in magnitude to the association between cholesterol and vascular disease, thus implicating hyperhomocysteinemia as a significant public health concern. Currently, large-scale intervention trials are being conducted to determine if supplements of the B vitamins folate, vitamin B_{12}, and vitamin B_6, each of which plays an integral role in homocysteine metabolism, reduce the incidence of vascular disease. If successful, B vitamin supplements may prove to be an inexpensive and safe prophylactic to reduce the risk of heart attacks and strokes.

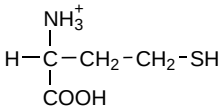
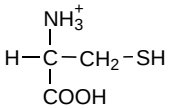
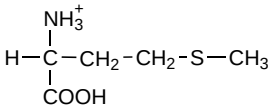
Structure and Forms

The structure of homocysteine is shown in **Table 1** along with the related structures of cysteine and methionine. The most prominent features of homocysteine and cysteine are the free sulfhydryl groups located at the end of the side-chains of both amino acids. These sulfhydryl groups are highly susceptible to oxidation and the formation of disulfide linkages with other sulfhydryl compounds. The primary forms of homocysteine found in the blood (**Table 1**) consist of homocysteine in disulfide linkage with: (1) cysteine residues within the primary sequences of albumin and other plasma proteins (protein-bound); (2) free cysteines or cysteine-containing peptides (mixed disulfides); and (3) other homocysteine molecules (homocystine). Only about 1% of homocysteine in the blood is in the free-reduced form. Methionine, in contrast, does not have a free sulfhydryl group, and thus does not form disulfide compounds.

Biosynthesis and Metabolism

The biosynthesis and metabolism of homocysteine is presented in **Figure 1**. The ultimate source of homocysteine is dietary methionine. Methionine is first activated by addition of an adenosyl group (from ATP) to form *S*-adenosylmethionine (SAM). SAM is an important intermediate known as the universal methyl donor for its role as the methylating agent in a variety of essential reactions, including those involving DNA, RNA, proteins, membrane phospholipids,

Table 1 Structures and forms of homocysteine and related amino acids

Structures of homocysteine and related amino acids		*Forms of homocysteine found in blood*	
	Homocysteine	HCY-S—S-CYS-albumin	Protein-bound
	Cysteine	HCY-S—S-CYS	Mixed disulfide
		HCY-S—S-HCY	Homocystine
	Methionine	HCY-SH	Free reduced

HCY, homocysteine; CYS, cysteine.

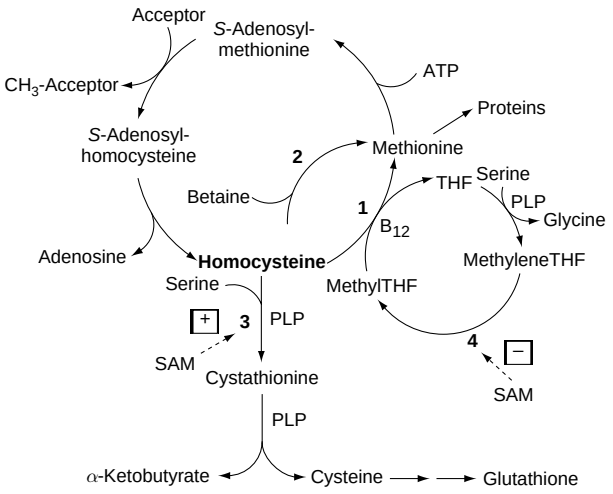


Figure 1 The biosynthesis and metabolism of homocysteine. Reactions that are regulated by *S*-adenosylmethionine (SAM) are indicated by positive and negative signs. Key enzymes: (1) methyltetrahydrofolate-homocysteine methyltransferase or methionine synthase; (2) betaine-homocysteine methyltransferase; (3) cystathionine β-synthase; (4) methylenetetrahydrofolate reductase. Abbreviations: THF, tetrahydrofolate; PLP, pyridoxal-5′-phosphate (vitamin B_6).

neurotransmitters, and the synthesis of creatine. A product of all SAM-dependent methylation reactions is *S*-adenosylhomocysteine (SAH), which in turn is metabolized to form adenosine and homocysteine. Homocysteine is then at a metabolic crossroad: it can be remethylated to form methionine or catabolized through cystathionine synthesis.

In remethylation, homocysteine reacquires a methyl group in a reaction catalyzed by methionine synthase (5-methyltetrahydrofolate-homocysteine methyltransferase) (EC 2.1.1.13) with methyltetrahydrofolate serving as the methyl donor and vitamin B_{12} serving as a cofactor. This reaction occurs in all mammalian cells. Alternatively, homocysteine can be remethylated in a folate- and vitamin B_{12}-independent reaction utilizing betaine as the methyl donor and catalyzed by betaine-homocysteine methyltransferase (EC 2.1.1.5). This reaction occurs primarily in the liver, and to a lesser extent in the kidney and possibly the brain.

Homocysteine catabolism occurs through cystathionine synthesis in a condensation reaction with serine. This reaction is catalyzed by cystathionine β-synthase (EC 4.2.1.22), which requires vitamin B_6 in the form of pyridoxal-5′-phosphate (PLP) as a cofactor. Cystathionine is then cleaved to form α-ketobutyrate and cysteine in a second PLP-dependent reaction catalyzed by cystathionase (EC 4.4.1.1). Further metabolism of cysteine leads to the formation of glutathione or inorganic sulfate.

Regulation of Metabolism

An important aspect of homocysteine metabolism is that it is subject to allosteric control. In addition to serving as the universal methyl donor, SAM also is an activator of cystathionine β-synthase and an inhibitor of methylenetetrahydrofolate reductase (MTHFR) (EC 1.7.99.5), the enzyme responsible for the synthesis of methyltetrahydrofolate (**Figure 1**). These allosteric functions serve to control whether homocysteine is recycled to form methionine or catabolized to form cystathionine. When dietary supply of methionine is high, i.e., after a protein meal, intracellular SAM levels increase. The high concentration of SAM activates cystathionine β-synthase and inhibits MTHFR, thus promoting homocysteine catabolism and limiting homocysteine remethylation. This serves to reduce the recycling of homocysteine when there is an adequate dietary supply of methionine. Conversely, under fasting conditions when there is no dietary influx of methionine, intracellular SAM levels go down. Cystathionine β-synthase is then not activated and the inhibition of MTHFR is relieved, thus promoting homocysteine remethylation over catabolism. Consequently, this maintains intracellular methionine levels during times of limited dietary supply.

An additional level of control of homocysteine metabolism is exerted by oxidative stress. Though the biochemical mechanism remains undefined, oxidative stress tends to divert homocysteine toward cystathionine synthesis away from methionine synthesis. This serves to increase synthesis of glutathione, a product of homocysteine metabolism through the trans-sulfuration pathway and an important intracellular antioxidant.

As discussed below, alterations in homocysteine metabolism also occur after menopause, in diabetes, and in hypothyroidism. These observations suggest that hormones, including estrogen, insulin, thyroxine, and thyroid-stimulating hormone, may directly or indirectly affect homocysteine metabolism. As for oxidative stress, the mechanisms by which these hormones affect homocysteine metabolism are poorly understood.

Hyperhomocysteinemia

Under conditions of maximal metabolic efficiency, plasma levels of homocysteine range from 4 to $10\,\mu mol\,l^{-1}$. Metabolic blocks in homocysteine metabolism lead to accumulation of intracellular homocysteine with subsequent export into the blood. Depending on the magnitude of the metabolic impairment, plasma homocysteine levels can rise to varying degrees, as defined in **Table 2**.

Table 2 Degrees of hyperhomocysteinemia

Total plasma homocysteine	*Designation*
4–10 μmol l^{-1}	Normal
11–25 μmol l^{-1}	Mild to moderate
26–50 μmol l^{-1}	Intermediate
>50 μmol l^{-1}	Severe

Genetic Defects

Severe elevations in plasma homocysteine (concentrations as high as several hundred μmol l^{-1}) are observed in individuals with homozygous genetic defects affecting cystathionine β-synthase, MTHFR, or any of several enzymes responsible for the conversion of vitamin B_{12} to its methionine synthase-associated cofactor form. These autosomal recessive genetic disorders, collectively termed homocystinuria because homocysteine accumulates in the urine as well as the blood, are associated with severe premature vascular disease, including thrombosis and atherosclerosis, mental retardation, dislocation of the eye lens (ectopia lentis), and skeletal malformations. Premature death (often in childhood) usually results from a major thrombotic or embolic event. Notably, one of the genetic defects that afflicts a significant proportion of homocystinuria patients reduces the affinity of cystathionine β-synthase for its vitamin B_6 cofactor, PLP. For these patients, the metabolic defect can be overcome to some extent with high-dose vitamin B_6 supplements, which significantly lower plasma homocysteine levels, reduce morbidity, and increase life expectancy. Interestingly, for other genetic defects involving cystathionine β-synthase that cause homocystinuria independent of the affinity of the enzyme for PLP, high-dose vitamin B_6 supplements nonetheless have a therapeutic effect despite having little or no influence on plasma homocysteine levels.

B Vitamin Deficiencies

Hyperhomocysteinemia is also caused by B vitamin deficiencies. Deficiencies of folate and vitamin B_{12} lead to impaired remethylation of homocysteine causing mild, moderate, or severe elevations in plasma homocysteine, depending on the severity of the deficiency, as well as coexistence of genetic or other factors that interfere with homocysteine metabolism (see below). Because riboflavin is required for the synthesis of flavin adenine dinucleotide (FAD), and because FAD serves as a cofactor for MTHFR, riboflavin deficiency can also affect homocysteine remethylation, and thus contribute to elevations in plasma homocysteine. Vitamin B_6 deficiency leads to impairment of homocysteine catabolism and thus also causes hyperhomocysteinemia. However, the nature of hyperhomocysteinemia caused by vitamin B_6 deficiency differs from that caused by folate and vitamin B_{12} deficiencies: In vitamin B_6 deficiency, fasting blood levels of homocysteine are usually not elevated or only slightly elevated. Only after a protein meal or after consumption of an oral methionine load (see below), does plasma homocysteine become abnormally elevated in vitamin B_6-deficient patients. In contrast, plasma homocysteine levels tend to be elevated regardless of prandial state in patients with folate or vitamin B_{12} deficiency. The basis for these different manifestations is likely due to differential effects of the vitamin deficiencies on intracellular SAM levels and consequent disruption of the allosteric control of homocysteine metabolism.

Recently, there has been growing interest in the concept of nutritional genomics. This refers to genetic variability among individuals and its effect on nutritional requirements. A prime example of this concept is a common polymorphism in MTHFR (677C→T) in which an alanine is replaced by valine at codon 222 in the primary sequence of the enzyme. Individuals with the homozygous variant (677TT) of this gene (10–15% of the general population; lower in blacks, higher in Latinos and in some parts of Europe, e.g., Southern Italy) have an enzyme that is thermolabile, with reduced affinity for its substrate (methylenetetrahydrofolate) and its cofactor (FAD). Consequently, 677TT individuals require a higher intake of folate and riboflavin to maintain optimal enzyme activity than those with the wild-type isoform of the enzyme (677CC). This is reflected by the fact that blood homocysteine levels are higher in people with the 677TT isoform than in those with the 677CC isoform, but only when overall folate and/or riboflavin status is low. When overall folate and riboflavin status is high, no difference in homocysteine levels is observed between the isoforms.

The clinical and public health importance of the MTHFR polymorphism is that women with the 677TT isoform are at increased risk of having a child with a neural tube defect (e.g., spina bifida, *sp*. anencephaly). This risk can be reduced by folic acid supplements, an observation that underlies the decision by the US government to mandate folic acid fortification of grain products as of January, 1998. This program has been highly successful, having reduced the prevalence of folate deficiency from over 20% to about 1%, the prevalence of hyperhomocysteinemia by about 50%, and the incidence of neural tube defects by at least 20%. The success of the folic acid fortification program in the US

spawned similar programs in several countries in the Americas, including Canada, Chile, and Costa Rica. Folic acid fortification has also been initiated in Hungary and Israel, but other European countries, most notably the UK, have been slow to adopt this intervention strategy. This is due to concerns about the feasibility of fortification, a hesitancy to impose manditory fortification on the population, lingering concerns over masking B_{12} deficiency, and the possibility of other unrecognized health consequences associated with excess folic acid intake.

Other polymorphisms in MTHFR and other enzymes involved in homocysteine metabolism (e.g., methionine synthase, methionine synthase reductase (EC 1.16.1.8), cystathionine β-synthase) have been identified and their overall influence on homocysteine metabolism, B vitamin requirements, and disease risk have been and continue to be evaluated.

Other Causes of Hyperhomocysteinemia

Other pathophysiological causes of hyperhomocysteinemia include renal dysfunction and hypothyroidism. The kidney is a major site of homocysteine metabolism and renal disease leads to a significant reduction in the body's overall capacity to metabolize this amino acid. The resulting moderate to severe hyperhomocysteinemia can be attenuated, in part, by high-dose B vitamin supplements, which putatively maximize the residual renal metabolism, as well as the metabolic capacities of the extrarenal organs. Mild elevations in homocysteine occur in patients with hypothyroidism, which resolve to normal with thyroid replacement therapy. This observation implies that thyroxine and/or thyroid-stimulating hormone influence homocysteine metabolism directly, perhaps through up- or downregulation of key homocysteine-metabolizing enzymes. Alternatively, homocysteine may become elevated in hypothyroid patients secondary to mild impairment of renal function that may accompany the disorder.

Patients with diabetes (both insulin dependent and insulin independent) tend to have mild hyperhomocysteinemia. However, this seems to be confined to those patients whose diabetic condition has progressed to involve renal insufficiency. Interestingly, in the absence of renal involvement, homocysteine levels in diabetic patients tend to be lower than normal. Insulin has been shown to inhibit homocysteine catabolism through cystathionine synthesis. Therefore, reduced insulin levels in diabetic patients may actually promote homocysteine catabolism, thus leading to lower plasma levels.

Premenopausal women tend to have lower plasma homocysteine than men of similar age, and homocysteine levels tend to rise in women after the menopause. Hormone replacement therapy reduces homocysteine back to premenopausal levels. Moreover, homocysteine decreases in male to female transsexuals, and increases in female to male transsexuals, primarily related to the estrogen and androgen regimens that such individuals respectively follow. Taken together, these observations strongly suggest an influence of sex hormones on homocysteine metabolism, though the mechanisms are not well understood.

Drugs can also affect homocysteine metabolism and lead to elevations of homocysteine in the blood. Certain anticancer drugs, such as methotrexate, and antiepilepsy medications, such as valproate and carbamazepine, are inhibitors of folate metabolism. The resulting functional folate deficiency leads to hyperhomocysteinemia. The anti-Parkinsonian drug, levodopa or L-dopa, causes elevations in blood homocysteine levels by a different mechanism: a significant proportion of an oral dose of L-dopa is methylated by SAM, leading to increased intracellular synthesis of SAH and homocysteine. The excess synthesis of homocysteine can overwhelm the capacities of the homocysteine metabolic pathways, particularly when B vitamin status is suboptimal, leading to hyperhomocysteinemia.

Homocysteine and Vascular Disease

The current interest in homocysteine is primarily related to its recognized status as an independent risk factor for cardiovascular, cerebrovascular, and peripheral vascular disease. This homocysteine theory of vascular disease comes directly from a seminal observation made by Kilmer McCully. In the early to mid-1960s, it was recognized that a prominent characteristic of patients with homocystinuria caused by defects in cystathionine β-synthase were very high elevations of both homocysteine and methionine in the blood. Therefore, it was not clear whether the vascular complications of this disorder were the consequence of hyperhomocysteinemia or hypermethioninemia. McCully observed that a patient with homocystinuria caused by a defect in a B_{12}-metabolizing enzyme had hyperhomocysteinemia, but not hypermethioninemia. Nonetheless, this patient had similar (though not identical) vascular pathology to that observed in patients with homocystinuria caused by cystathionine β-synthase deficiency. From this McCully concluded that the vascular culprit was homocysteine, and not methionine.

McCully's hypothesis, however, was not immediately embraced. The prevailing theory of atherosclerosis at the time centered on cholesterol, and it proved difficult for McCully to convince his peers and

national funding agencies of the potential importance of this new and competing hypothesis. Contributing to this was a lack of a reproducible animal model of homocysteine-induced vascular disease and a lack of a sensitive method to measure homocysteine in the blood. Consequently, McCully's hypothesis went into temporary obscurity.

In the mid-1970s, David and Bridget Wilcken reinvigorated McCully's hypothesis with their observation that a subset of patients with premature coronary artery disease had reduced ability to metabolize homocysteine. Notably, this association was observed in individuals who did not have any of the severe genetic defects that underly homocystinuria, suggesting that less severe or modest impairment of homocysteine metabolism may contribute to vascular disease risk. Subsequently, the advent of reliable, high-throughput assays for total plasma or serum homocysteine in the 1980s (see below) allowed for large-scale epidemiological assessment of associations between homocysteine and vascular diseases, both cross-sectionally and longitudinally. Through the 1990s, an explosion of population and case–control studies established that hyperhomocysteinemia is, indeed, a risk factor for heart attack, stroke, thrombosis, and peripheral atherosclerotic disease. Moreover, the risk associated with hyperhomocysteinemia is independent of other prominent risk factors, such as hypertension, hypercholesterolemia, hyperlipidemia, smoking, male gender, and others. Further indication of the importance of homocysteine with respect to vascular disease is the estimate that the relative risk of coronary artery disease associated with hyperhomocysteinemia is about equivalent to that associated with hypercholesterolemia. As the evidence mounted, McCully was vindicated and his contribution became widely recognized.

Homocysteine, Cognitive Function, and Dementia

As the relationship between homocysteine and vascular disease became increasingly apparent, researchers also addressed the hypothesis that hyperhomocysteinemia may affect cognitive function and the risk of dementia in older adults. This was based primarily on the recognized association between homocysteine and cerebrovascular disease, but also the observation that homocysteine and its metabolite, homocysteic acid, can induce excitotoxicity in neurons. Throughout the 1990s and into the new century, many cohort studies revealed significant inverse correlations between plasma homocysteine concentration and performance on a variety of cognitive function tests. Moreover, individuals with Alzheimer's disease were found to have higher plasma homocysteine than age- and gender-matched controls, while baseline homocysteine levels predicted the risk of incident dementia.

Homocysteine and Pregnancy Outcomes

Hyperhomocysteinemia has also been suspected as a risk factor for pregnancy complications and birth defects. Elevated plasma homocysteine levels have been associated with placental vasculopathy, pre-eclampsia, and placental infarction, as well as recurrent premature delivery, low birth weight, and spontaneous abortion. Birth defects associated with hyperhomocysteinemia in the mother include neural tube defects, orofacial clefts, clubfoot, and Down's syndrome. The protective effect of folic acid supplementation and fortification against neural tube defects, and perhaps the other abnormal birth outcomes cited, may be related to reduced homocysteine levels.

Mechanisms

In parallel with epidemiological studies, a significant amount of basic research has focused on the mechanism(s) by which homocysteine may induce atherosclerosis and thrombosis. A definitive answer has proven elusive. Potential mechanisms with significant experimental support include, but are not limited to, the following: (1) modification of the endothelial cell surface; (2) modification of plasma proteins by formation of disulfides; (3) activation of platelets; (4) modification of monocyte functions; (5) increased expression or activity of vascular adhesion molecules; and (6) oxidative damage induced by peroxides formed during disulfide bond formation.

A seventh potential mechanism relates to a known quirk of homocysteine synthesis and metabolism. The equilibrium of the interconversion between SAH and homocysteine (catalyzed by SAH hydrolase) actually favors SAH synthesis (**Figure 1**). *In vivo*, this reaction proceeds toward homocysteine synthesis because of product removal, i.e., the efficient metabolism of homocysteine back to methionine or through cystathionine synthesis. However, when there is a block in homocysteine metabolism, as occurs in the genetic defects, B vitamin deficiencies, and other causes delineated above, homocysteine accumulates intracellularly. Consequently, SAH also accumulates within cells. The significance of this phenomenon is that SAH is a feedback inhibitor of all SAM-dependent methylation reactions. Therefore, hyperhomocysteinemia may cause or contribute to vascular disease through SAH-mediated inhibition of methylation.

Another area that is receiving increasing attention is the relationship between homocysteine, nitric

oxide, and endothelial function. One of the roles of nitric oxide is as a vasodilator. Homocysteine has been shown to be an inhibitor of nitric oxide synthesis, and thus can inhibit vasodilatation. This has led to the hypothesis that hyperhomocysteinemia, by inhibiting nitric oxide synthesis, impairs the ability of the vascular endothelium to maintain homeostasis of vascular tone. This in turn may directly or indirectly increase susceptibility to vascular insults, thus promoting atherosclerosis and thrombosis.

The search for the definitive pathogenetic mechanism implicating homocysteine as a cause of vascular disease continues, and it is recognized that several mechanisms may contribute synergistically. However, some have questioned whether homocysteine is a cause of vascular disease, or simply a consequence.

Cause or Effect?

Though there is considerable evidence, both epidemiological and experimental, that homocysteine is a causative factor in vascular disease, there are data that contradict this conclusion. First, though cross-sectional and case–control studies fairly consistently demonstrate that hyperhomocysteinemia is associated with vascular disease, some prospective studies have found no relationship between baseline homocysteine levels and risk of incident vascular events. Second, several studies have found no relationship between the MTHFR 667C→T polymorphism and venous thrombosis, despite the association of this polymorphism with elevated plasma homocysteine levels.

With these observations in mind, a plausible alternative hypothesis has been put forward to explain the association between hyperhomocysteinemia and vascular disease. One of the organs that can be significantly affected by vascular disease is the kidney. Reduced kidney function caused by atherosclerosis may lead to renal insufficiency and reduced capacity to metabolize homocysteine. In this way, hyperhomocysteinemia may actually result from vascular disease. This hypothesis remains to be tested. The possibility of a vicious cycle, i.e., one in which vascular disease causes homocysteine to become elevated in the blood, which in turn induces further vascular damage, must also be considered.

B Vitamin Supplementation

Currently, several large-scale intervention trials are underway to determine if B vitamin supplements (folic acid, B_{12}, B_6), which effectively lower blood homocysteine levels, reduce the incidence of vascular disease (**Table 3**). If proven effective, such supplements would be an inexpensive and relatively innocuous means by which the risk of vascular disease may be reduced. However, it must be recognized that if these trials are successful, they will not serve as definitive proof that homocysteine is a vascular toxin. It may be the case that one or more of the B vitamins influences vascular disease risk through separate mechanisms. For example, several studies have shown that low B_6 status has an association with vascular disease independent of homocysteine. The uncertain relationship between hyperhomocysteinemia, B vitamins, and vascular disease is summarized in **Figure 2**. If homocysteine is not a vascular toxin, it may still serve as a marker of both vascular disease and as an indicator of the efficacy of B vitamin supplementation.

Table 3 Intervention trials to determine the effect of B vitamin supplements on homocysteine and the risk of vascular disease

Study	*Location*	*Start date*
Cambridge Heart Antioxidant Study 2 (CHAOS-2)	UK	1998
Heart Outcomes Prevention Evaluation 2 (HOPE-2)	Canada	1999
Norwegian Multi-Center B-Vitamin Intervention Study (NORVIT)	Norway	1998
Prevention with a Combined Inhibitor and Folate in Coronary Heart Disease (PACIFIC)	Australia	2000
Study of Effectiveness of Additional Reductions in Cholesterol and Homocysteine (SEARCH)	UK	1999
Vitamins in Stroke Prevention (VISP)	USA	1998
Vitamins to Prevent Stroke (VITATOPS)	Australia	1999
Western Norway B-Vitamin Intervention Trial (WENBIT)	Norway	1999
Women's Antioxidant and Cardiovascular Disease Study (WACS)	USA	1998

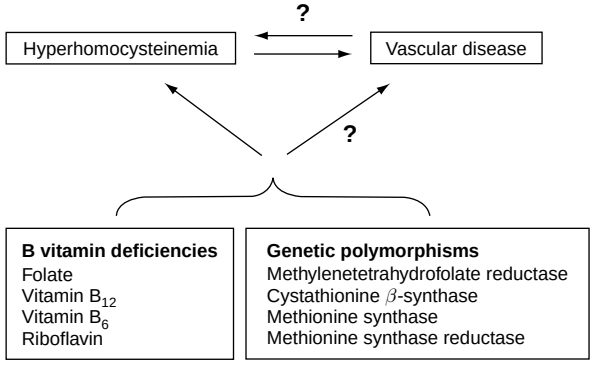


Figure 2 Hyperhomocysteinemia, B vitamins, genetic polymorphisms, and vascular disease. There is still some question whether elevated plasma homocysteine is a cause or consequence of vascular disease and whether there are influences of B vitamins and related polymorphisms on vascular disease risk that are independent of homocysteine.

Measurement of Blood Levels

A variety of assays have been developed to quantify blood homocysteine levels, with those employing high-pressure liquid chromatography perhaps being the most common. These assays have proven to be relatively accurate and precise (coefficients of variation less than 10%), and are relatively simple and quick to perform. The development of such assays in the 1980s was the technological breakthrough that spurred the exponential increase in homocysteine-related research over the last 15–20 years and the establishment of hyperhomocysteinemia as an independent risk factor for vascular disease.

As described above, homocysteine comes in several forms in the blood, including protein-bound, mixed disulfides, homocystine, and free-reduced. Assays for homocysteine usually measure the sum total of all these forms, i.e., total homocysteine. To accomplish this, the first procedure in homocysteine assays is a reduction step to break all disulfide bonds, thus converting all homocysteine to the free-reduced form. The free-reduced form is then quantified by one of various methods.

Blood sample collection and processing are critical factors in the determination of homocysteine concentrations. Typically, blood samples for homocysteine analysis are collected in tubes containing an anticoagulant (e.g., EDTA, heparin). Prompt separation of plasma from the blood cells after centrifugation is required to avoid excess release of intracellular homocysteine into the plasma or removal of homocysteine from the plasma by metabolically active leucocytes after blood draw. Keeping the blood sample cold until centrifugation and separation (ideally within 4 h of blood draw) minimizes this problem. Serum homocysteine concentrations typically exceed plasma concentrations by 20%. This is likely due to the fact that blood collected to isolate serum (i.e., without an anticoagulant) must clot at room temperature for 30–60 min before centrifugation and separation. Therefore, plasma is preferred for measurement of homocysteine. Once separated from the blood cells, the concentration of homocysteine in plasma or serum remains stable for years when stored frozen.

Another important issue in the measurement of homocysteine is the prandial state of the individual. For individuals with adequate B vitamin status, no genetic abnormalities, and no pathophysiological conditions that affect homocysteine metabolism, plasma homocysteine levels after an overnight fast are similar to levels after a meal (even high-protein meals containing methionine). However, for individuals with low vitamin B_6 status or heterozygous genetic defects in cystathionine β-synthase, postprandial homocysteine levels can be significantly higher than fasting levels. Because of the nutritional or genetic block in the conversion of homocysteine to cystathionine, there is decreased capacity to metabolize the influx of homocysteine synthesized from dietary methionine. This, in fact, is the basis for the methionine load test for detection of impaired cystathionine β-synthase activity. In this test, baseline blood is drawn after an overnight fast, and then again 4 h after consumption of a large dose of methionine dissolved in orange juice (100 mg methionine per kilogram body weight). Plasma homocysteine increases to a greater extent in individuals with low vitamin B_6 status or heterozygous genetic defects in cystathionine β-synthase than in individuals without these problems. Importantly, individuals with elevated fasting homocysteine and those with normal fasting levels, but elevated post-methionine load levels, are both at increased risk of vascular disease.

See also: **Amino Acids**: Chemistry and Classification; Metabolism; Specific Functions. **Cobalamins**. **Folic Acid**. **Riboflavin**. **Vitamin B_6**.

Further Reading

Blom H, Fowler B, Jakobs C, and Koch H-G (eds.) (1998) Disorders of homocysteine metabolism. *European Journal of Pediatrics* **157**(supplement 2): S39–S142.

Boushey CJ, Beresford SAA, Omenn GS, and Motulsky AG (1995) A quantitative assessment of plasma homocysteine as a risk factor for vascular disease: probable benefits of increasing folic acid intakes. *JAMA* **274**: 1049–1057.

Brattström L and Wilcken DEL (2000) Homocysteine and cardiovascular disease: cause or effect? *American Journal of Clinical Nutrition* **72**: 315–323.

Carmel R and Jacobsen DW (2001) *Homocysteine in Health and Disease*. Cambridge: Cambridge University Press.

Christen WG, Ajani UA, Glynn RJ, and Hennekens CH (2000) Blood levels of homocysteine and increased risks of cardiovascular disease: causal or casual? *Archives of Internal Medicine* **160**: 422–434.

Clarke R, Smith AD, Jobst KA, Refsum H, Sutton L, and Ueland PM (1998) Folate, vitamin B_{12}, and serum total homocysteine levels in confirmed Alzheimer disease. *Archives of Neurology* **55**: 1449–1455.

Finkelstein JD (1990) Methionine metabolism in mammals. *Journal of Nutritional Biochemistry* **1**: 228–237.

Finkelstein JD (2000) Homocysteine: a history in progress. *Nutrition Reviews* **58**: 193–204.

Graham I, Refsum H, Rosenberg IH, and Ueland PM (eds.) (1997) *Homocysteine Metabolism: From Basic Science to Clinical Medicine*. Boston, MA: Kluwar Academic Publishers.

Green R (1998) Homocysteine and occlusive vascular disease: culprit or bystander. *Preventive Cardiology* **1**: 31–33.

Green R and Jacobsen DW (1995) Clinical implications of hyperhomocysteinemia. In: Bailey LB (ed.) *Folate in Health and Disease*, pp. 75–122. New York: Marcel Dekker.
Green R and Miller JW (1999) Folate deficiency beyond megaloblastic anemia: hyperhomocysteinemia and other manifestations of dysfunctional folate metabolism. *Seminars in Hematology* **36**: 47–64.
Homocysteine Lowering Trialists' Collaboration (1998) Lowering blood homocysteine with folic acid based supplements: meta-analysis of randomised trials. *British Medical Journal* **316**: 894–898.
Jacques PF, Selhub J, Bostom AG, Wilson PW, and Rosenberg IH (1999) The effect of folic acid fortification on plasma folate and total homocysteine concentrations. *New England Journal of Medicine* **340**: 1449–1454.
McCully KS (1996) Homocysteine and vascular disease. *Nature Medicine* **2**: 386–389.
Meleady R and Graham I (1999) Plasma homocysteine as a cardiovascular risk factor: causal, consequential, or of no consequence? *Nutrition Reviews* **57**: 299–305.
Miller JW (2000) Homocysteine, Alzheimer's disease, and age-related cognitive decline. *Nutrition* **16**: 675–677.
Miller JW, Selhub J, Nadeau M, Thomas CA, Feldman RG, and Wolf PA (2003) Effect of L-Dopa on plasma homocysteine in PD patients. *Neurology* **60**: 1125–1129.
Mudd SH, Levy HL, and Skovby F (1995) Disorders of transsulfuration. In: Scriver CR, Beaudet AL, Sly WS, and Valle D (eds.) *The Metabolic and Molecular Bases of Inherited Disorders*, 7th edn, pp. 1279–1327. New York: McGraw Hill.
Pfeiffer CM, Huff DL, Smith SJ, Miller DT, and Gunter EW (1999) Comparison of plasma total homocysteine measurements in 14 laboratories: an international study. *Clinical Chemistry* **45**: 1261–1268.
Refsum H, Smith AD, Ueland PM *et al.* (2004) Facts and recommendations about total homocysteine determinations: an expert opinion. *Clinical Chemistry* **50**: 3–32.
Refsum H, Ueland PM, Nygard O, and Vollset SE (1998) Homocysteine and cardiovascular disease. *Annual Review of Medicine* **49**: 31–62.
Robinson K (ed.) (2000) *Homocysteine and Vascular Disease*. Norwell, Massachusetts: Kluwar Academic Publishers.
Selhub J, Jacques PF, Wilson PWF, Rush D, and Rosenberg IH (1993) Vitamin status and intake as primary determinants of homocysteinemia in an elderly population. *JAMA* **270**: 2693–2698.
Selhub J and Miller JW (1992) The pathogenesis of homocysteinemia: interruption of the coordinate regulation by S-adenosylmethionine of the remethylation and transsulfuration of homocysteine. *American Journal of Clinical Nutrition* **55**: 131–138.
Seshadri S, Beiser A, Selhub J *et al.* (2002) Plasma homocysteine as a risk factor for dementia and Alzheimer's disease. *New England Journal of Medicine* **346**: 476–483.
Stacey M (1997) The fall and rise of Kilmer McCully. *New York Times Magazine*, August **10**: 26–29.
Ueland PM, Refsum H, Beresford SAA, and Vollset SE (2000) The controversy over homocysteine and cardiovascular risk. *American Journal of Clinical Nutrition* **72**: 324–332.
Welch GN and Loscalzo J (1998) Homocysteine and atherothrombosis. *New England Journal of Medicine* **338**: 1042–1050.
Wilcken DEL and Wilcken B (1998) B vitamins and homocysteine in cardiovascular disease and aging. *Annals of the New York Academy of Sciences* **854**: 361–370.

HUNGER

J C G Halford, University of Liverpool, Liverpool, UK
A J Hill and J E Blundell, University of Leeds, Leeds, UK

This article is a revision of the previous edition article by A J Hill and J E Blundell, pp. 1015–1020, © 1999, Elsevier Ltd.

Hunger is a familiar but commonly misunderstood and mistrusted part of our eating behavior. This article will clarify the meaning of the term, describe the common procedures for measuring hunger, the ways in which hunger and satiety are interrelated, and the adaptability of hunger experience in a learning framework. The relationship between hunger and eating behavior will be examined at both a methodological and conceptual level, and putative disorders of hunger will be briefly examined.

Definition

The term hunger is used in more than one sense by both scientists and the lay public. World hunger is a widely used phrase to describe the shortage of food and state of malnutrition experienced by a substantial proportion of the world's population. Its use is emotive and largely descriptive. It is in the study of motivation that the term takes on a more precise and individual definition. In this context, hunger describes the drive or the motivational force that urges us to seek and consume food. It is the expression of a biological need to sustain growth and life. Hunger is therefore a purposeful experience that possesses a clear biological function.

There are two ways in which the term hunger is used within nutritional science. One is its use as a motivational construct in a scientific theory. Here, hunger is inferred from directly observable and measurable events. In this way, inferring increased or high levels of hunger from a long period of food deprivation or an increased willingness to expend effort in order to obtain food, hunger becomes a mediating concept or intervening variable. However, a more familiar use of the word is that collection of conscious feelings or sensations

that are linked to a desire to obtain and eat food. This is the sense in which lay people understand the term hunger and is what researchers attempt to capture by means of rating scales and other measurement devices.

The first serious investigation of the everyday experience of hunger used a questionnaire in which people were asked to note the presence of physical sensations in a number of bodily areas, together with moods, urges to eat, and preoccupation with thoughts of food. It was found that the observation, 'I feel hungry,' is typically based on the perception of bodily feelings, which at times are very strong. Gastric sensations, a hollow feeling or stomach rumbling, are frequent indicators of hunger, although people also report sensations in the mouth, throat, and head. These accompany more diffuse feelings of restlessness and excitability as well as an urge to eat. The consumption of food changes both the pattern of physical sensations and the accompanying emotional feelings, with unpleasant and aversive sensations becoming replaced by more pleasant ones. So, for example, an aching stomach becomes relaxed and the feeling of excitement and irritability is replaced by one of contentment.

Subsequent research has confirmed these general patterns of characteristic premeal sensations and feelings, particularly with regard to the salience of gastric sensations. However, it has also noted a great deal of variability both within and between individuals. In other words, hunger demands neither the consistent presence of single sensations prior to every act of eating nor in every person sitting down to eat. Despite this variability people are able to, and frequently do, make judgments regarding their state of hunger, partly through reference to these sensations.

The Measurement of Hunger

The process of measuring hunger is not as straightforward as it might seem. One reason is the frequently raised mistrust of subjective reports. Critics point to the variability in response between individuals and the absence of any objective 'standard' by which internal experience can be calibrated. However, as argued later, this issue of 'validity' is more complex than this criticism suggests. A second reason is the failure to appreciate the distinction made previously between an individual's assessment of his or her disposition to eat and inferring hunger from the amount of food consumed or from some part of the act of eating (e.g., eating speed). While in many circumstances they will be in accord, the subjective report and inferred construct can as easily diverge.

The two most common methods for quantifying hunger are fixed-point rating scales and visual analog scales (**Figure 1**). Fixed-point scales are quick and simple to use, and the data they provide are easy to analyze. Past examples of these scales show they vary greatly in complexity. In considering the appropriate number of points to be included in this type of scale, the freedom to make a range of possible responses must be balanced against the precision and reliability of the device. Research seems to indicate that scales with an insufficient number of fixed points can be insensitive to subtle changes in subjective experience. In addition, the fixed points are important determinants of the way people use the scales and distribute their ratings.

One way of overcoming some of these failings is to abolish the points completely. Thus, visual analog scales are horizontal lines (often 100 or 150 mm long), unbroken and unmarked except for word anchors at each end. The user of the scale is instructed to mark the line at the point that most accurately reflects the intensity of the subjective feeling at that time. The researcher measures the distance to that mark in millimeters from the negative end (no hunger), thus yielding a score of 0–100 (or 150). This is done either by hand or automatically if presented by computer screen. By doing away with all of the verbal labels except the end definitions, visual analog scales retain the advantages of fixed-point scales, while avoiding many of the problems with uneven response distributions.

An important aspect of these methods concerns the interpretation of differences between the fixed points or intervals on a visual analog scale. So, for example, it should not be assumed that the difference between 20 and 30 mm on a hunger scale is

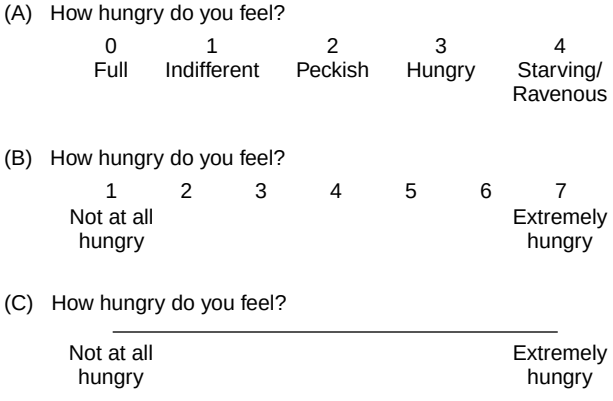


Figure 1 Examples of different types of scales used in the assessment of hunger: (A) fixed-point scale with points defined, (B) fixed-point scale, and (C) visual analog scale.

perceptually the same as the difference between 80 and 90 mm. Nor can a hunger rating of 80 mm be said to represent a feeling of hunger that is twice the intensity of that rated at 40 mm. Related to this is the problem of 'end effects.' This refers to the reluctance of a minority of subjects to make ratings away from the upper or lower end points of the scale, despite clear instructions. Despite these limitations, data from such scales are often analyzed using parametric statistical procedures, such as analysis of variance, and in general this appears to be a satisfactory approach.

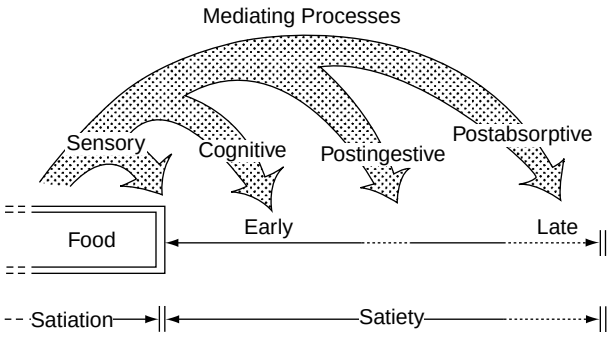


Figure 2 A representation of the satiety cascade showing the different phases of satiety and their associated mechanisms.

Hunger and Satiety

If hunger is that feeling that reminds us to seek food, then eating relieves hunger, albeit until the next snack or meal. The capacity of a food to reduce the experience of hunger is called 'satiating power' or 'satiating efficiency.' This power is the product of the body's handling of the nutritional composition and structure of the food eaten. It follows that some foods will have a greater capacity to maintain suppression over hunger than other foods.

The distinction between hunger and satiety is both conceptual and technical. As hunger diminishes, satiety rises. But it is useful to further separate those events that occur across the course of a meal from those between meals. In this way the process of satiation can be clearly distinguished from the state of satiety. Satiation can be regarded as the process that develops during eating and that eventually brings a period of eating to an end. Accordingly, satiation can be defined in terms of the measured size of an eating episode (such as its energy, weight, or volume). Hunger declines as satiation develops and usually reaches its lowest point at the end of a meal. Satiety is defined as the state of inhibition over further eating that follows at the end of a meal and that arises from the consequences of food ingestion. The intensity of satiety can be measured by the duration of time until eating starts once more, or by the amount consumed at the next meal. The strength of satiety is also measured by the time that hunger is suppressed. And as satiety weakens, hunger is restored.

In examining the mechanisms responsible for suppressing hunger and maintaining its low state, it is clear that they range from those that occur when food is initially sensed to the effects of metabolites on body tissues following the digestion and absorption of food (across the wall of the intestine and into the bloodstream). By definition, satiety is not an instantaneous event but occurs over a considerable time period. The different phases of satiety and their associated mechanisms are shown in **Figure 2**. Sensory effects are generated through the smell, taste, temperature, and texture of food, and it is likely that these factors have effects on eating in the very short term. Cognitive influences represent the beliefs held about the properties of foods, and these factors may also help inhibit hunger in the short term.

The category identified as postingestive processes includes a number of possible actions, such as gastric distension and rate of emptying, the release of hormones such as cholecystokinin, and the stimulation of certain receptors along the gastrointestinal tract. The postabsorptive phase of satiety includes those mechanisms arising from the action of metabolites after absorption into the bloodstream. These include the action of glucose and amino acids, which act directly on the brain after crossing the blood–brain barrier, and which influence the brain indirectly via neural inputs following stimulation of peripheral chemoreceptors. The most important suppression and subsequent control of hunger is brought about by postingestive and postabsorptive mediating processes.

It follows from this framework that foods of varying nutrient composition will have different effects on the mediating processes and will therefore differ in their effects on hunger, satiation, and satiety. There is considerable interest, for example, in whether protein, fat, and carbohydrate differ in their satiating power.

The balance of evidence shows that per unit energy, protein (within normal dietary limits) has the greatest satiating efficiency of all the macronutrients. This is particularly true in short-term studies and is observed in lean and obese subjects alike. Longer term evidence of this effect is currently lacking. However, of great practical and theoretical interest is the comparative effect of carbohydrate and fat since they form the majority of our routine energy intake. Research shows that carbohydrates are efficient hunger relievers. A variety of carbohydrates, including glucose, fructose, sucrose, and maltodextrins, all suppress

later test meal energy intake. This suppression is roughly equivalent to their energy value, although the time course of this effect varies according to the rate at which they are metabolized. In contrast, high-fat foods appear to stimulate energy intake (in contrast to low-fat, high-carbohydrate foods), or at least have a disproportionately weak action on satiety. The mechanisms responsible for this may include the effect of fat-promoting food palatability, the high-energy density of fat, and the absence of inhibitory feedback from body fat stores. Taken together, these findings show why diets high in fat can promote weight gain and lead to obesity.

Hunger: Physiological Determinants

Stomach distension and the detection of macronutrients such as fat or protein within the gut are all powerful satiety cues. They bring a meal to an end and for a time inhibit further consumption. Eventually, hunger again prevails and food intake follows. The flux between hunger and satiety is episodic and underpins the expression of our eating behavior throughout the day. However, it is not just the absence of episodic satiety cues (e.g., stomach distension and intestinal or absorbed nutrients) that influence the expression of hunger. Reduction in blood glucose levels or in levels of the circulating adipose tissue hormone leptin indicates a deficit in available energy and in energy reserves. Fluctuation of these factors indicates the metabolism and storage of the body's energy reserves. These are a tonic class of physiological signals that also influence the expression of appetite. Like episodic satiety signals, these tonic signals normally act on inhibitory mechanisms with the hypothalamus (anorexogenic circuits). Their absence elicits an active feeding response. Other tonic factors that indicate the body's energy status, such as adiponectin, cytokines, and gonadal hormones, also appear to act on energy regulator centres within the brain, particularly the hypothalamus, mainly to suppress hunger.

However, not all physiological signals, episodic or tonic, inhibit hunger. For instance, blood levels of the recently discovered gut hormone ghrelin have been shown to increase prior to a meal. Subsequent intake has been shown to suppress ghrelin release. Further studies have shown that ghrelin infusions increase food intake. Thus, this is a hormone that acts to promote food intake. Interestingly, ghrelin receptors are found in various hypothalamic locations that form part of the orexogenic circuits promoting food intake. These circuits contain many neuropeptides, such as neuropeptide Y, orexins, melanocortin concentrating hormone, and galanin, which all stimulate food intake. The precise nature of the physiological and neurobiological regulation of appetite is discussed elsewhere in this encyclopedia. Finally, it should be noted that the biological mechanisms critical to the expression of hunger are not independent of psychological ones. Indeed, the sensory and cognitive cues that stimulate hunger produce physiological changes that anticipate the ingestion and metabolism of energy and subsequently aid these processes. This brings on the psychological factors critical in the expression of appetite.

Conditioned Hunger

One of the essentials for an omnivore faced with a variety of new and different foods is the capacity to learn. It is not possible for an inborn preference or aversion to guide the choice of every possible food. Therefore, we learn which foods are beneficial (and which are not) by eating them. This learning involves the association between the sensory and the postabsorptive characteristics of foods. In this way the sensory characteristics of foods act as cues and come to predict the impact that foods will later have. Consequently, these cues should suppress hunger according to their relationship with subsequent physiological events.

It is possible to demonstrate experimentally how human beings adapt their eating to a food's energy content. A distinctively flavored food which contains 'extra' hidden energy, presented on several occasions, will result in a change in eating and in preference. When deprived of food, subjects' preference for the taste increases with gained experience. If presented when satiated, preference for the taste decreases. This process is also observable in young children, who eat smaller meals following a taste previously associated with a high-energy snack, and larger meals following a taste previously associated with a low-energy snack.

The idea that we can have conditioned hunger for specific nutrients is far more contentious. The concept of conditioned hunger suggests that the organism, faced with a diet deficient in a single important nutrient, will seek an alternative food source that contains the missing nutrient. However, earlier evidence from animals has largely been reinterpreted from the standpoint of conditioned aversions. Indeed, conditioned aversions are far more potent examples of the impact of learning on eating behavior than any examples of conditioned hunger. A conditioned aversion that will be familiar to many readers is the profound dislike that occurs in response to a food or drink that was eaten prior to vomiting or illness. An example of a conditioned taste aversion was famously described by learning theorist Martin Seligman. Steak with sauce

Béarnaise was Seligman's last meal before a bout of gastric flu. Yet knowing that it was the flu rather than the food that made him sick did not prevent the subsequent aversion to sauce Béarnaise. In fact, surveys show that conditioned taste aversions are commonplace and reported by 40–60% of people.

Conditioned taste aversions are important in the present context not because they represent a special form of one-trial learning that we are biologically pre-prepared to acquire. Rather, they show that the strength of cue–consequence learning in the area of food intake depends on the stability and reliability of the relationship between tastes (sensory cues) and physiological effects (metabolic consequences) of food. When there is distortion, variation, or extreme complexity in the relationship between sensory characteristics and nutritional properties, then the conditioned control of hunger is weakened or lost. In many respects, the variety of foods available to us represents a cacophony of different sensory characteristics and has the added complication of ingredients that preserve the sensory qualities while altering their nutritive value. Learned hunger therefore is a relatively less important factor when the food supply contains many food items with identical tastes but differing metabolic properties.

Hunger and Eating Behavior

If hunger is biologically useful and a subjective experience that indicates a depleted nutritional state, then a close correspondence between hunger and eating would be expected. So hunger should be either a necessary or a sufficient condition for eating to occur. However, this is not invariably the case. Instances of people deliberately refraining from eating in spite of hunger (fasting for moral or political conviction) show hunger not to be a sufficient condition. And examples in research and daily experience, of eating a tempting food when otherwise satiated, show hunger not to be necessary for eating to take place. But while the relationship between hunger and eating is not based on biological inevitability, in many circumstances they are closely linked.

Unfortunately, the lack of a one-to-one correspondence between hunger and eating has been used as another way to question the validity of hunger ratings. But should a high correlation between hunger ratings and subsequent food intake be expected in all circumstances? The previous examples show that in certain circumstances the two can be disengaged. So, for example, eating can occur when hunger is low (such as when highly palatable food is offered unexpectedly) and not at other times when hunger is high (when food is unavailable or other activities have priority). In addition, many experimental analyzes of the correlational relationship between hunger and food intake report the relationship only when subjects are hungry. In other words, the correlation is only examined for a small portion of the available scale. Very few studies have looked at the association between hunger and food intake when hunger has been represented in all its possible degrees.

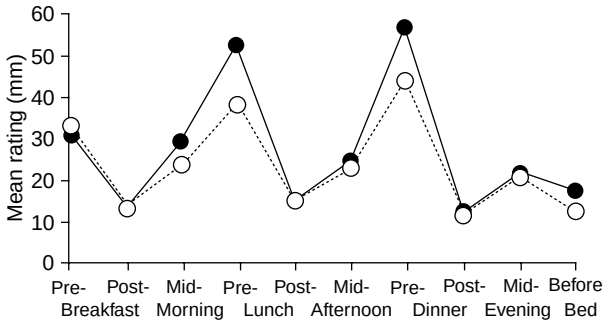


Figure 3 Ratings of hunger made across the day by a group of obese women taking an appetite suppressant drug (dotted lines) or placebo (solid lines).

It is clear that hunger ratings cannot be used simply as a proxy measure for food intake. Equally, there is good evidence that in most circumstances self-report ratings of hunger correlate statistically and meaningfully with eating. This association exists not simply across single meals, but across the entire day as shown in **Figure 3**. The rhythmic oscillation of hunger is tied closely to the overall pattern of food intake in this group of individuals. As such, it presents an elegant and experimentally useful way of examining diurnal variations in the experience of hunger.

In questioning the relationship between hunger and eating, we are also forced to place the action of hunger within a broader context of social and psychological variables that moderate food choice and eating behavior. Eating patterns are maintained by enduring habits, attitudes and opinions about the value and suitability of foods, and an overall liking for them. These factors, derived from the cultural ethos, largely determine the range of foods that will be consumed and sometimes the timing of consumption. The intensity of hunger experienced may also be determined, in part, by the culturally approved appropriateness of this feeling and by the host of preconceptions brought to the dining table. Hunger is therefore only one portion of the range of determinants of eating in any given situation.

Disorders of Hunger

The clinical eating disorders, anorexia nervosa and bulimia nervosa, are commonly believed to

encompass major disturbances of hunger. Yet the role that hunger may play is not entirely clear. Contrary to the literal meaning of the term, 'anorexia' is not experienced as a loss of appetite. Rather, clinicians recognize that anorexics may endure intense periods of hunger during their self-restricted eating. For some, their strength in resisting intense episodes of hunger provides a feeling of self-mastery and control that is absent in other areas of their lives. Research suggests that restricting anorexics (compared with those who binge) have the greatest blunting of hunger response, and that this disturbance in hunger is not a product of other areas of perceptual confusion.

There is evidence that in conditions of total starvation hunger may become temporarily diminished. This circumstance is extremely rare and obviously relatively brief. Once eating is recommenced, hunger returns rapidly and with extreme intensity. The accounts of the male volunteers who submitted to a 6-month period of semistarvation during World War II (the 'Minnesota Experiment') are a testament to the extreme power of hunger. Referred to as semistarvation neurosis, these men's activities were shaped by their need for food. And their hunger experience was extreme. Nearly two-thirds reported feeling hungry all the time and a similar proportion experienced physical discomfort due to hunger. Participants described a marked increase in what was referred to as 'hunger pain.' For some this was mildly discomforting and vaguely localized in the abdomen. For others, it was extremely painful. This account is especially useful in reminding why energy-reduced diets aimed at achieving weight loss are often difficult to maintain and easy to abandon.

Like anorexia, bulimia finds its literal meaning in changed hunger — 'ox hunger.' Again, however, the term is imprecise. Close analysis of the precursors of binge episodes show hunger to be lower than it is prior to a normal meal. In addition, while the urge to eat may be strong during a binge, the large amount of food consumed implies some defect in satiation rather than in hunger. And binging is often a well-practised behavior that develops and changes with time. As with anorexics, it is likely that a stable eating pattern is necessary in order to normalize the experience of hunger, a process that may take a long time to establish.

The question of whether obesity reflects a disorder of hunger is now regarded as largely redundant. Obesity is strictly a disorder of weight, and as such reflects potentially long-term failure in the regulation of energy balance. There is hardly any evidence of heightened levels of hunger contributing to excessive energy input. However, an exception to this is the rare disorder Prader–Willi syndrome. Genetically determined and characterized mainly by intellectual disability, obesity is a well-recognised feature of the syndrome. Emerging research suggests that the excessive levels of food intake are associated with both a delayed reduction in hunger while eating and a more rapid return to premeal states when eating has finished. Clearly, a better understanding of the biological events that accompany such aberrant eating patterns will strengthen understanding of the psychobiological framework that supports hunger.

See also: **Appetite**: Physiological and Neurobiological Aspects; Psychobiological and Behavioral Aspects. **Carbohydrates**: Requirements and Dietary Importance. **Eating Disorders**: Anorexia Nervosa; Bulimia Nervosa. **Famine**. **Food Choice, Influencing Factors**. **Obesity**: Definition, Etiology and Assessment. **Starvation and Fasting**. **Weight Management**: Approaches.

Further Reading

Blundell JE (1980) Hunger, appetite and satiety—Constructs in search of identities. In: Turner M (ed.) *Nutrition and Lifestyles*, pp. 21–42. London: Applied Sciences Publishers.

Booth DA (1977) Satiety and appetite are conditioned reactions. *Psychosomatic Medicine* **39**: 76–81.

Cornell CE, Rodin J, and Weingarten H (1989) Stimulus-induced eating when satiated. *Physiology and Behaviour* **45**: 695–704.

Flint A, Raben A, Blundell LE, and Astrup A (2000) Reproducibility, power and validity of visual analogue scales in assessment of appetite sensations in single test meal studies. *International Journal of Obesity* **24**: 38–48.

Friedman MI, Ulrich P, and Mattes RD (1999) A figurative measure of subjective hunger sensations. *Appetite* **32**: 395–404.

Halmi KA and Sunday SR (1991) Temporal patterns of hunger and fullness ratings and related cognitions in anorexia and bulimia. *Appetite* **16**: 219–237.

Hill AJ, Rogers PJ, and Blundell JE (1995) Techniques for the experimental measurement of human eating behaviour and food intake: A practical guide. International Journal of Obesity **19**: 361–375.

Keys A, Brozek J, Henscher A *et al.* (1950) In *The Biology of Human Starvation.* Minneapolis: University of Minnesota Press.

Kirkmayer SV and Mattes RD (2000) Effects of food attributes on hunger and food intake. *International Journal of Obesity* **24**: 1167–1175.

Kissileff HR (1984) Satiating efficiency and a strategy for conducting food loading experiments. *Neuroscience and Biobehavioural Reviews* **8**: 129–135.

Monello LF and Mayer J (1967) Hunger and satiety sensations in men, women, boys and girls. *American Journal of Clinical Nutrition* **20**: 253–261.

Ogden J (2002) The Psychology of Eating. *From Healthy to Disordered Behaviour*. Oxford: Blackwell.

Womble LG, Wadden TA, Chandler JM, and Martin AR (2003) Agreement between weekly vs. daily assessment of appetite. *Appetite* **40**: 131–135.

HYPERACTIVITY

M Wolraich, Vanderbilt University, Nashville, TN, USA

To discuss the issues of hyperactivity and diet, it is first important to understand the issues related to the diagnosis of hyperactivity, or what is now called the 'attention deficit/hyperactivity disorder' (ADHD). Since most of the recommendations for dietary changes have been for children who have been diagnosed with ADHD, this article will first review the historic and current changes in the diagnosis of ADHD and then review the diets that have been recommended for treatment and the evidence as to their efficacy.

Diagnostic Issues

Hyperactivity, or ADHD, is a condition that has been recognized for many years and has been quite extensively researched, but the diagnostic criteria and treatment continue to be controversial. The symptoms of ADHD were first described by a German physician, Heinrich Hoffman, in a children's book written in 1848. The symptoms were represented by two children, Harry, who looks in the air (inattention), and Fidgety Phil (hyperactivity). In 1902, George Still presented a lecture in England about 20 children who were aggressive, defiant, excessively emotional and lacking inhibitory volition, and who were also noted to have impaired attention and overactivity. A more etiological conceptualization of the condition did not occur until after World War I.

Symptoms of hyperactivity and inattention were suspected to be caused by the influenza epidemic that occurred after World War I, when postencephalitic behavior manifestations in children included extreme examples of hyperactivity and inattention. This led to the suggestion that these symptoms were due to organic brain damage. The concept of inattention and hyperactivity being part of a spectrum with less intense manifestations secondary to subtle injuries became known as the syndrome of 'minimal brain damage' in the 1960s. However, the lack of clear evidence for brain damage eventually resulted in a shift to a more descriptive labeling of the disorder. This is reflected in the American Psychiatric Association classification system (DSM) defining the 'hyperkinetic reaction of childhood.' The same disorder was similarly described in the United Kingdom, as reflected in the World Health Organization (WHO) classification. However, the conditions described differed in that the British disorder included more severe symptomatology and required that the symptoms had to be present in all settings.

In 1980, the US characterization of inattention and hyperactivity was changed in several ways. It was conceptually defined by three symptom dimensions: inattention, impulsiveness, and hyperactivity, with inattention playing a more prominent role. In addition, to address the heterogeneity within the disorder, two subtypes ('attention deficit disorder with hyperactivity' and 'attention deficit disorder without hyperactivity') were defined. Again, different from the British criteria, the symptoms were only required to be present in one setting such as school. Retaining the concept that the major contributions to the symptoms were related to innate characteristics in the child rather than to environmental influences, the symptoms were required to have been present before the age of 7 years and to have lasted for at least 6 months. The British system continued to use the term 'hyperkinetic syndrome of childhood' and to include the pervasive nature of the symptoms.

The most recent changes in diagnostic criteria used by the American Psychiatric Association (DSM-IV) and the WHO have moved the definitions closer to agreement. Considering the most recent studies, there is evidence to support two dimensions. In DSM, the first dimension, inattention, is characterized by the 'often' occurrence of at least six of nine of the inattentive behaviors presented in **Table 1**. The second dimension consists of both hyperactivity and impulsiveness and is characterized by the 'often' occurrence of at least six of nine of the hyperactive and/or impulsive behaviors presented in **Table 1**. The WHO definitions are similar but do not attempt to quantify the specific behaviors and do not include impulsiveness in the hyperactivity dimension.

In DSM, the two dimensions define three subtypes: predominantly inattentive type (meeting criteria on the inattentive dimension), predominantly hyperactive/impulsive type (meeting criteria on the hyperactive/impulsive dimension), and combined type (meeting criteria on both dimensions). In addition, there are other general criteria including the onset of symptoms before 7 years of age, the presence of symptoms for at least 6 months, the presence of symptoms in two or more settings (e.g., home, school, or work), and evidence that

Table 1 DSM-IV behaviors for ADHD

Inattention
Careless mistakes
Difficulty sustaining attention
Seems not to listen
Fails to finish tasks
Difficulty organizing
Avoids tasks requiring sustained attention
Loses things
Easily distracted
Forgetful
Hyperactivity
Fidgeting
Unable to stay seated
Moving excessively (restless)
Difficulty engaging in leisure activities quietly
'On the go'
Talking excessively
Impulsiveness
Blurting answers before questions completed
Difficulty awaiting turn
Interrupting/intruding upon others

the symptoms cause significant clinical impairment in social, academic, or occupational functioning. The WHO condition has been renamed 'disturbances of activity and attention.'

Treatments Other Than Diet

In considering dietary interventions, it is important to note that there are two other forms of treatment with proven efficacy. These are stimulant medications and behavior modification. Considerations about dietary interventions have to be considered in the context of these other interventions. The nature of the main beneficial treatments, stimulant medication and behavioral interventions, makes the issue of diagnostic criteria for ADHD extremely important. Both of these treatments are not specific for the disorder so that the determination about which children are treated is very dependent on who is diagnosed.

The stimulant medications consist of methylphenidate (Ritalin), dextroamphetamine (Dexedrine), and pemoline (Cylert). They are particularly popular in the United States because they represent safe, effective, and low-cost treatment. A review of numerous studies has shown that stimulants improve the core behaviors of inattention, impulsiveness, and hyperactivity for the duration of action of the medication, as well as providing temporary improvement of associated features including aggression, social interaction, and academic productivity. The margin of safety is very high, and the side effects on appetite, sleep, and, infrequently, tics or bizarre behavior are all reversible when the medication is stopped. The concern about growth has proved to be insignificant, and although abused by adults, the stimulants are rarely abused by the children who take them because they usually do not find taking the medication pleasurable. While there is no long-term evidence that the use of stimulant medication or behavioral interventions on their own have any long-term benefits, there is evidence of long-term benefits when they are used in combination.

Effective behavioral interventions have generally consisted of direct contingency management programs (e.g., point or token programs or a response cost program) and social skills training. Like stimulant medication, these interventions are not specific to ADHD and have no proven long-term benefit when used in isolation. Other approaches, such as traditional psychotherapy and play therapy, have not been found to be effective with this group of children. Likewise, cognitive behavioral techniques, where a therapist teaches a child to control his or her behavior, have usually not been effective for children with ADHD because of the difficulty these children experience in generalizing the techniques beyond the therapeutic sessions.

Dietary Interventions

The concept that specific dietary components may adversely affect behavior has rested on three hypotheses:

1. Oligoallergenic diet
2. Sugar restriction
3. Feingold diet.

The idea that food might have an adverse effect on behavior was first raised in 1922 by Shannon. This concept was further elaborated in 1947 by Randolph in his description of the 'tension fatigue syndrome,' a behavioral extension of the vomiting reaction to milk proteins, and was also promoted by Speer. Their theory suggested that some children have atypical allergic reactions to various foods, consisting of subtle and behavioral effects. Their treatment entailed placing a child on a restricted diet and then adding foods one at a time to determine which foods caused an adverse reaction. This has been referred to as the oligoallergenic diet by a recent clinical/research group.

A specific focus on sugar as a nutrient adversely affecting behavior first appeared in the 1970s, with a study reported by Langseth and Dowd. Among 271 hyperactive children, these authors found a large number of children who, during glucose

tolerance tests, had patterns of blood glucose levels similar to the pattern seen in adults with functional reactive hypoglycemia. Similar results have also been found in aggressive criminal offenders. A subsequent study showed that the patterns that Langseth and Dowd found can be normal variations in childhood, but the Langseth and Dowd study was followed by two correlational studies that suggested an association between sugar intake and hyperactivity. The hyperactive children who consumed more sugar displayed more hyperactive and aggressive behavior.

The third dietary intervention suggested to improve behavior was proposed by Dr. Benjamin Feingold in 1975. He reported that at least 50% of hyperactive and learning-disabled children improved when placed on diets that were salicylate and additive free. Over time, the three dietary interventions have been combined so that proposed dietary restrictions now tend to incorporate all three in their recommendations. However, it is useful to examine the scientific evidence for each of these three dietary interventions.

Objective Standards

In discussing the evidence for the efficacy of dietary interventions in improving behavior in children, it is first important to review the concepts important to prove efficacy. The major point to emphasize is that it is impossible to prove the null hypothesis. It is virtually impossible to prove definitively that no relationship exists between dietary constituents and behavior or cognitive function. This is because it is impossible to test every possible variation or type of child. Therefore, a realistic approach needs to be similar to that taken by the US Food and Drug Administration (FDA) for the criteria it requires to license a new medication. Basically, pharmaceutical companies are required to demonstrate that a new medication is both efficacious and does not cause significant harm. It is not the role of the FDA to disprove the efficacy of a drug treatment. With dietary interventions, they should not be recommended as a primary intervention for behavioral problems until there is clear evidence of their efficacy.

The main criteria required to evaluate objectively the efficacy of psychotropic medications are presented in **Table 2**. It is useful to use these criteria to evaluate the scientific merit of any studies on interventions affecting behavior. It is also important to examine the pattern of results of multiple studies from different research groups. Ideally, where other efficacious therapies are available (e.g., stimulant medication and behavior modification for children with ADHD), the proposed therapy should be compared with existing therapies. This latter examination, by and large, has not been undertaken with any of the three dietary interventions.

Table 2 Objective study criteria

Uniformity of subjects
Standard doses
Objective verifiable dependent measures
Control group
Placebo
Double-blind

Study Designs

There are two designs that can be employed to study the effects of nutrients on behavior. The most commonly employed design is the challenge study. This first places the children on the diet under study for a period of time and then challenges them with a food containing the offending agent (e.g., sucrose or tartrazine) or a food that does not contain the offending agent but looks and tastes identical to the offending agent, referred to as a placebo. This is the most commonly employed design because it is the easier and less expensive type to complete. The other design develops diets that appear similar, but the diets differ in what they contain (e.g., sugar or artificial sweeteners). In both designs the children, their families, and the researchers need to be blind about which diet or challenge food the children receive at any given time. In most studies, the children are used as their own controls (crossover studies). They are able to receive both diets or challenges in a sequence because the diets are not believed to result in permanent changes lasting once the diet is stopped.

The measures used to assess the effects (dependent measures) are then completed within the few hours after a challenge or repeatedly while the children are on diets. While parents, clinicians, and teachers are utilized as observers (completing behavior rating scales), ideally multiple measures are employed including some that are by independent observers or include objective assessments (e.g., performance on a continuous performance test and measuring activity level). Finally, it is important not to base results on one study. There need to be multiple studies performed by different groups of researchers, and a clear pattern of effects should emerge. When a number of studies have been completed, it is possible to combine them statistically with such techniques as meta-analysis to gain a more definitive picture.

Oligoallergenic Diet

While this is the oldest of the three dietary interventions, few controlled studies meeting the objective criteria outlined previously have been undertaken. Five investigations have studied the effects of placing children on restricted diets. These studies all included restricting the dietary intake of additives and simple sugars. The studies found beneficial effects from placing children on restricted diets compared with a placebo diet, or they found worsening behaviors in children on the restricted diets when they were challenged with offending foods compared with placebo challenges. In all but one of the studies, the only successfully completed dependent measure was behavior rating scales completed by the parents. While these are important measures and are collected in most studies, the raters are not independent of the children's behaviors. One study had multiple measures, but only those of the parents and physician found a significant difference between the offending agent and placebo challenges. More extensive research by additional research groups and additional independent measures are required to document the efficacy of this intervention before a decision can be made about its efficacy. Since the initial diet is extremely restrictive, care must be taken to make sure that the diet is adequately balanced and contains adequate nutrition.

Sugar Restriction

Sugar restriction usually refers to limiting the amount of sucrose in the diet. While most of the studies examined sucrose restriction, some also examined restriction of fructose or glucose. The artificial sweetener employed as a placebo was most frequently aspartame, but several studies used saccharin or both aspartame and saccharin as separate conditions. The type of sweetener used did not seem to affect the results.

Sugar restriction has been studied as a treatment for children since 1982. There have been a total of 23 appropriate objective studies contained within 16 reports employing a wide variety of types of children, including children with ADHD and aggression as well as normal children, and varying in age from preschool children to adolescents. All of the studies with two exceptions were challenge crossover studies in which children were challenged with drinks containing either sugar (sucrose in most studies) or an artificial sweetener (mostly aspartame). The other two studies consisted of giving the children diets that were high in sucrose content or low in sucrose and sweetened with aspartame or saccharin. A meta-analysis of the 23 studies did not find any significant behavioral or cognitive effects from sugar. There were not enough studies to reach a definitive conclusion, and there was insufficient statistical power to detect small effects or to detect effects on a small subset of children. To date, there is not enough evidence to warrant the recommendation to restrict a child's sugar intake for the purpose of improving the child's behavior or cognitive functioning.

Feingold Diet

The Feingold diet restricts foods with dyes, preservatives, and salicylate compounds. Investigations of this diet, which were reviewed in 1986, generally involved children with ADHD. In most of the studies the children were kept on an additive-free diet and then challenged with a food containing an additive or an additive-free food as placebo. Two studies used additive-containing and additive-free diets. A problem in comparing studies was the variation in type and dose of additives used. There were a total of 13 controlled studies. The summation of the findings found little, if any, effect. At best, there was some suggestion that a small percentage of children (1%) were adversely affected by additives. However, a recent study found that 24 of 34 children referred for hyperactivity (no formal diagnosis was established) who responded in an open clinical trial to an additive-free diet responded adversely to challenges with varying doses of tartrazine compared with placebo, whereas all except 2 of 20 in a comparison group did not. The dependent measures were two behavior rating scales completed by the parents. There appeared to be a dose response that would be contrary to a usual allergic response. This is a much higher rate of response than found in any previous study including those using tartrazine. Further study is required to substantiate these results since they run contrary to most of the previous research. Overall, the evidence to date does not confirm the efficacy of the Feingold diet to warrant its promotion as a treatment for most children with behavioral problems. In addition, if the diet is strictly maintained including foods containing salicylate compounds, the diet may be deficient in vitamin C.

Potential Side Effects of Diets

With all the diets, maintaining compliance may be difficult. Children who have behavioral problems are generally less likely to be compliant, and it can require a major effort to maintain the diet, detracting from efforts to control other areas of behavior. Diets are also problematic because they require the children to eat foods different from their peers. In

children who are already singled out as different, this can further reduce their self-esteem. Care has to be taken to weight the benefits of diets with as yet objectively unproved effects against the potential harm and difficulties in administering them.

Conclusions

ADHD is a mental disorder and its diagnosis is based on a child manifesting the symptoms of inattention, hyperactivity, and impulsiveness to the extent that the symptoms impair the child's ability to function. The main beneficial treatments are two nonspecific treatments, stimulant medication and behavioral interventions. While neither alone has any proven long-term benefits, there is evidence that the combination of both treatments does have some long-term benefits.

Dietary interventions have included (1) restriction of allergenic foods starting with a generally restricted diet and adding those foods that do not worsen the child's behavior, (2) restriction of food additives and preservatives referred to as the Feingold diet, and (3) restriction of sugar. These dietary interventions have not been proved to be efficacious and more study is required to determine their effects.

See also: **Food Allergies**: Diagnosis and Management. **Sucrose**: Nutritional Role, Absorption and Metabolism.

Further Reading

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Washington, DC: American Psychiatric Association.

Baumgaertel A, Copeland L, and Wolraich ML (1996) Attention deficit hyperactivity disorder. In: Wolraich ML (ed.) *Disorders of Development and Learning*, 2nd edn. St Louis: Mosby-Yearbook.

Egger J, Stolla A, and McEwen LM (1992) Controlled trial of hyposensitisation in children with food induced hyperkinetic syndrome. *Lancet* **339**: 1150–1153.

Pelham WE and Sams SE (1992) Behavior modification. *Child and Adolescent Psychiatric Clinics of North American* **1**: 505–517.

Sprague RL and Werry JS (1971) Methodology of psychopharmacological studies with the retarded. *International Review of Research into Mental Retardation* **5**: 147–157.

Swanson JM, McBurnett K, Wigal T *et al.* (1993) Effects of stimulant medication on children with ADD: A review of reviews. *Exceptional Children* **60**: 154–162.

Wender EH (1986) The food additive-free diet in the treatment of behavior disorders: A review. *Journal of Developmental and Behavioral Pediatrics* **7**: 35–42.

Wolraich ML, Lindgren SD, Stumbo PJ *et al.* (1994) Effects of diets high in sucrose or aspartame on the behavioral and cognitive performance of children. *New England Journal of Medicine* **330**: 301–307.

Wolraich ML, Wilson DB, and White JW (1995) The effect of sugar on behavior or cognition in children: A meta-analysis. *Journal of the American Medical Association* **274**: 1617–1621.

World Health Organization (1992) *The ICD-10 Classification of Mental and Behavioural Disorders*. Geneva: World Health Organization.

HYPERLIPIDEMIA

Contents

Overview

T R Trinick and E B Duly, Ulster Hospital, Belfast, UK

Normal Lipid Metabolism

Lipids are a heterogeneous group of substances soluble in organic solvents but insoluble in water. They are largely intracellular but circulate in blood as lipoprotein particles. There are four general functions for lipids:

- Structural components of membranes
- Storage forms of metabolic fuel
- Transport forms of metabolic fuel
- Protective functions as an outer coating of the organism

Lipids consist of cholesterol and its derivatives, fatty acids, triacylglycerols, phospholipids, and

apolipoproteins. The lipoprotein particle has a core of neutral lipids (cholesterol esters and triacylglycerol) and a surface coat of polar lipids (unesterified cholesterol and phospholipids) and apolipoproteins. They are classified in terms of density. The following are the main lipoproteins:

- Chylomicrons
- Very low-density lipoprotein (VLDL)
- Immediate-density lipoprotein (IDL)
- Low-density lipoprotein (LDL)
- High-density lipoprotein (HDL)

Synthesis of lipoproteins occurs in the intestine or liver. They are then modified by enzymes and taken up by cell surface receptors in processes largely regulated by the apolipoproteins. A series of receptors, transporters, and enzymes are important in lipoprotein metabolism and function as detailed later. The physicochemical characteristics of the main lipoprotein classes are shown in **Table 1**.

Interest in lipids lies in circulating lipid concentrations and their relationship to atherosclerosis, particularly coronary heart disease, stroke, and peripheral vascular disease.

Cholesterol

Cholesterol is a sterol with the structure shown in **Figure 1**. Daily cholesterol intake is 0.5–1.0 g, half of which is absorbed. On a low-cholesterol diet (<300 mg/day) the body synthesizes approximately 800 mg of cholesterol per day, mainly in the liver and, to a lesser extent, in the intestine.

The rate-limiting step in synthesis is highly sensitive to cellular levels of cholesterol, themselves sensitive to circulating levels of cholesterol. This feedback regulation occurs through changes in the amount and activity of an enzyme called 3-hydroxy-3-methylglutaryl CoA reductase (HMG-CoA reductase), which catalyzes the formation of mevalonate, the rate-limiting step in cholesterol biosynthesis. The rate of synthesis of HMG-CoA reductase mRNA is controlled by the sterol regulatory element binding protein (SREBP). SREBP in its inactive state is attached to the endoplasmic reticulum or nuclear membrane, but when cholesterol levels decline the amino-terminal domain is released from its association with the membrane by proteolytic cleavage; it migrates to the nucleus and binds to the sterol regulatory element (SRE) on the 5′ side of the reductase gene to enhance transcription. As cholesterol levels increase, the proteolytic release of SREBP is blocked, SREBP in the nucleus is rapidly degraded, and cholesterol synthesis is switched off.

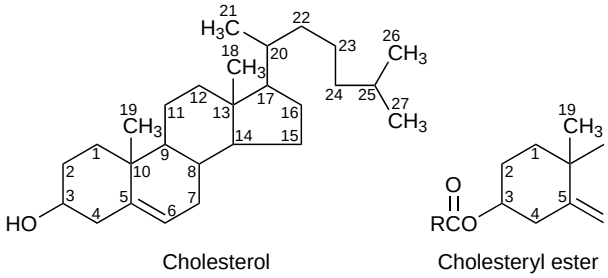


Figure 1 Structure of cholesterol and cholesteryl ester.

Cholesterol is found in the body largely as free cholesterol in membranes, but in the plasma it is two-thirds esterified, mainly as cholesterol linoleate and cholesterol oleate. Free cholesterol in plasma exchanges freely with cholesterol in membranes. The major route of cholesterol excretion is through the bile, directly as cholesterol or after conversion to bile salts, some of which are reabsorbed from the terminal ileum in the enterohepatic circulation.

Triacylglycerol

Triacylglycerols are glycerol molecules esterified with three fatty acid molecules (**Figure 2**). Diacylglycerols and monoacylglycerols have two and one fatty acid molecules, respectively. Triacylglycerols constitute the main energy storage form in mammals and are the main storage form of fatty acids.

Table 1 Physicochemical characteristics of the major lipoprotein classes

Lipoprotein	*Density (g/ml)*	*Molecular weight ($Da \times 10^6$)*	*Diameter (nm)*	*Triacylglycerol (% lipid)*	*Cholesterol (% lipid)*	*Phospholipid (% lipid)*	*Source*
Chylomicrons	0.95	>400	75–1200	80–95	2–7	3–9	Intestine
VLDL	0.95–1.006	10–80	30–80	55–80	5–15	10–20	Liver
IDL	1.006–1.019	5–10	25–35	20–50	20–40	15–25	Catabolism of VLDL
LDL	1.019–1.063	2.3	18–25	5–15	40–50	20–25	Catabolism of IDL
HDL	1.063–1.21	1.7–3.6	5–12	5–10	15–25	20–30	Liver, intestine

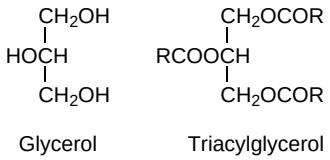


Figure 2 Structure of glycerol and triacylglycerol. 'R' denotes the position of a fatty acid within the triacylglycerol.

Fatty Acids

Fatty acids can be present as triacylglycerol, as part of lipoprotein particles, and as free fatty acids (bound to albumin). Common fatty acids and their sources are listed in **Table 2**.

Fatty acids are straight-chain compounds of differing lengths connecting a hydrocarbon group to a hydroxyl group. With only single bonds in the straight chain, the fatty acid is saturated; with one or more additional double bonds, the fatty acid is unsaturated. Fatty acids with only one double bond are said to be monounsaturated (e.g., oleic acid, C18:1), whereas fatty acids with two or more double bonds are said to be polyunsaturated (e.g., arachidonic acid, C20:4). The presence of a double bond allows there to be two isomers, depending on whether the hydrogen atoms attached to the carbon atoms on either side of the double bond lie on the same side (*cis*) or opposing sides (*trans*). *Cis* isomers are the only naturally occurring isomers and form kinks in the fatty acid chain. *Trans* isomers occur as part of food processing and maintain the straight direction of fatty acid chains. The common saturated fatty acids are palmitic (C16:0) and stearic (C18:0) acids.

Table 2 Fatty acids and their sources

Fatty acid	*Structure*	*Source*	*Melting point (°C)*
Saturated			
Lauric	C12:0	Coconut oil, palm kernel oil	44
Palmitic	C16:0	Palm oil, milk, butter, cocoa, butter, beef, pork, lamb	63
Stearic	C18:0		69
Behanic	C22:0	Some seed oils, especially peanut	80
Lignoceric	C24:0		84
Unsaturated			
Oleic	C18:1	Olive oil, most commonly occurring fatty acid	11
Linoleic	C18:2	Corn oil, soya bean oil, sunflower oil and sunflower seed oil	−5
Linolenic	C18:3	Linseed oil	−11
Arachidonic	C20:4	Fish oils	−50
Eicosapentenoic	C20:5	Cod, salmon, pilchard, mussel, oyster	−54
Docosahexenoic	C22:6		

From Durrington PN (2004) *Hyperlipidaemia: Diagnosis and Management*. London: Hodder Arnold.

Diets rich in omega-3 polyunsaturated fatty acids (n-3 PUFAs), such as α-linoleic acid, eicosapentanaenoic acid, and decosahexaenoic acid, are associated with less coronary heart disease, and conjugated linoleic acids have beneficial effects against atherosclerosis. n-3 PUFAs function mainly by changing membrane lipid composition, cellular metabolism, signal transduction, and regulation of gene expression. It is postulated that receptors exist for fatty acids or their metabolites that are able to regulate gene expression and affect metabolic or signalling pathways associated with coronary heart disease. Three nuclear receptors are thought to be fatty acid receptors that respond to dietary and endogenous ligands: peroxisome proliferator activated receptors, retinoid X receptors, and liver X receptors.

Phospholipids

The common phospholipids in plasma are derived from glycerol and consist of triacylglycerol containing phosphate and a nitrogenous base (glycerophospholipids). The phosphate group is usually attached at position 3 of the glycerol molecule, and the nitrogenous base is usually an amino acid or an alcohol. The phosphatidyl cholines (lecithins) are the most common phospholipid and are found in plasma and in cell membranes. Lecithin–cholesterol acyl transferase (LCAT) catalyzes the transfer of a fatty acyl group at position 2 on glycerol to cholesterol to produce cholesteryl ester and leaves monoacyl glycerophosphate (lysolecithin). Another class of phospholipids, the cephalins, includes phosphatidyl ethanolamine, phosphatidyl serine, and phosphatidyl inositol.

Phospholipids are able to bridge nonpolar lipids and water and act to allow lipids to mix with water in an emulsion. The nonpolar hydrocarbon end of the phospholipid is attracted to lipid, whereas the polar phosphate group is attracted to water. In a lipid droplet, the inner oily centre is surrounded by phospholipid, which has its outer phosphate group attracted to the surrounding water environment, to form a micelle.

Apolipoproteins A, B, C, and E

The lipoprotein particle (VLDL, LDL, and HDL) is composed of lipid and protein molecules. Among the protein molecules are a group of proteins found at the surface of the lipoprotein particle called apolipoproteins. Their function is integral to the metabolism of lipoproteins. They interact with phospholipids to solubilize cholesterol esters and triacylglycerol, regulate the reaction of enzymes (LCAT, lipoprotein lipase, and hepatic lipase) with lipid, and bind with cell surface receptors to determine the metabolism of lipoproteins.

Apolipoprotein A This is the main protein of HDL and has two forms, apoA-I and apoA-II. ApoA-I is the main protein component in HDL, and the production and catabolism of apoA-I determine the plasma concentration of HDL cholesterol. It acts as an activator of LCAT, which is responsible for esterification of free cholesterol in plasma, and allows the binding of HDL to many cell surfaces. ApoA-II is a structural component of HDL.

Apolipoprotein A-I Milano ApoA-I Milano is a specific form of apoA-I seen in some Italian families, which appears to protect against the development of atherosclerosis.

Apolipoprotein B ApoB-100 is the main protein component of LDL and is synthesized in the liver. It is also found in chylomicrons and VLDL. ApoB-48 is synthesized from the intestine and is the amino-terminal half of apoB-100 synthesized from the same gene. ApoB-100 is the receptor ligand for the LDL receptor.

Apolipoprotein C ApoC is composed of three separate apolipoproteins. ApoC-I is mainly found in VLDL but also in chylomicrons and HDL. ApoC-II is present in a circulating reservoir of HDL, transferring to chylomicrons and VLDL, where it acts as an activator of lipoprotein lipase, allowing the lipolysis of triacylglycerols from circulating triacylglycerol-rich lipoproteins. ApoC-III is the most abundant form of apoC and may act as a modulator of lipoprotein lipase.

Apolipoprotein E ApoE is a glycoprotein with several isoforms designated as apoE-2, -E-3, and -E-4. ApoE-3 is the most common isoform. It is present in VLDL, IDL, and HDL (mainly HDL_2). ApoE facilitates chylomicron remnant metabolism through the chylomicron remnant and VLDL receptors of the liver. ApoE-3 and -E-4 bind avidly with hepatic receptors, whereas apoE-2 is poorly bound. Patients with only apoE-2 isoform clear chylomicron remnants and IDL slowly, and apoE-2 is associated with dysbetalipoproteinemia (type III hyperlipoproteinemia). ApoE also facilitates metabolism through the LDL receptor (particularly the apoE-4 isoform). A large number of tissues express mRNA for apoE, including the brain, although the reason for this is unclear.

Apolipoprotein (a) Apo(a) joined together with one LDL particle, which contains apoB, constitutes a lipoprotein called Lp(a). Interest in Lp(a) arose because apo(a) shows close sequence homology with plasminogen, suggesting that a high level of Lp(a) would impair thrombolysis. Lp(a) is an independent risk factor for developing vascular disease, with levels above a cutoff value of 300 mg/l placing individuals at risk, especially if combined with other risk factors.

Lipoproteins

The main function of the lipoproteins is to transport lipids from one organ to another. Their main characteristics are shown in **Table 1**.

Chylomicrons These are the largest lipoproteins, consisting mainly of triacylglycerol with apoB-48 and apoA, -C, and -E. Triacylglycerol is hydrolyzed with endothelial-bound lipoprotein lipase, changing the chylomicron into a chylomicron remnant rich in cholesteryl ester. These remnants are removed from the circulation by interaction with the remnant receptors mainly present on hepatocytes. Peak chylomicronemia occurs 3–6 h after a meal, with a half-life of less than 1 h, and is cleared from the circulation after a 12-h fast.

Very low-density lipoproteins These triacylglycerol-rich lipoproteins are secreted mainly by the liver, with apoB-100 and apoE on their surface, whereas some VLDLs are synthesized by the gut. They are transformed into mature VLDLs by accumulating cholesterol ester, apoC, and apoE from HDLs. They then either interact with lipoprotein lipase to convert into IDLs, which can be taken up by the liver, or convert to LDLs by interacting with hepatic triglyceride lipase.

VLDL particles vary in size. Small VLDL is converted into LDL, via IDL, to a greater extent than large VLDL, which is converted to a form of IDL that appears to be removed from the plasma before conversion to LDL.

Intermediate-density lipoproteins IDLs are intermediate particles formed from the conversion of VLDL to LDL. Also known as VLDL remnants, some are removed directly from plasma, whereas some convert into LDL.

Low-density lipoproteins LDL is the major cholesterol-carrying particle in the plasma. The core is cholesterol ester and has one apolipoprotein, apoB-100, per LDL particle. There are different sizes of LDL. Approximately one-third of the intravascular pool is catabolized per day and three-fourths of the circulating LDL is cleared through the liver, mainly through the LDL receptor. Small, dense LDL is more common in some dyslipidemias and may be more easily oxidized than larger LDL. Normal LDL does not cause foam cell formation, but lipid peroxidation of LDL makes the LDL a ligand for certain receptors (the scavenger receptor and perhaps a specific receptor for oxidized LDL) and results in the formation of cholesterol-laden foam cells. In addition, oxidized LDL in the cell wall stimulates the production of cytokines and growth factors, resulting in monocyte recruitment and the proliferation of smooth muscle cells. This mechanism underlies one model of atherogenesis.

High-density lipoproteins Nascent HDL is secreted by the liver and gut. It acquires unesterified cholesterol in the circulation, catalysed by LCAT to cholesteryl ester. HDL can pass cholesteryl ester to VLDL in exchange for triacylglycerol, facilitated by cholesterol ester transfer protein (CETP), or HDL can be taken up by the liver directly. The idea that HDL protects against coronary heart disease (CHD) comes from epidemiological studies. A 0.026 mmol/l increase in plasma HDL cholesterol decreases CHD risk by 2% in men and 3% in women.

Enzymes and Transfer Proteins

Acylcoenzyme A Cholesterol acyltransferase (ACAT; EC 2.3.1.26) ACAT-1 and ACAT-2 are membrane-bound proteins responsible for cholesterol ester formation, metabolizing excess cholesterol within cells to cholesterol ester, which is allosterically activated by cholesterol.

Adenosine-binding cassette transporter In peripheral tissues, adenosine-binding cassette transporter (ABCA-1) protein facilitates transfer of intracellular cholesterol out of cells to lipid-poor apoA-1 or pre-β HDL particles. When it is deficient or inactive, cholesterol accumulates in peripheral tissues as in Tangier disease or familial HDL deficiency.

Cholesterol Ester transfer protein CETP mediates the exchange of cholesteryl ester from HDL with triacylglycerol from VLDL or chylomicrons.

Fatty acid binding protein Fatty acid binding proteins (FABPs) play a role in the solubilization of long-chain fatty acids (LCFAs) and their CoA-esters to various intracellular organelles. FABPs serve as intracellular receptors of LCFAs and are involved in ligand-dependent transactivation of peroxisome proliferator-activated receptors (PPARs) in trafficking LCFAs to the nucleus.

Hepatic lipase (EC 3.1.1.3) Hepatic lipase (HL) is an endothelial-bound enzyme that removes triacylglycerol from lipoproteins in the metabolism of chylomicrons, VLDL, and HDL. HL hydrolyzes HDL triacylglycerol and phospholipids to form HDL_3 from HDL_2, contributing to the process of HDL regeneration in the reverse cholesterol transfer process.

Lecithin-cholesterol acyltransferase (EC 2.3.1.43) LCAT mediates the esterification of cholesterol by transferring a fatty acid from lecithin to cholesterol to form cholesteryl ester.

Lipoprotein lipase (EC 3.1.1.34) Lipoprotein lipase and hepatic lipase are endothelial-bound enzymes that remove triacylglycerol from lipoproteins. Lipoprotein lipase is activated by apoC-II and is involved in catabolism of chylomicrons and VLDL. Endothelial lipase, lipoprotein lipase, and hepatic lipase belong to the same gene family.

Microsomal triglyceride transfer protein Microsomal triglyceride transfer protein is present in enterocytes and hepatocytes, and it is responsible for adding neutral lipid to apoB to protect it from ubiquitinylation and degradation.

Phospholipid transfer protein Phospholipid transfer protein transfers phospholipids from other lipoproteins to HDL, contributing to the functionality of HDL.

Sterol regulatory element binding protein SREBP is a protein that binds with part of the LDL receptor promoter to increase cholesterol synthesis.

Receptors

A large number of lipoprotein receptors have been identified. Some of the more important receptors are discussed here. Lipoprotein uptake at the cell

membrane may be non-receptor-mediated, perhaps by pinocytosis, where 'binding' is of low affinity but is not saturable.

LDL receptor The LDL receptor (LDLR) is a transmembrane glycoprotein present on most cell surfaces, encoded on chromosome 19. Free cholesterol, building up in the cell through the receptor, reduces both cell synthesis of cholesterol and cell uptake of more LDL cholesterol.

LDL receptor-related protein The LDL receptor-related protein (LRP) is a multifunctional receptor (binding VLDL/chylomicron remnants and other nonlipid ligands such as bacterial toxins) present in nearly all tissues. It has a high affinity for apoE and a low affinity for apoB-100.

VLDL receptor This receptor binds VLDL, β-VLDL, and IDL. It recognizes apoE and is located mainly in adipose tissue and muscle.

Scavenger receptors These receptors are found on macrophages and hepatic endothelium. They bind and degrade chemically modified LDL, such as oxidized or acetylated LDL. They are not downregulated by intracellular cholesterol accumulation. Hepatocellular uptake of HDL and/or its cholesteryl ester content is facilitated by a scavenger receptor and a HDL receptor.

Other remnant receptors The lipolysis-stimulated receptor found on fibroblasts recognizes surface apoE and takes up VLDL, chylomicrons, and LDL. Two membrane-binding proteins (MBP 200 and MBP 235) have been described on macrophages and appear to bind VLDL. Remnants from both chylomicrons and VLDL (after hydrolysis of more than 70% of their triacylglycerol content) appear to be removed by both the LDL and the LRP receptors.

Peroxisome proliferator-activated receptors PPARs are a family of intranuclear receptors, including PPARα and PPARδ, that regulate a variety of genes involved in lipid metabolism, thrombosis, and inflammation.

Exogenous (Dietary) Lipid Pathways

Ingestion of food containing fat (triacylglycerol) and cholesterol results in absorption into the enterocyte of fatty acids, monoacylglycerols, free cholesterol, and lysolecithin. In the enterocyte, reesterification of fatty acids into triacylglycerol and cholesterol into cholesteryl ester occurs to form chylomicrons, to which is added a surface layer of apoB-48, -A-I, -A-II, and -A-IV, phospholipid, and free cholesterol. This allows secretion of the chylomicron into the intestinal lymphatics. ApoB-48 is required for secretion of the chylomicron. ApoB-48 is a truncated form of apoB-100, synthesized in the liver but missing the LDL receptor-binding domain of apoB-100. The action of the apoB-editing enzyme in enterocytes changes a nucleotide base in apoB mRNA to a stop codon. There is one apoB-48 per intestinal triglyceride-rich particle.

Chylomicrons in the circulation take up apoC from HDL (releasing it back to HDL later) and acquire apoE. ApoC-II allows the chylomicron to activate lipoprotein lipase on capillary endothelial cells of muscle and fat. This allows hydrolysis of triacylglycerol, releasing glycerol and fatty acids to be taken up by local tissue. Surface phospholipids, free cholesterol, and apoC transfer to HDL as the particle shrinks. This small chylomicron is called a chylomicron remnant and is catabolized through the LDL receptor and other remnant receptors on the liver. This transport of dietary lipid from the intestinal to the peripheral tissues is shown in **Figure 3**.

Endogenous Lipid Pathways

The liver is the main source of endogenous lipid (**Figure 4**). In particular, the liver secretes the triacylglycerol-rich lipoprotein VLDL. Triacylglycerol, which is formed from fatty acids either newly synthesised or taken up from plasma, together with free cholesterol, synthesised from acetate or delivered to the liver in chylomicron remnants, join with apoB and phospholipids to form VLDL. ApoC and apoE are added in the circulation. Triacylglycerol is progressively removed from VLDL in the same way as occurs with chylomicrons. Free cholesterol transfers to HDL and is esterified with LCAT and transferred back to VLDL, using a protein called cholesteryl ester transfer protein (CETP), in exchange for triacylglycerol transfer from VLDL to HDL. In this way, VLDL becomes smaller and transforms to become IDL, although some small VLDLs may be removed directly. IDL is further changed through interaction with hepatic lipase to LDL. In this way, most VLDL is transformed to LDL.

Reverse Cholesterol Transport

Lipids are transported to the peripheries from the gut and the liver. They return to the liver via HDL in a process known as reverse cholesterol transport (**Figure 5**). HDL particles arise in the liver and gut from a coalescence of apoA-I and phospholipid to

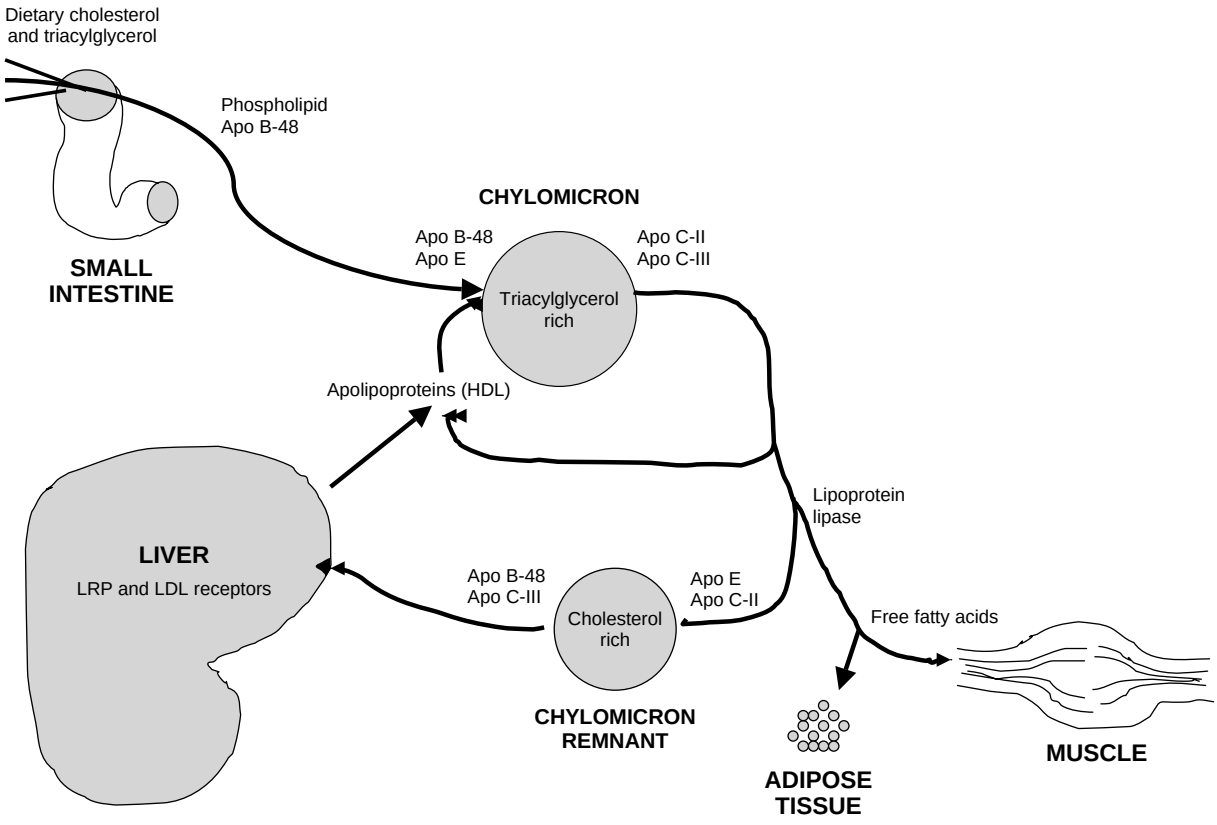


Figure 3 Exogenous (dietary) lipid pathway. This shows the transport of dietary lipid from intestine to peripheral tissues and liver. Movement of apolipoprotein between high-density lipoprotein (HDL) and chylomicrons is shown. LRP, low-density lipoprotein (LDL) receptor-related protein.

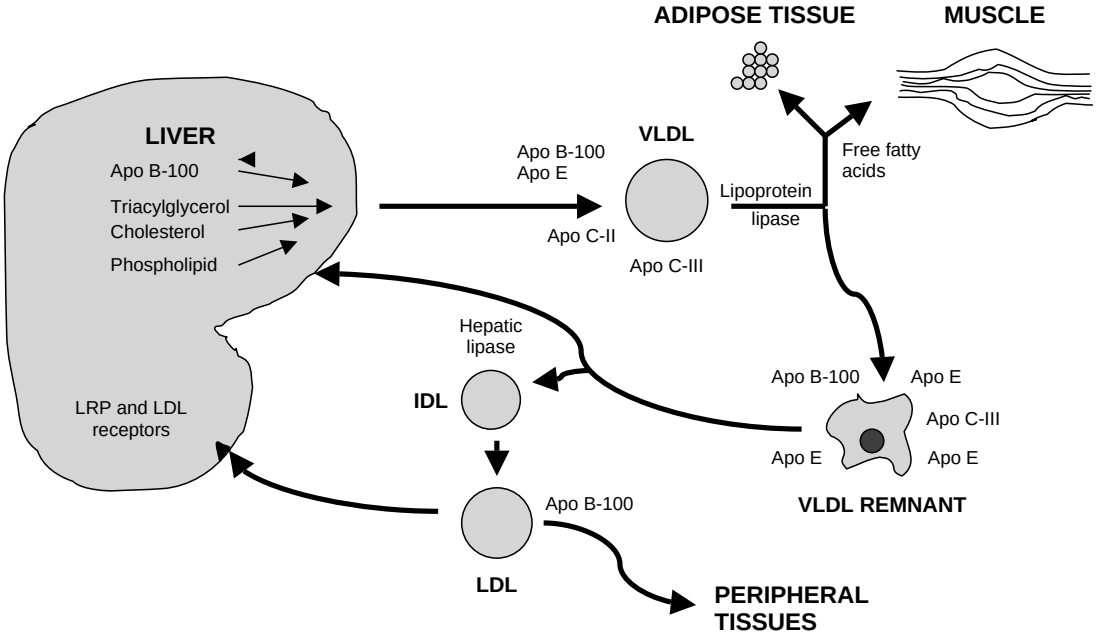


Figure 4 Endogenous lipid pathway. This shows the formation of very low-density lipoprotein (VLDL) lipid particles ($VLDL_1$ and $VLDL_2$) in the liver with the interconversion, through the action of lipoprotein lipase, to VLDL remnant and through immediate-density lipoprotein (IDL) to LDL. Lipids are taken up from LDL both peripherally and in the liver. LRP, low-density lipoprotein receptor-related protein.

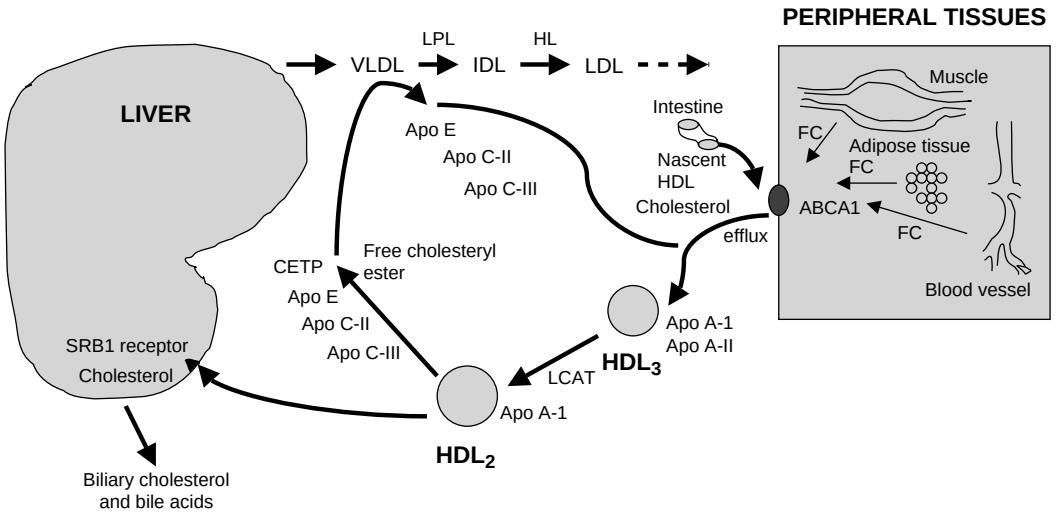


Figure 5 Reverse cholesterol transport. Nascent high-density lipoprotein (HDL_3) picks up free cholesterol from the peripheries to become HDL_2, by a lecithin–cholesterol acyl transferase (LCAT)-mediated conversion. Cholesterol is then transported to the liver with uptake by the SRBI receptor. A second method of transport to the liver involves CETP-mediated esterification of HDL and conversion into immediate-density lipoprotein (IDL) and low-density lipoprotein (LDL), which is then taken up by the LDL receptor. This transfer of lipid between HDL_2 and VLDL/IDL maintains a cycle within HDL, and IDL/LDL deliver cholesterol from the peripheries to the liver. HDL may also deliver cholesterol directly to the liver. FC, free cholesterol.

form cholesterol-deficient bilayered discs in the form of HDL_3. Circulating HDL particles, particularly a subset of HDL_3 called pre-β HDL or lipid-poor apoA-I, come into contact with cells, and ABCA-1 acts to move free cholesterol from the cell surface and out of the cells. This cholesterol is converted by LCAT to cholesteryl ester and moves into the core of the HDL, forming mature cholesterol-rich HDL. After accumulating cholesterol, the HDL starts to accept other apolipoproteins and becomes HDL_2. In turn, HDL_2 appears to pass cholesteryl ester to triglyceride-rich lipoproteins such as chylomicrons, chylomicron remnants, and VLDL under the influence of CETP. The cholesterol then finds its way back to the liver in the form of chylomicron remnants, IDL, and LDL. Some of the HDL_2 particles may lose cholesterol directly to the liver and some may be taken up directly by the liver.

Consequences of Hyperlipidemia

Clear evidence exists that as serum cholesterol rises, the risk of CHD rises, and as serum cholesterol falls, the risk of developing CHD falls. The epidemiological evidence comes from within-country studies, between-country studies, and migration studies. Support comes from animal studies and there is evidence of the beneficial effects of reducing serum cholesterol in both primary and secondary prevention of ischemic heart disease.

The within-country studies include the Multiple Risk Factor Trial Intervention (MRFIT) study, which followed 360 000 middle-aged men screened and followed up for CHD mortality. MRFIT showed a strong positive correlation between cholesterol levels at initial screening and later death from CHD. The Framingham Heart Study, started in 1949, is another prospective survey that followed a large cohort of Americans and examined lipid levels and risk of CHD, particularly the relationships between lipoprotein fractions and CHD. It showed a strong association between elevated LDL cholesterol and increased incidence of CHD and an inverse association between HDL cholesterol and CHD risk. Framingham has drawn attention to the value of the ratio of total cholesterol to HDL cholesterol, where a ratio of 3 or less suggests the disease is static and a ratio of 4 or higher suggests the disease is progressive. Framingham also drew attention to the incremental effect of additional risk factors in the development of CHD, such as hypertension, hyperglycemia, and smoking. Combinations of risk factors occur in the metabolic syndrome, in which insulin resistance appears to be the common denominator.

The best known between-country study is the Seven Countries Study by Ancel Keys linking diet, hypercholesterolemia, and CHD. He showed that a plot of each country's median total cholesterol against deaths from CHD was highly correlated. The variations in serum cholesterol were highly

correlated with the ratio of saturated to unsaturated fats in the diet.

Studies of migration and CHD include the Ni-Hon-San Study, in which cholesterol levels and CHD rates were compared in Japanese living in Japan, Honolulu, and San Francisco. There was a rise in both cholesterol levels and CHD rates across these groups, suggesting that as Japanese adopted a Western lifestyle their cholesterol increased and their risk of CHD increased.

Evidence that treatment of hyperlipidemia influences CHD is substantial. The methods of treating hyperlipidemia have varied from diet to drugs, surgery, meditation, and multiple risk factor reduction. The conclusion is that treatment of hyperlipidemia improves CHD morbidity and mortality. The Lipid Research Clinics Coronary Primary Prevention Trial, started in the 1970s, examined 4000 men without evidence of CHD but with hypercholesterolemia, randomized to receive cholestyramine or placebo. After 7 years, despite a relatively minor difference in cholesterol levels, there was a 20% decrease in CHD in the drug-treated group.

In the Oslo study, high-risk Norwegian men were given antismoking and dietary advice, resulting in a significant reduction in the incidence of CHD. The effect of partial ileal bypass surgery has been studied in patients who had experienced a myocardial infarction and were hypercholesterolemic (POSCH study). This surgical procedure improved blood lipids and reduced morbidity caused by CHD. In the Scandinavian Simvastatin Survival Study (4S study), patients with CHD and hypercholesterolemia were randomized to receive the HMG-CoA reductase inhibitor simvastatin or placebo. After a 4-year follow-up period, both morbidity and mortality were significantly reduced in the treatment group. This secondary prevention study was followed by a primary prevention study using pravastatin in men with hypercholesterolemia. This study (WOSCOPS study) randomized men without evidence of CHD to treatment with the HMG-CoA reductase inhibitor pravastatin or to placebo and followed them for 4.9 years. Treatment with the drug significantly reduced the incidence of myocardial infarction and death from cardiovascular causes. In the Air Force/Texas Coronary Atherosclerosis Prevention Study, a primary prevention study, subjects with low levels of LDL cholesterol and HDL cholesterol showed a reduced risk of CHD with statins. The Helsinki Heart Study and the Veterans Affairs-HDL Intervention Study used fibrate drugs in patients with low LDL cholesterol and showed impressive increases in HDL cholesterol and reductions in CHD risk. A 1% increase in HDL was equivalent to a 3% decrease in CHD risk.

Studies such as the Cholesterol Lowering Atherosclerosis Study, in which patients were allocated to drug therapy or placebo, used coronary angiography to follow the effect of drugs on disease. A small reduction in cholesterol results in a disproportionately larger reduction in cardiovascular events.

These studies show that it is possible to arrest progress of the disease and, in some cases, bring about regression of atherosclerosis. The extent to which this happens seems to depend on the underlying disease and the degree of cholesterol lowering.

Atherosclerosis has a complex and multifactorial etiology characterized by inflammation. Clinical markers of inflammation include C-reactive protein, modified LDL, homocysteine, lipoprotein (a), and fibrinogen, which are emerging risk factors and may give prognostic information for patient management. Folate may be beneficial by reducing plasma homocysteine, enhancing endothelial nitric oxide, and showing antiinflammatory properties. Other antiinflammatory agents, such as IL-10, may be of benefit.

Classification of Hyperlipidemia

There are a number of classification systems available. In 1967, Fredrickson, Levy, and Lees introduced the first classification as a method of reporting that lipoproteins were raised. The World Health Organization adopted this classification (**Table 3**).

In 1987, the European Atherosclerosis Society recommended a five-group classification of primary hyperlipidemia (**Table 4**), and the National Cholesterol Education Program Adult Treatment Panel III published guidelines in 2002 for normal and elevated lipid levels (**Table 5**).

Clinically, the most important step is to determine if the lipid abnormality is primary or secondary to another condition. **Table 6** shows the lipid changes seen in some common conditions. In practice, it is often easiest to classify lipid abnormalities into three

Table 3 Fredrickson/WHO classification of hyperlipoproteinemia

Type	*Lipids increased*	*Lipoprotein increased*
I	Triacylglycerol	Chylomicrons
II-a	Cholesterol	LDL
II-b	Cholesterol and triacylglycerol	LDL and VLDL
III	Cholesterol and triacylglycerol	Chylomicron remnants and IDL
IV	Triacylglycerol	VLDL
V	Cholesterol and triacylglycerol	Chylomicrons and VLDL

Table 4 European Atherosclerosis Society classification of hyperlipoproteinemia

Group	*Total cholesterol (mmol/l)*	*Triacylglycerols (mmol/l)*
Normal	<5.2	<2.3
A (mild hypercholesterolemia)	5.2–6.5	and <2.3
B (moderate hypercholesterolemia)	6.5–7.8	and <2.3
C (isolated hypertriglyceridemia)	<5.2	and 2.3–5.6
D (combined hyperlipidemia)	5.2–7.8	and 2.3–5.6
E (severe hypercholesterolemia and/or hypertriglyceridemia)	>7.8	and/or >5.6

Table 5 Adult Treatment Panel III levels for blood lipids

Classification	*Total cholesterol (mmol/l)*	*LDL cholesterol (mmol/l)*	*Triacylglycerols (mmol/l)*
Normal	<5.2	<2.59	<1.7
Above optimal	—	2.6–3.3	—
Borderline high	5.2–6.2	3.4–4.1	1.8–2.2
High	>6.2	4.2–4.8	2.3–5.6
Very high	—	>4.9	>5.6

Table 6 Lipid changes in some common conditions

Condition	*Total cholesterol*	*HDL cholesterol*	*Triacylglycerol*
Diabetes mellitus	Normal or ↑	↓	↑
Hypothyroidism	↑	↑	Can be ↑
Chronic renal failure	Normal or ↑	↓	↑
Nephrotic syndrome	↑	Often ↓	Often ↑
Cholestasis[a]	↑	↓	Can be ↑

[a]An abnormal lipoprotein called LpX is present.

categories: raised total cholesterol, raised triacylglycerol, mixed hyperlipidemia.

It is becoming clear that certain lipoprotein patterns are particularly atherogenic. Elevated IDL with increased small, dense LDL particles and low HDL is one such pattern. Classifications based on these patterns may emerge.

Causes of Hypercholesterolemia

Serum cholesterol at birth does not exceed 2.5 mmol/l and is rarely above 4.0 mmol/l in children. The values for adults are given in **Table 5**. A raised cholesterol level, with little or no elevation of triacylglycerol, is usually a result of raised LDL level. Occasionally, a raised HDL level is responsible for high cholesterol, as seen in the familial condition of primary hyper-α-lipoproteinemia. Secondary causes given in **Table 6** include hypothyroidism, nephrotic syndrome, some cases of diabetes mellitus, and cholestasis. Primary causes include polygenic familial hypercholesterolemia, in which several gene abnormalities together with environmental effects serve to raise serum cholesterol. Several genetic loci contribute to increased plasma LDL levels, but there are five specific monogenic disorders that increase LDL: familial hypercholesterolemia (LDL receptor gene), familial ligand-defective apoB-100 (apoB gene), autosomal recessive hypercholesterolemia (ARH gene), sitosterolen (*ABCG5* or *ABCG8* genes), and cholesterol 7α-hydroxylase deficiency (*CYP7A1* gene).

Much less common, but more clearly defined, are the two autosomal conditions of familial combined hyperlipidemia (FCH) and monogenic familial hypercholesterolemia (FH). In FCH, there appears to be an increase in apoB production and thus an increase in serum LDL. Serum VLDL levels are raised in one-third of these subjects with an associated triacylglycerol increase, one-third show increases in LDL, and one-third show increases in LDL and VLDL. Monogenic FH is caused by a defect in the LDL receptor (LDLR). The consequent reduced LDL uptake by cells, particularly in the liver, results in raised LDL and cholesterol levels. There are 683 mutations in the LDLR gene. Of these, 58.9% are missense mutations, 21.1% are minor rearrangements, 13.5% are major rearrangements, and 6.6% are splice site mutations. The majority of mutations are found in two functional domains of the LDLR, the ligand binding domain (42%) and the epidermal growth factor precursor-like domain (47%).

Predominant hypertriglyceridemia may result from raised VLDL or chylomicron levels. Secondary causes include excess alcohol ingestion, obesity and excess carbohydrate intake, diabetes mellitus, renal failure, and pancreatitis. Primary hypertriglyceridemia can be a result of familial combined hypertriglyceridemia, familial endogenous hypertriglyceridemia, or hyperchylomicronemia.

Familial endogenous hypertriglyceridemia results from increased hepatic triacylglycerol production with increased VLDL production. It is associated with obesity, glucose intolerance, and hyperuricemia. Hyperchylomicronemia is a result of inherited or acquired impairment of lipoprotein lipase activity.

Reduced insulin levels in diabetes mellitus impair the activity of lipoprotein lipase, and hyperchylomicronemia can occur. Inherited deficiency of the lipase enzyme

is rarely seen, as is deficiency of the apolipoprotein (apoC-II) required to activate the enzyme.

Mixed hyperlipidemia is often a secondary condition. Primary causes include familial combined hyperlipidemia and type III hyperlipidemia (dys-β-lipoproteinemia or broad β disease). Type III hyperlipidemia is associated with the apoE 2/2 phenotype, resulting in impaired recognition of apoE by hepatic receptors and an accumulation of IDL.

Dyslipoproteinemia is a central feature of the metabolic syndrome, which is associated with accelerated atherosclerosis. Visceral obesity, dyslipidemia, insulin resistance, hypertension, and a proinflammatory and prothrombotic state are the main characteristics of this condition. It has been defined by the World Health Organization and the National Cholesterol Education Programme. The worldwide increase in levels of obesity in the developed world may presage an increase in CHD.

Dietary Effects

Principles of Treatment

Treatment of hyperlipidemia is part of the management of CHD risk. This encompasses lifestyle changes, such as stopping smoking, increasing exercise, and modifying diet, as well as management of hypertension. Diet is the cornerstone of treating hyperlipidemia, best delivered by qualified dieticians, involving the whole family.

The main aims of diet are to correct excess calorie intake and to reduce the cholesterol and saturated fat content. Patients with hyperlipidemia can expect to see benefits from diet after 6 weeks and are reviewed every 4 months.

Diet can reduce total cholesterol 8–12%, with 60–80% of this change attributed to reductions in saturated fatty acid intake. The remaining change comes from reduced dietary cholesterol and changes in the intake of fiber and monounsaturated and polyunsaturated fatty acids. Dietary modification may not be successful in some primary hyperlipidemias. The Diet and Reinfarction Trial and the Mediterranean Diet Study in postmyocardial infarction survivors showed that dietary modification, not necessarily accompanied by plasma cholesterol lowering, can improve short-term prognosis.

Fat Most of the saturated fats in the diet come from just four fatty acids: lauric acid (C12:0), myristic acid (C14:0), palmitic acid (C16:0), and stearic acid (C18:0). The first three fatty acids reduce LDL receptor activity, raising LDL and total cholesterol by approximately 0.25 mmol/l per 10 g of saturated fat ingested. Watts and coworkers showed that total dietary fat (mainly saturated) increases hepatic VLDL-apoB secretion, so decreasing total fat intake should decrease hepatic apoB secretion.

Monounsaturates are being recommended more often. The most common is oleic acid (C18:1), found in the Mediterranean diet as olive oil. Animal fats are rich in monounsaturates but are also rich in saturated fats. The *trans* isomers of monounsaturates may raise total and LDL cholesterol and are best avoided.

In both type I and type V hyperlipidemia, the dietary management is to reduce fat intake to 20–40 g/day. Medium-chain triacylglycerols are used and fish oils can be tried, but the mainstay of therapy is reduced fat intake. Dietary β-sitosterol can block cholesterol absorption to a limited extent but is not used therapeutically.

Carbohydrate and calories Obesity is a common cause of hypertriglyceridemia due to raised VLDL levels in the obese subject. This may be because of an increase in insulin resistance resulting from obesity with concomitant hyperinsulinemia and elevation in hepatic VLDL synthesis. Some hypertriglyceridemic patients experience a further increase in triacylglycerol levels with an increase in carbohydrate intake, known as carbohydrate induction. This situation is accompanied by an increase in serum insulin levels. With weight reduction, the hypertryglyceridemia reduces and HDL cholesterol increases after 24 months.

Mild alcohol ingestion increases HDL cholesterol. Excess alcohol ingestion can precipitate hypertriglyceridemia of a type IV phenotype due to increased hepatic synthesis and secretion that, in subjects who cannot clear triacylglycerols efficiently, can progress to a type V phenotype. Serum LDL levels are usually low in alcoholics, although in some individuals they can be elevated.

Protein Changes in dietary protein intake have minimal effects on lipid levels. Vegetarians have lower serum lipids than nonvegetarians, but it is not clear how much of this is the result of a change from animal to vegetable protein.

Fiber Soluble fiber such as oat bran and guar lower cholesterol levels, perhaps by reducing bile acid absorption.

Recommendations

The National Food Survey 1999 showed that the total amount of fat in the British diet decreased

from 93 g/day in the1980s to 75 g/day in 1998 and so fat now contributes approximately 40% of calories, of which 15–20% comes from saturated fat. Cholesterol intake in the diet is approximately 500 mg/day. The American Heart Association (AHA) has recommended a two-step approach to dietary change, outlined in **Table 7**, and European recommendations for the diet of the population are shown in **Table 8**. The central approach of dietary therapy is to reduce cholesterol-raising fatty foods, reduce cholesterol intake, and achieve a desirable body weight. The AHA step 1 diet can reduce total cholesterol by 0.5–1.0 mmol/l and the step 2 diet can provide a further 0.2–0.4 mmol/l reduction. Saturated fat in the diet is best replaced by increasing complex carbohydrates, with modest increases in monounsaturated and ω-6 polyunsaturated fatty acids. Increased fish oil intake giving additional ω-3 fatty acids will reduce triacylglycerol levels (but increase LDL cholesterol in certain patients).

Although a low-fat, high-carbohydrate, energy-deficient diet may be used for weight reduction in obese subjects, increasing evidence suggests that increased carbohydrate may not be desirable. Recently, a low-carbohydrate, high-protein, high-fat diet (the Atkins diet) has become popular. Although current studies are promising, the long-term effects of this diet are unknown and it is not currently recommended. Fresh fruit, vegetables, and fiber are encouraged.

Table 7 American Heart Association dietary recommendations

Nutrient	*Recommendations (% of total calories)*	
	AHA step 1	*AHA step 2*
Total fat	<30	<30
Fatty acids		
Saturated fat	<10	<7
Polyunsaturated fatty acid	<10	<10
Monounsaturated fatty acids	10–15	10–15
Carbohydrates	50–60	50–60
Protein	10–20	10–20
Cholesterol	<300 mg/day	<200 mg/day
Reduce total calories to achieve and maintain desirable weight		

From Denke MA (1994) Diet and lifestye modification and its relationship to atherosclerosis. *Medical Clinics of North America: Lipid* Disorders **78**: 197–223.

Table 8 Intermediate and ultimate nutrient goals for Europe

	Intermediate goals		*Ultimate goal*
	General population	*Cardiovascular high-risk group*	
Percentage of total energy[a] derived from			
Complex carbohydrates[b]	>40	>45	45–55
Protein	12–13	12–13	12–13
Sugar	10	10	10
Total fat	35	30	20–30
Saturated fat	15	10	10
P:S ratio[c]	≥0.5	≥1.0	≥1.0
Cholesterol (mg/day)	<300	<300	<300
Fiber (g/day)	30	>30	>30
Salt (g/day)	7–8	5	5

[a]All values given refer to alcohol-free total energy intake.
[b]The complex carbohydrate data are implications of the other recommendations.
[c]The ratio of polyunsaturated to saturated fatty acids.
From Pyorala K, De Backer G, Graham I, Poole-Wilson P, and Wood D (1994) Prevention of coronary heart disease in clinical practice. Recommedations of the Task Force of the European Society of Cardiology, European Atherosclerosis Society and the European Society of Hypertension. *European Heart Journal* **15**: 1300–1331.

See also: **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Dietary Fiber**: Physiological Effects and Effects on Absorption; Potential Role in Etiology of Disease; Role in Nutritional Management of Disease. **Fats and Oils**. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Hyperlipidemia**: Overview. **Lipids**: Chemistry and Classification; Composition and Role of Phospholipids. **Vitamin E**: Metabolism and Requirements; Physiology and Health Effects.

Further Reading

De Backer G, Ambrosioni E, Borch-Johnsen K *et al.* (2003) European guidelines on cardiovascular disease prevention in clinical practice. Third Joint Task Force of European and Other Societies on Cardiovascular Disease in Clinical Practice. *European Journal of Cardiovascular Prevention and Rehabilitation* **10**(supplement 1): S1–S78.

Denke MA (1994) Diet and lifestyle modification and its relationship to atherosclerosis. *Medical Clinics of North America: Lipid Disorders* **78**: 197–223.

Durrington PN (2004) *Hyperlipidaemia: Diagnosis and Management*. London: Hodder Arnold.

Ginsberg HN (1994) Lipoprotein metabolism and its relationship to atherosclerosis. *Medical Clinics of North America: Lipid Disorders* **78**: 1–20.

Grundy SM (1992) Etiologies and treatment of hyperlipidemia. In: Willerson JT (ed.) *Treatment of Heart Disease*, pp. 4.1–4.79. London: Gower Medical.

Marais AD (2004) Familial hypercholesterolaemia. *Clinical and Biochemical Reviews* **25**: 49–68.

National Cholesterol Education Program (2002) *Third Report of the Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III)*, NIH Publication No. 02-5215. Bethesda, MD: National Heart, Lung and Blood Institute.

Pyorala K, De Backer G, Graham I, Poole-Wilson P, and Wood D (1994) Prevention of coronary heart disease in clinical practice. Recommendations of the Task Force of the European Society of Cardiology, European Atherosclerosis Society and the European Society of Hypertension. *European Heart Journal* **15**: 1300–1331.

Scandinavian Simvastatin Survival Study Group (1994) Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: The Scandinavian Simvastatin Survival Study (4S). *Lancet* **344**: 1383–1389.

Shepherd J, Cobbe SM, Ford I *et al.* (1995) Prevention of coronary heart disease with pravastatin in men with hypercholesterolaemia. *New England Journal of Medicine* **333**: 1301–1307.

Steinberg D, Parthasarathy S, Carew TE, Khoo JC, and Witztum JL (1989) Beyond cholesterol. Modifications of low-density lipoproteins that increase its atherogenicity. *New England Journal of Medicine* **320**: 915–924.

Sullivan DR, Celermajer DS, Le Couteur DG, and Lam CWK (2004) The vascular implications of post-prandial lipoprotein metabolism. *Clinical and Biochemical Reviews* **25**: 19–30.

Watts GF and Burnett JR (2004) HDL revisited: New opportunities for managing dyslipoproteinaemia and cardiovascular disease. *Clinical and Biochemical Reviews* **25**: 7–18.

Watts GF, Moroz P, and Barrett PHR (2000) Kinetics of very-low-density lipoprotein apolipoprotein B-100 in normolipidaemic subjects: Pooled analysis of stable-isotope studies. *Metabolism* **49**: 1204–1210.

Nutritional Management

A H Lichtenstein, Tufts University, Boston, MA, USA

Introduction

There is a wide range of dietary approaches purported to decrease the risk of developing cardiovascular disease (CVD). Some were first identified early in the twentieth century while others have been recognized more recently. None are without controversy as to their absolute and relative efficacy. Some of this controversy is more likely attributable to biological variation among individuals (e.g., genetics, gender), interaction among putative dietary factors (e.g., per cent of energy as fat relative to carbohydrate), and differences in environmental factors (e.g., body weight, level of physical activity) than actual differences in whether they are effective modalities or not. This chapter will present current trends in dietary approaches to the prevention and management of CVD.

Surrogate Markers of CVD Risk

It is difficult, if not impossible, to directly assess the effect of any dietary intervention on CVD risk because the natural course of the disease frequently is as long or longer than the productive research life span of the scientists designing and implementing the studies. Hence, most dietary interventions aimed at reducing CVD risk are evaluated on the basis of surrogate markers of disease. However, as with dietary variables thought to be efficacious in decreasing CVD risk, the number of surrogate markers purported to be predictive of CVD has likewise multiplied in the past two decades. The relative importance of each has yet to be sorted out.

Traditionally, total cholesterol, high-density lipoprotein (HDL) cholesterol, and triacylglycerol (triglyceride) levels were measured directly. Low-density lipoprotein (LDL) levels were calculated using the 'Friedewald formula' (LDL cholesterol = total cholesterol − HDL cholesterol − (triglyceride/5)). Reliable automated direct assays for LDL cholesterol are now available and obviate the need for this calculation. Also potentially important in estimating changes in CVD risk as a function of dietary intervention are changes in the levels of lipoprotein (a) (Lp(a)), homocysteine, C-reactive protein (and other markers of inflammation), LDL particle size, hematologic factors, apolipoprotein genotypes, insulin and glucose levels, HDL subspecies, and remnant-like particles. No doubt in the near future the relative importance of each will be clarified and others will emerge.

Dietary Lipid: Approaches to the Prevention and Management of CVD

Level of Dietary Fat

Dietary fat serves as a major energy source for humans. One gram of fat contributes 9 cal, a little more than twice that contributed by protein or carbohydrate ($4\,\mathrm{cal\,g^{-1}}$) and somewhat more than that contributed by alcohol ($7\,\mathrm{cal\,g^{-1}}$). When considering the importance of the level of dietary fat with respect to CVD prevention and management there are two major factors to consider; the impact on plasma lipoprotein profiles and body weight. The potential relationship with body weight is important because of secondary effects on plasma lipids, blood pressure, dyslipidemia, and type 2 diabetes, all potential risk factors for CVD.

With respect to the effect of the level of dietary fat on plasma lipoprotein profiles, the focus is usually on triglyceride and HDL cholesterol levels or total

cholesterol to HDL cholesterol ratios. Evidence indicates that when body weight is maintained at a constant level, decreasing the total fat content of the diet, expressed as a per cent of total energy, and replacing it with carbohydrate results in an increase in triglyceride levels, decrease in HDL cholesterol levels, and a less favorable (higher) total cholesterol/HDL cholesterol ratio. Low HDL cholesterol levels are an independent risk factor for CVD. Low fat diets are of particular concern in diabetic or overweight individuals who tend to have low HDL cholesterol levels.

With respect to the effect of the level of dietary fat on body weight two reviews of the long-term data published on the relationship between per cent of energy from fat and body weight have concluded that even a relatively large downward shift in dietary fat intake (approximately 10% of energy) resulted in only modest weight loss of 1.0 kg over a 12-month period in normal weight subjects and 3 kg in overweight or obese subjects. Some evidence suggests that dietary fiber content may be a mitigating factor. That is, substituting fruits, vegetables, and whole grains for fat instead of fat-free cookies, cakes, and snack foods may be more efficacious in promoting weight loss within the context of low-fat diets. The area of dietary fat and obesity is clearly complex. However, it is important to note that in those long-term studies where patients achieved a drastic reduction in dietary fat intake, in no case was weight gain reported.

Type of Fat

Studies done in the mid-1960s demonstrated that changes in the dietary fatty acid profiles altered plasma total cholesterol levels in most individuals. As analytical techniques became more sophisticated, data on lipid, lipoprotein, and apolipoprotein levels routinely became available. Although many studies have confirmed these early observations, inconsistencies among the more recent results are not rare. These inconsistencies, when they do occur, are attributable to differences among the experimental diets, such as the magnitude or type of dietary perturbation, length of study period, habituation to nutrient intakes prior to the start of the study period, and the background diet on which the dietary variable was superimposed, as well as differences among experimental subjects, such as in age, sex, genetics, efficiency of cholesterol absorption, and initial blood lipid concentrations.

Saturated Fatty Acids

Early evidence demonstrated that the consumption of foods relatively high in saturated fatty acids (SFAs) increased plasma total cholesterol levels and that not all SFAs had identical effects. Subsequent work confirmed the hypercholesterolemic effect of SFAs, demonstrated that SFA intake results in an increase in both LDL and HDL cholesterol levels, and reaffirmed that not all SFAs have the same effect. Short-chain fatty acids (6:0 to 10:0) and stearic acid (18:0) produce little or no change in blood cholesterol levels, whereas SFAs with intermediate chain lengths (lauric (12:0) to palmitic (16:0) acids) appear to be the most potent in increasing blood cholesterol levels (**Table 1**). It has been postulated that stearic acid (18:0) is not absorbed or is rapidly converted to oleic acid (18:1), and for this reason has a relatively neutral effect on blood cholesterol levels. The underlying mechanism by which fatty

Table 1 Dietary fatty acids

Code	*Common name*	*Formula*
Saturated		
12:0	Lauric acid	$CH_3(CH_2)_{10}COOH$
14:0	Myristic acid	$CH_3(CH_2)_{12}COOH$
16:0	Palmitic acid	$CH_3(CH_2)_{14}COOH$
18:0	Stearic acid	$CH_3(CH_2)_{16}COOH$
Monounsaturated		
16:1n-7 *cis*	Palmitoleic acid	$CH_3(CH_2)_5CH{=}(c)CH(CH_2)_7COOH$
18:1n-9 *cis*	Oleic acid	$CH_3(CH_2)_7CH{=}(c)CH(CH_2)_7COOH$
18:1n-9 *trans*	Elaidic acid	$CH_3(CH_2)_7CH{=}(t)CH(CH_2)_7COOH$
Polyunsaturated		
18:2n-6,9 all *cis*	Linoleic acid	$CH_3(CH_2)_4CH{=}(c)CHCH_2CH{=}(c)CH(CH_2)_7COOH$
18:3n-3,6,9 all *cis*	α-Linolenic acid	$CH_3CH_2CH{=}(c)CHCH_2CH{=}(c)CHCH_2CH{=}(c)CH(CH_2)_7COOH$
18:3n-6,9,12 all *cis*	γ-Linolenic acid	$CH_3(CH_2)_4CH{=}(c)CHCH_2CH{=}(c)CHCH_2CH{=}(c)CH(CH_2)_4COOH$
20:4n-6,9,12,15 all *cis*	Arachidonic acid	$CH_3(CH_2)_4CH{=}(c)CHCH_2CH{=}(c)CHCH_2CH{=}(c)CHCH_2CH{=}(c)CH(CH_2)_3COOH$
20:5n-3,6,9,12,15 all *cis*	Eicosapentenoic acid	$CH_3(CH_2CH{=}(c)CH)_5(CH_2)_3COOH$
22:6n-3,6,9,12,15,18 all *cis*	Docosahexenoic acid	$CH_3(CH_2CH{=}(c)CH)_6(CH_2)_2COOH$

acids with 10 or fewer carbon atoms have different effects from those with 12–16 carbons is unknown.

When SFAs displace carbohydrate in the diet, total cholesterol levels increase (**Figure 1**). SFAs tend to be solid at room temperature. Notable exceptions are the tropical oils (palm, palm kernel, and coconut), which are liquid at room temperature because they have high levels of short-chain SFAs. Efforts to reduce dietary SFA intakes should include use of lean meat, the trimming of excess fat and skin from poultry, limiting portion size, and the substituting of non-fat and low-fat dairy products for their full-fat counterparts. The judicious use of ingredient listings and nutrient labels on processed foods will also help achieve the goal of reducing the SFA intakes.

Unsaturated Fatty Acids

Unsaturated fatty acids are fatty acids that contain one or more double bonds in the acyl chain. As the name implies, monounsaturated fatty acids (MUFAs) have one double bond and polyunsaturated fatty acids (PUFAs) have two or more double bonds. The majority of double bonds in fatty acids occurring in food are in the *cis* configuration, that is, the hydrogen atoms attached to the carbons forming the double bond are on the same side of the acyl chain. Alternatively, some double bonds occur in the *trans* configuration, that is, the hydrogen atoms attached to the carbons forming the double bond are on the opposite side of the acyl chain. This part of the discussion of unsaturated fatty acids will be restricted to those containing *cis* double bonds.

Relative to SFAs, MUFAs and PUFAs lower both LDL and HDL cholesterol levels. The absolute magnitude of the change is greater for LDL cholesterol than HDL cholesterol. Most of the data suggest that MUFAs have a slightly smaller effect than PUFAs in lowering both LDL and HDL cholesterol levels so that the change in the total cholesterol/HDL

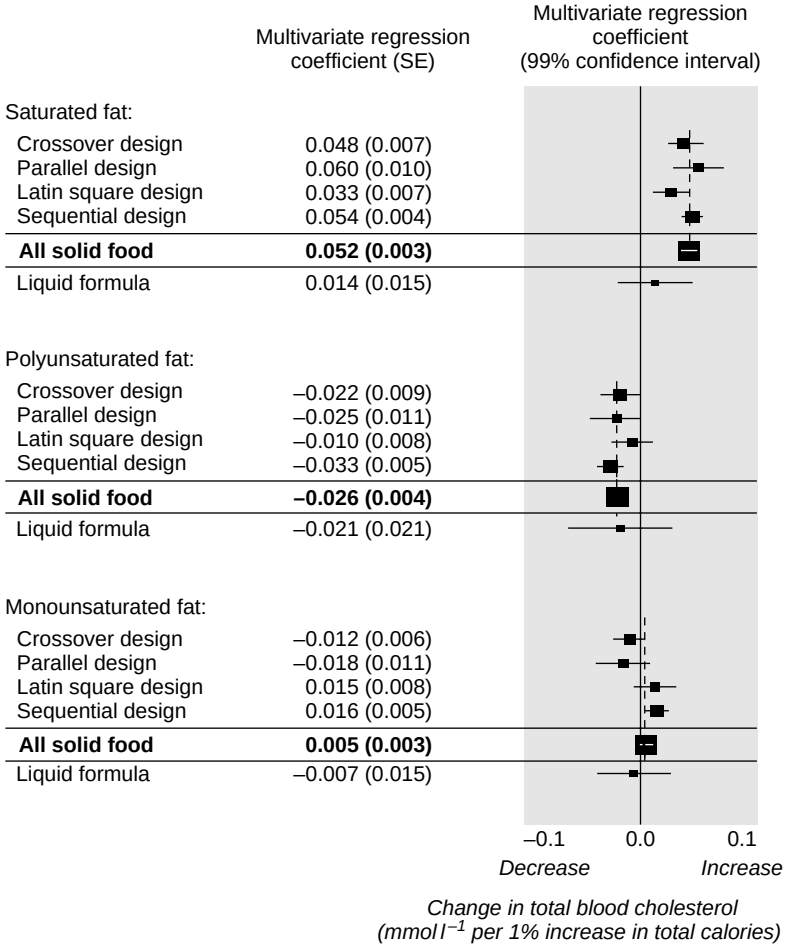


Figure 1 Change in total cholesterol when each fatty acid class displaces carbohydrate from the diet. (Reproduced from Clarke R, Frost C, Collins R, Appleby P, and Peto R (1997) Dietary lipids and blood cholesterol: quantitative meta-analysis of metabolic ward studies. *British Medical Journal* **314**: 112–117.)

cholesterol ratio (decrease) is similar. Because of the changes in plasma lipids and lipoproteins caused when unsaturated fat displaces SFAs from the diet, such a shift should be encouraged in the prevention and management of CVD.

MUFAs The major MUFA in the diet is oleic acid (18:1) (**Table 1**). Vegetable oils high in MUFAs include canola (rapeseed) and olive oils. Fat from meats are also relatively high in MUFAs but unlike vegetable oils, they also contain relatively high levels of SFA, hence would not be recommended as good sources of MUFAs. When MUFAs displace carbohydrate in the diet, there is little effect on total cholesterol levels (**Figure 1**). When MUFAs displace SFA in the diet, total cholesterol levels tend to decrease.

PUFAs There is a wider range of PUFAs than MUFAs in the diet. Dietary PUFAs vary on the basis of chain length, degree of saturation (number of double bonds), and position of the double bond(s) (positional isomers). Two positional isomers of interest with respect to diet and CVD risk are n-6 and n-3 (**Table 1**). The distinction is made on the basis of the location of the first double bond counting from the methyl end of the fatty acyl chain (as opposed to the carboxyl end). If the first double bond is six carbons from the methyl end, the fatty acid is classified as an n-6 fatty acid. If the first double bond is three carbons from the methyl end the fatty acid is classified as an n-3 fatty acid. When PUFAs displace carbohydrate in the diet, total cholesterol levels decrease (**Figure 1**). Vegetable oils high in PUFA include soy bean, corn, sunflower, and safflower oils. The major n-6 PUFA in the diet is linoleic acid (18:2n-6); other n-6 PUFAs, such as γ-linolenic acid (18:3 n-6) and arachidonic acid (20:4n-6), occur in smaller amounts but are important biologically.

n-3 fatty acids Quantitatively, the major n-3 PUFA in the diet is α-linolenic acid (18:3n-3). Major dietary sources include soy bean and canola oils. Two other n-3 PUFAs are eicosapentenoic acid (EPA, 20:5n-3) and docosahexenoic acid (DHA, 22:6n-3) and are sometimes referred to as very long-chain n-3 fatty acids (**Table 1**). The major source of these fatty acids is marine oils found in fish. Dietary intakes of very long-chain n-3 fatty acids are associated with decreased risk of heart disease and stroke (**Figure 2**). Interventions studies have substantiated these findings. The beneficial effects of EPA and DHA are attributed to decreased ventricular fibrillation resulting in decreased sudden death, and decreased triglyceride levels, platelet aggregation, and blood pressure.

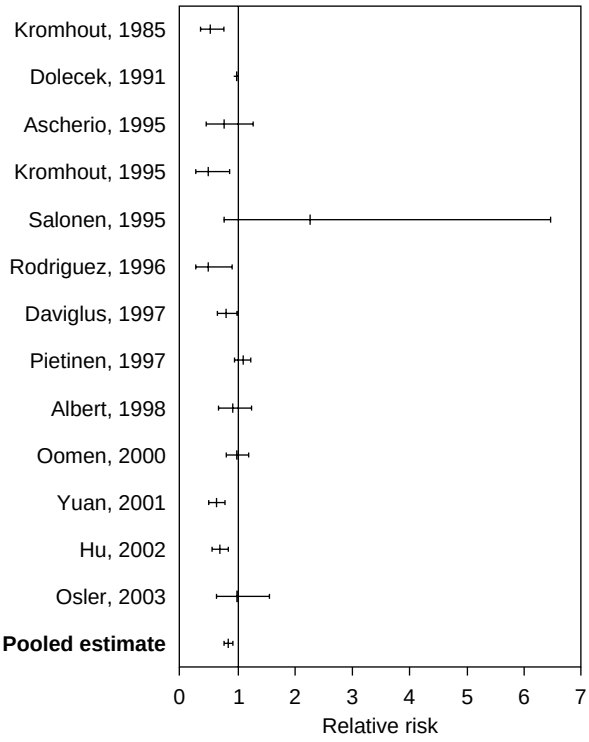


Figure 2 Mean relative risk of coronary heart disease for those consuming any amount of fish versus those reporting none. (Reproduced from Whelton SP, He J, Whelton PK, and Muntner P (2004) Meta-analysis of observational studies on fish intake and coronary heart disease. *American Journal of Cardiology* **93**: 1119–1123.)

Trans-Fatty Acids

Trans-fatty acids, by definition, contain at least one double bond in the *trans* configuration (**Figure 3**). Dietary *trans*-fatty acids occur naturally in meat and dairy products as a result of anaerobic bacterial fermentation in ruminant animals. *Trans*-fatty acids are also introduced into the diet as a result of the consumption of hydrogenated vegetable or fish oils. Hydrogenation results in a number of changes in the fatty acyl chain: the conversion of *cis* to *trans* double bonds, the saturation of double bonds, and the migration of double bonds along the acyl chain, resulting in multiple positional isomers. Oils are primarily hydrogenated to increase viscosity (change a liquid oil into a semiliquid or solid) and extend shelf life (decrease susceptibility to oxidation). The major source of dietary *trans*-fatty acids worldwide is from hydrogenated fat, primarily in products made from this, such as commercially fried foods and baked goods.

Since the early 1990s attention has been focused on the effects of *trans*-fatty acids on specific lipoprotein fractions. The findings of this work have suggested that, similar to saturated fatty acids,

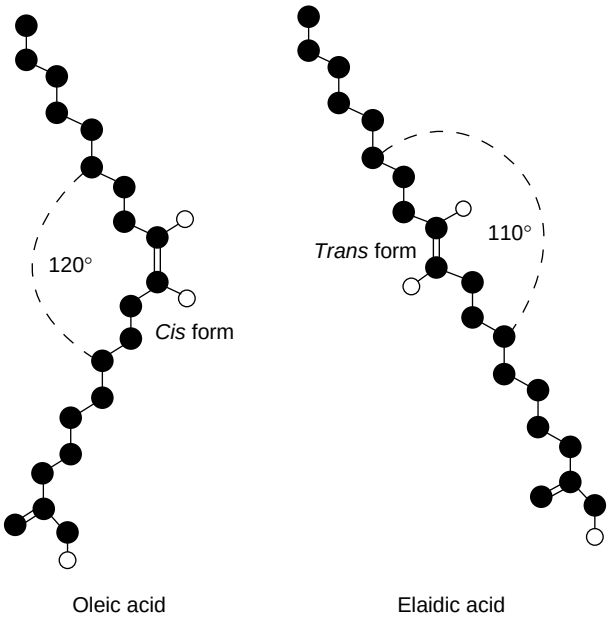


Figure 3 Geometric isomers of 18:1n-9 (oleic and elaidic acids).

trans-fatty acids result in increased LDL cholesterol levels. In contrast to saturated fatty acids, they do not raise HDL cholesterol levels. The changes result in a less favorable LDL cholesterol:HDL cholesterol ratio, with respect to CVD risk (**Figure 4**). A trend towards increased triglyceride levels is frequently reported. Some research has also suggested that *trans*-fatty acids may increase Lp(a) levels. Levels of Lp(a) tend to be positively correlated with risk of developing CVD. However, at this time it appears that the magnitude of increase in Lp(a) levels reported is not within the physiological range that would be predicted to increase CVD risk.

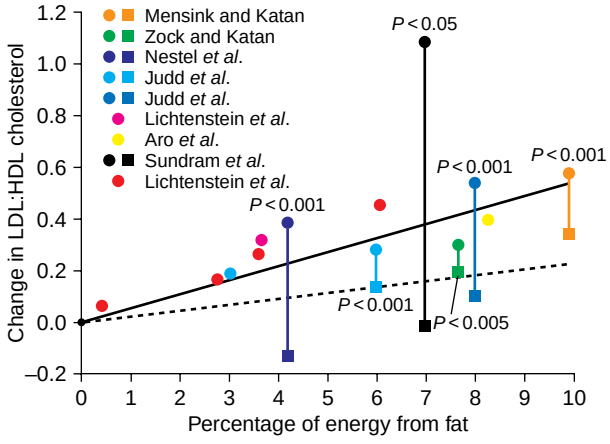


Figure 4 Change in LDL:HDL cholesterol ratio in response to *trans*-fatty acids (solid line) and saturated fatty acids (dashed line). (Reproduced from Ascherio A, Katan MB, Zock PL, Stampfer MJ, and Willett WC (1999) *Trans* fatty acids and coronary heart disease. *New England Journal of Medicine* **340**: 1994–1998.)

Recent estimates from 14 Western European countries report *trans*-fatty acid intakes ranging from 0.8% (Greece) to 1.9% (Iceland) of energy in women and 0.5% (Greece and Italy) to 2.1% (Iceland) of energy in men. Data collected in the US and Canada suggest average *trans*-fatty acid intakes ranging from 1% to 2.5% of energy. By way of contrast, estimates of saturated fat intake range from 10% to 19% per cent of energy.

Dietary Cholesterol

The observation that dietary cholesterol increased blood cholesterol levels and was associated with the development of arteriosclerosis was originally made early in the 20th century in rabbits. In humans, a positive correlation has been repeatedly observed between dietary cholesterol and both blood cholesterol levels and CVD risk, although relative to SFA, the effect is modest. Whether the increase in plasma cholesterol levels induced by dietary cholesterol is linear or curvilinear, or whether there is a break point or threshold/ceiling relationship beyond which individuals are no longer responsive, remains to be determined. With few exceptions, dietary cholesterol is present in foods of animal origin. Therefore, restricting saturated fat intake is likely to result in a decrease in dietary cholesterol.

Other Dietary Approaches for the Prevention and Management of CVD

Very Low-Fat/High-Carbohydrate Diet and High-Protein/Low-Carbohydrate Diet

When considering diets very low in fat and high in carbohydrates ('very low-fat' diets), it is important to separate the effects of the composition of the diet from confounding factors associated with intentional weight loss. For the purposes of this discussion, a very low-fat diet will be defined as less than 15% of energy as fat. Consumption of a very low-fat diet without a decrease in energy intake frequently decreases blood total, LDL, and HDL cholesterol levels and increases the total cholesterol:HDL cholesterol ratio (less favorable) and triglyceride levels. A mitigating factor may be the type of carbohydrate providing the bulk of the dietary energy: complex (whole grains, fruits, and vegetables) or simple (fat-free cookies and ice cream). The reason for this later observation has yet to be investigated. Notwithstanding these considerations, for this reason moderate fat intakes, ranging from <30% to 25 to 35% of energy

as fat, are currently recommended to optimize lipoprotein profiles with respect to decreasing CVD risk.

Current interest in the area of weight loss is centered on high-protein/low-carbohydrate (high protein) diets. Recently, high-protein diets were shown to result in significantly more weight loss than standard reduced energy diets and were accompanied by more favorable blood lipid profiles (lower triglyceride, higher HDL cholesterol levels). However, by 1 year the advantage in terms of weight loss attributed to the high-protein diet did not persist (**Table 2**). The major concern with high-protein diets is that in the absence of steady weight loss the higher intakes of saturated fat and cholesterol can ultimately have an adverse effect on LDL cholesterol levels. Ongoing work will most likely resolve some of these issues.

Fiber

Dietary soluble fiber, primarily β-glucan, has been reported to have a modest independent effect on decreasing blood total and LDL cholesterol levels. A meta-analysis concluded that 3 g of soluble fiber (equivalent of three servings of oatmeal) reduced both total and LDL cholesterol levels approximately $0.13\,\mathrm{mmol\,l^{-1}}$ (**Figure 5**). Most evidence suggests that soluble fiber exerts its hypocholesterolemic effect by binding bile acids and cholesterol in the intestine, resulting in an increased fecal loss and altered colonic metabolism of bile acids. The fermentation of fiber polysaccharides in the colon yields short-chain fatty acids. Some evidence suggests that these compounds may have hypocholesterolemic effects via alterations in hepatic metabolism. At this time there is no evidence to suggest that insoluble fiber has an effect on blood lipid levels.

Soy Protein

The potential relationship between soy protein and the risk of developing CVD has a long history dating back to the 1940s. Despite this relatively protracted lead-time attempts at more precisely defining this relationship have been slow in coming and somewhat inconsistent. Renewed interest developed in the relationship between soy protein and blood lipid levels after a meta-analysis was published in the mid-1990s suggesting that soy protein resulted in significant reductions in total and LDL cholesterol levels, with the most pronounced effect in hypercholesterolemic individuals. Changes in HDL cholesterol levels were not significant. Whether the effect on total and LDL cholesterol levels was attributable to the soy protein *per se* or other soybean derived factor(s), the most likely being the constitutive isoflavones, had yet to be determined. Since then a number of well-controlled studies have re-examined the effect of soy protein and/or isoflavones on blood lipid levels in humans. The results of more recent studies are variable. Declines in LDL cholesterol levels attributable to the substitution of 25–50 g of soy protein for animal protein range from null to small (3–6%) in normocholesterolemic and hypercholesterolemic individuals. Changes in HDL cholesterol levels were highly variable, ranging from -15% to $+7\%$. Soy-derived isoflavones do not appear to have an independent effect on blood lipid levels. On the basis of the most recent data it can be concluded that, although helpful when used to displace products containing animal (saturated) fat from the diet, despite the current claims, individuals should be cautioned against an overreliance on the casual use of soy protein containing foods or the use of isolated isoflavones to control serum lipid levels.

Plant Sterols

Sterols compare for a group of compounds that are essential constituents of cell membranes in animals and plants. Cholesterol is the major sterol of mammalian cells. Phytosterols, such as beta-sitosterol, campesterol, and stigmasterol, are the major sterols of plant cells. In humans, plant sterols are not synthesized, are poorly absorbed, and appear to interfere with cholesterol absorption. It is this later property that has been exploited in the use of these compounds as blood cholesterol-lowering agents. Maximal LDL cholesterol lowering attributable to plant sterols occurs at a dose of about $2\,\mathrm{g\,day^{-1}}$ (**Figure 6**). Although a relatively wide range of responses has been reported, the majority of work suggests an expected LDL cholesterol lowering of about 10% in hypercholesterolemic subjects. Plant sterol-enriched margarines and other foods are currently available in some countries. Few side effects of plant sterols have been reported with the exception of decreased levels of circulating carotenoids; the long-term effect of this is unclear at this time but should continue to be monitored carefully.

Antioxidant Nutrients

Considerable interest had been generated in the potential benefit of dietary supplementation with vitamin E and other antioxidant nutrients in reducing CVD risk. Support for this hypothesis came from two sources. First from the epidemiological observations suggesting that vitamin E supplement use was associated with decreased risk of CVD.

Table 2 Summary mean outcomes

	Carbohydrates in diet (g day^{-1})							
	Lower (≤60)				*Higher (>60)*			
	No. of diets	*No. of participants*	*Summary mean change (SD)*	*95% CI*	*No. of diets*	*No. of participants*	*Summary mean change (SD)*	*95% CI*
Weight change (kg)								
All studies, all participants	34	668	−16.9 (0.2)	−16.6, −17.3	130	2092	−1.9 (0.2)	−1.6, −2.2
RCT and R-Cross only	7	132	−3.6 (1.2)	−1.2, −6.0	75	1122	−2.1 (0.3)	−1.6, −2.7
Caloric content of the diet (kcal day^{-1})								
<1500	18	614	−17.5 (0.2)	−17.1, −17.8	45	870	−3.1 (0.4)	−2.4, −3.8
≥1500	16	53	−5.7 (0.2)	−5.4, −6.0	84	1222	−1.5 (0.2)	−1.2, −1.9
Diet duration (days)								
<15	14	72	−13.6 (0.1)	−13.5, −13.8	25	198	−1.5 (0.2)	−1.1, −1.8
16–60	9	142	−5.3 (0.6)	−4.2, −6.4	52	827	−3.5 (0.4)	−2.9, −4.3
>60	10	447	−2.4 (2.1)	+1.8, −6.5	45	968	−1.1 (0.6)	−.01, −2.3
Participant age (years)								
<40	22	426	−17.7 (0.2)	−17.4, −18.1	59	642	−1.4 (0.2)	−1.0, −1.8
≥40	12	242	−5.0 (0.6)	−3.8, −6.2	62	1231	−2.9 (0.3)	−2.4, −3.5
Baseline weight (kg)								
<70	3	22	−19.6 (0.2)	−19.2, −20.0	19	230	−3.2 (0.6)	−1.9, −4.4
70–100	13	365	−0.8 (1.6)	+2.4, −4.0	77	1357	−2.4 (0.4)	−1.3, −0.4
>100	7	138	−6.6 (0.7)	−5.2, −8.0	18	301	−8.1 (0.8)	−6.5, −9.7
BMI (kg/m^{-2}) in all studies, all participants	1	113	−1.4 (4.6)	+7.6, −10.3	27	739	−0.4 (0.4)	+0.3, −1.1
Body fat (%) in all studies, all participants	5	66	−1.0 (5.6)	+4.0, −6.0	27	536	−1.0 (0.6)	+0.1, −2.1

Adapted from Bravata DM, Sanders L, Huang J, Krumholz HM, Olkin I, and Gardner CD (2003) Efficacy and safety of low-carbohydrate diets: a systematic review. *JAMA* **289**: 1837–1850.

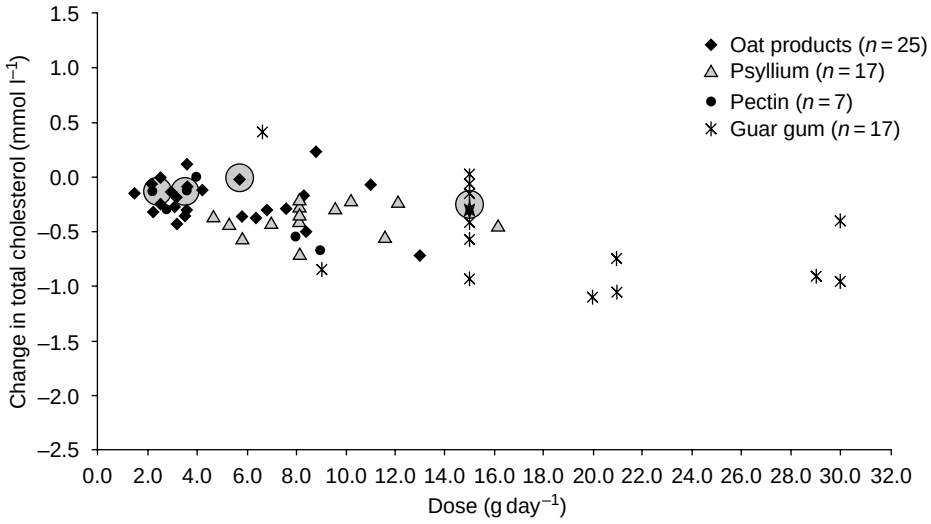


Figure 5 Relationship between fiber intake and change in total cholesterol levels. (Reproduced from Brown L, Rosner B, Willett WW, and Sacks FM (1999) Cholesterol-lowering effects of dietary fiber: a meta-analysis. *American Journal of Clinical Nutrition* **69**: 30–42.)

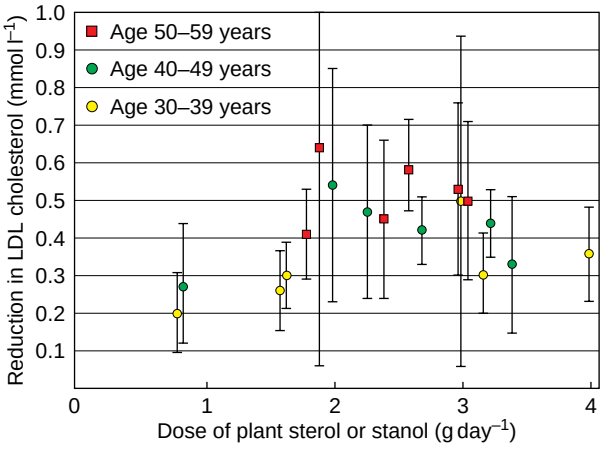


Figure 6 Effect of plant sterols or stanols on LDL cholesterol levels. (Reproduced from Law M (2000) Plant sterol and stanol margarines and health. British Medical Journal **320**: 861–864.)

Second from the *in vitro* work demonstrating that vitamin E in LDL was correlated with decreased susceptibility of the lipoprotein particle to oxidation and that in cell culture oxidized LDL resulted in foam cell formation. A number of recent intervention studies have failed to demonstrate a benefit of vitamin E or other antioxidant vitamins. At this time the data do not support a recommendation to use antioxidant vitamins for the prevention or management of CVD.

Conclusions

The relationship between diet and blood lipid levels has clearly been established. Current data support dietary recommendations to decrease CVD risk that include restrictions in saturated and *trans*-fatty acids, and cholesterol and to include a source of long-chain n-3 fatty acids. Other dietary approaches to prevent and manage CVD include consuming fiber-rich diets such as that founds in fruits, vegetables, and whole grains. Attainment or maintenance of optimal body weight should be emphasized. All individuals should be encouraged to engage in physical activity daily. These recommendations are the culmination of nearly a century of work. They have evolved slowly. No doubt this evolution, frequently accompanied by debate, will continue into the future. It is important for nutrition scientists to implement current recommendations aimed at optimizing blood lipid levels and favorably affecting newer surrogate markers of CVD risk, and to reassess these recommendations as new findings emerge.

See also: **Antioxidants**: Diet and Antioxidant Defense; Observational Studies; Intervention Studies. **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Dietary Fiber**: Physiological Effects and Effects on Absorption; Potential Role in Etiology of Disease; Role in Nutritional Management of Disease. **Fats and Oils**. **Fatty Acids**: Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated. **Hyperlipidemia**: Overview. **Lipids**: Chemistry and Classification; Composition and Role of Phospholipids. **Lipoproteins**.

Further Reading

Expert Panel on Detection Evaluation and Treatment of High Blood Cholesterol in Adults (2001) Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). *Journal of the American Medical Association* **285**: 2486–2497.

Krauss RM, Eckel RH, Howard B *et al.* (2000) AHA Dietary Guidelines: revision 2000: A statement for healthcare professionals from the Nutrition Committee of the American Heart Association. *Circulation* **102**: 2284–2299.

Brown L, Rosner B, Willett WW, and Sacks FM (1999) Cholesterol-lowering effects of dietary fiber: a meta-analysis. *American Journal of Clinical Nutrition* **69**: 30–42.

Willett WC (1998) Is dietary fat a major determinant of body fat? *American Journal of Clinical Nutrition* **67**: 556S–562S.

Yao M and Roberts SB (2001) Dietary energy density and weight regulation. *Nutrition Reviews* **59**: 247–58.

Kris-Etherton P, Lichtenstein AH, Howard B, and Steinberg D, W JL (2004) Antioxidant vitamin supplements and cardiovascular disease. *Circulation* **110**: 637–641.

Kris-Etherton PM, Harris WS, and Appel LJ (2002) Fish consumption, fish oil, omega-3 fatty acids, and cardiovascular disease. *Circulation* **106**: 2747–57.

Ascherio A, Katan MB, Zock PL, Stampfer MJ, and Willett WC (1999) *Trans* fatty acids and coronary heart disease. *Journal of Medicine* **340**: 1994–1998. New England.

2005 US Dietary Guidelines http://WWW.health.gov/dietaryguidelines/dga2005/document/

HYPERTENSION

Contents
Etiology
Dietary Factors
Nutritional Management

Etiology

T Morgan, University of Melbourne, Melbourne, VIC, Australia
H Brunner, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland

Blood pressure (BP) is determined by cardiac output (CO) and total peripheral resistance (TPR):

$$BP = CO \times TPR$$

These variables are controlled in turn by the activity of the autonomic nervous system, regulated by a variety of nuclei in the brain. There is a complex interaction between plasma volume, blood pressure, and a variety of humoral and neural mechanisms that determine blood pressure.

Blood pressure is not, however, a static value. It varies markedly in response to a variety of stimuli. Change of posture activates a variety of controls which keep the pressure relatively constant. Physical and mental activity may be associated with alterations in blood pressure, and there is a marked fall in blood pressure during sleep. Thus, there is no such value as a normal blood pressure based on a single measurement, as blood pressure needs to be related to the circumstances under which it is measured. Likewise, there is no single blood pressure level that means a person is hypertensive. The present convention is that a blood pressure greater than 140 mmHg systolic or 90 mmHg diastolic on clinic recording makes a person hypertensive. Recently the JNC VII report has stated that patients with a blood pressure >120/80 are prehypertensive. However, blood pressure has a marked circadian variation (**Figure 1**), and an individual could have a blood pressure of

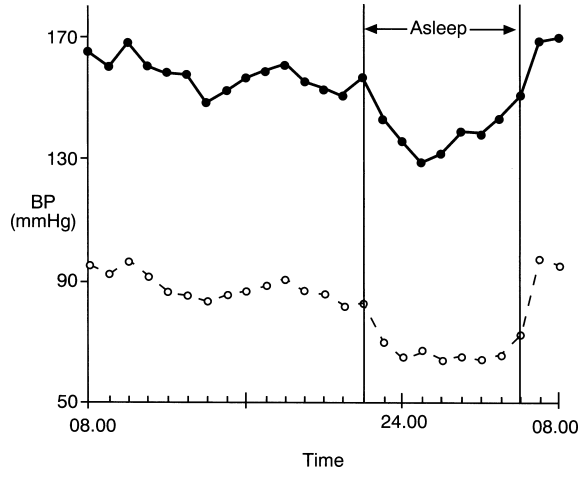


Figure 1 Hourly ambulatory blood pressure (BP) measurements in a 58-year-old man with borderline hypertension. Solid circles, systolic pressure; open circles, diastolic pressure.

150/90 mmHg at 09:00 h and be classified as hypertensive, while at 14:00 h it might be 137/85 mmHg and would be classified as normotensive. Thus in a normal person blood pressure may vary markedly during a day associated with reactive events, but in some people the baseline blood pressure eventually rises to a level that is defined as 'hypertension.' In this person with hypertension there will be fluctuations in blood pressure associated with the same controls as in normal people, but the fluctuations may be exaggerated, leading to high blood pressure levels.

At different times of the day blood pressure is regulated by different systems. Thus, during the day the cardiovascular sympathetics activated by the baroreceptors are important controls of blood pressure. When asleep the cardiovascular sympathetics turn off and blood pressure then appears to be maintained more by the renin angiotensin system. The variability in the activity of systems controlling blood pressure means that in hypertensive patients the response to drugs that act on these systems may have a circadian variation.

The etiology of essential hypertension is unknown; however, the condition is believed to result from an interaction of environmental and genetic factors. Environmental factors are undoubtedly of major importance, because in certain communities hypertension is virtually nonexistent; however, when such a community alters its life style, hypertension becomes common and may exist in 30% of the population. Not all people develop hypertension, and the ones who do are determined by their genetic composition (**Figure 2**). Investigations are under way to attempt to determine which individuals are more likely to develop hypertension and its complications, so that life style and environmental alterations can be initiated to prevent the disease occurring in such people. Certain specific genetic abnormalities have been identified and these cases are then removed from the classification of essential hypertension. It is of interest that the disorders that have been found in general alter sodium handling by the body. These have been either abnormalities in channels or transporters in the nephron that alter sodium excretion, or alternatively defects in circulating hormones that regulate the activity of the renal transporters. Hypertension is not seen in hunter-gatherer communities where sodium intake is low and potassium intake is high, and thus the genetic abnormality is not expressed phenotypically even though the genotype is probably present.

In established hypertension the defect is an increased peripheral resistance rather than an increased cardiac output. However, in people with minor blood pressure elevations and prehypertensive people cardiac output is increased, and it has been postulated that increased cardiac output in response to the retention of sodium

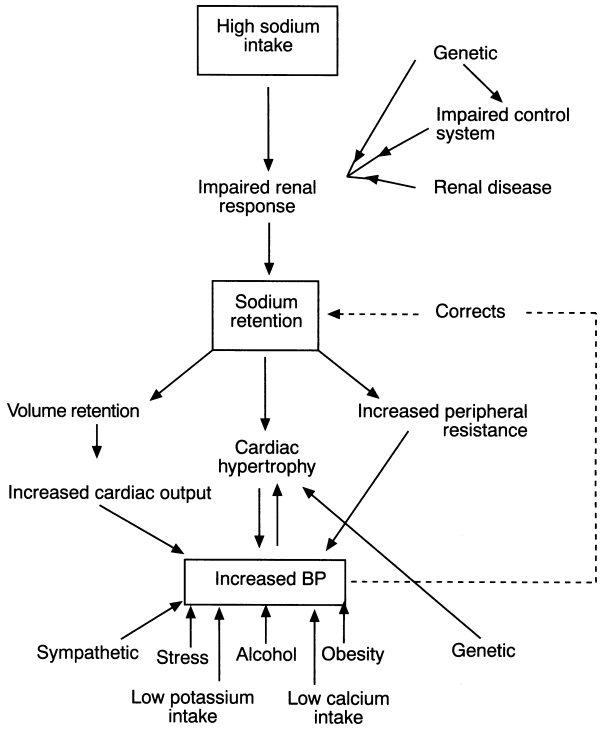


Figure 2 The interrelationships between sodium intake, renal function, hormonal control systems, and genetic inheritance in the etiology of hypertension and cardiac hypertrophy. BP, blood pressure.

is the initial hemodynamic change that leads to hypertension (**Figure 2**). However, experimentally hypertension can be produced without a stage of increased cardiac output, and increased peripheral resistance can result without an antecedent high cardiac output. It is likely that there is heterogeneity in the way people respond. The concept of an increased cardiac output leading to hypertension has been extensively developed by Guyton in a variety of computer and experimental models. However, in carefully conducted studies in which blood volume was measured in hypertensive patients, blood volume was decreased rather than increased, making this theory probably not applicable to all people. The relationship with sodium is also complicated. In young hypertensive subjects there is a better inverse correlation with total body potassium rather than a direct correlation with total body sodium content. In older people the correlation with body sodium content becomes more pronounced. The lack of a direct correlation between body sodium and hypertension in the young casts doubt on the absoluteness of the link between sodium and hypertension, and clearly potassium has an important effect modulating the response.

It has been suggested that the prime defect leading to increased peripheral resistance is the presence of a circulating factor that inhibits $(Na^+\text{-}K^+)$-ATPase

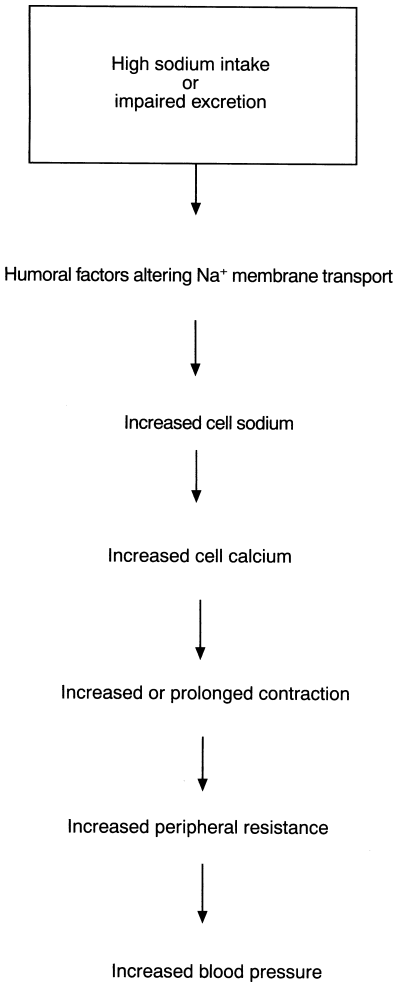


Figure 3 Mechanistic approach indicating how at the cellular level a high sodium intake may initiate the series of events leading to increased peripheral resistance and a high blood pressure.

activity, thereby increasing the sodium content of cells (**Figure 3**). This increased sodium content decreases the rate at which calcium can be removed from the cell by the Na^+-Ca^{2+} countertransport. In skeletal and cardiac muscle cells the contractile response is triggered by a small influx of Ca^{2+} that releases Ca^{2+} from the endoplasmic reticulum. The response is terminated by reuptake of Ca^{2+} into the endoplasmic reticulum and thus the Na^+-Ca^{2+} countertransport is not of critical importance, though in all cells the basal level and total content of calcium may be increased. In smooth muscle cells, including the arteriolar (resistance vessels) cells, contraction is initiated by entry of calcium across the cell membrane. If there is a defect in calcium removal by the Na^+-Ca^{2+} countertransport contraction will be prolonged, and if the basal level of cellular calcium is higher contraction may be more intense. Thus, peripheral resistance rises and hypertension results. All the physiological factors to support the above have been identified. However, despite intensive research it is unclear if a true circulating physiological factor capable of inhibiting (Na^+-K^+)-ATPase has been identified. Claims have been made for an ouabain-like factor in plasma and the hypothalamus, but there is skepticism whether this is the important physiological variable. In hypertensive patients cell sodium levels are elevated. This elevation need not necessarily be due to inhibition of (Na^+-K^+)-ATPase but could result from an increased entry of sodium into the cell down its electrochemical gradient by a variety of channels or transporters. There is evidence that abnormalities of these exist and are more prevalent in hypertensive people. There is also evidence that the rate of entry of Na^+ can be increased by a high sodium intake and that a circulating but unidentified factor may be increased. The signal for release of such a factor is unclear and does not appear to be plasma sodium concentration and probably not total plasma volume. It may be modulated by the kidney and related to 'turnover' of sodium.

The body can control plasma sodium concentration (by antidiuretic hormone) and plasma volume and total body sodium within well-defined limits, despite large variations (20–400 mmol) in daily sodium chloride intake. This control involves a variety of humoral factors (**Table 1**). Renin-angiotensin, aldosterone, atrial natriuretic peptide, sympathetic activity, and other variables are all altered by changes in sodium chloride and/or potassium intake. The capacity of these systems to respond maintains blood pressure in the 'normal' range. It is only when this capacity is exceeded that blood pressure becomes elevated. The increase in blood pressure will also correct the body sodium because the kidney has a sensitive 'pressure natriuresis response'. Thus, in most people as blood pressure rises sodium is excreted; this self-correction ensures that blood pressure does not rise to excessive levels. It has been suggested that in addition to high sodium intake and abnormalities of sodium handling by cells there must be a defect in the pressure natriuresis response. This could be due to

Table 1 Factors altering sodium balance

Variable	*Site of action*
Increases Na^+ retention	
Angiotensin II	Proximal tubule
	Increases aldosterone
Aldosterone	Distal nephron
Sympathetic	Proximal tubule
	Hemodynamics
Increases Na^+ excretion	
Atrial peptide	Proximal tubule
	Distal nephron
Parathyroid hormone	Proximal nephron
Natriuretic hormone (?)	Loop of Henle, plus others
Elevated blood pressure	Hemodynamics

excessive amounts of circulating hormones (aldosterone) or defective control systems in the kidney. The pressure natriuresis response may also be defective owing to reduction in nephron number following developmental problems or disease, or associated with the aging process. An association has been found between the weight of children at birth and subsequent development of hypertension and cardiovascular disease. The low birthweight could be due to defective nutritional intake of the mother or to diseases that affect fetal and placental growth. It has been suggested that the total nephron number is reduced, and that this alters sodium handling and causes hypertension.

Much research has focused on the importance of dietary sodium chloride, but there needs to be an associated genetic defect which may be a subtle defect in the systems controlling sodium excretion. Thus, the defect may be an inability of the renin–angiotensin–aldosterone system to suppress adequately or appropriately for that level of sodium intake. There is evidence from twin studies that the suppressibility of renin secretion is genetically determined and thus in some people there are inappropriate levels of angiotensin II for their level of sodium intake, resulting in hypertension. There are changes in secretion or response of renin, aldosterone, adrenaline, sympathetic activity, atrial peptide, and nitric oxide with increase in sodium intake. In most cases these responses are appropriate and prevent the unfettered rise in blood pressure. However, the ability to respond may be exceeded and blood pressure then rises.

Hemodynamics

As discussed above, it is unlikely that all people go through an increased cardiac output stage. Well-established hypertensive patients have high peripheral resistance and a normal cardiac output, but there are exceptions. The hypertension process itself causes significant alterations in hemodynamics affecting both the heart and the blood vessels, and reversal of these effects may be as important as reducing blood pressure (**Figure 4**).

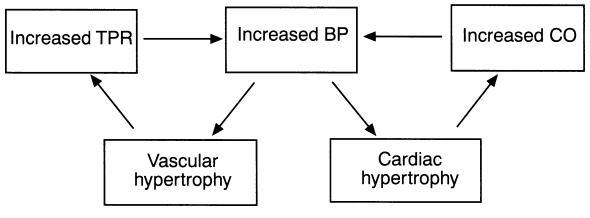


Figure 4 Interaction between the various parameters that control blood pressure, showing how they set up a positive feedback leading to worsening blood pressure. BP, blood pressure; CO, cardiac output; TPR, total peripheral resistance.

Early in the hypertensive process there is an increase in the thickness of the arteriolar muscle wall. This is probably a compensatory process which returns the wall tension to normal. Contrary to expectations, compliance of larger arteries is normal or increased in young hypertensive patients. However, the thickening of the resistance vessels, depending on the way it takes place, has certain consequences, and for a similar degree of muscle contraction there is a greater increase in vascular tone and thus peripheral resistance rises more, leading to a higher blood pressure, greater wall tension, and a further increase in vessel thickness. This is a positive feedback response and a vicious cycle may result (**Figure 4**). In the early hypertensive process the systolic and diastolic blood pressures rise more or less in parallel. However, in the older hypertensive patient the pulse pressure widens, due probably to increased stiffness of the arteries. This increased stiffness, which is associated with a loss of elastin and an increase in collagen, has important effects on the heart.

The endothelium of blood vessels is a major regulator of vascular tone and an important mechanism is the production of nitric oxide. If nitric oxide is removed, peripheral resistance rises and hypertension results. However, it is unlikely that defects in nitric oxide production are the cause of high blood pressure. In fact in early hypertension the nitric oxide production may be increased as a compensatory event modulating the rise in pressure, and this may explain why dynamic compliance is normal (**Figure 5**). However, when hypertension is established and there is vessel disease the nitric oxide response and endothelial control become impaired. This is probably an important factor leading to stiffness of the arteries and atherosclerosis.

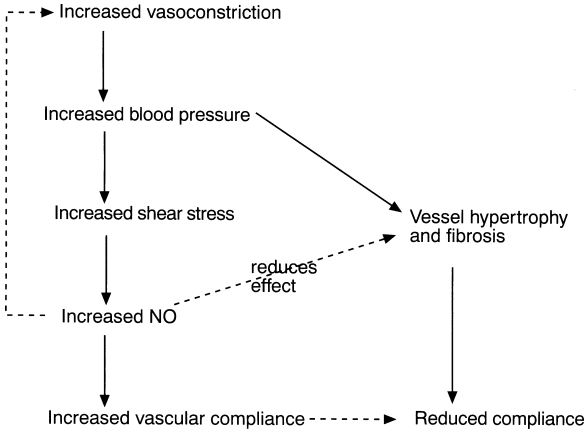


Figure 5 An outline indicating how the initial response of the endothelium is to prevent the rise of blood pressure by releasing nitric oxide (NO). This increases vessel compliance, reducing the adverse effects. If this system's capacity is exceeded the arterial damage process is accelerated. The dotted lines represent negative feedback attempting to restore the status quo.

Table 2 Factors determining extent of reflection and site where the reflected wave meets the flow wave

Poor arterial compliance	Increased pulse wave velocity Reflected wave closer to heart
Arterial branch points	Reflective site
Peripheral resistance	Increased reflection

The stiffness of blood vessels in older hypertensive patients has a number of important consequences. The pulse wave velocity is increased and thus reflected waves arrive back at the heart while the ventricle is still contracting, thereby augmenting the central systolic pressure (**Table 2**). In normotensive people the place at which the reflected wave and the oncoming flow meet is near the brachial artery, and thus central systolic pressure is lower than brachial artery systolic pressure (**Figure 6**). This increased central systolic blood pressure means that the heart contracts against a greater load and thus performs more work, leading to hypertrophy greater than might be predicted from the brachial artery pressure. The extent of the augmentation due to the pressure wave depends upon the degree of reflection, which is controlled in part by the peripheral resistance. The site at which augmentation is highest depends on the pulse wave velocity. The deterioration in the elastic properties of the large blood vessels with loss of elastin and more collagen leads to increased pulse pressure, increased augmentation of central systolic pressure and a decrease in the peripheral diastolic pressure, all of which are common in the elderly hypertensive patient.

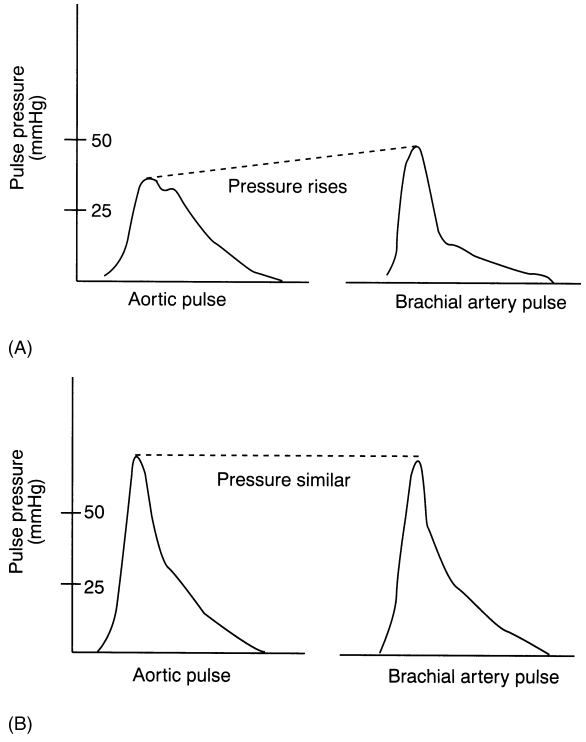


Figure 6 The central aortic and brachial artery pulse wave forms in normotensive (A) and hypertensive (B) subjects. In (B) the heart pumps blood out against a higher pressure leading to cardiac hypertrophy. See O'Rourke (1995) for discussion of how central aortic pressure is higher than brachial artery pressure due to reflected waves.

The Heart

In hypertensive patients the left ventricle is frequently enlarged and this is associated with an increased risk of cardiovascular death. When assessed by electrocardiography left ventricular hypertrophy (LVH) is relatively uncommon, but if assessed by echocardiography LVH is present in up to 50% of mild hypertensive patients and in adolescents not classified as hypertensive, but in the upper 10 percentile of blood pressure there is a 10–15% prevalence of LVH (**Table 3**). The cause of the LVH is not certain (**Table 4**). There is a better correlation with 24 h blood pressure than with clinic values, but the r value is about 0.14 indicating considerable variability. It is possible that acute elevations of blood pressure sustained for 1–2 h may have a potent effect by increasing wall stress and activating the processes that lead to myocyte hypertrophy. This may be of particular importance if it occurs at a time when plasma levels of potential growth factors such as angiotensin II and growth hormone are elevated. These hormones are elevated during sleep and thus blood pressure elevation at that time may be particularly detrimental. This is supported by observations that people who do not have the usual night-time (sleep) fall in blood pressure are more likely to have cardiac and renal complications. There is a significant genetic influence on cardiac hypertrophy and it has been proposed that cardiac enlargement may be antecedent to and the cause of hypertension. High blood pressure can undoubtedly cause cardiac enlargement, but independent of blood pressure elevation angiotensin II and salt can probably enlarge the heart.

The strongest predictor in some studies of cardiac size was the salt intake. In animals a high salt intake

Table 3 Prevalence of left ventricular hypertrophy

Subjects	*Prevalence (%)*
Normotensive	1–2
Adolescent, upper 10%	10–15
Borderline hypertensive	
by echo	20–50
by ECG	3–5
Severe hypertensive	90

ECG, electrocardiogram; echo, echocardiogram.

Table 4 Factors affecting cardiac hypertrophy

Factors leading to hypertrophy	*Factors reducing or preventing hypertrophy*
24 h cardiac work	Nitric oxide
Ventricular wall stress	Bradykinin
Sodium intake	
Sympathetic activity	
Angiotensin II	
Insulin-like growth factor	
Growth hormone	
Genotype	

can cause cardiac hypertrophy and a low salt intake allows resolution. The increased size of the heart is a response that decreases the wall stress of the ventricle and is a compensatory phenomenon. The cardiac hypertrophy with hypertension is concentric in nature, with sarcomeres laid down in the myocytes in parallel (**Figure 7**). The increased thickness of the myocytes together with associated fibrosis of the interstitium means that the oxygen diffusion pathway is increased and this may lead to precipitation of arrhythmias and sudden death. In cardiac hypertrophy associated with exercise the sarcomeres are laid down in series, and with this 'eccentric' hypertrophy there is no increased mortality.

In addition to cardiac hypertrophy in hypertensive patients there is significant impairment of diastolic relaxation. This may result from poor oxygen delivery to the mitochondria and thus a retarded reuptake of calcium into cell organelles. Thus, there is a dynamic aspect to diastolic dysfunction which may be reversible. However, in addition the laying down of fibrous tissue in the heart contributes to stiffness and poor diastolic filling. The poor diastolic function may occur prior to any increase in cardiac size. The poor diastolic filling due to reduction in left ventricular compliance may explain the subnormal stroke volume seen in hypertensive patients during exercise. In these circumstances the increased pulse rate means that there is insufficient time for a stiff left ventricle to fill adequately.

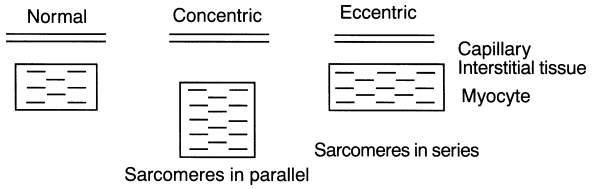


Figure 7 The diffusion distance in normal and eccentric hypertrophy is not increased. In concentric hypertrophy there is often associated fibrosis; this leads to a longer extracellular diffusion distance as well as a longer intercellular pathway. Thus, oxygen delivery to mitochondria is poor, the reuptake of calcium (an energy-dependent process) is sluggish, and 'functional' relaxation is slow, leading to impaired diastolic filling.

Early in the development of hypertension in spontaneously hypertensive rats and in humans total peripheral resistance is elevated. In rats changes in the resistance vessels are seen early. In borderline and mild hypertension in humans there may be little increase in total peripheral resistance at rest, but the total peripheral resistance does not fall to normal levels during conditions when maximal vasodilatation would be expected (e.g., exercise, heating, autonomic blockade). This probably indicates that structural changes occur early in the disease and the failure to dilate adequately may in part explain the excess rise in blood pressure seen in hypertensive patients during exercise.

Increased peripheral resistance is not evenly distributed across all regional vascular beds and the resistance in the kidney frequently appears to be increased, resulting in a reduction of about 10% in renal blood flow. In contrast, in prehypertensive people an increase has been reported in renal blood flow. Whether this has any pathogenic significance is not known. However, the reduced blood flow could result from activation of the tubuloglomerular feedback response due to altered sodium reabsorption in the proximal tubule.

The coronary flow in hypertensive patients is of importance. These people already may have an increased oxygen demand. The flow at rest is usually normal but even in patients with no evidence of coronary artery disease the flow reserve is impaired. In normal people the coronary artery rapidly dilates to meet the increased oxygen demand but in hypertensive patients this dilation is sluggish and does not reach the same maximal flow. The reason is complex and is possibly a combination of structural change and an impaired endothelial response.

The Sympathetic Nervous System

Many investigators have postulated that hypertension may result from impaired central control and this is mediated via the sympathetic nervous system. The increased cardiac output and heart rate seen in many people with early or incipient essential hypertension could be explained by excess sympathetic activity. However, it has been difficult to demonstrate that there is increased sympathetic activity because many of the techniques are relatively crude. It has been reported that plasma noradrenaline levels correlate with cardiac index and peripheral resistance in mildly hypertensive patients. It is difficult to know if increased sympathetic activity is primary, but in adolescents who later develop hypertension there is an increased blood pressure rise associated with mental or physical stress, which

supports the concept of a dysregulatory neurogenic component. Sympathetic activity may also be altered by changes in sodium or potassium intake, and thus the 'prime' cause of hypertension remains to be elucidated.

Renal Function

There are undoubtedly subtle abnormalities in renal function in most hypertensive people. It is unclear if this is a cause or effect of hypertension. In spontaneous hypertensive rats (SHR) early in life the proximal tubule cells are very responsive to angiotensin II and this could cause sodium retention and initiate the development of the hypertensive process. However, in mature rats the responsiveness to angiotensin II of the proximal tubule is lost.

In hypertensive patients there is a reduced renal blood and plasma flow associated with an increased filtration fraction and hence a normal glomerular filtration rate. These changes would result in an increased fractional absorption of sodium by the proximal tubule and potential difficulty in excreting sodium by a pressure natriuresis. The pressure natriuresis curve is shifted with less sodium being excreted for a given pressure at rest, but exaggerated when pressure is acutely increased. It is not clear what is cause or effect, but it is tempting to assume that resetting of the pressure natriuresis response takes place, because if it operated normally the increased pressure should cause salt loss and correct the hypertensive process.

In some but not all people blood pressure falls with sodium restriction. Patients with salt-sensitive hypertension tend not to have a nocturnal fall in blood pressure; they have a greater prevalence of cardiac hypertrophy, microalbuminuria, and a worse prognosis.

Conclusions

Hypertension is not a disease but a sign of some underlying disturbance in the usual control systems for blood pressure. It is thus difficult to have a single description of the physiology of essential hypertension as it will depend upon the cause. There are, however, certain features common to many people. In people with certain (at present unknown) abnormalities in their genotype, exposure to a high-sodium, low-potassium diet together with other alterations in their life style leads to an elevation in blood pressure. In some people there is an initial stage of high cardiac output, but when hypertension is established peripheral resistance is elevated and is the explanation for the high blood pressure. The genetic abnormalities may relate to impairment of the control systems for excreting sodium chloride or a deficit in the ability of the kidney to excrete sodium. There are associated abnormalities in the sympathetic nervous system and the central regulation of blood pressure. When blood pressure is elevated a series of compensatory events are activated, particularly cardiac and vascular hypertrophy, which are initially appropriate responses but lead to the creation of a positive feedback loop which eventually becomes a vicious cycle leading to malignant hypertension.

Essential hypertension in some ways is a misnomer. It is caused by alterations in nutrition and life style in people with a susceptible genotype. The challenge is to identify such people and remove the appropriate environmental factor.

See also: **Coronary Heart Disease**: Hemostatic Factors; Lipid Theory. **Hypertension**: Nutritional Management. **Potassium**. **Sodium**: Physiology.

Further Reading

Avolio AP, Deng FQ, Li WQ *et al.* (1986) Improved arterial distensibility in normotensive subjects on a low salt diet. *Arteriosclerosis* **6**: 166–169.

Barker DJ, Winter PD, Osmond C, Margetts B, and Simmonds SJ (1989) Weight in infancy and death from ischaemic heart disease. *Lancet* **ii**(8663): 577–580.

Dampney RAL (1994) Functional organisation of central pathways regulating the cardiovascular system. *Physiological Reviews* **74**: 323.

Draaijer P, Kool MJ, Maessen JM *et al.* (1993) Vascular distensibility and compliance in salt-hypertensive and salt-resistant borderline hypertension. *Journal of Hypertension* **11**: 199–1207.

Folkow B (1982) Physiological aspects of primary hypertension. *Physiological Reviews* **62**: 347–504.

Guyton A (1980) *Arterial Pressure and Hypertension*. Philadelphia: WB Saunders.

Hayoz D, Rutschmann B, Perrett F *et al.* (1992) Conduit artery compliance and distensibility are not necessarily reduced in hypertension. *Hypertension* **20**: 1–6.

Lund-Johansen P and Omvik P (1990) Haemodynamic patterns of untreated hypertensive disease. In: Laragh J and Brenner BM (eds.) *Hypertension: Pathophysiology, Diagnosis and Management*, pp. 305–327. New York: Raven Press.

Morgan TO and Anderson A (2003) Different drug classes have variable effects on blood pressure depending on the time of day. *American Journal of Hypotension* **16**: 46–50.

Morgan T, Aubert J-F, and Brunner H (2001) Interaction between sodium intake, angiotensin II and blood pressure as a cause of cardiac hypertrophy. *American Journal of Hypertension* **14**(9): 914–920.

Morgan TO, Brunner HR, Aubert J-F, Wang Q, Griffiths C, and Delbridge L (2000) Cardiac hypertrophy depends upon sleep blood pressure: A study in rats. *Journal of Hypertension* **18**: 445–451.

O'Rourke M (1995) Mechanical principles in arterial disease. *Hypertension* **26**: 2–9.

Dietary Factors

L J Appel, Johns Hopkins University, Baltimore, MD, USA

Worldwide, elevated blood pressure is an extraordinarily common and important risk factor for cardiovascular and kidney diseases. As blood pressure rises, so does the risk of these diseases (**Figure 1**). The relationship is strong, consistent, continuous, independent, and etiologically relevant. Accordingly, the adverse consequences of elevated blood pressure are not just restricted to individuals with hypertension (a systolic blood pressure $\geq$140 mmHg or a diastolic blood pressure $\geq$90 mmHg). Those with prehypertension, namely, a systolic blood pressure of 120–139 mmHg or diastolic blood pressure of 80–89 mmHg, have a high probability of developing hypertension and carry an excess risk of cardiovascular disease compared to those with a normal blood pressure (systolic blood pressure $<$120 mmHg and diastolic blood pressure $<$90 mmHg). In fact, almost one-third of blood pressure-related deaths from coronary heart disease occur in individuals with blood pressure in the nonhypertensive range.

In Western countries and most economically developing countries, systolic blood pressure rises with age. As a consequence, the lifetime risk of developing hypertension is extremely high, approximately 90% among US adults older than age 50 years. However, the rise in blood pressure with age is not inevitable. There are numerous isolated populations in which the rise in blood pressure is blunted or even flat. These populations are typically characterized by extremely low intakes of salt, relatively high intakes of potassium, and a lean body habitus.

Lifestyle modification, which includes dietary changes and increased physical activity, has important roles in both nonhypertensive and hypertensive individuals. In nonhypertensive individuals, including those with prehypertension, lifestyle modifications have the potential to prevent hypertension, reduce blood pressure, and thereby lower the risk of blood pressure-related cardiovascular disease. Even an apparently small reduction in blood pressure, if applied to an entire population, could have an enormous beneficial impact. It has been estimated that a 3 mmHg reduction in systolic blood pressure could lead to an 8% reduction in stroke mortality and a 5% reduction in mortality from coronary heart disease (**Figure 2**). In hypertensive individuals, lifestyle modifications can serve as initial

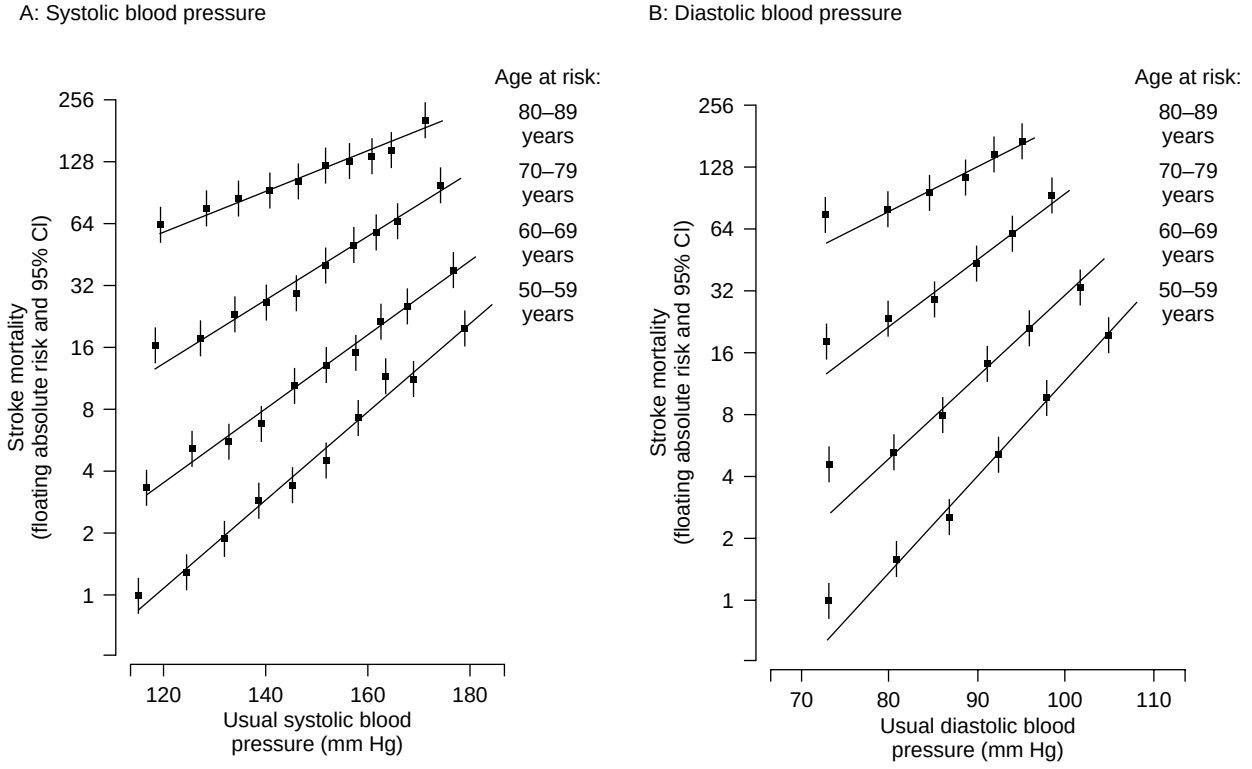
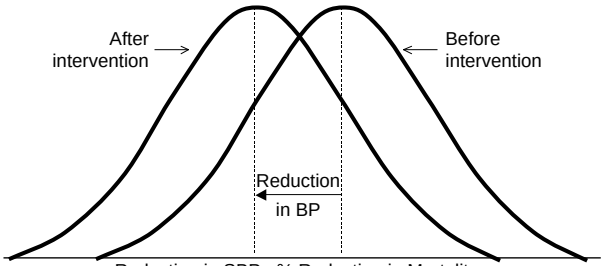


Figure 1 Stroke mortality rate by decade of age versus systolic blood pressure (A) and diastolic blood pressure (B): meta-analysis of 61 prospective studies with 2.7 million person-years. (Reprinted with permission from Lewington S, Clarke R, Qizilbash N, Peto R, and Collins R (2002) Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* **360**: 1903–13.



Reduction in SBP mm Hg	% Reduction in Mortality Stroke	CHD	Total
2	−6	−4	−3
3	−8	−5	−4
5	−14	−9	−7

Figure 2 Estimated effects of populationwide shifts in systolic blood pressure (SBP) on mortality. (Reprinted with permission from Stamler R (1991) Implication of the INTERSALT study *Hypertension* **17**: I-16–I-20.)

treatment before the start of drug therapy and as an adjunct to medication in people already on antihypertensive drug therapy. In hypertensive individuals with medication-controlled blood pressure, lifestyle therapies can facilitate drug step-down and potentially drug withdrawal in individuals who sustain lifestyle changes.

Dietary Factors That Lower Blood Pressure

Weight Loss

On average, as weight increases, so does blood pressure. The importance of this relationship is reinforced by the high and increasing prevalence of overweight and obesity throughout the world. With rare exception, clinical trials have documented that weight loss lowers blood pressure. Importantly, reductions in blood pressure occur before and without attainment of a desirable body weight. In one meta-analysis that aggregated results across 25 trials, mean systolic and diastolic blood pressure reductions from an average weight loss of 5.1 kg were 4.4 and 3.6 mmHg, respectively. Greater weight loss leads to greater blood pressure reduction.

Additional trials have documented that modest weight loss can prevent hypertension by approximately 20% among overweight, prehypertensive individuals and can facilitate medication step-down and drug withdrawal. Lifestyle intervention trials have uniformly achieved short-term weight loss, primary through a reduction in total caloric intake. In some instances, substantial weight loss has also been sustained over 3 or more years.

In aggregate, available evidence strongly supports weight reduction, ideally attainment of a body mass index less than 25 kg/m^2, as an effective approach to prevent and treat hypertension. Weight reduction can also prevent diabetes and control lipids. Hence, the beneficial effects of weight reduction in preventing cardiovascular–renal disease should be substantial. Finally, in view of the well-recognized challenges of maintaining weight loss, efforts to prevent weight gain among those with a normal body weight are critical.

Reduced Salt Intake

On average, as dietary salt (sodium chloride) intake rises, so does blood pressure.[1] To date, more than 50 randomized trials have tested the effects of salt on blood pressure, including several dose–response trials. Approximately 10 meta-analyses have aggregated data across these trials. In a recent meta-analysis that focused on moderate reductions in salt intake, a reduced sodium intake of 1.8 g/day (77 mmol/day) led to average systolic/diastolic blood pressure reductions of 5.2/3.7 mmHg in hypertensives and 1.3/1.1 mmHg in nonhypertensives.

One of the most important dose–response trials is the DASH-Sodium trial, which tested the effects of three different salt intakes separately in two distinct diets—the DASH (Dietary Approaches to Stop Hypertension) diet and a control diet more typical of what Americans eat. As displayed in **Figure 3**, the rise in blood pressure with higher salt intake was evident in both diets. Of note, the blood pressure response to salt intake was nonlinear. Specifically, decreasing salt intake caused a greater lowering of blood pressure when the starting sodium intake was less than 2.3 g/day (100 mmol/day) than when it was above this level.

The blood pressure response to changes in salt intake is heterogeneous. Despite the use of the terms 'salt sensitive' and 'salt resistant' to classify individuals in research studies, the change in blood pressure in response to a change in salt intake is not binary. Instead, the change in blood pressure from a reduced salt intake has a continuous distribution, with individuals having greater or lesser degrees of blood pressure reduction. Genetic factors influence the response to salt reduction. Concomitant diet also modifies the effects of salt on blood pressure. The rise in blood pressure for a given increase in salt intake is blunted in the setting of either the DASH diet or a high potassium intake (**Figure 3**). In general, the effects of salt on blood pressure tend to be greater in blacks, middle-aged and older people, and individuals with hypertension, diabetics, or chronic kidney disease. Although it is possible to identify groups that tend to be salt sensitive, it is impossible, given currently available diagnostic tools, to identify individuals who are salt sensitive.

[1]In view of the format of published data and of dietary recommendations, data are presented as g/day (mmol/day) of sodium rather than g/day of salt.

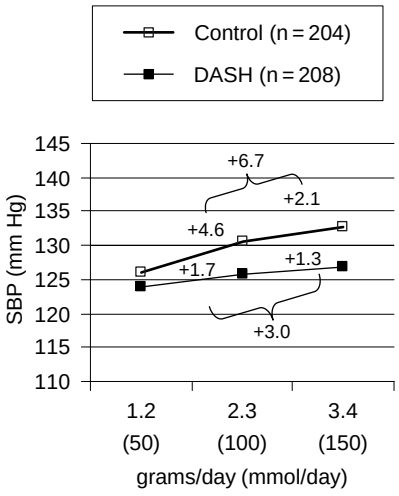


Figure 3 Mean systolic blood pressure (SBP) change in the DASH-Sodium trial from salt reduction in two diets and from the DASH diet at three salt levels. (Adapted with permission from Sacks FM, Svetkey LP, Vollmer WM *et al.* (2001) A clinical trial of the effects on blood pressure of reduced dietary sodium and the DASH dietary pattern (The DASH-Sodium Trial). *New England Journal of Medicine* **344**: 3–10.)

In addition to lowering blood pressure, clinical trials have documented that a reduced salt intake can prevent hypertension by approximately 20% (with or without concomitant weight loss) and can lower blood pressure in the setting of antihypertensive medication. Evidence from observational studies suggests that a reduced salt intake can blunt the age-related rise in systolic blood pressure (**Figure 4**) and can potentially prevent atherosclerotic cardiovascular events and heart failure. A reduced salt intake may also reduce the risk of left ventricular hypertrophy, osteoporosis, and gastric cancer.

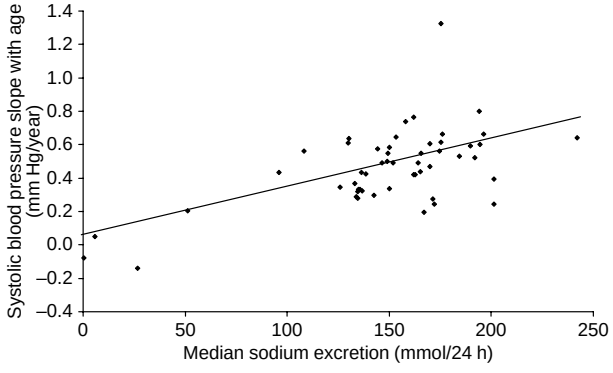


Figure 4 Slope of systolic blood pressure increase with age plotted by median sodium excretion in 52 communities worldwide: results from INTERSALT. (Adapted with permission from Rose G, Stamler J, Stamler R *et al.* (1988) INTERSALT: An international study of electrolyte excretion and blood pressure. Results for 24 hour urinary sodium and potassium excretion. *British Medical Journal* **297**: 319–328.)

Still, the effects of salt on health have been debated. Some have argued that the increases in plasma renin activity and perhaps insulin resistance that occur as a result of a reduced salt intake mitigate the beneficial effects of salt reduction on blood pressure. However, in contrast to blood pressure, the clinical relevance of increased plasma renin activity is uncertain, especially because antihypertensive medications that raise plasma renin levels actually lower cardiovascular disease risk. It has also been argued that a reduced salt intake has little or no effect on blood pressure in many individuals and that other aspects of diet (e.g., increased potassium intake or adoption of a mineral-rich diet) mitigate the harmful effects of salt on blood pressure. Although one cannot guarantee that all individuals will achieve a lower blood pressure from salt reduction, the fraction of individuals who will benefit is substantial.

In view of the progressive dose–response relationship between salt intake and blood pressure, it is difficult to set specific levels for dietary recommendations. Recently, an Institute of Medicine committee set 1.5 g/day (65 mmol/day) of sodium as an adequate intake level and 2.3 g/day (100 mmol/day) as an upper limit. Western-type diets that provide 1.5 g/day (65 mmol/day) have been shown to provide adequate levels of other nutrients. This level of salt intake also allows for excess sweat salt loss among unacclimatized individuals who become physically active or who become exposed to high temperatures. Numerous policymaking organizations have recommended an upper limit of 2.3 g/day (100 mmol/day) for sodium intake.

In most Western counties, average intake of sodium is high, greatly exceeding 2.3 g/day (100 mmol/day). In the United States, the median intake of sodium from foods, not including salt added at the table, varies by age and, according to a recent survey, ranges from 3.1 to 4.7 g/day (135 to 204 mmol/day) in adult men and 2.3 to 3.1 g/day (100 to 135 mmol/day) in adult women. Worldwide, there is greater variation in sodium intake, ranging from an estimated mean intake of 0.02 g/day (1.0 mmol/day) in Yanomamo Indians to more than 10.3 g/day (450 mmol/day) in northern Japanese.

In aggregate, available data strongly support current populationwide recommendations to lower salt intake. To reduce salt intake, consumers should choose foods low in salt and limit the amount of salt added to food. However, even motivated individuals find it difficult to reduce salt intake because more than 75% of consumed salt comes from processed foods (**Figure 5**). Hence, any meaningful strategy to reduce salt intake must involve the efforts of food manufacturers, who should reduce the amount of salt added during food processing.

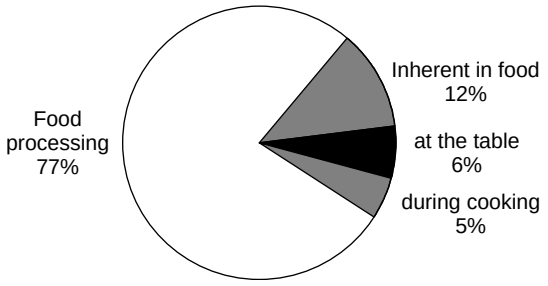


Figure 5 Sources of dietary sodium. (Data from Mattes RD and Donnelly D (1991) Relative contributions of dietary sodium sources *Journal of the American College of Nutrition* **10**: 383–393.)

Increased Potassium intake

High levels of potassium intake are associated with reduced blood pressure. Observational data have been reasonably consistent in documenting this inverse relationship, whereas data from individual trials have been less consistent. However, three meta-analyses of these trials have each documented a significant inverse relationship between potassium intake and blood pressure in nonhypertensive and hypertensive individuals. In one meta-analysis, average net systolic/diastolic blood pressure reductions from increased potassium intake were 4.4/2.4 mmHg. Available studies have documented greater blood pressure reductions from potassium in African Americans compared to non-African Americans. A high potassium intake has been shown to blunt the rise in blood pressure in response to increased salt intake. Potassium has greater blood pressure lowering in the context of a higher salt intake and lesser blood pressure reduction in the setting of a lower salt intake. Conversely, the blood pressure reduction from a reduced salt intake is greatest when potassium intake is low. These data are consistent with subadditive effects of reduced salt intake and increased potassium intake on blood pressure.

Most trials that tested the effects of potassium on blood pressure used pill supplements, typically potassium chloride. However, in foods, the conjugate anions associated with potassium are mainly citrate and other bicarbonate precursors. The latter is important because other potential benefits of foods rich in potassium (i.e., reduced risk of kidney stones and reduced bone turnover) likely result from effects of the conjugate anion. Because a high dietary intake of potassium can be achieved through diet rather than pills and because potassium derived from foods also comes with a variety of other nutrients, the preferred strategy to increase potassium intake is to consume foods, such as fruits and vegetables, rather than supplements.

On the basis of available data, an Institute of Medicine committee set an Adequate Intake for potassium of 4.7 g/day (120 mmol/day) for adults. This level of dietary intake should maintain lower blood pressure levels, reduce the adverse effects of salt on blood pressure, reduce the risk of kidney stones, and possibly decrease bone loss. Currently, dietary intake of potassium is considerably lower than this level. In recent surveys, the median intake of potassium by adults in the United States was approximately 2.9–3.2 g/day (74–82 mmol/day) for men and 2.1–2.3 g/day (54–59 mmol/day) for women. Because African Americans have a relatively low intake of potassium and a high prevalence of elevated blood pressure and salt sensitivity, this subgroup of the population would especially benefit from an increased potassium intake.

In the generally healthy population with normal kidney function, a potassium intake from foods higher than 4.7 g/day (120 mmol/day) poses no potential for increased risk because excess potassium is readily excreted in the urine. However, in individuals whose urinary potassium excretion is impaired, a potassium intake of less than 4.7 g/day (120 mmol/day) is appropriate because of adverse cardiac effects (arrhythmias) from hyperkalemia. Common drugs that impair potassium excretion are angiotensin converting enzyme inhibitors, angiotensin receptor blockers, and potassium-sparing diuretics. Medical conditions associated with impaired potassium excretion include diabetes, chronic renal insufficiency, end stage renal disease, severe heart failure, and adrenal insufficiency. Elderly individuals are at increased risk of hyperkalemia because they often have one or more of these conditions or take one or more of the medications that impair potassium excretion.

Moderation of Alcohol Intake

The relationship between alcohol intake and blood pressure is direct and progressive, particularly at an alcohol intake above approximately two drinks per day (~1 oz. or ~28 g of ethanol per day). A meta-analysis of 15 trials reported that decreased consumption of alcohol (median reduction in self-reported alcohol consumption of 76%) lowered systolic and diastolic blood pressure by 3.3 and 2.0 mmHg, respectively. In nonhypertensives and hypertensives, blood pressure reductions were similar. In aggregate, evidence supports moderation of alcohol intake (among those who drink) as an effective approach to lower blood pressure. It is recommended that alcohol consumption be limited to no more than 1 oz. (30 ml) of ethanol (e.g., 24 oz. (720 ml) beer, 10 oz. (300 ml) wine, or 2 oz. (60 ml) 100-proof whiskey) per day in most men and to no

more than 0.5 oz. (15 ml) ethanol per day in women and lighter weight people.

Whole Dietary Patterns

Vegetarian diets Vegetarian diets have been associated with low blood pressure. In observational studies, vegetarians also experience a markedly lower, age-related rise in blood pressure. Aspects of a vegetarian lifestyle that might affect blood pressure include nondietary factors (e.g., physical activity), established dietary risk factors (e.g., salt, potassium, weight, and alcohol), and other aspects of a vegetarian diet (e.g., high fiber and no meat). To a very limited extent, observational studies have controlled for the well-established determinants of blood pressure. Hence, it is unclear whether blood pressure reductions result from established dietary risk factors that affect blood pressure or from other aspects of a vegetarian diet.

The DASH diet The DASH trial tested whether modification of whole dietary patterns might affect blood pressure. In this trial, participants were randomized to eat one of three diets: (i) a control diet, (ii) a diet rich in 'fruits and vegetables' but otherwise similar to control, or (iii) the DASH diet. The DASH diet emphasizes fruits, vegetables, and low-fat dairy products; includes whole grains, poultry, fish, and nuts; and is reduced in fats, red meat, sweets, and sugar-containing beverages. Accordingly, it is rich in potassium, magnesium, calcium, and fiber and reduced in total fat, saturated fat, and cholesterol; it is also slightly increased in protein.

Among all participants, the DASH diet significantly lowered mean systolic blood pressure by 5.5 mmHg and mean diastolic blood pressure by 3.0 mmHg. The fruits and vegetables diet also reduced blood pressure but to a lesser extent—approximately half of the effect of the DASH diet. The effect was relatively rapid; the full effect was apparent after 2 weeks (**Figure 6**). In subgroup analyses, the DASH diet significantly lowered blood pressure in all major subgroups (men, women, African Americans, non-African Americans, hypertensives, and nonhypertensives). However, the effects of the DASH diet were especially prominent in African Americans, who experienced net systolic/diastolic blood pressure reductions of 6.9/3.7 mmHg, and hypertensive individuals, who experienced net blood pressure reductions of 11.6/5.3 mmHg.

Results from the DASH trial have important clinical and public health implications. The effect of the DASH diet in hypertensive individuals was similar in magnitude to that of drug monotherapy. From a public health perspective, the DASH diet could potentially shift the population distribution of blood pressure downward, thereby reducing the risk of blood pressure-related cardiovascular disease (**Figure 2**).

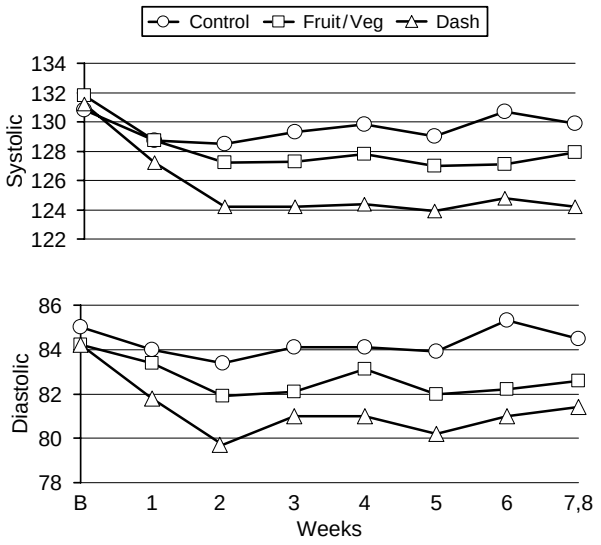


Figure 6 Blood pressure by week during the DASH feeding study in three diets (control diet, fruits and vegetables diet, and DASH diet). (Adapted with permission from Appel LJ, Moore TJ, Obarzanek E *et al.* (1997) The effect of dietary patterns on blood pressure: Results from the Dietary Approaches to Stop Hypertension (DASH) clinical trial. *New England Journal of Medicine* **336**: 1117–1124.)

Fish Oil Supplementation

High-dose, omega-3 polyunsaturated fatty acid (commonly termed 'fish oil') supplements can lower blood pressure in hypertensive individuals. In a meta-analysis of trials, average systolic and diastolic blood pressure reductions in hypertensive individuals were 5.5 and 3.5 mmHg, respectively. The effect of fish oil appears to be dose dependent, with blood pressure reductions only occurring at relatively high doses, namely 3 g/day or more. In nonhypertensive individuals, blood pressure reductions were nonsignificant and small. Side effects, including belching and a fishy taste, are common. In view of the side effect profile and the high dose required to lower blood pressure, fish oil supplements are not routinely recommended.

Dietary Factors with Limited or Uncertain Effect on Blood Pressure

Fiber

Evidence from observational studies and several clinical trials suggests that increased fiber intake may reduce blood pressure. A meta-analysis

documented that supplemental fiber (average increase of 14 g/day) was associated with net systolic/diastolic reductions of 1.6/2.0 mmHg, respectively. Still, high-quality epidemiologic studies and clinical trials are needed before one can recommend increased fiber intake as a means to lower blood pressure.

Calcium and Magnesium

Evidence that increased calcium intake might lower blood pressure comes from a variety of sources, including animal studies, observational studies, clinical trials, and meta-analyses. Meta-analyses of trials documented modest reductions in systolic and diastolic blood pressure of 0.89–1.44 and 0.18–0.84 mmHg, respectively, with calcium supplementation (400–2000 mg/day). There is also evidence that calcium intake may affect the blood pressure response to salt. Overall, data are insufficient to recommend supplemental calcium alone as a means to lower blood pressure.

The body of evidence implicating magnesium as a major determinant of blood pressure is inconsistent. In observational studies, often cross-sectional in design, a common finding is an inverse association of dietary magnesium with blood pressure. However, in pooled analyses of clinical trials, there is no clear effect of magnesium intake on blood pressure. Hence, data are insufficient to recommend increased magnesium intake alone as a means to lower blood pressure.

Fats (Other Than Fish Oil) and Cholesterol

Numerous studies, including both observational studies and clinical trials, have examined the effects of fat intake on blood pressure. Overall, there is no apparent effect of saturated fat and n-6 polyunsaturated fat intake on blood pressure. Although a few trials suggest that an increased intake of monounsaturated fat may lower blood pressure, evidence is insufficient to make recommendations. Likewise, few studies have examined the effect of dietary cholesterol intake on blood pressure. Hence, although modification of dietary fat and cholesterol intake can be recommended as a means to prevent and treat hyperlipidemia and dyslipidemia, evidence is insufficient to recommend these changes alone as a means to lower blood pressure.

Protein Intake

A large and generally consistent body of evidence from observational studies has documented that higher protein intake, particularly protein from plant-based sources, is associated with lower blood pressure. In contrast to the large volume of evidence from observational studies, comparatively few trials have examined the effects of protein intake on blood pressure. Recent trials have tested the effects of soy-based interventions on blood pressure. In several but not all of these trials, soy supplementation reduced blood pressure. Although it is reasonable to speculate that an increased intake of protein from plant sources can lower blood pressure, this hypothesis has not been adequately tested in a clinical trial of sufficient size and rigor.

Vitamin C

Laboratory studies, depletion–repletion studies, and epidemiological studies suggest that increased vitamin C intake or status is associated with lower blood pressure. However, few trials have addressed this issue, and results of these trials have been inconsistent. Overall, it remains unclear whether an increased intake of vitamin C lowers blood pressure.

Gene–Diet Interactions

A rapidly increasing body of evidence indicates that genetic factors affect blood pressure levels and the blood pressure response to dietary changes. Most of the evidence relates to genetic factors that influence the blood pressure response to salt. Several genotypes that influence blood pressure have been identified. Most of these genotypes influence the renin–angiotensin–aldosterone axis or renal salt handling.

Special Populations

Children

Elevated blood pressure begins well before adulthood, during the first two decades of life and perhaps earlier during gestation. Numerous observational studies have documented that blood pressure tracks with age from childhood into the adult years. Hence, efforts to reduce blood pressure and to prevent the age-related rise in blood pressure in childhood are prudent.

Direct empiric evidence from rigorous, well-controlled trials in children and adolescents is sparse. There is some direct evidence from studies conducted in children that the dietary determinants of blood pressure in children and adults are similar. In this setting, the effect of diet on blood pressure in children and adolescents is, in large part, extrapolated from studies of adults. Such extrapolations are reasonable because elevated blood pressure is a chronic condition resulting from the insidious

rise in blood pressure throughout childhood and adulthood.

Pregnant Women

Hypertension during pregnancy is a constellation of diverse clinical conditions, some of which can be extremely serious. Of substantial concern are preeclampsia and eclampsia. Both are multisystem disorders that are manifest by the onset of hypertension and proteinuria during the second half of pregnancy. Convulsions occur in the setting of eclampsia but not preeclampsia. The cause of these disorders is unknown. Several dietary interventions, including salt restriction, fish oil supplementation, and calcium supplementation, have been tested as a means to prevent preeclampsia, but none is considered effective. Although a meta-analysis of small trials suggested that calcium supplementation has some benefit in high-risk women, a large trial of calcium supplementation documented no benefit, either overall or in high-risk subgroups.

Older People

Because of the age-related rise in systolic blood pressure and because of the high prevalence of blood pressure-related cardiovascular disease in middle-aged and older people, dietary strategies should be especially beneficial as adults age. It is well documented that older people can make and sustain dietary changes, specifically weight loss and dietary salt reduction. Furthermore, salt sensitivity increases as individuals age. Lastly, because of the high attributable risk associated with elevated blood pressure in older people, the beneficial effects of dietary changes on blood pressure should translate into substantial reductions in cardiovascular risk in this age group.

Populations Defined by Race/Ethnicity or Geography

Worldwide, there is substantial variation in blood pressure among populations. In certain primitive societies, such as the Yanomamo Indians in Brazil, blood pressure does not rise with age, and hypertension is absent. In rural Africa and southern China, the prevalence of hypertension is less than 20%. Among urbanized populations, the prevalence of hypertension is high, especially among African Americans, a population in which the prevalence of hypertension approaches 40%. Other groups, such as Australian Aborigines, Eastern Europeans, and Russians, also have a high prevalence of hypertension.

Understanding the causes of geographic variation is difficult. However, migration studies provide strong evidence that modifiable environmental factors (e.g., diet and physical activity) rather than genetic factors or geographic factors account for this variation. Furthermore, as noted previously, trials have documented that compared to non-African Americans, African Americans achieve greater blood pressure reduction from several nonpharmacologic therapies, specifically a reduced salt intake, increased potassium intake, and the DASH diet. The potential benefits of these dietary therapies is amplified because US survey data indicate that African Americans consume less potassium than non-African Americans. On average, salt intake is high and similar in African Americans and non-African Americans. Hence, changes in diet should provide a means to reduce racial and perhaps geographic disparities in blood pressure.

Conclusion

In view of the continuing epidemic of blood pressure-related cardiovascular disease, efforts to reduce blood pressure in both nonhypertensive and hypertensive individuals are warranted. Such efforts will require individuals to change behavior and society to make substantial environmental changes. The current challenge to health care providers, researchers, government officials, and the general public is to develop and implement effective clinical and public health strategies that lead to sustained dietary changes among individuals and more broadly among populations.

See also: **Alcohol**: Absorption, Metabolism and Physiological Effects; Disease Risk and Beneficial Effects. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements; Deficiency States. **Calcium**. **Fish**. **Hypertension**: Etiology; Nutritional Management. **Magnesium**. **Obesity**: Complications. **Older People**: Physiological Changes. **Potassium**. **Pregnancy**: Energy Requirements and Metabolic Adaptations. **Sodium**: Physiology; Salt Intake and Health. **Vegetarian Diets**.

Further Reading

Appel LJ, Moore TJ, Obarzanek E *et al.* (1997) The effect of dietary patterns on blood pressure: Results from the Dietary Approaches to Stop Hypertension (DASH) clinical trial. *New England Journal of Medicine* **336**: 1117–1124.

Chobanian AV, Bakris GL, Black HR *et al.* (2003) National High Blood Pressure Education Program Coordinating Committee.

Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension* **42**: 1206–1252.

Institute of Medicine (2004) *Dietary Reference Intakes for Water, Potassium, Sodium, Chloride and Sulfate*. Washington, DC: National Academy of Sciences.

Izzo JL and Black HR (eds.) (2003) *Hypertension Primer: The Essentials of High Blood Pressure*, 3rd edn. Washington, DC: American Heart Association.

Krauss RM, Eckel RH, Howard B *et al.* (2000) AHA dietary guidelines: Revision 2000: A statement for healthcare professionals from the nutrition committee of the American Heart Association. *Circulation* **102**: 2284–2299.

Lewington S, Clarke R, Qizilbash N, Peto R, and Collins R (2002) Age-specific relevance of usual blood pressure to vascular mortality: A meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* **360**: 1903–1913.

Rose G, Stamler J, Stamler R *et al.* (1988) INTERSALT: An international study of electrolyte excretion and blood pressure. Results for 24 hour urinary sodium and potassium excretion. *British Medical Journal* **297**: 319–328.

Sacks FM, Svetkey LP, Vollmer WM *et al.* (2001) A clinical trial of the effects on blood pressure of reduced dietary sodium and the DASH dietary pattern (The DASH-Sodium Trial). *New England Journal of Medicine* **344**: 3–10.

Simons-Morton DG and Obarzanek E (1997) Diet and blood pressure in children and adolescents. *Pediatric Nephrology* **11**: 244–249.

Stamler J, Stamler R, and Neaton JD (1993) Blood pressure, systolic and diastolic, and cardiovascular risks: U.S. population data. *Archives of Internal Medicine* **153**: 598–615.

Vasan RS, Beiser A, Seshadri S *et al.* (2002) Residual life-time risk for developing hypertension in middle-aged women and men: The Framingham Heart Study. *Journal of the American Medical Association* **287**: 1003–1010.

Whelton PK, He J, Appel LJ *et al.* (2002) Primary prevention of hypertension. Clinical and public health advisory from the National High Blood Pressure Education Program. *Journal of the American Medical Association* **288**: 1882–1888.

Whelton PK, He J, and Louis GT (eds.) (2003) *Lifestyle Modification for the Prevention and Treatment of Hypertension*. New York: Marcel Dekker.

Nutritional Management

C M Champagne, Pennington Biomedical Research Center, Baton Rouge, LA, USA

Hypertension or high blood pressure affects more than 25% of adult Americans (50 million) and Canadians, and the rate of hypertension is reportedly as much as 60% higher in some European countries. African Americans typically have higher rates of hypertension compared to whites. Hypertension is also more common among older people, occurring in two-thirds of those older than 65 years of age—a population that is often untreated. The lifetime risk of developing hypertension for middle-aged and elderly individuals is 90%.

Recently, new guidelines suggest that the previous values of 120 mmHg/80 mmHg should not be considered normal. The revised categories are shown in **Table 1**. The recommendations for management of hypertension include lifestyle modification for all categories (even normal), with drugs not routinely promoted until the patient presents with at least stage 1 hypertension or other compelling indications.

The most definitive trials directed toward the nutritional management of hypertension are the DASH (Dietary Approaches to Stop Hypertension) and DASH-Sodium trials. DASH focused on establishing dietary patterns that would lower blood pressure while keeping sodium content constant. DASH-Sodium was designed to test the effects of varying levels of sodium in conjunction with the DASH diet in order to determine whether lowering sodium intake would have additional beneficial effects. Both trials were metabolic feeding trials, and each enrolled more than 400 participants at four sites in the United States. As a follow-up to the DASH and DASH-Sodium trials, essentially the same group of investigators conducted the PREMIER trial to test whether individuals could lower blood pressure by implementing established guidelines for treating hypertension and included the DASH diet in addition to the established recommendations. PREMIER used a lifestyle counseling approach and randomized more than 800 subjects to one of three treatment arms: advice only, established recommendations, and established recommendations plus DASH.

Table 1 Classification of blood pressure for adults 18 years of age or older

Category	*Systolic blood pressure (mmHg)*	*Diastolic blood pressure (mmHg)*
Normal or desirable	<120	<80
Prehypertensive	120–139	80–89
Stage 1 hypertension	140–159	90–99
Stage 2 hypertension	≥160	≥100

U.S. Dept of Health and Human Services, National Institutes of Health, National Heart, Lung and Blood Institute. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. NIH Publication No. 04-5230, August 2004.

Recommended Lifestyle Modifications

Traditionally, the following lifestyle modifications have been recommended for the treatment of hypertension:

Lose weight, if one is overweight.
Reduce intake of salt, sodium, and foods containing them as much as possible.
Increase physical activity.
Limit intake of alcohol.
Stop smoking.
Control stress.

The established recommendations for lifestyle modification used in one arm of the PREMIER clinical trial were weight loss, increasing physical activity, reducing sodium intake, limiting alcohol consumption, and reducing total and saturated fat intake to that of an American Heart Association step 1 diet with 30% of energy from total and 10% from saturated fat. A second arm in PREMIER included essentially the same lifestyle modifications but a lower fat diet comparable to an American Heart Association step 2 diet with 25% of energy from total and 7% from saturated fat and also adherence to the DASH diet (emphasizing consumption of fruits, vegetables, and low-fat dairy products).

Nutritional Considerations

Reduction of Sodium Intake

Sodium reduction typically results in lower blood pressure in industrialized societies. Current guidelines in the United States suggest reducing the daily intake of sodium to approximately 100 mmol, or approximately 2.4 g of sodium or less per day.

The DASH-Sodium trial demonstrated that reduction of sodium intake from 100 to 50 mmol per day (approximately 1.5 g) significantly reduced blood pressure in individuals following either the common US diet or the DASH diet. In addition, TOHP2 (phase 2 of the Trials of Hypertension Prevention) and TONE (Trials of Nonpharmacologic Interventions in the Elderly) documented that reducing sodium can either prevent hypertension or facilitate hypertension control. It should also be noted that salt sensitivity increases with age, so those who demonstrate this sensitivity should maintain a reduced salt diet.

Consumers should either eliminate or limit salt added to foods in cooking and at the table as a means of reducing sodium intake. Nutrition facts labels require sodium content to be listed so that consumers can be more prudent about their diets. The amount of sodium in processed food, such as convenience foods (e.g., boxed products one would prepare at home), soups, and processed meats (e.g., sausage, ham, and other meat products), is often alarming. If there is not a nutrition label on a processed food product, one should assume sodium content is high. Canned products generally contain more sodium than fresh or frozen items, unless a product is specifically labeled as 'no salt added.' The consumption of fresh, unprocessed foods should be promoted.

Moderation of Alcohol Intake

The relationship between high consumption of alcohol (typically three or more drinks per day) and elevated blood pressure has been shown in numerous epidemiologic studies. A drink is defined as 12 oz. of beer, 5 oz. of wine, or 1.5 oz. of distilled spirits. Most evidence indicates that alcohol should be limited to two drinks per day for men and one drink per day for women. Ideally, daily alcohol consumption should be avoided. Whenever possible, alcohol, if consumed, should be done so with meals.

Consumption of a DASH Diet (Increasing Potassium, Magnesium, Calcium, and Fiber Intakes by Increasing Intakes of Fruits, Vegetables, and Low-Fat Dairy Foods)

The contribution of minerals, particularly potassium, magnesium, and calcium, and fiber was identified by contributions from fruits, vegetables, low-fat dairy products, whole grains, and nuts in the DASH eating plan. The DASH diet effectively used these components through an ideal dietary pattern to lower blood pressure.

Increased intakes of potassium have been associated with lower blood pressure. A meta-analysis of several trials suggested that 60–120 mmol per day of supplemental potassium reduces systolic and diastolic blood pressure by 4.4 and 2.5 mmHg, respectively, in hypertensive individuals. In normotensive individuals, systolic and diastolic blood pressure was reduced by 1.8 and 1.0 mmHg, respectively. Dietary intake of potassium can be easily achieved through consumption of various foods.

The DASH diet, while promoting dietary patterns, was developed very carefully with particular attention paid to the use of specific foods within categories that contribute more to the intakes of desired nutrients. As an example, consider the rank-ordered listing of potassium content of fruits and fruit juices presented in **Tables 2** and **3**. Dried fruits typically have the highest potassium content, followed by raw fruits and frozen fruits. Canned fruit products generally do not contain as high potassium content as other forms. There is less potassium contained in fruit juices and generally the

Table 2 Fruits, ranked by potassium content (mg/100 g)

Fruit	*K (mg)*
Dried	
Apricots, dehydrated (low moisture)	1850
Bananas, dehydrated, or banana powder	1491
Peaches (low moisture)	1351
Apricots	1162
Litchis	1110
Prunes (low moisture)	1058
Peaches	996
Currants, zante	892
Persimmons, Japanese	802
Raisins	746–825
Plums (prunes)	732
Dates, medjool	696
Figs	680
Longans	658
Dates, deglet noor	656
Apples (low moisture)	640
Pears	533
Jujube	531
Apples	450
Raw	
Tamarinds	628
Plantains	499
Breadfruit	490
Avocados	485
Durian, raw or frozen	436
Custardapple (bullock's heart)	382
Bananas	358
Passion-fruit, (granadilla), purple	348
Sapotes (marmalade plum)	344
Currants, European black	322
Kiwi fruit, (Chinese gooseberries)	312
Persimmons, native	310
Abiyuch	304
Jackfruit	303
Rhubarb	288
Guavas, common	284
Elderberries	280
Soursop	278
Currants, red and white	275
Cherimoya	269
Melons, cantaloupe	267
Longans	266
Loquats	266
Carissa (natal-plum)	260
Apricots	259
Pomegranates	259
Papayas	257
Jujube	250
Sugar apples (sweetsop)	247
Figs	232
Melons, honeydew	228
Cherries, sweet	222
Prickly pears	220
Pummelo	216
Roselle	208
Nectarines	201
Gooseberries	198
Quinces	197
Crabapples	194
Mulberries	194
Sapodilla	193
Grapes	191
Peaches	190
Kumquats	186
Melons, casaba	182
Cherries, sour, red	173
Litchis	171
Oranges	166–196
Carambola (starfruit)	163
Blackberries	162
Persimmons, Japanese	161
Plums	157
Tangerines (mandarin oranges)	157
Mangos	156
Feijoa	155
Strawberries	153
Raspberries	151
Acerola, (West Indian cherry)	146
Lemons	138–145
Rowal	131
Grapefruit	127–150
Rose apples	123
Pears, asian	121
Pears	119
Limes	117
Pineapple	115–125
Watermelon	112
Apples, with skin	107
Pitanga, (surinam cherry)	103
Apples, without skin	90
Cranberries	85
Java plum (jambolan)	79
Blueberries	77
Mammy apple (mamey)	47
Oheloberries	38
Fruits, frozen	
Strawberries	148
Loganberries	145
Boysenberries	139
Cherries, sour, red	124
Raspberries, red	114
Rhubarb	108

From U.S. Department of Agriculture, Agricultural Research Service (2003) *USDA National Nutrient Database for Standard Reference*, Release 16. Available at www.nal.usda.gov/fnic/foodcomp.

fresh forms of the juices have incrementally more than the processed forms. Fruits and juices in general contain some magnesium, another mineral of interest to the DASH investigators. Most fruits contain 2–30 mg of magnesium per 100 grams, but dried fruits contain much more (30–90 mg) and the amounts vary greatly. Fruit juices contain less than 20 mg of magnesium per 100 grams, with most containing less than 10 mg. Fiber content of fruit ranges from approximately 7 to 14 g of fiber for dried fruits on a per 100 gram basis and between 1 and 5 g for other fruits per 100 grams. Generally, fruit juices contribute less than 1 g of dietary fiber per 100 grams, but high-pulp varieties of juices provide slightly more dietary fiber.

Table 3 Fruit juices, ranked by potassium content (mg/100 g)

Fruit juice	*K (mg)*
Passion fruit juice, fresh	278
Prune juice, canned	276
Orange juice, fresh	200
Orange juice, from concentrate	190
Tangerine juice, fresh or canned	178
Orange juice, canned	175
Grapefruit juice, white or pink, fresh or canned	162
Pineapple juice, canned or from concentrate	134
Grape juice, canned or bottled, unsweetened	132
Apple juice, from frozen concentrate, unsweetened	126
Lemon juice, fresh	124
Apple juice, canned or bottled, unsweetened	119
Apricot nectar	114
Lime juice, fresh	109
Lemon juice, canned or bottled	102
Acerola juice, fresh	97
Cranberry juice, unsweetened	77
Lime juice, canned or bottled	75
Peach nectar	40
Papaya nectar	31
Grape juice, from frozen concentrate, sweetened	21
Pear nectar	13

From U.S. Department of Agriculture, Agricultural Research Service (2003) *USDA National Nutrient Database for Standard Reference*, Release 16. Available at www.nal.usda.gov/fnic/foodcomp.

Table 4 contains a rank-ordered listing of vegetables (including beans) by content of potassium. Magnesium content is also shown. The data are presented for vegetables and beans in the raw form generally. It is important to remember that many fresh forms are concentrated in terms of weight when cooked, especially spinach and other greens, and it is thus possible to obtain a higher mineral content from cooked vegetables (especially in the case of potassium). The magnesium content differs less from the fresh to the cooked state for most vegetables and beans. Most vegetables contain approximately 1–3 g of dietary fiber per

Table 4 Potassium and magnesium content of vegetables (including beans), rank ordered by potassium content (mg/100 g; presented for raw vegetables unless otherwise specified)

Vegetable	*K (mg)*	*Mg (mg)*
Tomatoes, sun-dried	3427	194
Palm hearts	1806	10
Arrowhead	922	51
Yam	816	21
Beet greens	762	70
Lemon grass (citronella)	723	60
Butterbur (fuki)	655	14
Taro leaves	648	45
Epazote	633	121
Soybeans, green	620	65
Amaranth leaves	611	55
Cress, garden	606	38
Taro, tahitian	606	47
Yautia (tannier)	598	24
Taro	591	33
Winged bean tuber	586	24
Waterchestnuts, Chinese (matai)	584	22
Wasabi, root	568	69
Chrysanthemum, garland	567	32
Chrysanthemum leaves	567	32
Jute, potherb	559	64
Spinach	558	79
Lotus root	556	23
Parsley	554	50
Pigeonpeas, immature seeds	552	68
Bamboo shoots	533	3
Coriander (cilantro) leaves	521	26
Sweetpotato leaves	518	61
Mushroom, oyster	516	20
Vinespinach (basella)	510	65
Purslane	494	68
Fireweed, leaves	494	156
Mushrooms, portabella	484	11
Soybeans, mature seeds	484	72
Borage	470	52
Lima beans, immature seeds	467	58
Horseradish tree, pods	461	45
Corn salad	459	13
Squash, zucchini, baby	459	33
Cowpeas, leafy tips	455	43
Potatoes, red, flesh and skin	455	22
Arrowroot	454	25
Lambsquarters	452	34
Kale, scotch	450	88
Mustard spinach (tendergreen)	449	11
Mushrooms, brown, Italian or Crimini	448	9
Kale	447	34
Pumpkin leaves	436	38
Cowpeas (blackeyes), immature seeds	431	51
Jerusalem artichokes	429	17
Potato, flesh and skin	421	23
Chicory greens	420	30
Mountain yam, Hawaii	418	12
Potatoes, russet, flesh and skin	417	23
Ginger root, raw	415	43
Fennel, bulb, raw	414	17
Potatoes, skin	413	23
Potatoes, white, flesh and skin	407	21
Garlic	401	25
Cardoon	400	42
Dandelion greens	397	36
Dock	390	103
Brussels sprouts	389	23
Peas, mature seeds, sprouted	381	56
Mushrooms, enoki	381	16
Salsify (vegetable oyster)	380	23
Chard, Swiss	379	81
Parsnips	375	29
Artichokes (globe or French)	370	60
Fiddlehead ferns	370	34
Arugula	369	47
Seaweed, laver	356	2
Mustard greens	354	32
Squash, winter, butternut	352	34

Continued

Table 4 Continued

Vegetable	*K (mg)*	*Mg (mg)*
Kohlrabi	350	19
Squash, winter, all varieties	350	14
Pumpkin	340	12
Eppaw	340	32
Peppers, hot chili, green	340	25
Horseradish tree leafy tips	337	147
Rutabagas	337	23
Sweet potato	337	25
Shallots	334	21
Taro shoots	332	8
Beans, fava, in pod	332	33
Celtuce	330	28
Watercress	330	21
Beets	325	23
Broccoli, leaves	325	25
Broccoli, flower clusters	325	25
Broccoli, stalks	325	25
Lentils, sprouted	322	37
Peppers, hot chili, red	322	23
Carrots	320	12
Squash, winter, hubbard	320	19
Broccoli	316	21
Endive	314	15
Mushrooms	314	9
Swamp cabbage (skunk cabbage)	312	71
Burdock root	308	38
Beans, navy, mature seeds	307	101
Beans, pinto, mature seeds	307	53
Pepper, Serrano	305	22
Cauliflower	303	15
Okra	303	57
Radicchio	302	13
Celeriac	300	20
Cauliflower, green	300	20
Balsam pear (bitter gourd), pods	296	17
Chives	296	42
Turnip greens	296	31
Chicory roots	290	22
Radishes, white icicle	280	9
Onions, spring or scallions	276	20
Grape leaves	272	95
Cassava	271	21
Corn, sweet, yellow or white	270	37
Tomatillos	268	20
Squash, summer, all varieties	262	17
Celery	260	11
Tomatoes, yellow	258	12
Nopales	257	52
Pepper, banana	256	17
Cabbage, Chinese (pak-choi)	252	19
Hyacinth beans, immature seeds	252	40
Broadbeans, immature seeds	250	38
Broccoli, frozen	250	16
Lettuce, cos or romaine	247	14
Cabbage	246	15
Peas, green	244	33
Cabbage, red	243	16
Pokeberry shoots (poke)	242	18
Yardlong bean	240	44
Cabbage, Chinese (pe-tsai)	238	13
Lettuce, butterhead	238	13
Tomatoes, red, ripe	237	11
Carrots, baby	237	10
Radishes	233	10
Cabbage, savoy	230	28
Eggplant	230	14
Radishes, oriental	227	16
Seaweed, agar	226	67
Winged beans, immature seeds	223	34
Cowpeas, young pods with seeds	215	58
Peppers, jalapeno	215	19
Onions, welsh	212	23
Squash, summer	212	21
Tomatoes, orange	212	8
Peppers, sweet, yellow	212	12
Chicory, witloof	211	10
Peppers, sweet, red	211	12
Beans, snap, green	209	25
Beans, snap, yellow	209	25
Tomatoes, green	204	10
Asparagus	202	14
Peppers, Hungarian	202	16
Peas, edible, podded	200	24
Broccoli raab	196	22
Turnips	191	11
Beans, kidney, mature seeds	187	21
Lettuce, red leaf	187	12
Sesbania flower	184	12
Poi	183	24
Squash, summer, scallop	182	23
Leeks (bulb and lower leaf portion)	180	28
Winged bean leaves	176	8
Peppers, sweet, green	175	10
Pumpkin flowers	173	24
Lettuce, iceberg	152	8
Gourd, white flowered (calabash)	150	11
Yambean (jicama)	150	12
Mung beans, mature seeds	149	21
Cucumber, with peel	147	13
Onions	144	10
Gourd, dishcloth (towelgourd)	139	14
Cucumber, peeled	136	12
New Zealand spinach	130	39
Seaweed, spirulina	127	19
Chayote, fruit	125	12
Onions, sweet	119	9
Squash, winter, spaghetti	108	12
Seaweed, kelp, raw	89	121
Radish seeds, sprouted, raw	86	44
Alfalfa seeds, sprouted, raw	79	27
Seaweed, irishmoss, raw	63	144
Seaweed, wakame, raw	50	107
Jew's ear (pepeao), raw	43	25

Data source: U.S. Department of Agriculture, Agricultural Research Service. 2003. USDA National Nutrient Database for Standard Reference, Release 16. Nutrient Data Laboratory Home Page, http://www.nal.usda.gov/fnic/foodcomp.

100 grams; beans and legumes offer approximately 5 g of dietary fiber, and some dried vegetables offer more than double this amount.

Nuts were also an important part of the DASH diet, contributing potassium, magnesium, fiber, and protein. They contain fat, mostly monounsaturated, and thereby contribute energy to the diet. **Table 5**

Table 5 Potassium, magnesium, and fiber content of nuts and seeds per 100 grams, ranked by potassium content

Description	*K (mg)*	*Mg (mg)*	*Fiber (g)*
Nuts			
Pistachio nuts, dry roasted	1042	120	10.3
Pistachio nuts, raw	1025	121	10.3
Ginkgo nuts, dried	998	53	9.3
Chestnuts, European, dried, unpeeled	986	74	11.7
Almonds, dry roasted	746	286	11.8
Almonds	728	275	11.8
Almonds, oil roasted	699	274	10.5
Almonds, blanched	687	275	10.4
Hazelnuts or filberts	680	163	9.7
Cashew nuts, raw	660	292	3.3
Brazil nuts, dried, unblanched	659	376	7.5
Hazelnuts or filberts, blanched	658	160	11.0
Cashew nuts, oil roasted	632	273	3.3
Pine nuts, pinyon, dried	628	234	10.7
Pine nuts, pignolia, dried	597	251	3.7
Chestnuts, European, roasted	592	33	5.1
Cashew nuts, dry roasted	565	260	3.0
Walnuts, black, dried	523	201	6.8
Chestnuts, European, raw, unpeeled	518	32	8.1
Ginkgo nuts, raw	510	27	9.3
Walnuts, English	441	158	6.7
Hickorynuts, dried	436	173	6.4
Pecans, dry roasted	424	132	9.4
Butternuts, dried	421	237	4.7
Pecans	410	121	9.6
Pecans, oil roasted	392	121	9.5
Macadamia nuts, raw	368	130	8.6
Macadamia nuts, dry roasted	363	118	8.0
Ginkgo nuts, canned	180	16	9.3
Seeds			
Breadnuttree seeds, dried	2011	115	14.9
Cottonseed kernels, roasted (glandless)	1350	440	5.5
Breadfruit seeds, roasted	1082	62	6.0
Breadfruit seeds, raw	941	54	5.2
Sunflower seed kernels, dry roasted	850	129	9.0
Pumpkin and squash seed kernels, dried	807	535	3.9
Pumpkin and squash seed kernels, roasted	806	534	3.9
Sunflower seed kernels, dried	689	354	10.5
Flaxseed	681	362	27.9
Sunflower seed kernels	491	129	11.5
Sunflower seed kernels, oil roasted	483	127	6.8
Sesame seeds, whole, dried	468	351	11.8
Sesame seed kernels, dried (decorticated)	407	347	12.7
Sesame seed kernels, toasted	406	346	16.9

From U.S. Department of Agriculture, Agricultural Research Service (2003) *USDA National Nutrient Database for Standard Reference*, Release 16. Available at www.nal.usda.gov/fnic/foodcomp.

includes the potassium, magnesium, and fiber content of some common nuts and seeds. Although this is presented based on a rank-ordered content of potassium, it is easy to see that some nuts and seeds are a significant source of magnesium and dietary fiber and their consumption was therefore encouraged in the DASH diet.

Low-fat dairy products were also an important part of the DASH diets. These were used primarily to increase the calcium content of the diets from a low content of approximately 450 mg on the control and fruit and vegetable diets to approximately 1250 mg on the DASH diet at the 2000 kcal (8368 kJ) level. Calcium has frequently been reported to have an inverse relationship with blood pressure, but studies utilizing supplemental calcium have been inconsistent. With supplements, effects on blood pressure reduction have been negligible. Nonetheless, the blood pressure lowering effect of the DASH diet has been suggested to be in part related to the calcium content of the diet. It should be noted that the DASH diet also was lower in fat and higher in protein, and therefore it is not easily attributable to one factor alone but rather a combination of several factors, as depicted in **Figure 1**.

The final point regarding composition of the DASH diet is that it included specific food choices. The diet contained whole grains, poultry, and fish (in addition to the fruits, vegetables, low-fat dairy, and nuts previously mentioned). Although it was reduced in total and saturated fat, it was also reduced in meats, sweets, and sugar-containing beverages. Food was consumed as an overall pattern in which it is quite possible that the interaction between food items is as important as the specific foods in reducing blood pressure. Thus, the DASH diet contained dietary patterns promoted by the National Institutes of Health, National Heart, Lung, and Blood Institute. The dietary patterns of DASH are presented at three energy levels in **Table 6**.

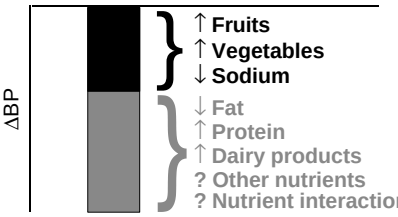


Figure 1 Assessing the effects of the DASH-Sodium diet on blood pressure. The figure depicts the fact that there is some certainty associated with the increases in fruits and vegetables and reduction in sodium. The gray areas represent the other components of the diet and the possible contribution of each, alone or in combination with other factors.

Table 6 Food group servings for the DASH diet at three energy levels

Food group	*Daily servings (except as noted)*		
	1600 kcal or 6694 kJ	*2000 kcal or 8368 kJ*	*3100 kcal or 12970 kJ*
Grains and grain products	6	7–8	12–13
Vegetables	3–4	4–5	6
Fruits	4	4–5	6
Low-fat or fat-free dairy foods	2–3	2–3	3–4
Meats, poultry, fish	1–2	2 or less	2–3
Nuts, seeds, dry beans	3 per week	4–5 per week	1
Fats and oils	2	2–3	4
Sweets	0	5 per week	2

Dietary Protein Consumption

Results of meta-analyses from several investigators indicate an inverse association between dietary protein and blood pressure levels. However, data have not been conclusive. The DASH diet contained approximately 18% of energy from protein compared to 15% of energy from protein in the other diets tested. Because of the addition of low-fat dairy foods and the reduced emphasis on high-fat meats, it can be assumed that this elevation in protein was brought about by foods that contributed protein from perceived beneficial sources. This is an area of dietary intake that requires more research.

Fish Oil Supplementation

Studies suggest that an intake of fish oil at a level of approximately 4 g per day reduces systolic blood pressure by approximately 1.7–2.1 mmHg and diastolic blood pressure by 1.5–1.6 mmHg. These effects tend to be larger in individuals older than 45 years of age and in populations with blood pressure readings greater than 140/90 mmHg. Generally, there have been differences associated with fish oil capsules compared to naturally occurring sources of EPA and DHA from fatty fish, again indicating dietary pattern rather than consumption of individual items may be crucial. The DASH diet had a relatively high fish content (compared to animal meats) and this may be yet one more factor contributing to the lowering of blood pressure in individuals on the DASH diet.

The American Heart Association recommends eating two servings of fish per week and emphasizes that the choice should be a fatty fish (such as salmon, herring, or mackerel). Not all fish have the same content of omega-3 fatty acids. **Table 7** provides a listing of amounts of combined EPA/DHA in fish and other seafood sources and the amount of consumption (in ounces of product) necessary to provide a 4-g intake. Descriptors include common raw and canned products, but the intakes given are rough estimates due to potential variability in oil content within species, season, and diet. Cooking methods and other preparation techniques may affect the final concentrations in raw fish.

Other Fatty Acid Effects

Monounsaturated fatty acids, particularly olive oil, may help lower blood pressure. Olive oil has typically been associated with the popularized Mediterranean diet, which has been promoted as a treatment for cardiovascular disease. Other oils (e.g., canola and peanut oil) have a high monounsaturated fat content. Nuts, which are part of the DASH diet, contain significant amounts of monounsaturated fats and fit well in the Mediterranean diet.

Caffeine

Although a link between caffeine consumption (particularly coffee) and hypertension may exist, effects of coffee drinking on blood pressure appear to be dependent on the time of consumption and subsequent determination of blood pressure values. Generally, a role for caffeine intake and development of hypertension is not believed to be significant.

Weight Reduction

Obesity and overweight are considered independent risk factors for cardiovascular disease and are closely associated with hypertension. This linkage was demonstrated in the 1960s by the Framingham Heart Study investigators in the United States. Obesity in the industrialized world has been increasing at epidemic proportions. The relationship between increasing body weight and increasing blood pressure has been termed obesity hypertension, and treatment requires consideration of physiologic changes related to this disorder. Although efforts have been under way in the United States to reduce overweight and obesity, it is estimated that the age-adjusted prevalence of overweight and obesity (body mass index (BMI) $\geq$25.0) among adults aged 20 or older is 64%; for those considered obese (BMI $\geq$30.0) it is 30%. During a 25-year period in the United States, this reflects approximately a 36% increase in the combined levels of overweight and obesity and essentially a doubling of obesity rates.

Table 7 Eicosapentaenoic (EPA, 20:5 n-3) and docosahexaenoic (DHA, 22:6 n-3) acid in fish/seafood (per 100 grams) and the amount of consumption (in ounces) required to provide ~4 g of EPA + DHA per day (ranked in order of content)

Fish/seafood	*EPA (g)*	*DHA (g)*	*EPA + DHA (g)*	*Oz. to provide ~4 g EPA + DHA*
Fish, caviar, black and red, granular	2.74	3.80	6.54	2.2
Fish, mackerel, salted	1.62	2.97	4.58	3.1
Fish, roe, mixed species, cooked	1.26	1.75	3.01	4.7
Fish, shad, American, raw	1.09	1.32	2.41	5.9
Fish, roe, mixed species, raw	0.98	1.36	2.35	6.0
Fish, mackerel, Atlantic, raw	0.90	1.40	2.30	6.1
Fish, anchovy, European, canned in oil	0.76	1.29	2.06	6.9
Fish, salmon, chinook, raw	1.01	0.94	1.95	7.2
Fish, salmon, Atlantic, farmed, raw	0.62	1.29	1.91	7.4
Fish, herring, Pacific, raw	0.97	0.69	1.66	8.5
Fish, salmon, pink, canned	0.85	0.81	1.65	8.5
Fish, herring, Atlantic, raw	0.71	0.86	1.57	9.0
Fish, anchovy, European, raw	0.54	0.91	1.45	9.7
Fish, mackerel, Pacific and jack, raw	0.51	0.93	1.44	9.8
Fish, salmon, Atlantic, wild, raw	0.32	1.12	1.44	9.8
Fish, sablefish, raw	0.68	0.72	1.40	10.1
Fish, mackerel, spanish, raw	0.33	1.01	1.34	10.5
Fish, whitefish, mixed species, raw	0.32	0.94	1.26	11.2
Fish, salmon, coho, farmed, raw	0.39	0.82	1.21	11.7
Fish, salmon, chum, canned	0.47	0.70	1.18	12.0
Fish, tuna, fresh, bluefin, raw	0.28	0.89	1.17	12.0
Fish, salmon, sockeye, raw	0.52	0.65	1.17	12.0
Fish, salmon, sockeye, canned	0.49	0.66	1.16	12.2
Fish, salmon, coho, wild, raw	0.43	0.66	1.09	13.0
Fish, salmon, pink, raw	0.42	0.59	1.01	14.0
Fish, sardine, Atlantic, canned in oil	0.47	0.51	0.98	14.4
Fish, trout, rainbow, farmed, raw	0.26	0.67	0.93	15.2
Fish, halibut, Greenland, raw	0.53	0.39	0.92	15.4
Fish, tuna, white, canned in water	0.23	0.63	0.86	16.4
Fish, shark, mixed species, raw	0.32	0.53	0.84	16.7
Fish, bluefish, raw	0.25	0.52	0.77	18.3
Fish, bass, striped, raw	0.17	0.59	0.75	18.7
Fish, trout, mixed species, raw	0.20	0.53	0.73	19.3
Fish, smelt, rainbow, raw	0.28	0.42	0.69	20.4
Mollusks, oyster, Pacific, raw	0.44	0.25	0.69	20.5
Fish, swordfish, raw	0.11	0.53	0.64	22.1
Fish, spot, raw	0.22	0.41	0.63	22.4
Fish, salmon, chum, raw	0.23	0.39	0.63	22.5
Fish, wolffish, Atlantic, raw	0.31	0.32	0.62	22.6
Fish, bass, freshwater, mixed species, raw	0.24	0.36	0.60	23.7
Fish, sea bass, mixed species, raw	0.16	0.43	0.60	23.7
Fish, trout, rainbow, wild, raw	0.17	0.42	0.59	24.0
Fish, pompano, florida, raw	0.18	0.39	0.57	24.8
Mollusks, oyster, eastern, wild, raw	0.27	0.29	0.56	25.2
Fish, drum, freshwater, raw	0.23	0.29	0.52	27.3
Mollusks, squid, mixed species, raw	0.15	0.34	0.49	28.9
Crustaceans, shrimp, mixed species, raw	0.26	0.22	0.48	29.4
Fish, sucker, white, raw	0.19	0.29	0.48	29.5
Mollusks, mussel, blue, raw	0.19	0.25	0.44	32.0
Fish, tilefish, raw	0.09	0.35	0.43	32.8
Fish, pollock, Atlantic, raw	0.07	0.35	0.42	33.5
Mollusks, oyster, eastern, farmed, raw	0.19	0.20	0.39	36.1
Fish, pollock, walleye, raw	0.15	0.22	0.37	37.9
Fish, seatrout, mixed species, raw	0.17	0.21	0.37	37.9
Crustaceans, crab, queen, raw	0.26	0.11	0.37	37.9
Fish, catfish, channel, wild, raw	0.13	0.23	0.36	38.8
Fish, halibut, Atlantic and Pacific, raw	0.07	0.29	0.36	38.9
Crustaceans, crab, blue, canned	0.19	0.17	0.36	38.9
Fish, carp, raw	0.24	0.11	0.35	40.1
Fish, cisco, raw	0.10	0.26	0.35	40.1

Continued

Table 7 Continued

Fish/seafood	*EPA (g)*	*DHA (g)*	*EPA + DHA (g)*	*Oz. to provide ~4 g EPA + DHA*
Fish, rockfish, Pacific, raw	0.14	0.20	0.35	40.9
Fish, mullet, striped, raw	0.22	0.11	0.33	43.4
Crustaceans, crab, blue, raw	0.17	0.15	0.32	44.1
Fish, mackerel, king, raw	0.14	0.18	0.31	45.1
Fish, pike, walleye, raw	0.09	0.23	0.31	45.4
Fish, snapper, mixed species, raw	0.05	0.26	0.31	45.4
Crustaceans, crab, dungeness, raw	0.22	0.09	0.31	46.0
Fish, ocean perch, Atlantic, raw	0.08	0.21	0.29	48.5
Fish, sturgeon, mixed species, raw	0.19	0.09	0.29	49.2
Fish, catfish, channel, farmed, raw	0.07	0.21	0.27	51.5
Fish, tuna, light, canned in water	0.05	0.22	0.27	52.3
Fish, sheepshead, raw	0.14	0.12	0.26	54.1
Fish, tuna, fresh, skipjack, raw	0.07	0.19	0.26	55.1
Fish, perch, mixed species, raw	0.08	0.17	0.25	55.8
Fish, grouper, mixed species, raw	0.03	0.22	0.25	57.1
Fish, whiting, mixed species, raw	0.09	0.13	0.22	63.0
Fish, croaker, Atlantic, raw	0.12	0.10	0.22	64.1
Fish, tuna, fresh, yellowfin, raw	0.04	0.18	0.22	64.7
Fish, cod, Pacific, raw	0.08	0.14	0.22	65.6
Fish, flatfish (flounder and sole species), raw	0.09	0.11	0.20	70.9
Mollusks, scallop, mixed species, raw	0.09	0.11	0.20	71.3
Fish, haddock, raw	0.06	0.13	0.19	76.3
Fish, cod, Atlantic, raw	0.06	0.12	0.18	76.7
Fish, burbot, raw	0.07	0.10	0.17	85.0
Mollusks, octopus, common, raw	0.08	0.08	0.16	89.9
Fish, eel, mixed species, raw	0.08	0.06	0.15	96.0
Crustaceans, crayfish, farmed, raw	0.12	0.03	0.14	98.0
Mollusks, clam, mixed species, raw	0.07	0.07	0.14	99.4
Crustaceans, crayfish, wild, raw	0.10	0.04	0.14	99.4
Mollusks, snail, raw	0.12	0.00	0.12	118.6
Fish, sunfish, pumpkin seed, raw	0.04	0.07	0.11	129.4
Fish, dolphinfish, raw	0.02	0.09	0.11	130.6
Fish, pike, northern, raw	0.03	0.07	0.11	131.9
Mollusks, cuttlefish, mixed species, raw	0.04	0.07	0.11	134.4
Turtle, green, raw	0.02	0.03	0.06	252.0
Mollusks, abalone, mixed species, raw	0.05	0.00	0.05	287.9
Frog legs, raw	0.01	0.02	0.03	415.0

From U.S. Department of Agriculture, Agricultural Research Service (2003) *USDA National Nutrient Database for Standard Reference*, Release 16. Available at www.nal.usda.gov/fnic/foodcomp.

The increase in obesity is seen in all ethnic, gender, and age groups. This problem is not confined to the average American; the US military reported that more than 50% of military personnel were overweight and more than 6% were obese in the late 1990s, despite high physical activity levels due to the rigors of basic training and regular field exercises. For the military, this reflects a trend that mirrors what is happening in the general population.

Globally, more than 1 billion adults are classified as overweight and approximately 300 million as clinically obese, ranging from less than 5% in China, Japan, and some African nations to more than 75% in urban Samoa. Alarmingly, this epidemic has spread to children, with 17.6 million children younger than 5 years of age estimated to be overweight worldwide. Data from the United States indicate that 15% of children and adolescents 6–19 years of age are overweight, a figure at least three times higher than that reported in the period from 1960 to 1970. Overweight children are at risk of becoming overweight adults but, more important, are likely to experience chronic health problems (including hypertension) associated typically with only adult obesity.

The World Health Organization has recommended an integrated, multifaceted, population approach be implemented to bring about effective weight management for those at risk of overweight and obesity. The key elements for developing such an environmental support include the following:

- Availability and access to a variety of low-fat, high-fiber foods
- Opportunities for physical activities

- Promotion of healthy behavior to encourage, motivate, and enable individuals to lose weight by
  - Eating more fruits, vegetables, nuts, and whole grains
  - Engaging in moderate physical activity for at least 30 minutes a day
  - Reducing the amounts of fat and sugar in the diet
  - Changing from a diet containing saturated animal fats to one emphasizing unsaturated vegetable oils
- Proper training of clinical personnel to ensure effective support for those trying to lose weight or avoid further weight gain

Obviously, it is essential to maintain a healthy body weight and thus necessary to keep a focus on energy intake in an effort to prevent overweight. Regarding hypertension, weight reduction appears to be the most promising answer in terms of potential impact on lowering blood pressure. Losing as few as 4.5 kg, or 10 pounds, of body weight can reduce blood pressure. Adopting healthy eating patterns yields additional benefits.

Strategies for Implementing Nutritional Changes to Control Blood Pressure

Self-Monitoring

A strategy undeniably praised for weight control is self-monitoring of one's food intake. Although a difficult undertaking for most individuals, the success of this technique is impressive, as evidenced by the successful long-term weight loss maintainers in the National Weight Control Registry.

The self-monitoring technique has been used to help people comply with other lifestyle recommendations (e.g., to increase physical activity by recording physical activity minutes). In the PREMIER clinical trial, participants in the 'established plus DASH' arm were required to monitor intake of energy, sodium, total fat, and saturated fat and servings of fruit, vegetables, and dairy to determine their compliance with the intervention. Those participants in the 'established' arm only recorded energy, sodium, and total fat intake. Most people find these recordings difficult but readily admit that they are successful in documenting dietary compliance if taken seriously.

Although a time-consuming and difficult task, successful diet compliers continue this behavior change strategy over the long term. Those who discontinue this technique often revert back to their old habits and relapse.

Working with a Dietetics Professional

Dietitians in North America and abroad typically have to meet national standards set by professional organizations such as the American Dietetic Association and the Canadian Dietetic Association. As such, these individuals are called 'registered dietitian' or other titles used only by those who have met these standards. Although one does not necessarily need to work with a professional, for some it is often easier to implement change when they can clear up confusing and often conflicting information by working with a dietitian who can provide credible nutrition information. Dietitians are taught to interpret the science into meaningful terms for the consumer. In addition, a well-trained professional will be equipped with a knowledge of motivational and behavior strategies to help effect change.

There are many important aspects of behavior change that are taught to the hypertensive client. Making lifestyle changes gradually so that one adjusts to one change before making another change is important. One should strive for short-term, attainable goals. Getting off track is not uncommon, but identifying what triggered the sidetrack and getting back on track are equally important. One should understand that slips are inevitable; it takes time to get used to the changes. In essence, lifestyle change is a long-term process, but it is worthwhile for good health.

Conclusions

Ultimately, blood pressure control will mandate lifestyle changes, even if hypertensive medications are prescribed. Most important, body weight has to be a key focus and the goal should be to work toward an ideal body weight and avoid gaining weight during the aging process. In addition to diet, physical activity factors into this scenario, and one should strive to be more physically active and less sedentary as one grows older. Diet, as described previously, plays a key role in blood pressure control. Increasing fruits, vegetables, and dairy products in the diet and reducing sodium should be the first objective. This will increase potassium, magnesium, and calcium by natural sources as much as possible. Increasing whole grains, nuts, and legumes will also improve mineral and fiber content of the diet. One should focus less on red meat; instead, one should choose fatty fish to increase consumption of omega-3 fatty acids. If one drinks alcohol, one should do so in moderation. These simple, but often difficult to accomplish, strategies will help to lower blood pressure and improve risk against cardiovascular disease.

See also: **Alcohol**: Effects of Consumption on Diet and Nutritional Status. **Caffeine**. **Fats and Oils**. **Fish**. **Fruits and Vegetables**. **Hypertension**: Etiology; Dietary Factors. **Magnesium**. **Nuts and Seeds**. **Potassium**. **Protein**: Synthesis and Turnover; Requirements and Role in Diet. **Sodium**: Physiology; Salt Intake and Health. **Weight Management**: Approaches. **World Health Organization**.

Further Reading

Appel LJ, Champagne CM, Harsha DW *et al.* Writing Group of the PREMIER Collaborative Research Group (2003) Effects of comprehensive lifestyle modification on blood pressure control: Main results of the PREMIER clinical trial. *Journal of the American Medical Association* **289**: 2083–2093.

Appel LJ, Moore TJ, Obarzanek E *et al.* for the DASH Collaborative Research Group (1997) A clinical trial of the effects of dietary patterns on blood pressure. *New England Journal of Medicine* **336**: 1117–1124.

Blumenthal JA, Sherwood A, Gullette ECD *et al.* (2002) Biobehavioral approaches to the treatment of essential hypertension. *Journal of Consulting and Clinical Psychology* **70**: 569–589.

Chobanian AV, Bakris GL, Black HR *et al.* and the National High Blood Pressure Education Program Coordinating Committee (2003) The seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: The JNC 7 report. *Journal of the American Medical Association* **289**: 2560–2572.

Ferrara LA, Raimondi S, d'Episcopa L *et al.* (2000) Olive oil and reduced need for anti-hypertensive medications. *Archives of Internal Medicine* **160**: 837–842.

Geleijnse JM, Giltay EJ, Grobbee DE *et al.* (2002) Blood pressure response to fish oil supplementation: Metaregression analysis of randomized trials. *Journal of Hypertension* **20**: 1493–1499.

Klag MJ, Wang NY, Meoni LA *et al.* (2002) Coffee intake and risk of hypertension: The Johns Hopkins precursors study. *Archives of Internal Medicine* **162**: 657–662.

Lindquist CH and Bray RM (2001) Trends in overweight and physical activity among U.S. military personnel, 1995–1998. *Preventive Medicine* **32**: 57–65.

Rosenthal T, Shamiss A, and Holtzman E (2001) Dietary electrolytes and hypertension in the elderly. *International Urology and Nephrology* **33**: 575–582.

Sacks FM, Svetkey LP, Vollmer WM *et al.* for the DASH-Sodium Collaborative Research Group (2001) Effects on blood pressure of reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. *New England Journal of Medicine* **344**: 3–10.

Suter PM, Sierro C, and Vetter W (2002) Nutritional factors in the control of blood pressure and hypertension. *Nutrition in Clinical Care* **5**: 9–19.

Trials of Hypertension Prevention Collaborative Research Group (1997) Effects of weight loss and sodium reduction intervention on blood pressure and hypertension incidence in overweight people with high-normal blood pressure. The Trials of Hypertension Prevention, Phase II. *Archives of Internal Medicine* **157**: 657–667.

U.S. Department of Health and Human Services, National Institutes of Health and National Heart, Lung, and Blood Institute (1998) *Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults: The Evidence Report*, NIH Publication No. 98-4083. Bethesda, Md.: National Institutes of Health.

U.S. Department of Health and Human Services, National Institutes of Health and National Heart, Lung, and Blood Institute (2003) *Facts about the DASH Eating Plan*, NIH Publication No. 03-4082. Bethesda, Md.: National Institutes of Health.

Vasan RS, Beiser A, Seshadri S *et al.* (2002) Residual lifetime risk for developing hypertension in middle-aged women and men: The Framingham Heart Study. *Journal of the American Medical Association* **287**: 1003–1010.

Whelton PK, Appel LJ, Espeland MA *et al.* for the TONE Collaborative Research Group (1998) Sodium reduction and weight loss in the treatment of hypertension in older persons: A randomized, controlled trial of nonpharmacologic interventions in the elderly (TONE). *Journal of the American Medical Association* **279**: 839–846.

Wing RR and Hill JO (2001) Successful weight loss maintenance. *Annual Review of Nutrition* **21**: 323–341.

Wofford MR, Davis MM, Harkins G *et al.* (2002) Therapeutic considerations in the treatment of obesity hypertension. *Journal of Clinical Hypertension* **4**: 189–196.

Wolf-Maier K, Cooper RS, Banegas JR *et al.* (2003) Hypertension prevalence and blood pressure levels in 6 European countries, Canada, and the United States. *Journal of the American Medical Association* **289**: 2363–2369.

World Health Organization (2003) *Diet, Nutrition and the Prevention of Chronic Diseases: Report of a Joint WHO/FAO Expert Consultation*, WHO Technical Report Series 916. Geneva: WHO.

Zimmerman E and Wylie-Rosett J (2003) Nutrition therapy for hypertension. *Current Diabetes Reports* **3**: 404–411.

HYPOGLYCEMIA

V Marks, University of Surrey, Guildford, UK

Introduction

Hypoglycemia is defined as a blood glucose concentration of 2.2 mmol l^{-1} (plasma glucose concentration of 2.5 mmol l^{-1}) or less. Its definition is necessarily arbitrary and owes its importance to the fact that hypoglycemia (literally low blood glucose) of this severity produces brain dysfunction by depriving its neurons of glucose.

Hypoglycemia is not a disease but a manifestation of it. It has, however, come to have a totally different meaning amongst certain sections of the

population that has very little to do with blood glucose concentration but a lot to do with their feelings of well being, discomfort, and attitudes to life but, above all, with the role of diet in the achievement and maintenance of good health. And whilst no discussion of the dietary treatment of hypoglycemia can be meaningful without reference to this concept – referred to, for want of a better term, as nonhypoglycemia, – hypoglycemia will, throughout this article, be used only to describe a measured low blood glucose concentration.

Brain Function and Hypoglycemia

The brain malfunction to which hypoglycemia gives rise will be referred to as neuroglycopenia.

The brain is often thought of as being incapable of using metabolites other than glucose as a source of energy. This is untrue. It has been known for more than 30 years to be able, under certain circumstances including prolonged fasting, to utilize the 'ketone bodies,' β-hydroxybutyrate and aceto-acetate. Under these circumstances the need for glucose and its supply through gluconeogenesis is drastically reduced. The survival value of this ability is immense as it permits fat stores rather than structural muscle and other tissue proteins to be utilized for maintenance of vital processes under these stressful conditions. Only when fat stores have become completely exhausted and plasma ketone levels fallen to below normal fasting levels does the brain's demand for glucose rise above the ability of gluconeogenesis to provide it. Only at this point does hypoglycemia intervene and portend death from starvation or inanition (see later).

The Blood Glucose Concentration

Failure to appreciate the differences between arterial and venous blood glucose is a major cause of the confusion that has surrounded the recognition and diagnosis of hypoglycemia and been responsible for nonhypoglycaemia becoming a common diagnosis amongst those whom Singer and coworkers refer to as, the folk sector.

In the fasting subject the concentration of glucose in arterial and venous blood is virtually identical but may differ by as much as $2.5\,\mathrm{mmol\,l^{-1}}$ following ingestion of a carbohydrate-rich meal. Because it is arterial blood glucose that determines glucose supply to the brain, regulates the secretion of insulin and other hormones, and is itself homeostatically controlled, it is necessary to define hypoglycemia in terms of glucose in arterial (or more realistically free flowing capillary) than in venous blood.

Mechanism of Hypoglycemia Glucose Pool in Fasting Subjects

Glucose is confined within the body to the extracellular fluid where it is referred to as the glucose pool: detailed discussion of its regulation is outside the scope of this article except to stress that it reflects the concentration of glucose in the blood. This remains remarkably constant despite huge changes in the rates of delivery and utilization of glucose, by meals and exercise (and fasting), respectively, and is described as glucose homeostasis (**Figure 1**). The main but far from sole regulator is insulin.

Insulin Release in Response to Eating and Fasting

Evidence for a 'cephalic phase' of insulin secretion in humans is scanty and conflicting. Most observers have found a minimal, if any, response to the prospect of eating, or the reality of drinking, a noncalorigenic sweet drink except in some obese individuals.

After a carbohydrate-containing meal, glucose derived from food enters the portal vein. From here it is conveyed to the liver where much of it is extracted and converted to glycogen. What remains unabsorbed passes into the systemic circulation, producing small and variable rises in arterial, capillary, and, initially, venous blood glucose concentrations. The modest rise in arterial blood glucose concentration perfusing the pancreas, augmented by nervous stimuli and insulinotrophic hormones, collectively called incretins, released from the gut in response to meals containing carbohydrate and/or fats, leads to the secretion of insulin in greater amounts than is occasioned by the rise in blood glucose concentration alone.

In the postprandial period, as the blood glucose concentration fall towards its homeostatically controlled level, insulin secretion declines to a level that is just sufficient to suppress unbridled lipolysis. Absence of this constitutive insulin secretion in patients with type 1 diabetes is the cause of diabetic ketoacidosis.

The Role of the Liver in Glucose Homeostasis

The liver, under the influence of insulin reaching it in high concentration in the portal vein after ingestion of a meal, switches from being a net exporter to net importer of glucose from the glucose pool. Any insulin not extracted and degraded by the liver passes through the heart and lungs to reach peripheral tissues, notably muscle, adipose tissue, and skin, where, providing the concentration of insulin in blood is sufficiently high, it promotes glucose uptake.

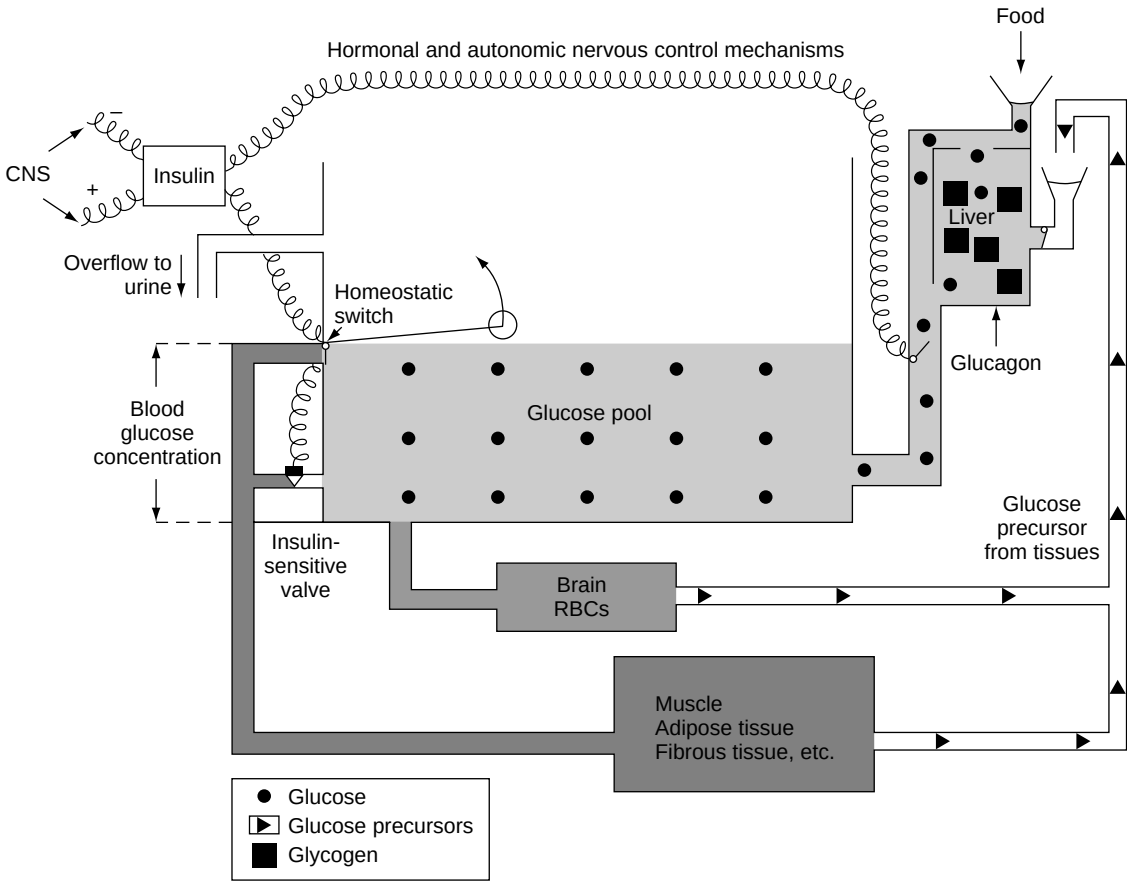


Figure 1 Schematic representation of homeostatic control of blood glucose level and mechanism of hypoglycemia. Hypoglycemia results whenever inflow of glucose from the gut and/or liver fails to meet the outflow of glucose from the glucose pool, which consists of glucose dissolved in the extracellular water only. Imbalance arises from: (1) excessive outflow into the tissues due to insulin (or very rarely IGF-II) overproduction or activity; or (2) in the fasting state, an inability of the liver to liberate or produce glucose at a rate sufficient to meet the non-insulin-dependent, and obligatory, requirements of the brain and red blood cells for glucose.

Except in disease, the glucose pool, amounting to just 5–15 g, rarely expands by more than 100% even after ingestion of a meal providing up to 300 g of carbohydrate as starch or glucose. Nor does it shrink to less than 4 g even after many days of fasting.

Entry of glucose into the glucose pool is limited by the rate at which it can be absorbed from the intestine. This is normally in the region of 25–50 g h^{-1}. In people with normal glucose tolerance, venous blood glucose levels generally return to overnight fasting values within 2 h of eating a meal regardless of how much carbohydrate it contains. Arterial blood glucose levels take somewhat longer to return to preingestion levels but they too are always within the normal fasting range by 3 h, even though the evidence provided by measurement of the gut hormones Glucose-Dependent Insulinotrophic Peptide (GIP), the main incretin, indicate that absorption of large meals continues for much longer. Absorption of a 200-g liquid glucose meal by normal healthy subjects, for example, is still incomplete after 5 h even though both their venous and arterial blood glucose levels have long since returned to normal.

The outflow of glucose into the tissues, on the other hand, depends upon many factors; the two most important are the plasma insulin concentration and the blood concentration itself. Under maximum insulin stimulation – and at 'normal' blood glucose levels – glucose can disappear from the glucose pool at a rate of up to 40–50 g h^{-1} but these conditions are rarely encountered except experimentally or in cases of gross insulin overdose.

Onset of insulin action is almost instantaneous and persists for as long as insulin remains bound to insulin receptors. This is generally slightly longer than insulin levels in the blood themselves remain elevated. In other words glucose continues to enter insulin-dependent cells for up to 30 min after plasma insulin levels have returned to 'fasting' levels. During this time the glucose pool may shrink sufficiently to produce hypoglycemia unless replenished by glucose continuing to enter from the intestine (or

experimentally/therapeutically by intravenous infusion) or from the liver, once it has switched from the glycogenic to glycogenolytic mode.

Small, and always temporary, imbalances between the rate at which insulin action declines and glucose enters the glucose pool can occur in healthy subjects after ingestion of a large dose of glucose in solution on an empty stomach, but is rare following the ingestion of an ordinary mixed meal.

A slight delay in stimulating insulin release in response to a meal is the earliest and most characteristic abnormality observed in patients with non-insulin-dependent diabetes mellitus (NIDDM) who may secrete more insulin in total than people of comparable age, though not of body mass index. They are, however, generally insulin resistant, which explains why, despite the larger amounts of insulin secreted in response to meals in the early stages of their illness, they do not suffer from meal-induced hypoglycemia.

Hypoglycemic Syndromes

Brain Malfunction from Hypoglycemia

The brain ordinarily requires a regular and plentiful supply of glucose, which gets to it from the blood by active transport utilizing the glucose transporter protein GLUT 1. Reduction of supply to below critical limits causes the brain to malfunction and this manifests itself subjectively as symptoms and objectively as neurological deficit. The blood glucose level at which impairment occurs varies. Symptoms are unusual at blood glucose levels above $3.0\,\text{mmol}\,\text{l}^{-1}$ except in diabetic and elderly subjects in whom they may occur at higher levels. Objective evidence of cerebral impairment can however often be discerned by an investigator at blood glucose levels around $3.5–4\,\text{mmol}\,\text{l}^{-1}$.

Causes of neuroglycopenia other than hypoglycemia, i.e., normoglycemic neuroglycopenia, are currently thought to be rare but include congenital or acquired reduction in GLUT 1. The possibility that such defects are more common than currently supposed and are responsible for some cases of 'nonhypoglycemia' cannot be dismissed at the present time, and would help explain why, under research conditions, some people diagnosed with this condition appear to develop symptoms at higher blood glucose levels than control subjects.

Neuroglycopenic Syndromes

Four more or less distinct neuroglycopenic syndromes (one of which is so rare that it will not be considered further here) can be recognized. They are not mutually exclusive, nor do they depend upon the ultimate cause of the hypoglycemia.

Acute Neuroglycopenia

This syndrome comprises a collection of vague symptoms such as feelings of alternating hot and cold, feeling unwell, anxiety, panic, inner trembling, unnatural feelings, blurring of vision, and palpitations, any or all of which may be accompanied by objective signs of facial flushing, sweating, tachycardia, and unsteadiness of gait. There is no particular order in which these features occur, nor are they constant. Nevertheless, patients on insulin therapy for diabetes, in whom they are common, rely upon them to warn of more severe neuroglycopenic impairment culminating in loss of consciousness. These patients can be taught to abort progression of symptoms by eating carbohydrate.

Many of the features of acute neuroglycopenia resemble those produced by adrenaline and consequently are often referred to as adrenergic.

Subacute Neuroglycopenia

This syndrome is more insidious and may go completely unrecognized unless or until the patient loses consciousness. Often, however, there is loss of spontaneous activity, impairment of cognitive function and the onset of somnolence that is more discernible to the bystander than to the patient and which, when it occurs *de novo* in an insulin-treated diabetic, is often referred to as 'hypoglycemia unawareness.'

Acute can proceed to subacute neuroglycopenia and both can progress to stupor or coma unless relieved by food or injection of glucagon. Even when this is not done, however, full recovery, under the influence of endogenous counter-regulatory mechanisms, is almost invariable and is the reason why treatment with insulin is so safe despite the potential dangers of hypoglycemia.

Chronic Neuroglycopenia

The third syndrome is exceedingly rare. It occurs only when the blood glucose concentration remains low, either due to the presence of an insulin-secreting tumor of the pancreas or overzealous treatment of diabetes with insulin for weeks or months on end. It is characterized by mental dysfunction resembling clinical depression, schizophrenia, or dementia, the symptoms of which are not relieved by restoring the blood glucose level to normal. Partial recovery may, however, take place over the

ensuing months or years if the cause of the hypoglycemia is remedied.

This condition might be confused with 'nonhypoglycaemia' were it not for the fact that the blood glucose concentration is invariably low ($<3.0\,\text{mmol}\,l^{-1}$) while the patient is fasting, does not rise normally in response to food, and evidence of underlying disease can always be found.

Diagnosis

Causes of Hypoglycemia

There is something in the region of 100 causes of hypoglycemia but all, apart from exogenous (or iatrogenic) insulin overdose, are uncommon. Some of the most important causes of recurrent hypoglycemia are listed and briefly described in **Table 1**. Simultaneous occurrence of symptoms, a measured low blood glucose concentration, and relief from intravenous glucose are a sine qua non for diagnosis. Differentiation is seldom simple and always rests heavily upon the results of laboratory data of which measurements of plasma insulin and C-peptide are the most important.

Endocrinological and other anatomico-pathological causes of hypoglycemia will not be considered further. Nor will iatrogenic or toxic causes, of which alcohol-induced fasting hypoglycemia is easily the most common. Instead, attention will be given to those conditions (including 'nonhypoglycaemia') that have a mainly or exclusively dietary etiology and which respond partially or completely to dietary measures.

Spontaneous Reactive Hypoglycemia

Within a year of the discovery of insulin, and the symptoms to which hypoglycemia can give rise, Seale Harris, an American physician, had proposed that spontaneous overproduction of endogenous insulin might produce a similar condition. Confirmation of this hypothesis soon followed. The seminal work of Whipple on the diagnosis of insulinoma and of Conn on diet-induced postprandial reactive hypoglycemia, both in 1936, distinguished between fast-induced (fasting) hypoglycemia and that which occurred only in response to feeding. The latter, reactive or postprandial hypoglycemia, could be reproduced by oral administration of large doses of glucose in solution and this became the standard criterion for its diagnosis – the 5-h glucose tolerance test.

Glucose Load Test

The observation that in a substantial percentage of normal healthy people glucose taken in solution on an empty stomach produces a rebound fall in venous blood glucose levels to below fasting levels was made soon after blood glucose measurements became possible and before the discovery of insulin. It attracted little attention at the time being considered to have only curiosity value and little pathological significance.

The situation changed dramatically during the early 1950s and, subsequently, particularly in the US, with the appearance of books written for lay consumption attributing a vast array of common symptoms to hypoglycemia, whether the blood glucose concentration was low at the time or not. Belief in the importance and prevalence of 'hypoglycemia' grew amongst fashionable medical practitioners and the general public alike to such an extent that, by the early 1970s, alarm bells began to ring amongst consumer action groups and the scientific medical community.

With the passage of time the original, well-defined syndrome of postprandial reactive hypoglycemia had become so distorted, and the criteria for its diagnosis so blurred, that anyone with vague symptoms could be, and often was, described as suffering from 'hypoglycemia,' without anyone bothering to measure their blood glucose concentration.

Not until a consensus 'Statement on 'Post Prandial' or 'Reactive' Hypoglycemia' was issued by the Third International Symposium on Hypoglycemia and generally recognized by medical practitioners throughout the world did scientific criteria for the diagnosis of reactive hypoglycemia gain universal acceptance and its purported incidence declined dramatically.

Definition

It is now accepted that some people exhibit, in the course of their everyday life, symptoms similar to those caused by acute neuroglycopenia and may, if accompanied by a capillary or arterialized venous blood glucose concentration of $2.8–2.5\,\text{mmol}\,l^{-1}$ or less, justify description as being of postprandial reactive hypoglycemic origin. Reactive hypoglycemia may itself be a consequence of any one of a large number of well-recognized but generally uncommon conditions that can also produce fast-induced hypoglycemia, and it is only after they have been excluded by appropriate laboratory investigations that a diagnosis of functional or dietary reactive hypoglycemia is justified.

Specifically, the prolonged oral glucose load (tolerance) test is now deemed inappropriate for the diagnosis of postprandial or reactive hypoglycemia since the incidence of false-positive results with this test is so high as to make it meaningless, especially if, as is so often the case, venous rather than arterial blood is sampled.

Table 1 The main causes of non-iatrogenic hypoglycemia

Description	*Mechanism*	*Diagnostic criteria*	*Dietary considerations*
Fasting hypoglycemia			
Insulin-secreting tumor (insulinoma) and nesidioblastosis	Abnormal B cells with failure to suppress insulin secretion in response to hypoglycemia	Inappropriate high plasma insulin ($>30\,pmol\,l^{-1}$) and C-peptide ($>100\,pmol\,l^{-1}$) concentrations in presence of hypoglycemia (BG $<2.2\,mmol\,l^{-1}$); Suppressed β-hydroxybutyrate levels ($<500\,\mu mol\,l^{-1}$)	High-carbohydrate intake orally or intravenously until curative surgical ablation or effective hyperglycemic therapy with diazoxide plus chlorothiazide can be instituted
Non-Islet cell tumor hypoglycemia (NICTH)	Abnormal tumor cells secreting big IGF-II	Low plasma insulin & C-peptide levels; low plasma IGF-I, normal or raised IGF-II levels; abnormal IGF-I:IGF-II ratio. Suppressed β-hydroxybutyrate levels ($<500\,\mu mol\,l^{-1}$)	High-carbohydrate intake orally or intravenously until curative surgical ablation of effective hyperglycemic therapy with growth hormone &/or prednisone can be instituted
Endocrine disease, e.g., Hypopituitarism, Addison's disease	Reduced availability of diabetogenic or hypoglycemia counterregulatory hormones	Clinical features of primary disease with subnormal levels of appropriate counter-regulatory hormones, e.g., cortisol, growth hormone. Appropriately raised β-hyroxybutyrate levels ($>500\,\mu mol\,l^{-1}$) during hypoglycemia	High-carbohydrate intake orally or intravenously until effective hormone replacement therapy has been established
Glycogen storage disease	Inability to release glucose from liver during fasting	Usually present in childhood; low blood glucose, high β-hydroxybutyrate levels, low insulin and C-peptide; high lactate; impaired or absent glucose response to glucagon	Avoid fasting: a constant intake of slowly absorbed carbohydrate may be required day and night in infants
Disorders of mitochondrial β-oxidation	Defective utilization of fat as fuel in tissues: compensatory increase in glucose utilization	Occurs in infancy; low glucose, low insulin & C-peptide, high FFA, normal lactate, low β-hydroxybutyrate, increased urinary organic acids. Hypocarnitinemia in some cases	Avoid fasting; frequent high-carbohydrate, low-fat feeding
Fasting alcohol-induced hypoglycemia	Alcohol impaired hepatic gluconeogenesis	Low blood glucose, raised blood alcohol, lactate and usually β-hydroxybutyrate: low plasma insulin and C-peptide	Avoid drinking alcohol whilst fasting or whilst on a low-energy diet
Idiopathic ketotic hypoglycemia of childhood	Varied: but always due to exhaustion of hepatic glycogen stores faster than cerebral adaptation to ketosis can occur	Low blood glucose; high plasma fatty acids and β-hydroxybutyrate; low insulin and C-peptide	High-carbohydrate feeding; avoidance of prolonged abstinence from food particularly during intercurrent illness, especially infections
Stimulative hypoglycemia			
Inborn errors of metabolism, e.g., hereditary fructose intolerance, galactosemia	Impaired release of glucose from liver in response to hepatotoxicity of food constituent	Hypoglycemia evoked by ingestion of foods containing appropriate noxious stimulus: galactose in galactosemia; fructose in hereditary fructose intolerance and fructose 1-6 bisphosphatase deficiency	Avoid foods containing provocative sugars, e.g., fructose, galactose as appropriate
Autoimmune insulin syndrome	Delayed release of insulin from antibody binding after all of meal has been absorbed	Profound hypoglycemia from 3 to 12 hs after last eating; total plasma insulin high; C-peptide high, normal or low proinsulin normal or high. Antibodies to insulin present. Common in Japan, infrequent elsewhere	Frequent small mixed meals, low in rapidly absorbed carbohydrate; rich in dietary fiber

Continued

Table 1 Continued

Description	*Mechanism*	*Diagnostic criteria*	*Dietary considerations*
Postgastrectomy and rapid gastric empying	Accelerated deposition of nutrients in duodenum and increased release of insulinotrophic hormones, e.g., GIP, GLP-1	Normal blood glucose during fasting; hypoglycemia only follows 1–3 hs after eating. History of gastrectomy or objective evidence of rapid gastric emptying. Exaggerated insulinemic response to food	Frequent small mixed meals rich in dietary fiber. May benefit from treatment with acabose or miglitol (α-glucosidase inhibitors)
Idiopathic reactive or functional hypoglycemia	Unknown: probably heterogeneous including increased insulin sensitivity, lowered cerebral threshold to neuroglycopenia	Hypoglycemia 3–5 hs after eating. Normal blood glucose during fasting: low capillary (arterial) blood glucose during spontaneous symptomatic neuroglycopenic episodes ($<3\,\text{mmol}\,l^{-1}$). All other objective tests of glucose homeostasis normal (including GIP and GLP-1 responses to food). Exclude noninsulinoma pancreatogenic hypoglycemia by intra-arterial calcium test	Frequent small mixed meals low in absorbed carbohydrates; rich in soluble dietary fiber. May benefit from treatment with acabose or miglitol (α-glucosidase inhibitors)

The Postprandial Syndrome

Typically, the patient is a normal-weight woman of 20–50 years whose main complaint is of vague feelings of distress occurring predominantly mid morning, about 11.00 a.m.–12.00 noon, but occasionally mid afternoon or evening and never before breakfast. In between attacks, characterized by feeling of faintness, anxiety, nervousness, irritability, inner trembling, rapid heart beat, headache, and sweatiness, either alone or in combination, they may be completely well. More often they describe themselves as suffering from increased tiredness, lacking in zest for life, and apathetic much, or all, of the time: symptoms often associated with depression or chronic alcohol abuse.

Patients seldom notice any fixed relationship to food unless, as so often happens nowadays, they have diagnosed themselves, on the basis of articles they may have read, as suffering from 'hypoglycemia.' Almost without exception they reject the possibility that their symptoms might have a contributory, or even large, psychogenic element.

Symptoms wax and wane during middle life but often remit completely for years or may never recur. They are not progressive and never cause severe neurological dysfunction such as coma, psychosis, or dementia. Hypoglycemia cannot be demonstrated during spontaneous symptomatic episodes in most people with the postprandial syndrome and some other explanation should be sought for them.

Differential Diagnosis

Studies using finger-prick blood sampling during spontaneous symptomatic episodes have shown that only a very small proportion of sufferers from the postprandial syndrome have hypoglycemia at the relevant time. Of those who do, a substantial proportion have an identifiable cause for it. The commonest is partial gastrectomy and rapid gastric emptying from any cause, in the West, and the autoimmune insulin syndrome in the Far East, i.e., Japan. Other more rare causes include insulinoma, the newly described condition of noninsulinoma pancreatogenic hypoglycemia, and abnormalities of GLP-1 secretion.

In some people reactive hypoglycemia occurs only in response to a specific dietary indiscretion: for example, ingestion of large amounts of gin (alcohol) and tonic (sugar and quinine) on an empty stomach. A hard core of subjects remains for whom no satisfactory pathogenic mechanism can be identified. Only in them is it justified to describe them as suffering from (idiopathic or functional) reactive hypoglycemia (**Figure 2**).

Dietary Management

Treatment of Attacks

Because of their short duration and modest severity, acute spontaneous neuroglycopenic episodes require no specific treatment beyond ingestion of a rapidly assimilable form of carbohydrate (e.g., a lump of sugar), exactly as for iatrogenic hypoglycemia.

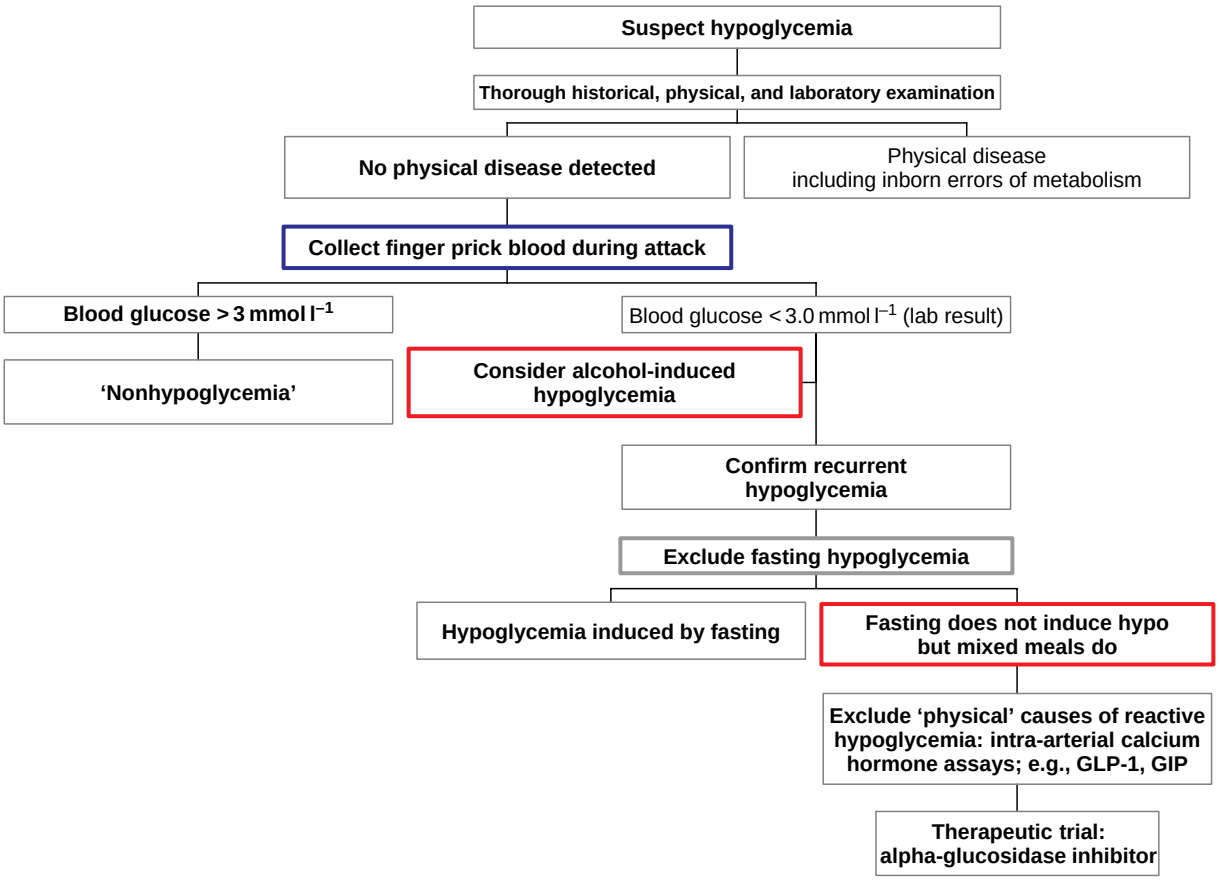


Figure 2 Investigation of reactive hypoglycemia. Steps in the diagnosis of reactive hypoglycemia of unknown etiology.

There is no evidence that this ever produces rebound hypoglycemia and should it do so the grounds for making a diagnosis of essential reactive hypoglycemia should be reviewed.

Prevention

Dietary prevention of reactive hypoglycemia, whether of the 'idiopathic', alimentary variety, or secondary to some other disease, is based on the premise that it is caused by imbalance between the timing and amount of insulin secreted in response to the ingestion of a meal and disposal of the glucose derived from it. Evidence for this supposition is small and disputed but provides the best explanation for the apparent breakdown in glucose homeostasis in patients with idiopathic reactive hypoglycemia.

Frequent small meals containing only modest amounts of sugars (glucose and sucrose) and refined starches but rich in poorly absorbed complex carbohydrates and containing dietary fiber have replaced the diets rich in proteins (and fats) previously advocated, but evidence of their unique efficacy is lacking. Avoidance of drinks rich in sucrose or glucose, especially with alcohol, may be helpful in subjects who are highly susceptible to this combination. There is no evidence that confectionery eaten in moderation is uniquely detrimental, though excessive use should be discouraged on general health grounds.

The long-term outcome of such dietary advice in patients in whom strict criteria for diagnosis were adopted are not available and most published studies on the subject have drawn attention to the need for supplementary pharmacological methods in order to achieve a satisfactory therapeutic outcome.

Pharmaceutical agents that have been used include guar, acarbose, and miglitol, all of which slow glucose absorption and decrease the insulinemic response to food, while others including phenytoin and propranolol do not. Paradoxically, diazoxide, which inhibits insulin secretion by direct action, has not been found effective except in patients with proven endogenous hyperinsulinism.

Nonhypoglycemia

No account of dietetic treatment of hypoglycemia would be complete without a brief description of 'nonhypoglycemia', which has been described as a

controversial illness and epidemic in the US. Clinically, the illness is indistinguishable from (idiopathic) reactive hypoglycemia, except that the blood glucose level is never pathologically low during symptomatic episodes. Moreover, although transient 'turns' are often a major feature of the illness, only rarely, if ever, does the patient consider their health, between turns, as normal.

The attribution of these patients' illness to hypoglycemia had its origins in the early 1950s with the appearance, in the US, of a book by Drs Abrahams and Pezet entitled 'Body, Mind and Sugar.' Other American practitioners, notably John Tintera, founder of the Hypoglycemia Foundation Inc., Stephen Gyland, Harry Saltzer and, others, including the medical writer Carlton Fredericks, publicized the concept. This led to 'hypoglycemia' being held, by a large section of the public, responsible for such diverse diseases as coronary artery disease, allergy, asthma, rheumatic fever, susceptibility to viral infections, epilepsy, gastric ulcer, alcoholism, suicide, and even homicide, as well as for a whole galaxy of symptoms in their own right. 'Hypoglycemia' was treated as though it were a disease entity and asserted by its advocates to be 'one of the most common illnesses in the United States' and that because of it 'thousands of Americans have forgotten, or perhaps never known, what it is like to feel completely healthy.' Diagnosis of 'nonhypoglycemia' generally depends upon the results of the 6-h oral glucose tolerance test, using venous blood, although some have dispensed even with this discredited formality in favor of just purely clinical criteria.

The appearance in the *New England Journal of Medicine* of an article entitled "Nonhypoglycemia is an epidemic condition" first drew international attention to the illness in 1974. It had previously been almost unknown outside the US and Australia, though known to a few fashionable medical practitioners in Britain.

Many patients with 'nonhypoglycemia' undoubtedly derive some benefit, probably through a powerful placebo effect, from severely restrictive dietary regimes. Although differing in details most of the diets emphasize the purported specifically detrimental effects of sugar (sucrose), salt, alcohol, and caffeine.

While the cause of illness in people with 'nonhypoglycemia' remains unknown, and is unlikely to be the same in all cases, in a tiny proportion it is due to caffeine intoxication, which can be confirmed by a dietary history and, above all, by measurement of plasma caffeine levels. Such patients do benefit specifically from reducing their intake of caffeinated beverages, though not necessarily from avoiding them completely. Ironically, and probably significantly, caffeine restores hypoglycemia awareness to diabetic patients on insulin who have become insensitive to it. The possibility exists, therefore, that a combination of reasonable or normal caffeine intake occurring in combination with the normal rebound in arterial blood glucose to just below fasting levels that sometimes occurs 3–5 h after a meal in someone with an unusually low threshold for neuroglycopenia, might precipitate symptoms. This explanation must, however, be considered no more than speculative.

On the other hand such diverse illnesses as hyperventilation, panic attacks, unadmitted alcohol or drug abuse, and genuine food intolerances are all established as capable of producing the 'nonhypoglyacemia' syndrome and should always be considered in the differential diagnosis.

Exercise-Induced Hypoglycemia

Previously only associated with marathon running, hypoglycemia is now recognized to be comparatively common in inadequately trained individuals undertaking strenuous exercise. Consumption of rapidly absorbed carbohydrate prior to taking exercise may encourage its appearance whilst consumption of slowly absorbed, low glycemic index foods may prevent it as does appropriate training.

Hepatic and Renal Failure

Considering the importance of the liver and kidney in the maintenance of blood glucose levels hypoglycemia is remarkably rare in both liver and kidney disease. In liver disease hypoglycemia is virtually confined to patients with acute toxic hepatic necrosis, whether due to overwhelming viral infection or specific hepatotoxins such as poisonous mushrooms, unripe akee fruit, and paracetamol in excess. Its appearance always portends an extremely poor prognosis. The association of hypoglycemia with primary cancer of the liver is comparatively common and due to overexpression and secretion of aberrant, or big IGF-II, and is not, as was once supposed, due to nonspecific destruction of hepatic tissue. Hypoglycemia is very rarely due to hepatic secondaries except from IGF-II secreting tumors.

Kidney failure is one of the commoner causes of hypoglycemia in nondiabetic hospital inpatients and does not carry as grave a prognostic significance as in patients with liver disease. It generally responds to appropriate dietary and other supportive treatments for end-stage kidney disease.

Inborn Errors of Metabolism

Hypoglycemia is a manifestation of many inborn errors of metabolism (see **Table 1**) especially in children but also occasionally in adults. It is particularly important in some varieties of liver glycogen storage diseases, especially types I and III, and in disorders of fatty acid metabolism in which it is often the presenting symptom.

Type I liver glycogen storage disease is due to a defect in glucose-6-phosphatase activity and produces a severe form of fasting hypoglycemia. Fortunately, this responds to dietary therapy in the form of continuous feeding with slowly absorbed starch solution through a nasal or gastrostomy tube, especially during the night when the body normally has to resort to glycogenolysis to maintain the supply of glucose to the brain. Hypoglycemia in untreated type I patients produces hypoinsulinemia and high to very high plasma ketone levels. Children with abnormalities of fatty acid metabolism, on the other hand, are characterized by hypoglycemia, hypoinsulinemia, and hypoketonemia. As with children with liver glycogen disease, treatment is to ensure that they are constantly supplied with carbohydrates and are never fasting for more than a very short period.

Starvation

Although average fasting blood glucose levels are lower in victims of famine than in well-fed populations, overt hypoglycemia is rare. Even in patients suffering from kwashiorkor, hypoglycemia is uncommon and is usually associated with infection, hypothermia, and coma. Patients with anorexia nervosa develop hypoglycemia only as an agonal phenomenon and its appearance generally portends imminent death. The characteristic clinical biochemistry findings are of low or undetectably low plasma insulin, proinsulin, C-peptide, and IGF-1 levels, grossly depressed plasma nonesterified fatty acids (NEFA) and β-hydroxybutyrate, and elevated growth hormone and cortisol levels. Relief of hypoglycemia by re-feeding is the only measure carrying any chance of preventing death, but it is rarely successful.

Hypoglycemia in the Elderly Sick

The high incidence of hypoglycemia in sick elderly patients has become apparent from the use of routine blood glucose measurements. The cause is seldom attributable to any of the well-recognized causes of hypoglycemia found in younger fitter people. It is probably due to chronic malnutrition that is so common in the elderly sick, compounded by coincident disease but which is not of itself sufficiently severe to produce hypoglycemia.

Conclusions

Symptoms due to documented spontaneous hypoglycemia are an unusual consequence of many different rare diseases and are sometimes the primary reason for a patient seeking medical help. In a minority of patients no pathological cause can be found to account for the hypoglycemia and no specific curative or palliative therapy can be instituted. Amongst these are a group of patients who only experience neuroglycopenic symptoms 2–5 h after eating a meal. They may benefit from eating small, frequent, slowly absorbed carbohydrate-rich meals. Usually, however, they also need addition of an α-glucosidase inhibitor to their diet.

Patients with self-diagnosed hypoglycemia in whom blood glucose levels are never pathologically low in everyday life and do not have any other known cause for their symptoms may also derive some benefit from a high-fiber, high complex carbohydrate diet, but how much of this is due to a placebo rather than specific dietary effect is still unknown.

See also: **Aging**. **Cancer**: Epidemiology and Associations Between Diet and Cancer; Epidemiology of Gastrointestinal Cancers Other Than Colorectal Cancers. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Exercise**: Beneficial Effects; Diet and Exercise. **Famine**. **Fatty Acids**: Metabolism. **Glucose**: Chemistry and Dietary Sources; Metabolism and Maintenance of Blood Glucose Level; Glucose Tolerance. **Liver Disorders**. **Starvation and Fasting**.

Further Reading

Brun JF, Dumortier M, Fedou C, and Mercier J (2001) Exercise hypoglycemia in non-diabetic subjects. *Diabetes and Metabolism* **27**: 92–106.

Brun JF, Fedou C, and Mercier J (2000) Postprandial reactive hypoglycemia. *Diabetes and Metabolism* **26**: 337–351.

Editorial (1974) Low blood sugar: Fact and Fiction. *Consumer Reports USA* **36**: 444–446.

Fonseca V, Ball S, Marks V, and Havard CWH (1991) Hypoglycaemia associated with anorexia nervosa. *Postgraduate Medical Journal* **67**: 460–461.

Hojlund K, Hansen T, Lajer M, Henriksen JE, Levin K, Lindholm J, Pedersen O, and Beck-Nielsen H (2004) A novel syndrome of autosomal-dominant hypersinsulinemic hypoglycemia

linked to a mutation in the human insulin receptor gene. *Diabetes* **53**: 1592–1598.

Kerr D, Sherwin RS, Pavalkis F, Fayad PB, Sikorski L, Rife F, Tamborlane WV, and During MJ (1993) Effect of caffeine on the recognition of and the responses to hypoglycemia in humans. *Annals of Internal Medicine* **119**: 799–804.

Klepper J and Voit T (2002) Facilitated glucose transporter protein type 1 (GLUT 1) deficiency syndrome: impaired glucose transport into brain – a review. *European Journal of Pediatrics* **161**: 295–304.

Lefebvre PJ, Andreani D, Marks V, and Creutzfeldt W (1988) Statement on postprandial hypoglycemia. *Diabetes Care* **11**: 439–440.

Marks V (1976) The measurement of blood glucose and the definition of hypoglycemia. In: Andreani D, Lefebvre PJ, and Marks V (eds.) *Hypoglycemia. Proceedings of the European Symposium, Rome. Hormone and Metabolic Research*, pp. 1–6. Suppl 6 Stuttgart: Georg Thieme.

Marks V (1987) Functional hypoglycaemia: fact or fancy. In: Andreani D, Marks V, and Lefebvre PJ (eds.) *Hypoglycaemia: Serono Symposia Publications*, vol. 38, pp. 1–17. New York: Raven Press.

Marks V (2003) Hypoglycaemia. In: Warrell DA, Cox TM, Firth JD, and Benz EJ (eds.) *Oxford Textbook of Medicine*, 4th edn, vol. 2, pp. 362–369. Oxford: Oxford University Press.

Mori S and Ito H (1988) Hypoglycemia in the Elderly. *Japanese Journal of Medicine* **27**: 160–166.

Palardy J, Havrankova J, Lepage R, Matte R, Belanger R, D'Amour P, and Ste-Marie L-G (1989) Blood glucose measurements during symptomatic episodes in patients with suspected postprandial hypoglycemia. *New England Journal of Medicine* **321**: 1421–1425.

Peter S (2003) Acabose and idiopathic hypoglycemia. *Hormone Research* **60**: 166–167.

Service FJ (1989) Hypoglycemia and the postprandial syndrome. *New England Journal of Medicine* **321**: 1472–1473.

Service FJ, Natt N, Thompson GB, Grant CS, van Heerden JA, Andrews JC, Lorenz E, Terzic A, and Lloyd RV (1999) Noninsulinoma pancreatogenous hypoglycemia: a novel syndrome of hyperinsulinemic hypoglycemia in adults independent of mutations in Kir6.2 and SUR1 genes. *The Journal of Clinical Endocrinology and Metabolism* **84**: 1582–1589.

Shilo S, Berezovsky S, Friedlander Y, and Sonnenblick M (1998) Hypoglycemia in hospitalized non-diabetic older patients. *Journal of the American Geriatric Society* **46**: 978–982.

Singer M, Arnold C, Fitzgerald M, Madden L, and Voight von Legat C (1984) Hypoglycemia: a controversial illness in US society. *Medical Anthropology* **8**: 1–35.

Snorgaard O, Lassen LH, Rosenfalck AM, and Binder C (1991) Glycaemic thresholds for hypoglycaemic symptoms, impairment of cognitive function, and release of counterregulatory hormones in subjects with functional hypoglycaemia. *Journal of Internal Medicine* **229**: 343–350.

Tamburrano G, Leonetti F, Sbraccia P, Giaccari A, Locuratolo N, and Lala A (1989) Increased insulin sensitivity in patients with idiopathic reactive hypoglycemia. *Journal of Clinical Endocrinology and Metabolism* **69**: 885–890.

Teale JD, Wark G, and Marks V (2002) The biochemical investigation of cases of hypoglycaemia: an assessment of the clinical effectiveness of analytical services. *Journal of Clinical Pathology* **55**: 503–507.

Yager J and Young RT (1974) Sounding board: non-hypoglycemia is an epidemic condition. *New England Journal of Medicine* **291**: 907–908.

Immune System *see* **Immunity**: Physiological Aspects; Effects of Iron and Zinc

IMMUNITY

Contents

Physiological Aspects

A T Borchers, C L Keen and M E Gershwin,
University of California at Davis, Davis, CA, USA

Immunity can be defined as the ability of an organism to resist or eliminate potentially harmful foreign organisms and materials or abnormal cells. Any substance capable of eliciting an immune response is called an antigen (antibody generator). Immune responses can be classified as innate or adaptive. Innate immune responses are also called nonspecific because they can be elicited by a wide range of foreign substances and are the same regardless of the exact nature of the substance and whether it had been encountered before. The major mechanisms of innate immunity include phagocytosis, inflammation, complement activation, and induction of cell death. Neutrophils and macrophages are the main cell types responsible for phagocytosis, and the chemical messengers that they and some other cell types produce play an important role in the initiation of an inflammatory response. The induction of apoptosis (programmed cell death) as part of the innate immune response is accomplished by natural killer (NK) cells.

In contrast to innate immune responses, adaptive immune responses are highly specific for a particular antigen and become stronger and more rapid over time. B cells and T cells represent the two types of lymphocytes responsible for adaptive immune responses. The main function of B cells is to produce antibodies, which neutralize pathogens or stimulate their elimination by other cell types through opsonization or complement activation. There are two major classes of T cells, namely helper T cells and cytotoxic T cells. One subclass of helper T cells provides help to macrophages in killing pathogenic microorganisms they have engulfed. The other subclass of helper T cells is vital for the induction of antibody production by B cells. Cytotoxic T cells directly eliminate infected cells by inducing them to undergo apoptosis. T cells also play a central role in self-tolerance (i.e., the ability not to respond to self antigens).

Initial exposure to pathogens (i.e., disease-producing microorganisms such as viruses and bacteria) most commonly occurs at the interfaces of host tissues and the external environment. Such tissues include the outer cells of the skin and, since the vertebrate body is essentially a 'tube within a tube,' the layers of cells and mucous lining the digestive, reproductive, and respiratory tracts. These cell layers and their secretions constitute nonimmunological physical and chemical barriers that provide a first line of defense against invasion by pathogenic microorganisms. Their barrier function is often reinforced by a variety of bacteria that generally do not harm the host but, on the contrary, provide additional protection from pathogens via competition, production of toxic substances, and stimulation of the immune system.

The main function of the immune system is to provide protection from invading pathogens, primarily viruses and bacteria but also fungi and parasites. For this purpose, the ability to discriminate between self or harmless non-self and potentially harmful non-self is absolutely crucial. Also important is the capability to recognize whether pathogens are extracellular (outside of the host's cells), such as fungi, certain bacteria, and some parasites, or intracellular, such as other bacteria and parasites and all viruses. Other activities of the immune system include the removal of worn-out cells, the identification and destruction of mutant or otherwise abnormal cells and also such inappropriate responses as allergies and autoimmune diseases, and graft rejection after organ transplantation.

Immune Cells and Organs

Immune cells are leukocytes (white blood cells) and, together with red blood cells, are ultimately derived from the same precursor or progenitor cells in the bone marrow. As illustrated in **Figure 1**, these stem cells give rise to either lymphoid or myeloid progenitors that subsequently differentiate into the different immune cells. A few other types of immune cells, including NK cells and mast cells, also arise from these pluripotent stem cells, but the pathways of their development are not fully known.

The differentiation of lymphocytes takes place in the central (also called primary) lymphoid organs—that is, bone marrow in the case of B cells and thymus in the case of T cells. After puberty, the thymus gradually atrophies and the production of new T cells decreases. After their maturation in the primary lymphoid organs, both types of lymphocytes migrate from these tissues through the bloodstream into the peripheral, or secondary, lymphoid tissues. These include the lymph nodes, spleen, and lymphoid tissues associated with mucosa, and they constitute the main sites at which the reaction of B and T lymphocytes with foreign antigens takes place. Lymph then carries circulating lymphocytes from the peripheral lymphoid tissues to the thoracic duct, where they reenter the bloodstream. The functions of the various immune cells are summarized in **Table 1**.

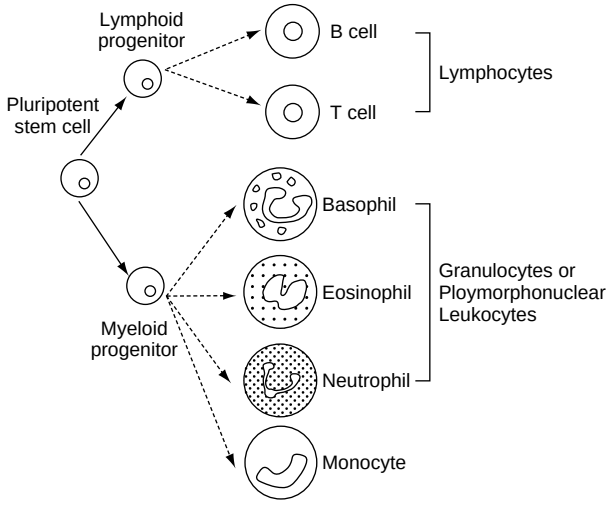


Figure 1 Immune cells give rise to either lymphoid or myeloid progenitors, which subsequently differentiate into the different immune cells.

Table 1 Functions of myeloid and lymphoid immune cells

Cell progenitor	*Cell type*	*Function*
Myeloid		
	Basophil	Unknown
	Eosinophil	Killing of antibody-coated parasites
	Neutrophil	Phagocytosis and activation of bactericidal mechanisms
	Macrophage[a]	Phagocytosis
	Mast cell	Release of histamine and other vasoactive and inflammatory mediators
Lymphoid		
	B cell	Antibody production
	T cell	
	Cytotoxic	Killing of virus-infected cells
	Helper (Th1 and Th2)	Activation of other cells such as B cells and macrophages
	Natural Killer cells	Killing of virus-infected cells and cancer cells

[a]This cell type circulates as monocytes in the bloodstream and matures into macrophages when taking up residence in a tissue.

Innate (Nonspecific) versus Adaptive (Specific) Immunity

Immune responses can be divided into two broad categories: innate and adaptive. Innate immune responses are also called nonspecific since they do not discriminate between most foreign substances. They are also not enhanced by previous exposure to a pathogen. In contrast, adaptive (also called acquired) immunity is highly specific to a particular pathogen and becomes more rapid and stronger with subsequent exposure to an antigen. Upon the initial encounter with an antigen, the adaptive immune response takes 4 or 5 days to become fully effective. During this period, the innate immune response plays a critical role in limiting and controlling infections. In addition, it is crucial in stimulating and directing the subsequent adaptive immune responses.

Innate or Nonspecific Immunity

Phagocytosis Pathogens can cause infection only after they have breached the nonimmunological

barriers of skin or mucosal surfaces. Generally, the first immune cells they come in contact with are macrophages. Among the many chemicals macrophages start to produce are chemokines, small proteins involved in the recruitment and activation of immune cells. The first cells to be recruited to the site of infection are neutrophils. Macrophages along with neutrophils are the major cell types involved in phagocytosis (i.e., the ingestion of foreign materials, including entire microorganisms). Phagocytosis is triggered via receptors on the surface of macrophages and neutrophils that recognize common cell wall components of bacteria. Killing of the ingested bacteria occurs via several different mechanisms involving the production of reactive oxygen and nitrogen species as well as the release of a variety of preformed antimicrobial substances.

Inflammation In addition to stimulating phagocytosis, the encounter of macrophages and neutrophils with bacteria frequently initiates an inflammatory response. The characteristics of inflammation are pain, redness, swelling, and heat. These symptoms are the consequence of the activities of cytokines and chemokines along with a variety of other vasoactive and inflammatory mediators, such as histamine, prostaglandins, and leukotrienes. They act mostly on local blood vessels, where their combined effect is to enhance blood flow, induce vasodilation, and increase the permeability of blood vessels. These changes allow leakage of fluids and plasma proteins, such as immunoglobulins, complement, and acute phase proteins, into the affected tissue. Cytokines also induce the expression of molecules that make it possible for immune cells to adhere to, and eventually pass between, the cells lining the blood vessels. Together, these alterations result in the infiltration of the site of inflammation by immune cells.

The major inflammatory cytokines are tumor necrosis factor-α (TNF-α), interleukin-1 (IL-1, IL-6, and IL-12, all of which are produced mostly by macrophages. Among these, TNF-α, IL-1, and IL-6 are central to mediating the acute phase response, which is characterized by elevation of the body temperature (fever) and a marked shift in the types of proteins secreted by the liver into the bloodstream. Whereas the synthesis of some liver proteins, called acute phase proteins, is dramatically increased, that of others is decreased. Among the acute phase proteins, C-reactive protein, mannose-binding protein, and serum amyloid P component undergo the most striking increase in synthesis, whereas only moderate increases are observed in a number of other acute phase proteins. It appears that the diverse activities of these proteins are ultimately beneficial to the host since they not only enhance the inflammatory response and other immune cell activities, thereby boosting host resistance, but also promote tissue repair.

Natural killer cells All viruses are intracellular pathogens since they lack the ability to replicate on their own and need to penetrate cells of the host in order to take over their replication machinery. The killing of such infected cells before the virus has had a chance to reproduce can be accomplished by a variety of cell types but is one of the major functions of NK cells. NK cells are large granular lymphocytes with a morphology and lineage that are distinct from those of B and T lymphocytes. They are known to be able to distinguish between normal cells and virally infected or tumorous cells, but the exact mechanisms by which they do so remain to be fully established. The granules of NK cells contain perforin, a protein that can polymerize and form transmembrane pores in the infected cell, possibly providing an entry route for a variety of enzymes also stored in the granules. As a result, the infected cell initiates an active suicide program called programmed cell death or apoptosis.

Interferons Interferon-α and interferon-β are proteins produced by many cells types in response to viral infection. They reduce the spread of viruses to uninfected cells by inhibiting protein synthesis and DNA replication in virus-infected cells and activating NK cells. In addition, they increase the expression of certain molecules and enhance certain cellular processes that are of great importance in activating components of the adaptive immune system involved in eliminating virally infected cells.

Complement The complement system consists of a group of proteins synthesized by the liver and released into the bloodstream in inactive form. It is part of the nonspecific immune response but can also be triggered by antigen–antibody complexes (i.e., it forms part of the humoral response in adaptive immunity). The latter pathway of activation is called the classical pathway and constitutes one of the three different pathways of complement activation. All three pathways involve a series of cleavage reactions converting inactive proteins into their active forms and ultimately converge at the formation of C3 convertase, an enzyme that cleaves complement component C3 into the large fragment C3b, on the one hand, and a group of smaller peptides consisting of C3a, C4a, and C5a, on the other hand. These smaller peptides mediate certain inflammatory processes and participate in the recruitment of

phagocytes. C3b binds to the surface of pathogens and, in the presence of simultaneous coating with antibodies, stimulates phagocytes to engulf and ultimately destroy the microorganism. In addition, further cleavage of C3b yields a group of terminal complement components that form a membrane attack complex able to damage the cell membrane and causing the lysis of certain pathogens.

Adaptive or Acquired Immunity

The cells of the innate immune system are vital as a first line of defense, but they are not always able to completely neutralize or eliminate infectious organisms. The adaptive immune system is thought to have evolved later in evolutionary history and now provides not only more specificity and versatility but also has added immunological memory as a further level of protection against reinfection with the same pathogen.

Adaptive immune responses can be classified as either antibody- or cell-mediated. Antibody-mediated, or humoral, responses are accomplished by plasma cells derived from B cells; cellular immune responses are mediated by activated T lymphocytes. Although they use vastly different effector mechanisms, the activation and subsequent differentiation of B and T lymphocytes nonetheless have many features in common.

In adaptive immunity, antigen alone is generally insufficient to activate naive antigen-specific lymphocytes. Naive T cells require a costimulatory signal from antigen presenting cells; naive B cells usually require accessory signals from an activated helper T cell, but in some cases the signal can be provided directly by microbial constituents. T cell help for B cells has to come from activated helper T cells that respond to the same antigen as the B cell, although the epitope—the specific part of the antigen that is recognized—is generally not identical. Upon recognition of its specific antigen in the context of the appropriate costimulatory signals, the previously small lymphocyte enlarges and undergoes a variety of changes in preparation for vastly increased RNA and protein synthesis. The activated cell is called a lymphoblast. This lymphoblast then begins to divide, duplicating every 6–12 h, thereby giving rise to ~1000 daughter cells, each exhibiting specificity that is identical to that of the parent. Thus, this group of cells constitutes a clone, defined as a population of identical cells that derive from the same ancestral line. Note that most antigens stimulate many different lymphocyte clones, making the resulting response polyclonal. The process of clonal expansion is followed by differentiation into either effector cells or memory cells.

Memory cells, unlike effector cells, do not participate in the initial immune response but can become activated cells when they encounter the same antigen at a later time point, in some cases years or even decades later. This, along with other changes in memory cells compared to virgin (or naïve) cells, accounts for the fact that the primary immune response is characterized by a lag phase of several days (the period during which lymphocytes undergo clonal expansion and differentiation) and is relatively weak, whereas a second exposure to the same antigen results in a much more rapid and stronger response.

B cells The primary function of B cells is the production of antibodies, or immunoglobulins. There are five major classes (isotypes) of antibodies: IgA, IgD, IgE, IgG, and IgM, with IgA and IgG having two and four subclasses, respectively. Resting B cells express IgM and IgD on their cell surface as antigen receptors. Their function is to capture antigen so that it can then be processed and displayed to helper T cells specific for peptide fragments of the same antigen. In response to binding antigen and receiving the necessary accessory signal from helper T cells in the form of cell–cell interactions along with secreted molecules, B cells start to produce a secreted version of IgM. Under the influence of certain cytokines produced predominantly by activated helper T cells, B cells undergo isotype switching, also called class switching, meaning that they start to produce other types of immunoglobulins. The types and combinations of immunoglobulin isotypes depend on the nature and relative amounts of these cytokines. In B cell activation and initiation of antibody production, a subclass of helper T cells called Th2 plays the major role; the Th1 subclass of helper T cells, however, participates in isotype switching via the production of interferon-γ, a cytokine that induces switching to specific subclasses of IgG, namely IgG2a and IgG3.

Antibody structure and diversity It is estimated that even in the absence of antigen stimulation, the human body contains B cells capable of producing approximately 10^{15} different antibody molecules. This enormous diversity is generated through a variety of mechanisms.

The basic structure of an antibody is a Y shape (**Figure 2**) consisting of two heavy chains and two light chains, with each arm containing a specific antigen binding site formed by parts of the respective heavy and light chain. The light and the heavy chain each have a constant region and a variable region. Within the variable region three small

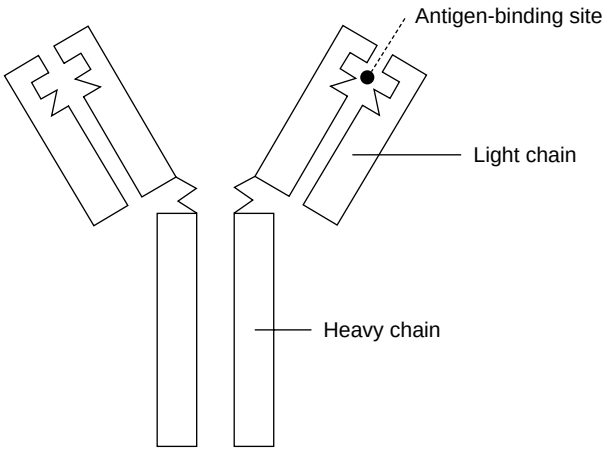


Figure 2 Basic structure of immunoglobulins.

hypervariable regions containing 5–10 amino acids form the antigen binding site. A different pool of gene segments encodes the constant and variable regions and, in addition, there is another pool for joining (J) segments for both heavy and light chains and a pool for diversity (D) segments in the case of heavy chains. The exact number of gene segments in these pools is not known, but the mouse genome is estimated to contain approximately 300 variable (V) segments for one of the two possible light chains, and these can be joined to any of 4 different J segments, yielding at least 1200 different V regions. In addition, there are approximately 500 V segments in the heavy-chain pool of the mouse, which can be combined with 4 J segments and at least 12 D segments to encode 24 000 different heavy-chain V regions. Thus, a total of at least 2.5×10^7 different antigen binding sites can be generated by combinatorial diversification, as the combining of V, J, and D segments is called.

The joining process further increases this diversity via two mechanisms. One operates in heavy- and light-chain segments and is the loss of 1 or more nucleotides from the ends of recombining gene segments. Heavy-chain gene segments can additionally be modified through the random insertion of up to 20 nucleotides. Although it is not uncommon for this junctional diversification to result in the production of nonfunctional genes, it nonetheless increases the number of different B cells in the mouse to an estimated 5×10^8.

After the assembly of functional antibody genes is completed, an additional mechanism for increasing diversity takes place when a B cell is stimulated by antigen. This process is called somatic hypermutation since it involves the insertion of point mutations at a rate that is approximately 1 million times greater than the spontaneous mutation rate in other genes. A few of these point mutations confer increased affinity for the antigen to the antigen receptors, ultimately resulting in the production of antibodies with progressively increasing affinity during the course of an immune response.

The function of antibodies The coating of pathogens and toxins with antibodies helps protect the host from infection in three main ways: neutralization, opsonization, and complement activation. Neutralization refers to the ability of antibodies to inhibit the adherence of pathogens to cells they might invade and destroy. Opsonization is defined as the coating of pathogens with antibodies in order to increase their susceptibility to ingestion by phagocytes. As discussed previously, antibodies complexed with antigen can also trigger the classical pathway of complement, thereby either enhancing opsonization or directly killing some bacteria through the formation of membrane attack complexes. Not all of the secreted antibody isotypes participate in all of these functions, and the extent to which they do so also differs. Certain additional functions are restricted to specific isotypes. For example, only some IgG subclasses can bind to certain viral proteins displayed on virally infected cells and, through interaction with specific receptors, signal NK cells to destroy these cells. Another example is IgE, which is the only isotype capable of sensitizing mast cells, resulting in a local inflammatory response mediated by the release of histamine and other inflammatory mediators. Allergic reactions are the consequence of such a response directed against innocuous antigens.

T cells Whereas B cells recognize and bind directly to extracellular antigens, generally native protein structures, T cells recognize partially degraded protein antigens—that is, peptide fragments that result from intracellular processing and are then carried to the cell surface for display there. The generation of peptide fragments is called antigen processing and the display is called antigen presentation.

Intracellular pathogens can be located in two different compartments of a cell, the cytosol or the vesicular compartment, which is separated from the cytosol by membranes. Depending on the cellular location of the microorganism, one of two different classes of T cells is activated, either cytotoxic T cells or helper T cells. Cells infected with cytosolic pathogens, such as viruses and some bacteria, are eliminated by cytotoxic T cells via mechanisms closely resembling those described for NK cells. Cells containing foreign material or microorganisms in their vesicular compartment stimulate helper T cells.

These do not kill cells but enhance the activity of the very cells stimulating them (i.e., macrophages and B cells). Although macrophages can phagocytose and kill many infectious agents without T cell help, there are certain situations in which such help is indispensable. For example, the mycobacteria responsible for tuberculosis and leprosy have developed mechanisms to survive the process of phagocytosis and can replicate inside vesicular structures. However, they can be eliminated when the macrophage is activated by a helper T cell. T lymphocytes providing help for macrophages belong to a subclass of helper T cells characterized mainly by the types of cytokines it produces and designated as Th1. Another subclass, Th2, activates B cells to make antibody.

Cytotoxic and helper T cells have the same kind of antigen receptors, designated as T cell receptors. This indicates that the ability of the different classes of T cells to distinguish between peptide fragments coming from the cytosolic or the vesicular compartment must involve other molecules. The most important of these are major histocompatibility complex (MHC) molecules.

The MHC is a cluster of genes encoding not only two different classes of MHC molecules but also a variety of other proteins that participate in immune responses. 'Histo' means tissue, and the name 'major histocompatibility complex' reflects the fact that the proteins encoded by this group of genes were first identified as the target of the immune reaction that can result in graft rejection after organ transplantation. In humans, MHC molecules are called human leukocyte-associated antigen (HLA) molecules. In addition to being polygenic (having several genes encoding proteins with the same function), the MHC genes are strikingly polymorphic, meaning that there are multiple alleles, or copies, of each gene.

T cells recognize antigen only when presented as peptide fragments by a MHC molecule. Furthermore, a T cell recognizing a peptide fragment bound by a MHC molecule encoded by a particular allele will not recognize the same peptide bound to another type of MHC molecule, an effect called MHC restriction. Together with the polymorphism of the MHC genes, this limits the ability of a pathogen to put entire populations or even species at risk since the individuals within the population will vary in their susceptibility to the pathogen.

There are two classes (I and II) of MHC molecules that are structurally similar, though distinct, but differ functionally. Class I MHC molecules present foreign peptides to cytotoxic T cells; class II MHC molecules present foreign peptides to helper T cells. Since viruses can infect any cell containing a nucleus, virtually all nucleated cells express MHC class I molecules, although the levels at which they do so can differ considerably. In contrast, the main function of helper T cells is to activate other cells of the immune system. Thus, MHC class II molecules are constitutively expressed on B lymphocytes, macrophages, and other antigen presenting cells, but they are inducible on many other cell types via certain cytokines.

Since MHC molecules insert themselves into the cell membrane once they have picked up processed antigen fragments inside the cell and the T cell receptor is also a cell surface molecule, T lymphocytes must make direct contact with their target cells. This cell–cell interaction is enhanced by so-called coreceptors, designated CD4 on helper T cells and CD8 on cytotoxic T cells. CD4 proteins recognize an invariable part of the class II MHC molecule, whereas CD8 proteins bind to a nonvariable region of the class I MHC molecule, and both play a vital role in ensuring that a T cell recognizes only those target cells bearing the correct type of MHC molecule.

As in the case of B cells, antigen alone is insufficient to activate T cells. For helper T cells, the accessory signal is provided either by a secreted signal such as IL-1 or by a specific plasma membrane molecule on the surface of an antigen presenting cell. The major cell type presenting antigen to T cells is the dendritic cell found in lymphoid organs, but macrophages and, under certain conditions, B cells can also act as antigen presenting cells. Antigen presenting cells with strong costimulatory activity also provide the secondary signal for cytotoxic T cells, but in some cases the presence of CD4 T cells seems to be required as well.

Tolerance

Many immune responses can be destructive to host tissue; hence, it is vital that they be restricted to pathogens and not be raised against innocuous substances. The devastating consequences of autoimmune diseases, which result from immune responses that are inappropriate in that they are directed against self, illustrate the crucial need for self-tolerance. It is now known that the immune system has the inherent capability of responding not only to foreign but also to self antigens. During development, it 'learns' not to respond to self antigens. The two major mechanisms for establishing self-tolerance are clonal deletion, the killing of self-reactive lymphocytes, and clonal anergy, the functional inactivation of self-reactive lymphocytes involving antigen stimulation in the absence of

accessory signals. These processes are focused mainly on T lymphocytes since most B cells require helper T cells to respond to antigen so that the elimination of self-reactive helper T cells also ensures the inactivation of self-reactive B cells.

Interactions between Nutrition and Immunity

There are considerable interactions between nutrition and immunity; that is, not only does nutrition affect immune functions but also immune responses have profound effects on metabolism and nutritional status. Essentially every aspect of immunity can be affected by undernutrition, and there is growing evidence that overnutrition also has detrimental effects on immune responses. Protein–energy malnutrition during childhood leads to stunted development of many of the immune organs. In subjects of all ages suffering from protein–energy malnutrition, almost all immune functions are considerably impaired, resulting in increased severity and prolonged duration of most infectious diseases. However, protein–energy malnutrition rarely occurs in the absence of inadequate intake of other essential nutrients, and deficiencies in virtually every vitamin and essential mineral or trace element are associated with reductions in one or more functions of innate and adaptive immunity. The amount of fat and types of fatty acids in the diet are also known to influence certain immune functions, such as phagocytosis, the ability of cells to move to the site of inflammation, and the production of proinflammatory cytokines and other inflammatory mediators. In addition, evidence is beginning to accumulate that many other dietary constituents (e.g., carotenoids, flavonoids, other plant-derived chemicals, and probiotics (beneficial bacteria)) can also modulate a variety of immune responses, but their effects in humans remain largely unexplored.

Infection, in turn, is associated with profound effects on nutritional status resulting from decreased nutrient intake due to loss of appetite, decreased nutrient absorption as a result of intestinal damage and malabsorption, and nutrient losses arising from diarrhea and increased urinary excretion. Moreover, the inflammatory processes following infection can cause oxidative damage to host cells, and the prevention of such damage increases the demand for antioxidant defenses, including the vitamins C and E and a variety of enzymes that depend on trace metals for their function. In addition to its effects on nutritional status, the acute phase response is accompanied by marked changes in a variety of metabolic processes, with priority being shifted to the synthesis of all the different proteins involved in protecting the host from the invading pathogen. Another protective mechanism involves the redistribution of iron away from the bloodstream into the cells that participate in the phagocytosis and killing of invading pathogens. The removal of iron from the blood is accomplished by lactoferrin, an iron-binding protein that is produced in greatly increased amounts during the inflammatory response. By sequestering iron from pathogens that require this trace element for growth, lactoferrin can prevent such organisms from multiplying. Note that supplementation of iron-deficient subjects may reduce the resistance to malaria and may increase the short-term risk of certain infections. In general, however, correction of nutritional deficiencies via supplementation results in the partial or even full restoration of the compromised immune functions.

See also: **Cytokines**. **Immunity**: Effects of Iron and Zinc. **Infection**: Nutritional Interactions. **Prostaglandins and Leukotrienes**.

Further Reading

Gershwin ME, German JB, and Keen CL (eds.) (2000) *Nutrition and Immunology: Principles and Practice*. Totowa, NJ: Humana Press.

Janeway CA Jr and Travers P (1996) *Immunobiology: The Immune System in Health and Disease*. London/New York: Current Biology Limited/Garland.

Roitt IM (1994) *Essential Immunology*. Oxford: Blackwell Scientific.

Weir DM and Stewart J (1993) *Immunology*. New York: Churchill Livingstone.

Effects of Iron and Zinc

C Doherty, MRC Keneba, The Gambia

Introduction

Iron and zinc have achieved prominence among the micronutrients due to the wealth of research detailing their fundamental importance to a multitude of basic cellular physiological mechanisms. The complexity of the immune system ensures that both divalent cations are necessary for normal function. The effects of deficiency and supplementation on the immune functioning of both deficient and replete individuals have

provided valuable clues to the specific immune processes in which they are involved. However, single nutrient deficiencies rarely occur alone and the effects of coexisting macro- and micronutrient deficiencies have contributed to the continuing debate on individual nutrient importance. Individuals and populations respond differently to supplementation. The basis of this variability is poorly understood but key to the improved targeting of supplementation. Specific mechanisms to explain the immune effects of zinc supplementation of deficient individuals have been particularly difficult to characterize as zinc is involved in so many cellular processes.

These nutrients are also important for prokaryotes and their acquisition by invading microbes is an important step in the development of a potential pathogen. The host action of micronutrient withdrawal is recognized as a mechanism of immune defense. In the era of genomic medicine the characterization of the molecular determinants of acquisition, storage, flux, and excretion of iron have increased our understanding and illustrated the complexity of iron homeostasis. In this article the evidence for the immune importance of both iron and zinc from *in vitro* experimentation and *in vivo* studies of human deficiency and supplementation is considered. The objective of supplementation, dose, route, preexisting level of deficiency, immunocompetence, coexistent deficiencies, genetic determinants, and the presence of infection should all be considered in the decision of who to supplement and when.

Iron and Zinc Homeostasis

Iron is the most abundant element on Earth. Despite this it is the most common micronutrient deficiency on Earth with up to 50% of all children under 5 years and pregnant women in developing countries affected. The ability of iron to both bind oxygen and to donate and accept electrons ensures that it has a central role in cellular energy metabolism. The utility of this redox potential is, however, counterbalanced by the propensity of iron to generate free radicals and damage cell membranes through lipid peroxidation. Genomic investment and redundancy in mechanisms to control iron availability at the cellular level illustrates both its importance and potential for toxicity.

Iron homeostasis depends on the regulation of iron absorption from the intestine as there are no pathways for iron excretion. On average, 1–2 mg enters the adult human body on a daily basis and a variable amount leaves via sloughing of skin and mucosal cells. Diets rich in heme iron and vitamin C promote iron absorption. Meat and nonanimal foods such as legumes and green leafy vegetables combine readily available heme iron with promoters of absorption and utilization of non-heme iron. Phytate-containing foods, e.g., cereals, inhibit absorption. Non-heme iron is reduced and solubilized to the ferrous form in the proton rich environment of the proximal duodenum and actively transported across the enterocyte. The transport of ferrous iron through the enterocyte represents the primary site of iron homeostasis – it can be stored as ferritin, lost through sloughing of intestinal cells, or exported systemically.

Transferrin binds and solubilizes ferric iron exported from the enterocyte with high affinity and transports it to cells. Uptake and internalization of iron-transferrin by endocytosis is followed by its dissociation at lower intracellular pH and storage of iron in cytoplasmic ferritin molecules. Iron absorption, however, does not fulfil the majority of daily hemopoietic requirement. Senescent red blood cells are phagocytosed by reticuloendothelial macrophages, which recycle the iron from heme – they load the ferric iron back onto transferrin for reuse and this recycling of heme iron accounts for 80% of hemopoietic requirement.

Hemoglobin and intracellular ferritin, in the liver, bone marrow and spleen, account for over 99% of total body iron. Iron is more readily available than zinc and we have developed strategies to manage large fluxes of iron. Conditions characterized by hemolysis demonstrate the complex adaptive mechanisms that protect cells from episodes of flux.

Zinc is the twenty-fifth most abundant element comprising less than 0.01% of the earth's crust. Its single oxidant state enables it to hydolyze bonds involving carboxyl and amino groups and its ability to form stable complexes with sulfur and nitrogen atoms is utilized in stabilizing proteins. It has structural and regulatory roles in numerous enzymes, signaling pathways, and gene transcription systems essential for growth, reproduction, and metabolism. Up to 2 g of zinc is present in an adult man but most (95%) is locked away in pools from which it cannot rejoin the circulation and influence plasma levels, e.g., muscle and bone. Small plasma and liver pools are accessible and labile and act as the only reserve available in dietary deficiency. Zinc homeostasis is thus dependent on dietary intake and the average man has an intake of 10 mg day^{-1}. Meat is a good source of zinc but plant sources (e.g., lentils and cereals) are often compromised by the presence of phytate, which inhibits absorption.

Plasma zinc is 99% bound to albumin and other low molecular weight proteins. Plasma zinc makes

up only a small percentage of body zinc. The control of zinc flux at the cellular level is much less well characterized than that of iron.

Metallothionens are a group of intracellular monomeric polypeptides that bind zinc and serve as homeostatic modulators of zinc availability. Relatively little is known of how zinc enters immune cells and how it influences function. Recently, a family of zinc transport genes (ZnT 1–4) has been cloned. ZnT 1 is associated with zinc efflux and expression of this gene is regulated by zinc intake. Further work is needed to clarify the role of this family in zinc transport and its possible regulatory influences, e.g., zinc status and inflammation. Zinc deficiency is difficult to identify both clinically and biochemically and only in the last decade has the widespread nature of this deficiency been recognized particularly in children in developing countries.

Deficiency

Iron and especially zinc deficiency are difficult to diagnose and differing diagnostic criteria contribute to the confusion surrounding deficiency and immune dysfunction. Plasma levels of either micronutrient are not adequate to define status, however they commonly have been used for such. Deficiency of iron can be quantified at individual and community levels using a combination of indices, e.g., hemoglobin/mean cell volume combined with an index of storage iron, ferritin plus an index of iron supply to tissues, or serum transferrin receptor concentration. Plasma zinc level decreases with inflammation and currently the best method of assessment for deficiency is response to supplementation. Alternatively, plasma zinc levels can be interpreted with caution in conjunction with a marker of the acute phase response. In developing countries iron and zinc deficiency are widespread and often occur together. Meat is the most important dietary source for both micronutrients; however, in many countries with predominately vegetarian diets phytate-containing cereals inhibit absorption of both.

Deficiency and Immunity

Proliferation of cells requires iron, as the DNA synthetic enzyme ribonucleotide reductase is iron dependent. Impaired T cell proliferation and impaired delayed type hypersensitivity have been consistently reported in iron deficiency. Phagocytosis is accompanied by the generation of toxic oxygen intermediates to kill ingested bacteria and both neutrophils and macrophages require iron for this process – nitroblue tetrazolium reduction and hydrogen peroxide production are reduced in neutrophils and macrophages if made iron deficient. In contrast, excessive iron results in decreased phagocytosis by neutrophils possibly due to increased free radical production and consequent lipid peroxidation damage of the phagosome membrane. Iron overload and saturation of transferrin also inhibits lymphocyte proliferation. When assessing the effect of iron on the immune system macrophages deserve special attention, as they are responsible for recycling heme iron back to the bone marrow. Iron thus fluxes through macrophages en route to fulfil hemopoietic need and is also utilized to kill intramacrophagal microbes. Macrophage activation and intracellular killing are dependent on the generation of toxic oxidant molecules such as the hydroxyl radical; however, these same free radicals can damage host cell membranes through lipid peroxidation. The control of intracellular iron in macrophages is thus important to hemopoiesis, microbial killing, and cell membrane stability within the macrophage itself. Iron overloading of macrophages impairs normal function and may increase risk of disease. Iron overloading of the macrophage causes oxidant damage to the phagocyte and an impaired ability to kill intracellular pathogens via IFNγ-mediated pathways. Iron has a direct inhibitory effect on the actions of IFN-γ, e.g., formation of tumor necrosis factor apha (TNF-α), expression of major histocompatibility factor (MHC) class II antigens, formation of neopterin, and nitric oxide synthesis.

Studies of immune function in iron-deficient human populations are frequently confounded by coexisting nutritional deficiencies, prevailing socioeconomic conditions, and differing diagnostic criteria, which make them difficult to interpret and compare. Increased morbidity from infectious disease has been reported in iron-deficient populations; however, it is not clear whether this is due to iron deficiency alone or at what level of deficiency the immune system is functionally compromised. In studies of children in predominately malarious areas a definite increase in mortality has only been demonstrated in anemic patients with less than $50\,g\,l^{-1}$ hemoglobin, whereas in those children with milder anemia the evidence for increased risk of mortality is inconclusive.

Zinc promotes mRNA stability, regulates gene expression, and influences DNA replication ensuring an essential role in cell division and activation. These processes are central to the immune response and zinc deficiency affects immune function at many levels both in the innate and specific arms. Zinc deficiency rarely occurs alone and has no pathognomonic clinical features. Cell-mediated immunity is profoundly affected in zinc deficiency. Lymphopenia

is common, as are defects in specific T and B lymphocyte function. Lymphoid atrophy, decreased delayed cutaneous hypersensitivity responses, reduction in numbers of CD4 helper cells, B cell dysfunction, impairment of phagocytosis, and deficient thymic hormone activity have all been described. Secretion and function of cytokines and the potentiation of apoptosis are affected by zinc deficiency. Gastrointestinal barrier function, polymorphonuclear and natural killer cell function, and complement activity are also affected. Zinc may also prevent free radical-induced injury through its antioxidant and cell membrane stabilizing properties. Mild zinc deficiency resulted in an imbalance between TH1 and TH2 functions in male volunteers – reduced serum thymulin activity, reduced CD4/CD8 lymphocyte ratio, and reduced interleukin-2 (IL-2) production but production of IL-4, IL-5, IL-6, and IL-10 was unaffected. Influencing TH1/TH2 balance is a potentially important pathway by which zinc deficiency affects cell-mediated immunity.

Supplementation

Studies of the effect of iron supplementation or food fortification on morbidity from infection have been inconsistent in showing evidence of benefit. Morbidity is difficult to measure accurately and populations differ in relative deficiency, dietary supply of iron, and prevalent infectious agents. Individuals will also differ in their response to supplementation and the causes of population and individual variation are not well defined. Whilst some populations have shown a decline in infectious disease morbidity some observational evidence that iron supplementation leads to increased morbidity and mortality from infection has been reported. Supplemental oral iron given to septic kwashiorkor children was associated with increased risk of death. Prophylactic iron dextran given intramuscularly to Polynesian newborns in New Zealand caused increased neonatal sepsis that declined on stopping supplementation. In a controlled trial of prophylactic iron given intramuscularly to 2-month-old infants in Papua New Guinea supplementation was associated with a significant increase in hospital admissions for pneumonia.

Malaria is associated with hemolysis and fluxing of iron through the reticuloendothelial macrophages back to the bone marrow. Malaria-associated anemia is therefore not necessarily associated with iron deficiency (though it commonly can coexist) but with iron delocalization in macrophages. The effect of supplemental iron on iron delocalization is unclear. An increase in the prevalence of malaria parasites in thick blood films, splenomegaly, and malaria-associated hospital admissions were reported in the Papua New Guinea study. Other trials of oral iron supplementation in children in areas where malaria is endemic have reported mixed findings but concerns persist as to possible increased morbidity from malaria due to iron supplementation. Current guidelines favor the continued use of iron supplementation in malarious regions as benefit to the host is felt to outweigh possible benefits to the parasite. Avoidance of parenteral administration and usage of lower supplementation doses has been advocated.

The study of conditions of chronic iron overload, e.g., hemochromatosis, chronic renal failure or hereditary hemolytic anemia requiring repeated blood transfusions, has provided further insights into how disturbed iron homeostasis affects immunity and infectious disease morbidity. Impaired phagocytic function inversely correlates with ferritin concentration and impaired natural killer cell function from iron overloaded thalassemic patients correlated with their degree of iron overload. Iron-overloaded patients are more susceptible to infections and the excess free iron encourages bacterial growth and pathogenicity. *Yersinia* infection has been reported in patients with hemochromatosis and those treated for acute iron toxicity with desferrioxamine. This bacterium can only acquire free iron and is therefore more likely to be invasive in its presence.

Zinc supplementation has been clearly demonstrated to improve immune function. Marasmic children given zinc supplementation demonstrate enlarged thymic shadows, increased conversion of delayed hypersensitivity skin reactions, enhanced lymphoproliferative response to PHA, and increased salivary IgA concentrations. Supplementation of malnourished children was associated with significantly larger delayed type hypersensitivity skin reactions and significantly decreased incidence of fever, cough, and upper respiratory tract infections. Zinc-supplemented infants also demonstrated significantly better serum IgA and significantly reduced incidence of pyoderma and anergy. Zinc supplementation decreased the percentage of children under 3 years who remained anergic to skin tests of delayed hypersensitivity associated with a significant rise in CD3, CD4, and the CD4/CD8 ratio.

The last decade has provided a wealth of studies detailing benefit from zinc supplementation in populations of under 5s in developing countries despite the difficulties in diagnosing zinc deficiency and understanding zinc homeostasis. By implication widespread deficiency exists. The evidence for immune benefit of zinc supplementation is much stronger than that for iron. There is now clear

evidence that malnourished children and children with chronic diarrhea benefit from therapeutic zinc supplementation and that prophylactic zinc supplementation reduces diarrheal disease morbidity. Studies continue on the use of zinc to prevent and treat acute respiratory tract infection but any benefit to malaria morbidity or mortality has not been clearly established. There is good evidence for zinc deficiency in certain population subgroups, e.g., stunted/wasted children and children with chronic diarrhea. Population-based supplementation rather than targeting specific patient groups has also demonstrated benefit but recognition of appropriate populations to supplement is problematic. Populations with high rates of stunting may well have a high likelihood of zinc deficiency and thus be considered as potential candidates for supplementation. Knowledge of local dietary supply of zinc and results of local intervention studies should guide targeting to zinc-deficient populations.

Population-based supplementation approaches assume the safety of supplementation of individuals within populations who are not deficient. In general zinc supplementation is safe; however, individual variability in response exists and supplementation of replete individuals is not beneficial. There may be subgroups within the population for which supplementation may be detrimental. Zinc supplementation during sepsis has caused clinical problems particularly in the presence of a compromised immune system. Marasmic infants have demonstrated reduced phagocytic and fungicidal monocytic activity and a significantly increased number and duration of episodes of impetigo. High-dose zinc supplementation ($6\,mg\,kg^{-1}\,day^{-1}$) given early in rehabilitation to severely malnourished children led to increased sepsis and mortality compared to low-dose supplementation ($1.5\,mg\,kg^{-1}\,day^{-1}$). *In vitro* evidence points to free zinc ion concentration as a determinant of effects on monocytes and lymphocytes. Excess zinc induces the release of IL-1β and IL-6 and TNF-α from monocytes and inhibits IL-1-dependent T cell stimulation. Granulocyte phagocytosis is also impaired by zinc in a concentration-dependent manner. Supplemental zinc in rats decreased mobilization of polymorphonuclear cells and macrophages into the peritoneal cavity and phagocytic function. In adult men administration of 150 mg of elemental zinc twice a week for 6 weeks was associated with a reduction in lymphocyte stimulation response to phytohemagglutinin as well as chemotaxis and phagocytosis of bacteria by polymorphonuclear leukocytes. *In vitro* evidence points to a concentration-dependent effect of zinc on immune function.

Inflammation and Micronutrient Flux

The acute inflammatory reaction is a coordinated and complex series of physiological and immune adaptations designed to optimize protection to an invading pathogen. Immune cells are activated and cytokines released (e.g., IL-1 and TNF) to increase endothelial permeability and chemotaxis and activate complement. Effector immune cells are thus brought to a site of tissue injury and activated. Cytokines mediate the dramatic changes in the micronutrient milieu that accompanies inflammation. Plasma levels of both iron and zinc fall as both micronutrients are withdrawn from readily available pools and diverted to the reticuloendothelial system so as to both optimize immune response and deny access to invading pathogens. Decreased iron absorption, decreased iron release from reticuloendothelial macrophages, and increased transferrin catabolism contribute to iron withdrawal. Macrophagal sequestration of iron is mediated by inflammatory cytokines. Cellular iron homeostasis is a post-transcriptional event and expression of the transferrin receptor limits acquisition of iron. Iron-response proteins (IRP) 1 and 2 bind to iron-responsive elements (IRE) in transferrin receptor and ferritin mRNA and thus regulate expression of these genes and uptake, utilization, and storage of intracellular iron. Inflammatory cytokines modulate intracellular iron status by regulating the IRP/IRE network. Proinflammatory cytokines released during inflammation, e.g., TNF-α and IL-1, increases ferritin transcription and induce a diversion of metabolically available iron into the storage compartment in macrophages thus limiting iron availability for erythropoiesis. This diversion of iron underlies the anemia of inflammation/chronic disease.

Zinc is also diverted to the reticuloendothelial system during inflammation and lower plasma zinc concentrations are associated with both optimal phagocytic function and decreased microbial virulence. Calprotectin is an acute phase zinc-binding protein produced by polymorphonuclear leucocytes that sequesters zinc from invading pathogens. IL-1 released in inflammation increases the expression of metallothionen 1 and 2 in the liver, bone marrow, and thymus, which accompanies the increased uptake of zinc in these organs.

Iron and zinc are essential for microbial survival as cofactors for both superoxide dismutase and catalase redox enzyme systems. These enzymes neutralize the reactive oxygen intermediates integral to phagosomal killing. Both iron and zinc are also cofactors for bacterial enzymes required for DNA

Table 1 Immune effects of iron and zinc deficiency.

Iron	*Zinc*
Effect of deficiency	**Effect of deficiency**
Impaired lymphocyte proliferation	Lymphopenia
Impaired delayed type hypersensitivity	Impaired T & B lymphocyte function
Impaired phagocytic function	Impaired phagocytic function
	Impaired gastrointestinal barrier function
	Impaired natural killer cell function
	Impaired complement function
	Impaired TH1/TH2 balance
Effect of overload	**Effect of overload**
Impaired lymphocyte proliferation	Impaired lymphocyte stimulation
Impaired phagocytic function	Impaired phagocytic function
Impaired natural killer cell function	

synthesis. Intracellular pathogens, e.g., tuberculosis, must acquire iron and polymorphic variants of host NRAMP1 (natural resistance-associated macrophage protein), which influence intramacrophagal flux of iron, can determine susceptibility to intracellular pathogens. The host macrophage uses iron to generate free radicals to kill *Mycobacterium tuberculosis* – the competition for essential micronutrients between the host and the invading microbe continues within immune cells. The control of iron flux during malaria may be important in determining the severity of malaria, the severity of postmalarial anemia, and the propensity to bacterial coinfection and may underlie the protective effect of hemoglobin, haptoglobin, and red cell enzyme variants. Oxidant stress accompanying the hemolysis of malaria is driven by free hemoglobin and is detrimental to both invading parasite and red blood cell membrane. The intraerythrocytic parasite degrades hemoglobin within its food vacuole and controls the resultant generation of free radicals by polymerizing free hemoglobin to hemozoin. Antimalarials, e.g., chloroquine, prevent this process and the parasite succumbs to its own waste. Hemoglobin, haptoglobin, or red cell enzyme variants that offer a more pro-oxidant environment can offer protection from severe malaria by ensuring earlier immune destruction of parasitized red cells. Strategies to manipulate hemolysis, oxidant stress, and iron flux during malarial episodes are key to the intraerythrocytic battle between host and parasite.

Correction of micronutrient deficiencies associated with defined functional consequences is a worthy goal. Iron and zinc deficiency are commonplace, particularly in children in developing countries, and have a significant effect on public health. Supplementation trials of zinc have been associated with significantly reduced infectious disease morbidity and mortality and there is good rationale for using targeted zinc supplementation to reduce infectious disease morbidity. Iron supplementation would not be advocated solely on the basis of its effect on infectious disease morbidity but this effect should be considered in supplementation programs. A recent meta-analysis showed

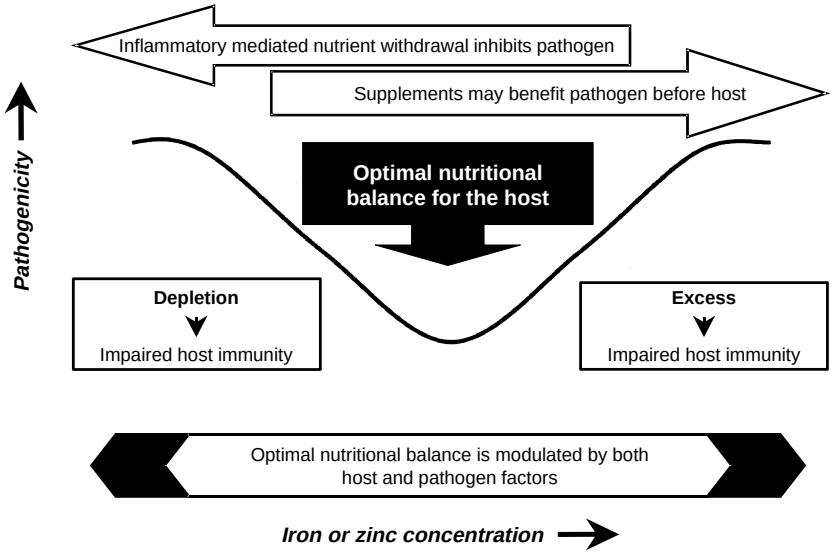


Figure 1 Micronutrient flux and the immune response to infection. (Adapted from Doherty CD, Weaver LT, and Prentice AM (2002) Micronutrient supplementation and infection: a double edged sword? *Journal of Pediatric Gastroenterology and Nutrition* **34**: 346–352.)

no harmful effect of iron supplementation on overall infectious disease incidence. However, if micronutrient withdrawal is a deliberate immune defense strategy and the control of micronutrient flux is worthy of such genomic investment then interference with blanket micronutrient supplementation will likely have adverse effects for subgroups within those populations. Understanding host variability in response to supplementation, the effect of supplementation during infection and nutrient–gene interactions in both host and potential pathogen is key to identifying these subgroups and improving micronutrient targeting.

The objective of supplementation, dose, route, pre-existing level of deficiency, immunocompetence, coexistent deficiencies, genetic determinants, and the presence of infection should all be considered in the decision of who to supplement and when.

See also: **Anemia**: Iron-Deficiency Anemia. **Bioavailability**. **Cytokines**. **Food Fortification**: Developed Countries; Developing Countries. **Infection**: Nutritional Interactions. **Iron**. **Supplementation**: Dietary Supplements; Role of Micronutrient Supplementation; Developing Countries; Developed Countries. **Zinc**: Physiology; Deficiency in Developing Countries, Intervention Studies.

Further Reading

Aggett PJ (1994) Zinc. *Annales Nestle* **52**: 94–106.

Andrews NC (2000) Iron homeostasis:insights from genetics and animal models. *Nature Reviews Genetics* **1**: 208–213.

Black RE (2003) Zinc deficiency, infectious disease and mortality in the developing world. *Journal of Nutrition* **133**: 1485S–1489S.

Destro Bisol G (1999) Genetic resistance to malaria, oxidative stress, and hemoglobin oxidation. *Parasitologia* **41**: 203–204.

Gera T (2002) Effect of iron supplementation on incidence of infectious illness in children: systematic review. *British Medical Journal* **325**: 1142–1151.

Ibs KH (2003) Zinc-altered immune function. *Journal of Nutrition* **133**: 1452S–1456S.

Oppenheimer SJ (2001) Iron and its relation to immunity and infectious disease. *Journal of Nutrition* **131**: 616S–635S.

Rink L (2000) Zinc and the immune system. *Proceedings of the Nutrition Society* **59**: 541–552.

Sherwood RA (1998) Iron homeostasis and the assessment of iron status. *Annals of Clinical Biochemistry* **35**: 693–708.

Weinberg E (2000) Modulation of intramacrophagal iron metabolism during microbial cell invasion. *Microbes and Infection* **2**: 85–89.

Weiss G (1995) Linkage of cell-mediated immunity to iron metabolism. *Immunology Today* **16**: 495–499.

Weiss G (2002) Iron and immunity: a double-edged sword. *European Journal of Clinical Investigation* **32**(supplement 1): 70–78.

Wyllie S (2002) The natural resistance-associated macrophage protein S1c11a1 (formerly Nramp1) and iron metabolism in macrophages. *Microbes and Infection* **4**: 351–359.

INBORN ERRORS OF METABOLISM

Contents

Classification and Biochemical Aspects

D L Marsden, Children's Hospital Boston, Boston, MA, USA

Introduction

Garrod identified the first inborn error of metabolism in 1902 when he described the symptoms that had been observed in patients with alkaptonuria as being due to an inherited enzyme deficiency. Since that time over 400 disorders have been described that are due to an enzyme deficiency in the catabolic pathways of proteins, fatty acids, and carbohydrates. The resulting accumulation of toxic intermediates and, in some cases, the depletion of a necessary end product cause a variety of metabolic derangements, often with significant neurological sequelae. The severity and the age of onset of symptoms usually, although not always, depend on the amount of residual enzyme activity.

The vast majority of these disorders are inherited in an autosomal recessive fashion. While the individual inborn errors of metabolism are rare, based on recent results of expanded newborn screening programs (in which over 30 disorders can be detected), the overall incidence is approximately 1 in 5000 live births

worldwide. The incidence of disorders may vary across populations because of the 'founder effect', where a specific mutation arises and is maintained in subsequent generations, and may be higher where there is a higher incidence of consanguinity.

With a few exceptions, infants are normal at birth because the placenta efficiently eliminates the toxic metabolites.

Newborn Screening

Mass population screening of newborns was introduced in the 1960s, initially for phenylketonuria (PKU), after the development of the bacterial inhibition assay (BIA) for phenylalanine by Robert Guthrie. This simple method, popularly referred to as the Guthrie test, is still the mainstay of screening for PKU in much of the world. Essentially, it entails the addition of a solution of *B. subtilis* to an agar well, to which is added a standardized punched sample from the newborn screening filter paper, from which the blood is then eluted. High levels of phenylalanine inhibit growth of the bacteria, and the laboratory technician can easily visually identify this 'no-growth' zone as abnormal. Quantification is necessary, using a follow-up method such as high-performance liquid chromatography. BIA has been adapted to screen for elevated levels of leucine (for maple syrup urine disease; MSUD) and for methionine (for homocystinuria).

The most significant advance in newborn screening since its inception has been the adaptation of tandem mass spectrometry (MS/MS). With this technology, multiple compounds can be identified (both amino-acids and acylcarnitine species) from the same dried blood filter-paper sample after a simple preparation. Over 30 different inborn errors of metabolism can now be identified. The major drawbacks, however, are the relative expense of the equipment and the lack of long-term outcome data on infants detected and treated presymptomatically. Further modification of MS/MS will enable future screening for many more disorders, for example steroid profiling for congenital adrenal hyperplasia and the identification of lysosomal storage disorders.

Disorders of Protein Metabolism

Amino-Acid Disorders

Amino-acidopathies are due to an enzyme deficiency early in the catabolic pathway of one or more amino-acids that results in the accumulation of the amino-acid(s); they are detected by amino-acid analysis of serum or plasma. Symptoms may be due to the chronic accumulation of toxic amino-acid(s) or due to acute metabolic decompensation, for which aggressive intervention is necessary to prevent death or severe morbidity. Treatment is dietary restriction of the toxic amino-acid by limiting the intake of whole protein and supplementing with special modular amino-acid formulas to provide the appropriate nutrients for normal growth and development. All disorders are inherited in an autosomal recessive fashion.

The classic example is PKU. In PKU, a deficiency of the phenylalanine hydroxylase (PAH) enzyme (**Figure 1**) results in a high level of phenylalanine, which, if not treated with dietary restriction of phenylalanine in the early newborn period, causes severe irreversible mental retardation. The diagnosis is confirmed by a phenylalanine level of more than $1200\,\mu mol\,L^{-1}$ in an infant on unrestricted protein intake. The incidence of PKU is approximately 1 in 20 000 in Caucasians. Although PKU is pan-ethnic, the incidence varies in certain populations. The level of phenylalanine can vary between individual patients because of variations in the amount of residual enzyme activity, which, in turn, depends on the specific mutations. There are currently over 400 known mutations. Most patients are compound heterozygotes (i.e., have one copy each of two different mutations). Prior to the introduction of newborn screening, PKU was the commonest cause of inherited mental retardation. Early recognition of presymptomatic infants allows for the institution of a phenylalanine-restricted diet, with the

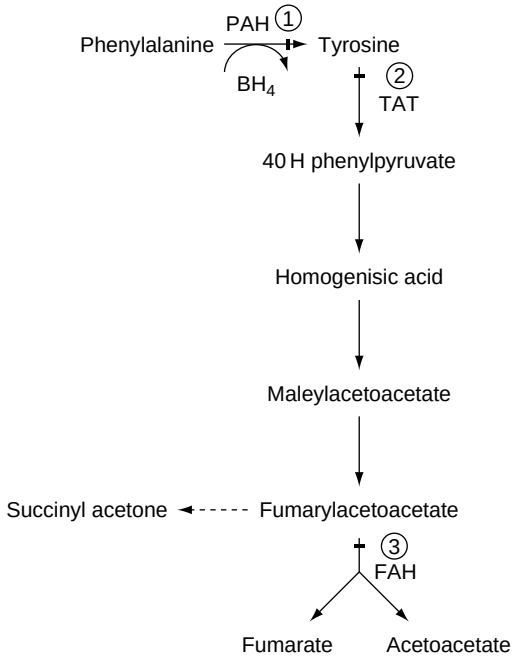


Figure 1 Catabolic pathway of phenylalanine. PAH, phenylalanine hydroxylase; TAT, tyrosine aminotransferase; FAH, fumarylacetoacetase; 1, PKU; 2, tyrosinaemia type II; 3, tyrosinaemia type I.

best outcomes achieved when recommended phenylalanine levels are attained by 2 weeks of age.

Untreated patients develop progressive severe mental retardation, often with seizures and Parkinson-disease-like neurological symptoms. The primary pathogenesis is due to the toxic effect of phenylalanine on the central nervous system; secondary symptoms may be due to a deficiency of tyrosine, which is an important precursor for the synthesis of some neurotransmitters. These symptoms include anxiety and depression.

Benign or mild hyperphenylalaninemia is due to allelic variants of PAH that result in greater residual enzyme activity. On an unrestricted diet, levels are typically in the range 120–360 $\mu mol\,L^{-1}$, and no dietary treatment is necessary.

Moderate elevation of phenylalanine is also present in patients with defects of tetrahydrobiopterin (BH_4), the cofactor for PAH. BH_4 is also the cofactor for other enzymes, tryptophan hydroxylase and tyrosine hydroxylase. These amino-acids are important precursors of the neurotransmitters 5-hydroxytryptophan and dopamine. A deficiency causes a neurological syndrome characterized by hypotonia, seizures, and movement disorder (dystonia).

MSUD has an incidence of approximately 1 in 185 000 births. It is due to a deficiency of the branched chain ketoacid dehydrogenase enzyme and the resulting accumulation of the branched chain amino-acids (BCAAs) leucine, isoleucine, and valine, which are detected by plasma amino-acid analysis. Elevation of alloisoleucine (a derivative of isoleucine) is pathognomonic. In classic MSUD, symptoms typically occur in the first week of life and, if untreated, rapidly progress to cerebral oedema, coma, and death. Toxicity is due primarily to high levels of leucine. The characteristic maple syrup (or burnt sugar) odour is due to the presence of sotolone, a metabolite of isoleucine or alloisoleucine. It is detectable only when the BCAAs are significantly elevated; the ester is concentrated in the urine and the earwax of affected patients.

Variant forms of MSUD also occur. Intermediate MSUD typically presents in infancy with developmental delay; seizures may occur. Moderate levels of the BCAAs (including alloisoleucine) are present. Intermittent MSUD is associated with intermittent symptoms during acute infections or periods of prolonged fasting. Typical symptoms include ataxia, vomiting, and seizures. Acute severe decompensation may occur, similar to the classic form of MSUD. The BCAAs are elevated only during the episode of acute symptoms. Other disorders are listed in **Table 1**.

Urea cycle defects are due to enzyme deficiencies associated with the elimination of waste nitrogen produced by the normal catabolism of protein. There are six enzymatic steps involved in this process (**Figure 2**): a deficiency in any of the first five enzymes causes accumulation of nitrogen, in the form of ammonia (NH_3), and increased levels of the amino-acids glutamine and glycine.

Symptoms typically occur in the newborn period, except in the case of arginase deficiency, but milder late-onset variants have been well described. Symptoms include lethargy, poor feeding, vomiting, tachypnea, and progressive encephalopathy. Routine biochemical testing shows respiratory alkalosis and hyperammonemia. The liver transaminases are usually elevated. Hypoglycemia is not typical.

Plasma amino-acid and urine organic acid analyses are necessary to make a presumptive diagnosis.

Table 1 Disorders of amino-acid metabolism

Disorder (Deficient enzyme)	*Elevated analyte*	*Clinical features*	*Treatment*
Tyrosinemia type I (fumarylacetoacetase)	Tyrosine SA	Cirrhosis Liver failure Failure to thrive Renal tubular acidosis Rickets Hepatocellular Carcinoma (late)	NTBC (inhibits SA production) Tyrosine restriction
Tyrosinemia type II (tyrosine aminotransferase)	Tyrosine (↑↑)	Keratoconjunctivitis Palmar keratosis Mental retardation	Tyrosine restriction
Homocystinuria (cystathionine β synthase)	Methionine Total homocysteine Free homocystine+ Mixed disulfides	Mental retardation Thromboembolism Lens dislocation Osteoporosis	Vitamin B_6 (50% respond) Methionine restriction
Nonketotic hyperglycinemia (glycine cleavage enzyme deficiency)	Glycine (↑↑) (plasma and CSF)	Seizures Developmental delay	Sodium benzoate (decreases glycine)

CSF, cerebrospinal Fluid; SA, Succinylacetone; NTBC, 2-(2-nitro-4-trifluoro-methylbenzoyl)-1,3-cyclohexanedione.

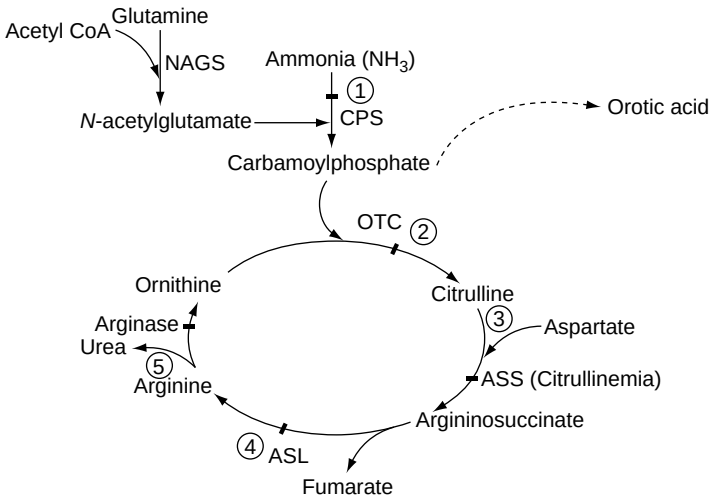


Figure 2 The urea cycle. NAGS, *N*-acetylglutamine synthetase; CPS, carbamoylphosphate synthetase; OTC, ornithine transcarbamoylase; ASS, argininosuccinate synthetase; ASL, argininosuccinate lyase; 1, CPS deficiency; 2, OTC deficiency; 3 citrullinemia; 4, argininosuccinic aciduria; 5, arginase deficiency.

In argininosuccinate synthetase (ASS) deficiency citrulline is elevated, in argininosuccinate lyase (ASL) deficiency argininosuccinic acid and citrulline are elevated, and in arginase deficiency, arginine is elevated. In ornithine transcarbamoylase (OTC) deficiency, on the other hand, citrulline is very low or absent. In each of these disorders, orotic acid is also present (found on urine organic acid analysis); it is produced via the pyrimidine cycle from the excessive carbamoylphosphate that accumulates owing to each enzyme defect. In carbamoylphosphate synthetase (CPS) deficiency, however, the amino-acid and orotic acid levels are normal, so the diagnosis is essentially one of exclusion in a patient who presents with the typical symptoms and severe hyperammonemia in which no other cause is determined. Confirmation of the diagnosis requires a liver biopsy for enzyme analysis of CPS and OTC. Skin fibroblasts can be assayed for ASS and ASL deficiencies, and red blood cells can be assayed for arginase deficiency; mutation analysis may be possible in some cases.

OTC deficiency is the most common urea-cycle defect. It is inherited in an X-linked fashion (all other disorders are autosomal recessive); symptomatic females can present with variable symptoms ranging from acute hyperammonemia to recurrent episodes of nausea, vomiting, and headache. The severity of the symptoms in female patients depends on the degree of lyonization (the normal random inactivation of one of the X chromosomes) and the resultant residual enzyme activity. Some women may remain asymptomatic, and a diagnosis is made only after the birth of a symptomatic son.

Patients with ASL deficiency also have progressive cirrhosis of the liver, possibly owing to the direct toxic effect of the argininosuccinic acid. In arginase deficiency, hyperammonemia is rare (most of the urea has already been eliminated), but arginine itself is toxic to the central nervous system, causing progressive spastic quadriplegia and developmental delay; seizures are common.

The toxicity of these disorders is primarily due to the accumulation of ammonia (NH_3) and glutamine, which is increased because of the transfer of excess ammonium ions (transamination). Acute severe hyperammonemia in the newborn period is catastrophic and often fatal. Survivors have variable neurological deficits.

Acute treatment of hyperammonemia due to a urea cycle defect involves the elimination of dietary protein, elimination of ammonia (by hemodialysis or peritoneal dialysis), administration of a high concentration of dextrose to reverse catabolism, arginine (except in arginase deficiency) to regenerate the cycle, and the nitrogen scavenging drugs sodium benzoate (which conjugates with glycine to form hippurate) and sodium phenylacetate (which conjugates with glutamine to form sodium phenylacetylglutamine). Early reintroduction of limited dietary protein is necessary to provide a substrate for anabolism and prevent further catabolism. This should consist of whole protein and a special formula to provide enough essential amino-acids to ensure normal weight gain, without producing excessive amounts of nitrogen for ammonia production. Chronic treatment involves similar dietary protein restriction, arginine, and an oral

Table 2 Organic acidemias

Disorder (Deficient enzyme)	*Elevated analyte(s)*	*Clinical features*	*Treatment*
Methylmalonic acidemia (methylmalonyl CoA Mutase)	Methylmalonic acid	Metabolic acidosis Hyperammonemia Failure to thrive Vomiting	Protein restriction Carnitine
Isovaleric acidemia (isovaleryl CoA dehydrogenase)	Isovalerylglycine	Metabolic acidosis Vomiting	Protein restriction Glycine
Glutaric aciduria type I (glutaryl CoA dehydrogenase)	3-Hydroxyglutaric acid Glutaric acid	Metabolic acidosis Vomiting Macrocephaly Developmental delay	Protein restriction Carnitine
3-Methylcrotonyl glycinuria (3-methylcrotonyl CoA carboxylase)	3-Hydroxyisovaleric acid 3-Methylcrotonylglycine	Metabolic acidosis Hypoglycemia Hyperammonemia Seizures (Some patients asymptomatic)	Protein restriction Carnitine
Mitochondrial acetoacetyl CoA thiolase deficiency	2-Methyl-3-hydroxybutyrate acid 2-Methylacetoacetic acid Tiglylglycine	Metabolic acidosis Vomiting	Protein restriction Carnitine

form of nitrogen-scavenging medication (sodium phenylbutyrate). For arginase deficiency, dietary protein restriction and formula is usually adequate.

Organic acidemias (**Table 2**) are due to enzyme deficiencies further along the catabolic pathway, usually of several amino-acids, resulting in the accumulation of the toxic products of intermediary metabolism (organic acids). In some cases, there is a functional defect of the enzyme owing to a deficiency of the enzyme cofactor, rather than of the enzyme itself. Examples of this are biotinidase deficiency and defects of cobalamin (vitamin B_{12}) metabolism.

The accumulation of large amounts of organic acids causes severe metabolic acidosis and ketosis. Hyperammonemia is often present, owing to secondary inhibition of the urea cycle. Hypoglycemia may be variably present, owing to secondary inhibition of fatty-acid oxidation. Symptoms are often present in the newborn period; recurrent episodes of metabolic decompensation can occur because of excessive protein intake or because of catabolism (and therefore an increased load of amino-acids endogenously released from muscle) associated with acute infections or prolonged periods of fasting. Morbidity and mortality is due to acute acidosis and the associated neurologic sequelae.

The diagnosis is made by finding high levels of the characteristic organic acids in the urine. Newer analytic methods, such as MS/MS, can detect even small elevations of characteristic plasma acylcarnitine and urine acylglycine conjugates of the intermediary metabolites. Confirmation is by enzyme analysis, usually in skin fibroblasts; DNA mutation analysis is available for many disorders.

Propionic acidaemia is a typical organic acidaemia. It is due to an isolated defect of the enzyme propionyl CoA carboxylase in the catabolic pathways of the amino-acids isoleucine, valine, methionine, and threonine as well as cholesterol and odd chain fatty acids (**Figure 3**). The resulting accumulation of the intermediary metabolites 3-hydroxypropionic acid, methylcitric acid, propionylglycine, and tiglylglycine can cause severe metabolic acidosis, ketosis, coma, and death. Other associated symptoms can include hyperammonemia, hypoglycemia, and pancytopenia, owing to bone marrow suppression by the accumulated toxic organic acids.

Symptoms can occur within days of birth in the classic disease or later in infancy or childhood in the milder variant forms. The later-onset form may be associated with persistent vomiting, failure-to-thrive,

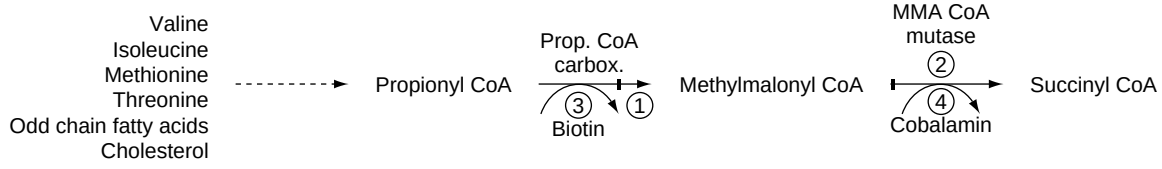


Figure 3 The catabolic pathways of isoleucine, valine, methionine, threonine, cholesterol, and odd chain fatty acids. 1, propionic aciduria; 2, methylmalonic aciduria; 3, multiple carboxylase deficiency; 4, cobalamin disorders.

and developmental delay, but often does not involve severe episodes of metabolic acidosis. Dystonia may occur owing to infarction of the basal ganglia.

Cofactor Deficiencies

Biotin is an essential cofactor for the four carboxylase enzymes propionyl CoA carboxylase, methylcrotonyl CoA carboxylase, pyruvate CoA carboxylase, and acetyl CoA carboxylase. It is endogenously derived from lysine and also present in its protein-bound form in small amounts in many foods. Holocarboxylase synthetase (HCS), which forms the inactive parent apoenzyme, is also biotin dependent. Enzyme activation requires free biotin, which is released by the action of biotinidase; this enzyme also plays an essential role in the recycling of biotin for further use. A deficiency of biotinidase, therefore, results in depletion of biotin and a functional defect of the carboxylases. Symptoms include hypotonia, lethargy, vomiting, and ataxia. Recurrent metabolic acidosis may occur. Alopecia and a generalized erythematous rash are common. The symptoms are more severe in HCS deficiency. The characteristic pattern of organic acids is present in both disorders. The diagnosis is made by measuring biotinidase activity in plasma or carboxylase-enzyme activity in leucocytes or fibroblasts. Treatment with pharmacologic doses of biotin is effective.

Multiple defects of cobalamin (vitamin B_{12}) metabolism can occur, involving the transport of vitamin B_{12} into the cell (defects of the transporter proteins transcobalamin I and II) or subsequent intracellular utilization of the different biologically active forms. These disorders are classified into complementation groups, depending on whether the defect is in adenosylcobalamin (Cbl A and B), methylcobalamin (Cbl G and E), or both (Cbl C and D).

Adenosylcobalamin is the cofactor for methylmalonyl CoA mutase; a defect results in a milder form of methylmalonic acidaemia than that found with a defect of the enzyme itself. Methylcobalamin is the cofactor for methionine synthase; a defect results in low methionine and homocystinuria (distinct from classic homocystinuria due to a defect of cystathionine β synthase). A defect of both adenosylcobalamin and methylcobalamin causes both methylmalonic acidaemia and homocystinuria.

Symptoms vary with the complementation group, but can include metabolic acidosis, hypotonia, developmental delay, macular degeneration, and megaloblastic anemia. Treatment with hydroxocobalamin corrects some of the biochemical derangements, especially in Cbl A and B. Treatment is less successful in the other groups.

A syndrome similar to Cbl C has been described in the breast feeding infants of strict vegetarian (vegan) mothers and in mothers with pernicious anemia, who are vitamin B_{12} deficient.

Disorders of Fatty-Acid Oxidation

Disorders of fatty-acid oxidation have been recognized only since the early 1980s, but as a group they represent the most common inborn errors of metabolism. Fat provides a significant source of energy in the form of glucose and ketone bodies during times of metabolic stress (such as febrile illness) or during prolonged fasting. Free fatty acids, released from the adipose tissue, are transported into the mitochondria via the carnitine shuttle system, where they undergo β-oxidation (**Figure 4**), the progressive cleavage from an 18-carbon very long-chain fatty acid to the two-carbon aceto-acetyl CoA, the substrate for glucose (via the TCA (Tricarboxylic acid) cycle) and ketones. A deficiency of any of the enzymes in this pathway can cause symptoms of hypoketotic hypoglycemia and hepatic encephalopathy, with hyperammonemia (due to secondary inhibition of the urea cycle) and sudden death. Many cases of what would previously have been diagnosed as Reye syndrome are now known to be due to defects of fatty-acid oxidation. Symptoms can occur at any time, from the newborn period to adulthood.

Carnitine has a dual role: in addition to its critical role in the transport of free fatty acids into mitochondria, it conjugates with the fatty acyl CoA intermediates that accumulate proximal to an enzyme block, forming acylcarnitine species that can be excreted by the kidneys. They can also be measured in plasma for diagnostic purposes and in the newborn-screening dried blood spot. Increased use of carnitine owing to an enzyme defect causes a secondary depletion, further impairing fatty-acid oxidation.

Long-chain fatty-acid (carnitine palmityl transferase (CPT) oxidation defects I and II, very long-chain acyl CoA dehydrogenase (VLCAD), TFP (Trifunctional protein), and long-chain 3-hydroxy acyl CoA dehydrogenase (LCHAD)) may present in the newborn period or later in infancy with severe hypoketotic hypoglycemia, cardiomyopathy, and hepatic encephalopathy, due to deposition of fat in the heart and liver. Rhabdomyolysis (lysis of muscle cells) is common. Pigmentary degeneration of the retina may be present in LCHAD and is thought to be due to impaired endogenous production of docosahexanoic acid (DHA), which is necessary for normal retinal function. Milder variant forms of CPT II and VLCAD may present in adolescence or adulthood with muscle cramping and rhabdomyolysis, which may be severe enough to cause

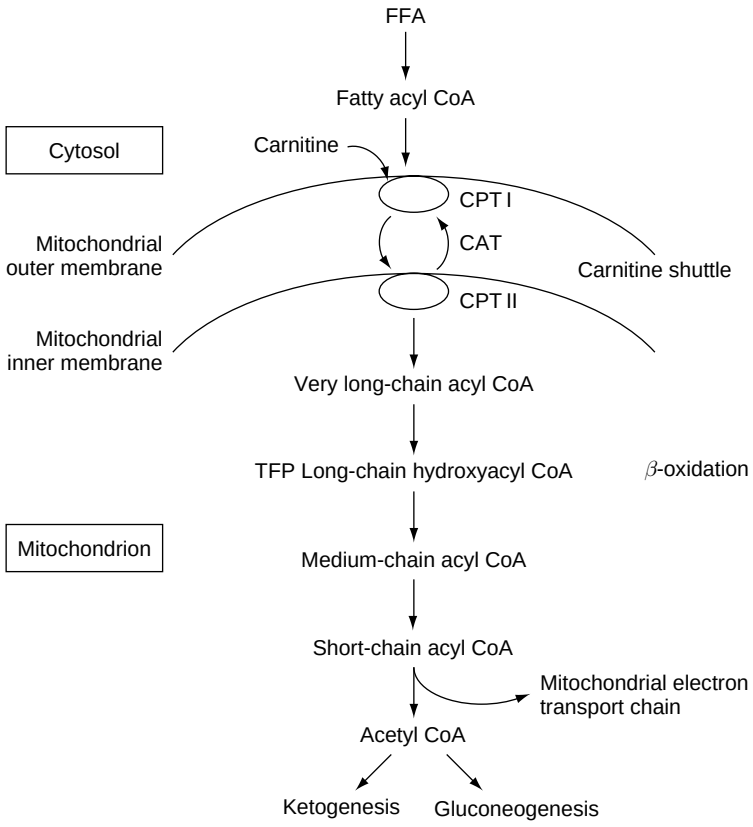


Figure 4 Fatty-acid oxidation. FFA, free fatty acids; CPT I, carnitine palmitoyltransferase type I; CPT II, carnitine palmitoyltransferase type II; CAT, carnitine acylcarnitine translocase; TFP, trifunctional protein.

acute renal failure owing to the deposition of the muscle pigment myoglobin in the renal tubules.

Treatment of these disorders, which can reverse the cardiac and liver disease, includes frequent feeding and avoidance of fasting, together with limitation of dietary fat and supplementation with medium-chain triglycerides (MCT), which bypass the metabolic block. There is no clear consensus on the amount of MCT needed; the general recommendation is to provide 20–40% of total calories from fat, with about half of these calories coming from MCT. Special formulas can provide the MCT requirements, but some are deficient in some essential fatty acids, such as linoleic and linolenic acids and DHA. Addition of oil, such as canola oil, provides most of the essential fatty acids. DHA is not currently commercially available, but fish oil may provide an alternative source. Uncooked cornstarch can provide an alternative source of complex carbohydrate (especially for overnight fasting) after the age of about 9 months. Normal pancreatic amylase activity is necessary and may not be adequate prior to this age.

Treatment of the medium-chain and short-chain defects is simpler, involving avoidance of fasting and early intervention during acute illness to prevent hypoglycemia. Carnitine supplementation is frequently used to prevent secondary depletion. The dietary-fat recommendation is approximately 30% of total calories, or a 'heart healthy' diet.

Disorders of Carbohydrate Metabolism

Galacatosemia

Galactose is derived primarily from dietary lactose, which is the major disaccharide in dairy products, human breast milk, and many fruits and vegetables. There is also a small contribution from endogenous production. There are three known enzyme deficiencies in the pathway that oxidizes galactose to glucose (**Figure 5**); all are autosomal recessive genetic disorders.

Classic galactosemia is due to the almost complete absence of galactose-1-phosphate uridyltransferase (GALT) activity. Symptoms generally occur in the first few weeks of life, with poor weight gain, lethargy, hypotonia, liver disease (hyperbilirubinemia, coagulopathy, and hepatomegaly), and renal tubular acidosis. Hypoglycemia can occur. *Escherichia coli* sepsis may also be a complication: elevated

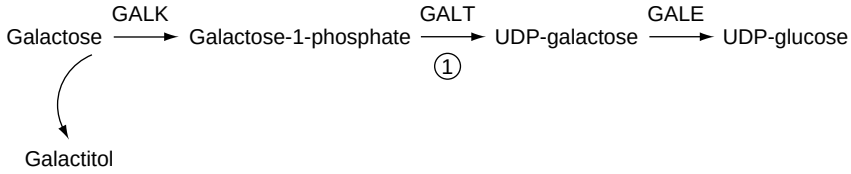


Figure 5 Galactose oxidation. GALK, galactokinase; GALT, galactose-1-phosphate uridyl transferase; GALE, uridine diphosphate galactose 4-epimerase; 1, galactosemia; UDP, uridine diphosphate.

galactose is thought to impair leucocyte bactericidal activity, allowing the bacteria to invade the red blood cells more easily with subsequent dissemination. Mental retardation is a long-term complication.

The underlying pathogenesis of galactosemia is not fully understood; despite compliance with a lactose-restricted diet, speech delay is almost universal, some patients have learning disorders, and female patients have ovarian failure.

Treatment is by restriction of lactose in the diet, primarily by eliminating dairy products and other foods known to be high in galactose.

Variant forms of galactosemia occur owing to mutations in the GALT gene that result in greater residual enzyme activity. The commonest variant is the Duarte variant, in which there is usually one copy of a classic galactosemia mutation (e.g., Q188R) and one copy of the variant N314D. This combination results in approximately 25% residual enzyme activity. There is varying opinion as to whether or not dietary treatment is necessary: some clinicians consider that the residual enzyme activity is sufficient to prevent the pathologic sequelae, others elect to treat the patient with lactose restriction for the first year of life. There are no long-term outcome data to support either approach.

Galactokinase deficiency causes an excessive accumulation of galactitol, which is oxidized from galactose by an alternative pathway. High levels of galactitol cause cataract formation, which is the only symptom of this disorder. Lactose restriction is necessary.

Epimerase deficiency is very rare. There are two isoforms of the enzyme, one isolated to red blood cells and one in the liver. The most common disorder is due to an isolated deficiency of the red-blood-cell isoform, which will be detected incidentally by newborn screening programs that measure total galactose. There are no clinical symptoms and no treatment is necessary. A defect of both isoforms will cause symptoms similar to those of classic galactosemia and should be treated similarly.

Glycogen Storage Disorders

Glycogen is a complex carbohydrate stored primarily in the liver and muscle. Liver glycogen provides glucose to maintain blood-sugar levels between normal feeding; defects of the liver enzymes for glycogen degradation lead to hypoglycemia and/or liver disease because of excessive accumulation of glycogen. Muscle glycogen is an important substrate for energy production for normal muscle function, so disorders are usually indicated by cramping with exercise.

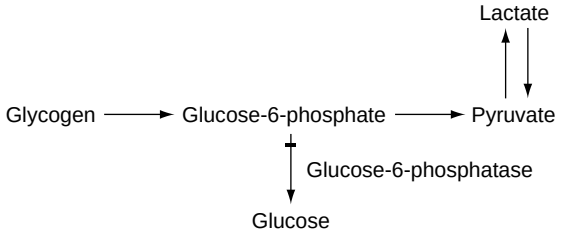


Figure 6 Glycogen storage disease type I.

Glycogen storage disease (GSD) type I (GSD I) (**Figure 6**), the most common disorder, is due to a deficiency of glucose-1-phosphatase in the liver, kidney, and intestinal mucosa. Symptoms typically occur in infancy when the frequency of feeding decreases. Profound hypoglycemia can occur; progressive hepatomegaly and liver dysfunction are due to storage of glycogen. Other metabolic derangements include lactic acidaemia, which is due to increased pyruvate production; increased fatty-acid synthesis causes hypertriglyceridemia and hypercholesterolemia (causing xanthomas); hyperuricemia (causing gout and renal calculi) is due to decreased renal excretion (lactate is preferentially excreted) and increased uric-acid production owing to phosphate depletion. Other long-term complications include progressive renal disease (proteinuria) and hepatocellular carcinoma. Treatment involves frequent meals and continuous nocturnal feeding (in infants); supplemental uncooked cornstarch provides exogenous glucose.

Other GSDs are summarized in **Table 3**.

Disorders of Fructose Metabolism

There are three disorders of fructose metabolism, all inherited in an autosomal recessive fashion. Fructose is widely distributed in the diet as the primary sugar in fruits, vegetables, and honey. It is also derived from sucrose and sorbitol, which are found in large variety of products, including infant formulas and intravenous fluids. The toxic effect of fructose is due to inhibition

Table 3 Glycogen storage disorders

Disorder	*Deficient enzyme*	*Primary affected tissue*	*Symptoms*	*Treatment*
GSD O	Glycogen synthase	Liver	Hypoglycemia	Uncooked cornstarch, frequent feeds
GSD I	Glucose-6-phosphatase	Liver, muscle	Hypoglycemia, hepatomegaly, growth retardation, proteinuria, lactic acidemia, hyperlipidemia, hyperuricemia (gout), hepatocellular carcinoma	Uncooked cornstarch, frequent feeds
GSD II (Pompe disease)	Acid maltase (α glucosidase)	Lysosomes of muscle (skeletal and cardiac)	Cardiomyopathy, skeletal myopathy, cardiorespiratory failure	Enzyme replacement (in clinical trial)
GSD III	Debranching enzyme (amylo-1, 6-glucosidase)	Liver, muscle	Hypoglycemia (mild), hepatomegaly, myopathy, hyperlipidemia	Uncooked cornstarch, frequent feeds
GSD IV (amylopectinosis)	Branching enzyme	Liver	Hepatomegaly, cirrhosis, liver failure, myopathy	Liver transplant, Uncooked cornstarch
GSD V (McArdle disease)	Myophosphorylase	Muscle	Muscle cramping (with exercise)	Oral glucose, high-protein diet
GSD VI (Hers disease)	Liver phosphorylase	Liver	Hepatomegaly, hypoglycemia, myopathy	Frequent feeds
GSD VII (Tarui disease)	Phosphofructokinase	Muscle	Fatigue exercise intolerance, cramping	Avoidance of strenuous exercise
GSD IX	Phosphorylase kinase	Liver, muscle	Hepatomegaly, growth retardation	Frequent feeds

GSD, glycogen storage disorder.

of gluconeogenesis by high levels of fructose-1-phosphate and subsequent depletion of inorganic phosphate and, thus, adenosine triphosphate.

Essential fructosuria is a benign disorder due to a defect of the enzyme fructokinase. Patients have increased urinary excretion of fructose, which is usually an incidental finding on routine testing for reducing substances.

Hereditary fructose intolerance is due to a deficiency in aldolase B, which splits fructose-1-phosphate into glyceraldehyde and dihydroxyacetone. Symptoms occur only after exposure to fructose, usually from dietary ingestion although they are more severe after intravenous infusion. The symptoms include gastrointestinal discomfort, vomiting, and hypoglycemia. Chronic exposure causes failure-to-thrive, liver disease, and renal tubular acidosis. Affected patients are often misdiagnosed as having behavioral problems or an eating disorder. Treatment involves elimination of fructose from the diet.

Fructose-1,6-bisphosphatase deficiency is a defect of gluconeogenesis and not dependent on exposure to fructose. Symptoms, including recurrent episodes of vomiting, lactic acidosis, tachypnea, seizures, and apnea, occur when dietary glucose and glycogen stores are depleted, for example during periods of fasting or febrile illness. Approximately 50% of patients are symptomatic in the newborn period. Treatment involves the prevention of fasting and supplementation with uncooked cornstarch to provide a source of complex carbohydrate. Acute episodes respond to intravenous infusions of dextrose.

Abbreviations

ASL	Argininosuccinate lyase
ASS	Argininosuccinate synthetase
BCAA	Branched chain amino-acids
BH_4	Tetrahydrobiopterin
BIA	Bacterial inhibition assay
CPS	Carbamoyl synthetase
CPT	Carnitine palmityl transferase
GALT	Galactose-1-phosphate uridyl transferase
GSD	Glycogen storage disorder
HCS	Holocarboxylase synthetase
LCHAD	Long-chain 3-hydroxy acyl CoA dehydrogenase
MCAD	Medium-chain acyl CoA dehydrogenase
MS/MS	Tandem mass spectrometry
MSUD	Maple syrup urine disease
OTC	Ornithine transcarbamylase

PAH	Phenylalanine hydroxylase
PKU	Phenylketonuria
SCAD	Short-chain acyl CoA dehydrogenase
VLCAD	Very long-chain acyl CoA dehydrogenase

See also: **Inborn Errors of Metabolism**: Nutritional Management of Phenylketonuria.

Further Reading

Burton BK (1998) Inborn errors of metabolism in infancy: a guide to diagnosis. *Pediatrics* **102**: E69.

Chace DH, Kalas TA, and Naylor EW (2003) Use of tandem mass spectrometry for multianalyte screening of dried blood specimens from newborns. *Clinical Chemistry* **49**: 1797–1817.

Fernandes J, Saudubray J-M, and van den Berghe G (eds.) (2000) *Inborn Metabolic Diseases*, 3rd edn. Berlin: Springer-Verlag.

Gillingham MB, Connor WE, Matern D *et al.* (2003) Optimal dietary therapy of long-chain-3-hydroxyacyl-CoA dehydrogenase deficiency. *Molecular Genetics and Metabolism* **79**: 114–123.

Holme E and Lindstedt S (1998) Tyrosinemia type I and NTBC (2-(2-nitro-4-trifluoromethylbenzoyl)-1,3-cyclohexanedione). *Journal of Inherited Metabolic Disease* **21**: 507–517.

Kahler SG and Fahey MC (2003) Metabolic disease and mental retardation. *American Journal of Medical Genetics* **117C**: 31–41.

Levy HL, Sepe SJ, Walton DS *et al.* (1978) Galactose-1-phosphate uridyl transferase deficiency due to the Duarte/galactosemia combined variation: clinical and biochemical studies. *Journal of Pediatrics* **92**: 390–393.

Podebrad F, Heil M, Reichert S *et al.* (1999) 4,5-Dimethyl-3-hydroxy-2[5H]-furanone (sotolone)—the odour of maple syrup urine disease. *Journal of Inherited Metabolic Disease* **22**: 107–114.

Vockley J, Singh RH, and Whiteman DA (2002) Diagnosis and management of defects of mitochondrial beta-oxidation. *Current Opinion in Clinical Nutrition and Metabolic Care* **5**: 601–609.

Walter JH, Collins JE, and Leonard JV (1999) Recommendations for the management of galactosemia. UK Galactosemia Steering Group. *Archives of Disease in Childhood* **80**: 93–96.

Zytkovicz TH, Fitzgerald EF, Marsden D *et al.* (2001) Tandem mass spectrometric analysis for amino, organic and fatty acid disorders in newborn dried blood spots: a two-year summary from the new England Newborn Screening Program. *Clinical Chemistry* **47**: 1945–1955.

Nutritional Management of Phenylketonuria

D L Marsden, F J Rohr and K C Costas, Children's Hospital Boston, Boston, MA, USA

Introduction

Phenylketonuria (PKU) is a disorder of amino acid metabolism caused by a deficiency in the enzyme phenylalanine hydroxylase, which converts the essential amino acid phenylalanine to tyrosine. High levels of phenylalanine are toxic to the central nervous system, resulting in severe irreversible mental retardation. Details of the biochemistry are discussed elsewhere in this encyclopedia.

PKU is often considered a paradigm for the nutritional therapy for metabolic disorders. It was the first inborn error of metabolism identified by newborn screening, thus allowing for early dietary treatment. Early treatment was successful in preventing the mental retardation associated with untreated PKU. Since the advent of successful dietary treatment of PKU four decades ago, the field has expanded greatly, but the principle of treating phenylketonuria remains the same – to control the intake of the amino acid that is not metabolized normally. This principle applies to all amino acidopathies, but PKU is used here as an example.

Dietary treatment is started as soon as the diagnosis is confirmed in a newborn. Outcomes are best when the diet is implemented and the phenylalanine levels are within the recommended guidelines by 2 weeks of age. Diet is now recommended to be life-long. Adult and adolescent patients who have resumed an unrestricted diet, while intellectually normal, have been shown to have an increased incidence of neuropsychiatric illness, such as increased anxiety and depression. Others report poor concentration, headaches, and sleep disturbance.

The pathophysiology of PKU is not well understood, although recent focus has been on the role of amino acids in the brain. Phenylalanine competes with other large neutral amino acids for transport across the blood–brain barrier, and it is theorized that high levels of brain phenylalanine and low levels of other amino acids, specifically tyrosine and tryptophan, may impede neurotransmitter synthesis in the brain and be responsible for the symptoms associated with untreated PKU. While the ideal brain level of phenylalanine has not been established, treatment guidelines have been established for blood levels at various ages, although these guidelines differ slightly in different countries. In the US, recommendations have been developed by an expert panel convened under the direction of the National Institutes of Health (NIH) and the American Academy of Pediatrics (**Table 1**).

The goal of nutritional therapy is to keep blood phenylalanine controlled while providing a nutritionally sound diet. This necessitates the use of a special medical food (most often as a formula) that provides amino acids other than phenylalanine. A medical food is required because the phenylalanine restriction needed to maintain blood levels within the desired range is so severe that the amount of natural protein

Table 1 Treatment goals for PKU

Age (years)	*Phenylalanine level (μmol l^{-1})*
0–12	120–360
12–adult	120–900
	(120–600 preferred in adolescents)
Maternal PKU	120–360

Adapted from NIH Consensus Development Conference Statement (2001) Phenylketonuria, screening and management. *Pediatrics* **108**(4): 972–982.

allowed in the diet would not support normal growth and development. Several medical foods are currently available. When PKU was first treated, only one medical food was commercially available: a protein hydrolysate from which most of the phenylalanine had been removed. Now, medical foods for PKU use synthetic L-amino acids (other than phenylalanine) as the protein source and are phenylalanine free. The medical foods vary in the amount of amino acids that they contain; in addition, most also provide carbohydrate and fat, vitamins and minerals, but others do not. The amount of medical food prescribed is intended to meet protein needs at various ages in the life cycle (see **Table 2**).

Introduction of Dietary Therapy

Infant formulas for PKU come in a powdered form and are mixed with water and taken as a substitute for regular infant formula or breast milk. In some clinics, only phenylalanine-free formula is given for a few days so that blood phenylalanine will quickly decrease to an acceptable level. A prescribed amount of breast milk or standard infant formula, however, should be shortly introduced into the diet. Whole protein is needed to meet phenylalanine requirements and prevent phenylalanine deficiency, which will lead to muscle protein catabolism and inadequate weight gain. For formula-fed infants, both standard infant formulas and PKU medical foods are used in prescribed amounts and are bottle fed. Breast-feeding of an infant with PKU is possible and, as with all infants, should be encouraged whenever possible. Mature breast milk contains approximately 46 mg 100 ml^{-1}of phenylalanine compared to approximately 59 mg 100 ml^{-1} in cows' milk protein-based formula and approximately 88 mg 100 ml^{-1} in soy-based formulas. Therefore, breast-fed infants may initially have slightly lower plasma phenylalanine levels. If a mother chooses to continue breast-feeding, she is advised about the proper ratio of breast milk to PKU medical food to feed her infant. The key to either method is frequent monitoring of blood phenylalanine and adjusting the diet based on phenylalanine intake, weight gain, and blood levels. Guidelines for the frequency of monitoring were also recommended by the NIH consensus panel (**Table 3**). The method used for monitoring varies

Table 2 Recommended daily nutrient intakes (ranges) for infants, children, and adults with PKU

Age	*Nutrient*				
	PHE	*TYR*	*Protein*	*Energy*	*Fluid*
Infants	(mg kg^{-1})	(mg kg^{-1})	(g kg^{-1})	(kcal kg^{-1})	(ml kg^{-1})
0 to <3 months	25–70	300–350	3.50–3.00	120 (145–95)	160–135
3 to <6 months	20–45	300–350	3.50–3.00	120 (145–95)	160–130
9 to <12 months	15–35	250–300	3.00–2.50	110 (135–80)	145–125
7 to <9 months	10–35	250–300	3.00–2.50	105 (135–80)	135–120
Girls and boys	(mg day^{-1})	(g day^{-1})	(g day^{-1})	(kcal day^{-1})	(ml day^{-1})
1 to <4 years	200–400	1.72–3.00	≥30	1300 (900–1800)	900–1800
4 to <7 years	210–450	2.25–3.50	≥35	1700 (1300–2300)	1300–2300
7 to <11 years	220–500	2.55–4.00	≥40	2400 (1650–3300)	1650–3300
Women					
11 to <15 years	140–750	3.45–5.00	≥50	2200 (1500–3000)	1500–3000
15 to <19 years	230–700	3.45–5.00	≥55	2100 (1200–3000)	1200–3000
≥19 years	220–700	3.75–5.00	≥60	2100 (1400–2500)	2100–2500
Men					
11 to <15 years	225–900	3.38–5.50	≥55	2700 (2000–3700)	2 000–3700
15 to <19 years	295–1100	4.42–6.50	≥65	2800 (2100–3900)	2100–3900
≥19 years	290–1200	4.35–6.50	≥70	2900 (2000–3300)	2000–3300

PHE, phenylalanine; TYR, tyrosine.

Reproduced with permission from Acosta PB and Yanicelli S (2001) *Nutrition Support Protocols*, 4th edn. Abbott Laboratories. Columbus, OH, USA.

Table 3 Monitoring for PKU

Age (years)	*Frequency of testing for phenylalanine*
0–1	Weekly
1–12	Twice monthly
12–adult	Monthly
Maternal PKU	Twice weekly

Adapted from NIH Consensus Development Conference Statement (2001) Phenylketonuria, screening and management. *Pediatrics* **108**(4): 972–982.

depending on the resources available at individual PKU clinics: either frequent visits to the clinic for blood drawing or filter paper samples (as used for newborn screening) that can be collected at home and then mailed to the clinic or, in some cases, to the newborn screening program, for analysis. Because of the time delay in the latter method, it is more suitable for use after the initial stabilization period.

When an infant with PKU is 4–6 months old, solid food is introduced. Since nearly all food contains some phenylalanine, it must be measured and counted. Lists of the phenylalanine content of foods are available and are essential to diet management. The phenylalanine content of foods is listed in milligrams; in some clinics, an exchange system is used where one exchange is equal to a given amount of phenylalanine (often 15 mg per exchange, but in some cases 20 or 50 mg). Since the total amount of phenylalanine taken daily remains the same (adjusted for weight gain), adjustments are made in the amount of regular infant formula or breast milk given to the infant once solid foods are started. This process continues until all of the phenylalanine requirement is provided as food. In general, infants with PKU begin with fruit and small amounts of infant cereal. As the infant's appetite increases, other foods are added, but the choices are limited to fruits, vegetables and, in some cases, small portions of bread and cereal products. For some individuals with PKU, the phenylalanine restriction is severe enough to preclude any regular grain products. Instead, specialty low-protein foods are available, often through mail order. A whole array of low-protein breads, cereals, crackers, bagels, pasta, cakes, cookies, and even low-protein cheeses and peanut butter are critical to proper diet management. These foods provide much needed variety and calories to the diet. High-protein foods such as meat, fish, poultry, dairy, nuts, eggs, and legumes are not allowed on a PKU diet. Thus, the phenylalanine-free medical food continues to be the main source of protein for life.

A wide variety of medical foods are now available for children and adults with PKU in order to meet different tastes and caloric needs. Some of the medical foods for children, teens, and adults are packaged in pouches or sachets for convenience, and several are available in bar, capsule, or tablet form to promote ease of use. Nevertheless, many individuals with PKU struggle with this aspect of the diet. If the full amount of medical food is not taken, nutritional intake is inadequate and may lead to catabolism of lean body mass, which in turn leads to poor control of blood phenylalanine.

Once established, the amount of dietary phenylalanine an individual is allowed remains the same, except for periods of rapid growth, when more phenylalanine may be necessary. A typical phenylalanine intake for a child with severe PKU is $250\,mg\,day^{-1}$, and for a child with moderate PKU is $400\,mg\,day^{-1}$. Thus, in addition to achieving the correct amount of medical food, the crux of the diet is to provide the prescribed amount of phenylalanine while making the diet taste and appear as appetizing and socially acceptable as possible. Families require a good deal of support in doing this. Internet-based support groups, newsletters, regional networks, family gatherings, as well as camps for children with PKU provide a link for families and a forum for exchange of practical information and emotional support. PKU clinic personnel are another source of support and reliable information on medical advances in treating PKU.

All patients with PKU should have their blood phenylalanine and other amino acids monitored regularly as long as they remain on the diet; they should have regular physical examinations, especially for assessment of growth parameters in children and adolescents, and review of the dietary intake since the previous visit. Extensive dietary counseling is an ongoing process. It is also recommended that adult patients who are not following phenylalanine-restricted diets with prescribed medical foods should be seen at least once a year for nutritional assessment, as they often tend to self limit their protein intake and may have inadequate diets.

Adequacy of Nutritional Therapy

Carefully executed diet therapy for individuals with PKU is widely considered to be safe as well as efficacious in preventing mental and neurological impairment. However, it cannot be assumed that largely synthetic diets supplemented with individual vitamins, minerals, and trace elements will confer the same benefits as diets composed of whole foods. Synthetic diets may have an inherent inability to supply all essential nutrients. In addition, patients who are noncompliant or partially compliant with their intake of medical food are at increased

nutritional risk. Formerly treated patients who are 'off diet' tend to select high-carbohydrate diets and continue their habit of avoiding high-protein foods such as meat, milk, and eggs. Micronutrients previously supplied by the medical food, such as vitamin B_{12}, zinc, and iron, may not be replaced in adequate amounts on such a self-selected diet.

Growth

A strict PKU diet supplies 80–90% of its prescribed protein via a phenylalanine-free medical food. Most of the nitrogen in medical foods is supplied via essential amino acids. Meals that supply most of the protein as L-amino acids result in more rapid absorption and oxidation than observed after consumption of whole-protein meals. L-amino acids also may not be as efficiently absorbed as whole protein. Owing to these reasons, protein requirements for patients with PKU are considered to be greater than those given in the WHO guidelines and recommended daily intakes (RDIs). Normal growth and protein status has been observed in infants consuming at least 3 g protein $kg^{-1}\,day^{-1}$. Long-term inadequate protein intake will result in impaired growth in infants and children, low plasma prealbumin concentrations, radiological bone changes (osteopenia), and reduced phenylalanine tolerance. Because phenylalanine is an essential amino acid, it is crucial to prevent its deficiency. Phenylalanine deficiency will result in catabolism of body protein stores and subsequent elevation of blood phenylalanine levels, anemia, and mental retardation as well as the above symptoms accompanying overall inadequate protein intake.

Fatty Acids

Diet-treated children and adults with PKU consume very small amounts of animal fats, including fish-derived oils and long-chain polyunsaturated fatty acids (LC PUFAs.) In infants, small amounts of cows' milk-based formula are typically used to supply phenylalanine requirements. The majority of fatty acids supplied are typically not longer than 18 carbons long. While many standard infant formulas in Europe and the US are now supplemented with docosohexanoic acid (DHA) and arachadonic acid (ARA), metabolic formulas are not. Fatty acids are a structural component of all cell membranes. Alpha linolenic acid-derived compounds are essential for proper development of the central nervous system and retina. Linoleic acid-derived compounds play a role in promoting normal growth, skin, and reproduction. Breast milk contains formed DHA and ARA, and some studies indicate that breast-fed infants have better visual and cognitive development than unsupplemented formula-fed infants. The diets of children with PKU provide similar energy, higher carbohydrate, and lower lipid (with high unsaturated/saturated ratio) and cholesterol content than controls. Circulating plasma lipid levels of treated PKU patients contain lower concentrations of arachadonic acid (ARA), docosahexanoic acid (DHA), and eicosapentanoic acid than controls. Erythrocyte membranes of patients contain relatively high amounts of ARA and relatively low amounts of DHA. In theory, patients receiving adequate amounts of the essential fatty acid precursors linoleic and alpha linolenic acids would be able to synthesize LC PUFAs via elongation and desaturation reactions. It is unclear whether the amount of LC PUFAs synthesized would be adequate for optimal tissue function. Especially in infants, DHA and ARA may be partially essential nutrients. Trials of LC PUFA supplementation in PKU patients are underway. A number of widely available PKU formulas for older children and adults do not supply fat. Patients prescribed these formulas are presently advised to regularly include good sources of linoleic and alpha-linolenic acids in their diets. Flax, canola, and walnut oils are good sources.

In theory, patients with poorly controlled phenylalanine levels cannot efficiently build reserves of DHA and ARA from precursors. Carnitine-dependent mitochondrial enzymes that also use a cofactor, alpha tocopherolquinone, perform elongation and desaturation reactions. Phenyllactate and phenylpyruvate, metabolic byproducts of phenylalanine, may inhibit the synthesis of the cofactor alpha tocopherolquinone. This process may be at least partly responsible for the mental retardation and microcephaly observed in untreated PKU patients and poorly controlled maternal PKU.

Iron, Zinc, Vitamin A, and selenium

Some diet-treated patients with PKU have exhibited altered iron, zinc, vitamin A, and selenium status. With the exception of selenium, aberrations have been demonstrated even when patients consumed close to or greater than the RDI levels of the vitamin/mineral in question. The mechanisms of these changes are unclear and may be multifactorial. The actual impact of these changes on the health of the individual patients is unknown.

Low serum ferritin but appropriate hemoglobin and mean erythrocyte volumes have been noted, even in patients consuming close to three times the recommended dietary allowance (RDA) for iron. Iron absorption or bioavailability may be inhibited by the presence of calcium and phosphorous salts, diets high in PUFAs, and dietary fiber. The presence of alterations in the PUFA composition of gut cell

membranes could affect iron absorption. In vitamin A-deficient rats, anemia occurred, which was not remedied by the administration of oral iron. This suggests that vitamin A deficiency in PKU patients could result in anemia unresponsive to iron therapy. The iron status of diet-treated patients should be serially monitored.

Low serum zinc has occurred in infants and children receiving greater than or equal to 70% of the RDA for zinc. Low serum zinc occurred more often in patients receiving casein hydrolysates than in patients receiving L-amino acids alone. Serum zinc may not be an accurate marker for assessment of zinc deficiency. Zinc absorption in general is inhibited by a PUFA-rich diet, fiber, phosphorous, and large amounts of iron. Competitive inhibition between calcium and zinc also occurs.

Low plasma retinol levels have been observed in infants and young children despite consumption of up to three times the RDA for vitamin A. Retinol is transported on retinol-binding protein (RBP); zinc is needed for the synthesis of RBP. Prealbumin is a carrier for RBP. RBP and zinc levels have been normal in nearly all patients with low retinol levels. Low prealbumin levels or abnormal release of RBP from prealbumin may be responsible for the low serum retinol levels. In fact, a number of children have low prealbumin levels despite receiving adequate protein and energy intakes.

Until recently, selenium was not routinely added to PKU formulas. In the past it was supplied to patients via contamination of foods grown in selenium-containing soil. Low serum, whole blood, urine, and hair levels of selenium have been observed in some patients with PKU on strict diet therapy. Low activity of the selenium-containing enzyme glutathione peroxidase also occurs. Clinical symptoms of selenium deficiency in the patients studied have not been reported.

Bone mineral density

Osteopenia is prevalent in diet-treated persons with PKU from early life. Reduced bone mineral density and/or bone mass has been detected in up to approximately 50% of patients screened by various methods. These methods have included DEXA (dual energy X-ray absorptiometry), pQCT (peripheral quantitative computed tomography), and SPA (single photon absorptiometry). The defect seems to be characterized by a reduction in the speed of bone mineralization, especially after 8 years of age. Osteoporosis is an important cause of morbidity and mortality in older adults in the general population. Reduction in bone mass increases the risk of fracture. A reduction of one standard deviation in spine bone mass is associated with a bone fracture rate of 2.0–2.5. Some authors have reported an increased fracture rate in children over 8 years of age with PKU.

The pathogenesis of osteopenia in PKU is under study. Discrepant associations have been reported between osteopenia and blood phenylalanine levels, serum vitamin and mineral levels, protein, vitamin and mineral intakes, serum markers of bone formation and PTH, and ratio of urinary minerals, to creatinine. One theory is that impaired mineralization is a direct effect of the lifelong disease process of PKU. The total and the bone-specific fraction of alkaline phosphatase are reduced in some patients. This reduction may affect osteoblast activity and impact bone formation and turnover. High blood phenylalanine levels have not been consistently correlated with osteopenia. High blood levels of phenylalanine and phenylalanine metabolites would result in their increased urinary excretion. Chelating of minerals with phenylalanine and phenylalanine derivatives could theoretically result in significant mineral losses.

Osteopenia may be an accumulated result of lifelong diet treatment or poor diet compliance at vulnerable stages of bone development. Compliant patients tend to have low variation in their lifelong intake of whole protein, as controlled amounts of whole protein are required to maintain good metabolic control. Compliant patients tend to have similar trends in overall intakes. Lack of adequate trace elements, whole protein, vitamins, and/or minerals may be culprits. Impaired absorption of the synthetic diet or the type of medical food used (hydrolysate versus elemental formulation) may exert an independent effect. Inadequate intakes of calcium and phosphorous are known risk factors for the development of osteoporosis in nonaffected persons. Tailoring medical foods to specifically deliver the amounts of calcium and phosphorous recommended in the new RDIs may help to prevent osteopenia.

Trials of calcitriol (1-25 (OH)2 D) supplementation in estrogenic patients with PKU are in progress. Calcitriol has been chosen as most patients already receive expected sun exposure from participating in normal outdoor activities, and their intakes of dietary vitamin D generally meet or exceed the RDA. Clairol has been shown to be a useful treatment; treated patients require close monitoring of urinary calcium excretion and blood calcium levels.

Maternal PKU

For women with PKU who intend to become pregnant, following a strict phenylalanine-restricted

diet and controlling blood phenylalanine to 120–360 mol l^{-1} is critical to offspring health. Women with PKU who have high blood phenylalanine levels are at high risk of having children with microcephaly, mental retardation, low birth weight, and congenital heart anomalies. In an International Study of Maternal PKU, women who had good metabolic control by 10 weeks' gestation had babies with good birth outcomes and development. In women with poor control, the degree of microcephaly and mental retardation was proportional to the level of blood phenylalanine. Congenital heart disease, on the other hand, was not directly related to the degree of metabolic control, suggesting that etiology is multifactorial, although in this study, no serious heart defects occurred when mothers were in good metabolic control by 10 weeks' gestation. The recommendation is for women to be on the diet for PKU and in good metabolic control before conceiving in order to prevent damage to the fetus. Nevertheless, many women come to medical attention during pregnancy, indicating the need for better strategies for keeping women on the diet for life or helping them return to the diet before pregnancy. While blood phenylalanine during pregnancy was the best predictor of outcome in maternal PKU in the Collaborative Study, other nutritional factors, including sufficient energy, protein, vitamin B_{12} and fat, also played an important role.

Alternative Therapies

Tetrahydrobiopterin (BH4)

Tetrahydrobiopterin (BH4) is the cofactor for phenylalanine hydroxylase. Some mutations for PAH are considered to be relatively milder than others; a number of studies have shown that certain of these mutations result in enzyme activity that may be improved by the addition of BH4 to the diet. In these cases, the degree of phenylalanine restriction needed to maintain good control could be liberalized, although not eliminated altogether.

Large Neutral Amino Acid Supplementation

The large neutral amino acids (LNAAs), phenylalanine, tyrosine, tryptophan, and the branched-chain amino acids share the same L-amino acid transport system across the blood–brain barrier. Therefore, high levels of phenylalanine in the blood impede the transport of these other amino acids into the central nervous system (CNS). Tyrosine and tryptophan are important neurotransmitter precursors, relative deficiency or imbalance of which may contribute to the neuropsychiatric symptoms seen in some adult PKU patients who have resumed an unrestricted diet. Treatment with supplemental LNAAs (in tablet form) theoretically will increase the competition with phenylalanine for transport into the CNS. A net reduction in phenylalanine and an increase in CNS tyrosine and tryptophan may result in improvement in symptoms. Long-term outcome data are not yet available. This treatment is not suitable for children or women in the childbearing years who might be contemplating pregnancy.

See also: **Bone**. **Brain and Nervous System**. **Breast Feeding**. **Osteoporosis**. **Selenium**. **Supplementation**: Role of Micronutrient Supplementation. **Vitamin A**: Physiology; Biochemistry and Physiological Role; Deficiency and Interventions. **Vitamin K**. **Zinc**: Deficiency in Developing Countries, Intervention Studies.

Further Reading

Acosta PB (1996) Nutrition studies in treated infants and children with phenylketonuria: vitamins, minerals, and trace elements. *European Journal of Pediatrics* **155**(supplement 1): S136–139.

Acosta PB, Matalon K, Castiglioni L *et al.* (2001) Intake of major nutrients by women in the maternal PKU (MPKU) study and effects on plasma phenylalanine concentrations. *American Journal of Clinical Nutrition* **73**: 792–796.

Al-Qadreh A, Schulpis K, and Athanasopoulou H (1998) Bone mineral status in children with phenylketonuria under treatment. *Acta Pediatrica* **87**(11): 1162–1166.

Giovanni M, Biasucci G, Agostini C *et al.* (1995) Lipid status and fatty acid metabolism in phenylketonuria. *Journal of Inherited Metabolic Diseases* **18**: 265–272.

Koch R, Fishler K, Azen C, Guldberg P, and Guttler F (1997) The relationship of genotype to phenotype in phenylalanine hydroxylase deficiency. *Biochemistry and Molecular Medicine* **60**(2): 92–101.

Koch R, Hanley W, Levy H *et al.* (2003) The Maternal Phenylketonuria International Study: 1984–2002. *Pediatrics* **112**(6 part 2): 1523–1529.

NIH Consensus Development Conference Statement (2001) Phenylketonuria, screening and management. *Pediatrics* **108**(4): 972–982.

Perez-Duenas P, Cambra F, and Vilaseca M (2002) New approach to osteopenia in phenylketonuric patients. *Acta Pediatrica* **91**(8): 899–904.

Przyrembel H and Bremer H (2000) Nutrition, physical growth, and bone density in treated phenylketonuria. *European Journal of Pediatrics* **159**(supplement 2): S129–S135.

INFANTS

Contents
Nutritional Requirements
Feeding Problems

Nutritional Requirements

S A Atkinson, McMaster University, Hamilton, ON, Canada

Optimal nutritional support of infants in the first year of life is essential to attain normal trajectories of growth and development. Additionally, evidence supports the thesis that during critical periods of early development nutrition may be key to 'programming,' possibly through modification of gene expression or differential cell proliferation, that subsequently impacts on risk for chronic diseases in later life. Information on early nutrition programming is not sufficient to be used as a basis to set dietary standards for infants. However, the importance of adopting the quantity and quality of nutrients in human milk as a gold standard in the determination of nutrient recommendations has been reinforced by several agencies worldwide, such as the pioneering partnership between the Food and Nutrition Board of the Institute of Medicine in the United States and Health Canada.

Recommended nutrient intakes or dietary standards are produced by many countries as well as key international agencies such as the FAO/WHO. For infants, the recommended intakes are usually intended for term-born, healthy, and normally growing infants who have a birth weight of more than 2500 g (and thus not small for gestational age). In this article, the nutrient requirements outlined reflect recent reports of the Dietary Reference Intakes (DRIs) for the United States and Canada as published by the Institute of Medicine. For a summary of the range of recommended nutrient intakes for infants that reflect a review of several international reports, the reader is referred to the March of Dimes document, *Nutrition Today Matters Tomorrow* (Appendix D-2).

The key changes in the derivation of the DRIs for infants compared to previous dietary standards from the United States and Canada include adoption of human milk as the reference model for setting recommended nutrient intakes for infants; simplification of age groupings within the first year; no specific provision of dietary recommendations for formula-fed infants; and gender-specific DRI values for fewer nutrients and only where data were available to support such gender specificity. Another major change from previous dietary standards is that upper levels (ULs) for nutrient intake were defined for the first time. Unfortunately, for infants, few ULs were established due to a paucity of pertinent knowledge; even for ages 1–18 years, the UL values were mostly extrapolated from adult values.

This article provides an overview of key concepts and examples of the DRIs specific for infants, future needs for additional research, and practical aspects of meeting the dietary recommendations for infants.

Dietary Reference Intakes for Infants

For infants, evaluation of evidence to establish the DRIs consistently revealed a paucity of appropriate studies on which to base an Estimated Average Requirement (EAR) or UL. A Recommended Dietary Allowance (RDA) could not be calculated if a value for the EAR was not established, in which case the recommended intake was based on an Adequate Intake (AI). The nutrient recommendations for infants from birth through 6 months of age for all nutrients except for energy and vitamin D were set as an AI, a value that represents "the mean intake of a nutrient calculated based on the average concentration of the nutrient in human milk from 2 to 6 months of lactation using consensus values from several reported studies," multiplied by an average volume (0.780 l/day) of human milk. The predicted daily volume of breast milk ingested by an infant was based on observational studies that used test weighing of full-term infants. For infants aged 7–12 months, the AI for many nutrients was based on mean observed nutrient intake from human milk in the second 6 months (0.6 l/day) in addition to published values for intake of nutrients from complementary or weaning foods if such data were available.

Assuming an adequate intake of milk for all infants was considered a valid approach since there is evidence that the volume of milk produced during the early months of lactation is very consistent

among women irrespective of racial, cultural, or nutritional diversity or variations in body size. The volume of milk produced increases with greater size of the infant, when twins are nursing, and in response to greater frequency of nursing.

Using consensus values for the nutrient content of human milk was deemed appropriate since for many nutrients—energy, macronutrients, and macrominerals—maternal diet does not influence the nutrient content of the milk. The exceptions to this include the fatty acid profile, selenium, iodine, and the water-soluble vitamins. Although human milk is known to contain many nonnutrient bioactive factors, such as immune and growth factors and live enzymes, these were not considered to impact on nutrient needs *per se*.

For nutrients for which intake data were not available for ages older than 6 months, the EAR or AI was derived by extrapolation from estimates of intakes from older children or adults using the formula with adjustments for metabolic body size, growth, and variability:

$$\mathrm{EAR}_{\text{infant or child}} = \mathrm{EAR}_{\text{child or adult}} \times (F)$$

where $F = (\text{weight}_{\text{infant or child}}/\text{weight}_{\text{child or adult}})^{0.75}$ (1 + growth factor); or occasionally by extrapolating up from intake of breast-fed infants with similar adjustments using the formula

$$\text{AI } 6-11 \text{ months} = \text{AI } 0-5 \text{ months} \times F$$

where $F = (\text{weight}_{6-11 \text{ months}}/\text{weight}_{0-5 \text{ months}})^{0.75}$

For a few nutrients, such as iron and zinc, sufficient metabolic data were available to derive an EAR using modeling or factorial methods.

Because no specific AIs were derived for formula-fed infants, it is incumbent upon industry to design formulas with a quantity and quality of nutrients that when fed will provide an amount of nutrients that meets the RDA or AI. An approach to establishing the amount of nutrient needed by formula-fed infants is addressed under the section titled "Special Considerations" in each DRI report.

When possible, a DRI called the tolerable upper level was defined as "the highest level of daily nutrient intake that is likely to pose no risk of adverse health effects for almost all individuals in the general population" (Institute of Medicine, 2002). Chronic consumption of nutrients above the UL increases the potential risk of adverse effects, the latter varying by nutrient. For infants, data were only available to reliably estimate ULs for vitamins A and D and the minerals fluoride, selenium, zinc, and iron. Although adequate data were not available to define a UL for infants for other nutrients, it is important to note that intake for nutrients for which a UL does not exist should only be consumed from food or formula and not from supplements. Also notable is that the UL for iron for infants is only relevant to intake from supplements and not foods.

Summary of DRIs for infants

Macronutrients: Energy, Carbohydrate, Fat, Protein, and Amino Acids

Energy The estimated energy requirement (EER) for infants was derived by summing predicted total energy expenditure (TEE) and energy deposition for growth. Because the energy needs for growth decelerate with advancing age, an equation for EER was established for three age intervals during the first year of life (**Table 1**). The TEE is calculated using an equation (**Table 1**) based on energy expenditure measured by doubly labeled water and adjusted for weight of the child. The EER is then the sum of TEE for an individual child plus the predicted energy deposition for age (**Table 1**). No adjustment for physical activity was included in the EER for infants. Examples of the EER for males and females using reference weights are shown in **Table 1** for infants at five ages during the first year. At most ages beyond the first 2 months of life, the values for EER exceed the average energy provided (500 kcal) by human milk assuming a volume of intake (0.780 l/day) from human milk.

Table 1 DRI estimated energy requirement (EER) for infants[a]

Equations	
0–3 months	(89 × weight of infant (kg) − 100) + 175 (kcal for energy deposition)
4–6 months	(89 × weight of infant (kg) − 100) + 56 (kcal for energy deposition)
7–12 months	(89 × weight of infant (kg) − 100) + 22 (kcal for energy deposition)

Calculated EER for age using reference weights for age

Age (months)	*Males (kcal/day)*	*Females (kcal/day)*
1	472	438
3	572	521
6	645	593
9	746	678
12	844	768

[a]From the Institute of Medicine (2002) *Dietary Reference Intakes for Energy, Carbohydrates, Fiber, Fat, Protein, and Amino Acids (Macronutrients)*. Washington, DC: National Academy Press.

Carbohydrate The AI for carbohydrate for infants through 1 year of age is based on the average carbohydrate intake from human milk and complementary foods for the 7- to 12-month age group (**Table 2**). Although the carbohydrate from human milk is almost exclusively lactose and that from infant formula may be lactose, sucrose, or glucose polymers alone or in combination, there is no evidence that non-lactose-containing formulas vary from lactose contained in human milk with regard to available energy.

Fat As for other nutrients, the AI for fat intake is based on the average intake of fat from human milk alone or in addition to complementary foods after 7 months of age (**Table 2**). Although infant formulas are designed to contain a percentage of energy as fat similar to human milk (approximately 50%), the type of fat in formulas varies widely, including such sources as safflower, sunflower, soybean oil, and coconut and palm oils, usually in some combination.

Linoleic acid (n-6) and α-linolenic acid (n-3) The n-6 fatty acids are essential for infants, and in extreme long-term deficiency skin lesions and delayed growth may develop. Linoleic acid serves as a precursor of arachidonic acid (AA), which is required for synthesis of prostaglandins and other eicosanoids. The n-3 fatty acids are also essential as a precursor of docosahexenoic acid (DHA), which comprises a large percentage of the fatty acids incorporated into developing brain and retina, and of eicosapentenoic acid, which is the substrate for eicosanoid synthesis. Human milk is a natural source of both fatty acid families, including the long-chain polyenoic derivatives DHA and AA. The pattern of all fatty acids in human milk, including the polyenoic fatty acids, is dependent on maternal diet. The AI established for infants for n-6 and n-3 fatty acids is based on the average content in human milk reported for North American women with the addition of that from complementary foods during months 7–12 (**Table 2**).

Feeding of mother's milk compared to cow milk-based infant formula has been associated with a positive benefit to developmental outcomes (cognitive, motor, and vision) in both retrospective and prospective studies (but not randomized trials for obvious ethical reasons). To date, investigations of the nutrient(s) possibly responsible for the observed benefits of mother's milk on neurodevelopment have focused on the long-chain polyenoic fatty acids DHA and AA. These fatty acids represent the greatest proportion of polyenoic fatty acids contained in phospholipids of neural and retinal tissues, and they are present naturally in human milk. Until very recently, DHA and AA were not provided in infant formulas, but such formula is now marketed globally.

A positive benefit of breast-feeding compared to formula feeding on short-term visual and developmental outcomes in term and premature infants has been observed in several studies. However, the evidence for a benefit is more consistently observed in premature than in term infants, perhaps due to a greater immaturity of their enzymatic pathway to convert α-linolenic and linoleic acids to the long-chain polyenoic derivatives. Due to the conflicting evidence, specific requirements for DHA and AA for term infants were not included in the recent DRI report.

Protein For infants age 0–6 months, the AI for protein is based on the intake from human milk (**Table 2**). For infants age 7–12 months, sufficient information was available from nitrogen balance studies and protein deposition to derive an EAR based on the factorial method. For both males and females, this averaged 1.1 g protein/kg body weight/day. The RDA was set as the EAR + 2 standard deviations (based on coefficients of variation observed in adults), which yielded a value for protein intake of 1.5 g/kg/day. Because the absorption and digestibility of protein contained in infant formula may be less efficient than from human milk, the quantity of protein contained in infant formulas may have to be adjusted depending on the protein source used.

Amino acids The DRI for the essential (indispensable) amino acids for infants was derived from the content of human milk for ages 0–6 months. For older infants, an EAR was derived for these amino acids using a factorial estimate that was based on

Table 2 DRI for macronutrients for infants—carbohydrate, protein, fat, and essential fatty acids[a]

Nutrient[b]	*0–6 months*	*7–12 months*
Carbohydrate, AI (g/day)	60	95
Protein		
AI (g/day)	9.1	—
RDA (g/day)	—	13.5
Total fat, AI (g/day)	31	30
Linoleic acid (n-6), AI (g/day)	4.4	4.6
α-Linolenic acid (n-3), AI (g/day)	0.5	0.5

[a]From the Institute of Medicine (2002) *Dietary Reference Intakes for Energy, Carbohydrates, Fiber, Fat, Protein, and Amino Acids (Macronutrients)*. Washington, DC: National Academy Press.
[b]No upper levels of nutrients were set for any macronutrients.
AI, Adequate Intake; RDA, Recommended Dietary Allowance.

Table 3 DRI for indispensible (essential) amino acids[a]

Amino acid[b]	*0–6 months*[c]	*7–12 months*
	AI (mg/kgday)	*RDA (mg/kg/day)*
Histidine	23	32
Isoleucine	88	43
Leucine	156	93
Lysine	107	89
Methionine + cysteine	59	43
Phenylalanine + tyrosine	135	84
Threonine	73	49
Tryptophan	28	13
Valine	87	58

[a]From the Institute of Medicine (2002) *Dietary Reference Intakes for Energy, Carbohydrates, Fiber, Fat, Protein, and Amino Acids (Macronutrients)*. Washington, DC: National Academy Press.
[b]No upper levels were set for any of the indispensable amino acids.
[c]AI values shown as amino acid in mg/kg/day can be converted to mg amino acid/day by multiplying by the reference weight of 6 kg for infants 0–6 months of age.

the amino acid needs for growth or protein deposition, with adjustments for efficiency of protein deposition and maintenance requirement. The RDA was determined by adding the coefficient of variation derived for maintenance and protein deposition to the value for the EAR. No values were set for UL for any of the amino acids. A summary of the AI and RDA for the indispensable amino acids of infants is provided in **Table 3**.

Other macronutrients For infants, no DRI was set for saturated fat, monounsaturated fat, *trans* fatty acids, and cholesterol or dietary fiber. Although some dietary fiber is present in the diet after solid foods are introduced, there are no data on fiber intakes in such young age groups and no theoretical basis exists on which to establish a need for fiber at less than 1 year of age.

Macrominerals: Calcium, phosphorus, magnesium, and fluoride The AI for infants for the "bone" minerals are summarized in **Table 4**. The content of human milk was used as the basis to derive the AI for calcium, phosphorus, and magnesium for infants age 0–6 months and with the addition of intake from complementary foods for those age 7–12 months. For fluoride, intake from human milk was the reference for the first 6 months only. After 6 months, the AI for fluoride was set at 0.05 mg/kg/day and adjusted to a reference weight for age, based on the well-documented evidence of the benefit of fluoride intake for the prevention of dental caries (**Table 4**).

Table 4 DRI for minerals for infants—calcium, phosphorus, magnesium, and fluoride[a]

Nutrient	*0–6 months*	*7–12 months*
Calcium		
AI (mg/day)	210	270
UL	ND	ND
Phosphorus		
AI (mg/day)	100	275
UL	ND	ND
Magnesium		
AI (mg/day)	30	75
UL	ND	ND
Fluoride		
AI (mg/day)	0.01	0.5
UL (mg/day)	0.7	0.9

[a]From the Institute of Medicine (1997) *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D and Fluoride*. Washington, DC: National Academy Press.
AI, Adequate Intake; ND, not determinable due to lack of data of adverse effects in infants; RDA, Recommended Dietary Allowance; UL, upper limit.

Microminerals/trace elements The AI for iron for ages 0–6 months is based on the concentration of iron in human milk albeit low (approximately 0.35 mg/l) but assumes the infant is born with maximal iron stores due to transplacental transfer of iron from an iron-replete mother. If the latter conditions do not apply, then an exogenous source of iron such as iron drops may be required. For infants age 7–12 months, an EAR and RDA were developed based on a factorial modeling method that summed basal loss of iron with needs for growth, increasing hemoglobin mass, and iron stores. This value was then adjusted for iron bioavailability using a factor of 10% for infants due to a medium bioavailability of iron from infant cereals, which are generally the major dietary source of iron in weaning foods before meats are introduced (**Table 5**). A UL was established (**Table 5**) for iron based on the risk of adverse gastrointestinal side effects from supplemental (not food) iron.

For zinc, an AI was based on the human milk model only for the 0- to 6-months age group (**Table 5**). The zinc content of human milk declines rapidly during the first 6 months (from 4 to 1.2 mg/l), so the AI was based on a milk zinc concentration of 2.5 mg/l. This value cannot be directly applied to infants being fed cow milk- or soy-based infant formulas because zinc absorption is significantly lower from these compared to human milk. The EAR for the 7- to 12-months age group was set using a factorial method that summed obligatory losses with requirements for growth and adjusted for fractional absorption of dietary zinc from human milk and complementary foods. The RDA

Table 5 DRI for micronutrient/trace minerals for infants—chromium, copper, fluoride, iodine, iron, manganese, molybdenum, selenium, and zinc[a]

Nutrient	*0–6 months*	*7–12 months*
Chromium		
AI (μg/day)	0.2	5.5
UL	ND	ND
Copper		
AI (μg/day)	200	220
UL	ND	ND
Iodine		
AI (μg/day)	110	130
UL	ND	ND
Iron		
AI (mg/day)	0.27	—
RDA (mg/day)	—	11
UL (mg/day)	40	40
Manganese		
AI (μg/day)	30	75
UL	ND	ND
Molybdenum		
AI (μg/day)	2	3
UL	ND	ND
Selenium		
AI (μg/day)	15	20
UL (μg/day)	45	60
Zinc		
AI (mg/day)	2	—
RDA (mg/day)	—	3
UL (mg/day)	4	5

[a]From the Institute of Medicine (2001) *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium and Zinc*. Washington, DC: National Academy Press.
AI, Adequate Intake; ND, not determinable due to lack of data of adverse effects in infants; RDA, Recommended Dietary Allowance; UL, upper limit.

was derived by adding twice the coefficient of variation of 10% to the EAR (2.5 mg/day of zinc) for infants 7–12 months (**Table 5**). A UL was set for zinc on the basis of the possibility of an adverse effect of high zinc intakes on copper status.

For the trace elements chromium, copper, iodine, manganese, molybdenum, and selenium, an AI was set for infants age 0–6 months based on the human milk model (**Table 5**). For the age group 7–12 months, data on intake from complementary foods were only available to set an AI for chromium, copper, and selenium (**Table 5**). For iodine and molybdenum, the AI represents an extrapolation up from the AI values for the age group 0–6 months based on differences in metabolic body weight ($kg^{0.75}$). For manganese, the AI represents an extrapolation down from the AI for adults as described previously (**Table 5**). Due to lack of relevant information no UL values for infants younger than 1 year of age were established for chromium, copper, iodine, manganese, or molybdenum, but intakes of these nutrients should be limited to foods and not supplements. A UL was established for selenium due to the known toxicity of chronic excess selenium ingestion, which presents clinically as brittleness and loss of nails and hair. The UL was set for infants based on the highest known intake of selenium from human milk and adjusting for a reference infant weight (**Table 5**). The UL value pertains to intake from both foods and supplements.

The trace elements arsenic, boron, nickel, silicon, and vanadium are recognized as having a role in human metabolism, but due to lack of information DRI values, including UL, could not be established for infants.

Globally, deficiencies of iron, iodine, and zinc in infants are still widespread despite international efforts to develop sustainable food fortification and supplementation programs. In North America, the prevalence of iron deficiency anemia is relatively low at 4–5% owing to the promotion of breast feeding and the widespread fortification of infant formulas and cereals with iron. In developing countries, prevalence of anemia can be 50% or more by one year of age. Premature and/or low birth weight (<2.5 kg) infants represent a particular risk group for iron deficiency owing to a major reduction in transplacental transfer of iron when birth occurs during the third trimester of pregnancy, the period when most iron is transferred to the fetus as long as the mother is not iron deficient. Infants of low birth weight require iron supplementation in a liquid form until complementary foods containing iron can be introduced. Use of weaning foods that are not iron-fortified and that often contain phytic acid, a strong inhibitor of iron absorption, is a key causative factor for high rates of anemia in many developing countries.

Fat-soluble vitamins A, K, E, and D For vitamins A, K, and E, the AI for infants 0–6 months of age was based on the human milk model as previously described (**Table 6**). For vitamins A, K, and E, the AI for infants 7–12 months of age was extrapolated up from the values for 0–6 months using a reference weight for infants at this age. There are two important points with respect to the AI established for vitamin K. First, the AI was set assuming infants had received a prophylactic injection of vitamin K just after birth. Since vitamin K is not readily transferred to the fetus while *in utero*, and human milk is relatively low in vitamin K, newborn infants, at least in North America, routinely receive an injection of vitamin K within a few hours after birth. Second, the AI set for ages 7–12 months may be lower than the actual intake of vitamin K once a child's diet of

Table 6 DRI for fat-soluble vitamins[a]

Nutrient	*0–6 months*	*7–12 months*
Vitamin A		
AI (μg/day)	400	500
UL (μg/day)	600	600
Vitamin D		
AI (μg/day)	5	5
UL (μg/day)	25	25
Vitamin E		
AI (mg/day)	4	5
UL	ND	ND
Vitamin K		
AI (μg/day)	2.0	2.5
UL	ND	ND

[a]From the Institute of Medicine (1997) *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D and Fluoride*. Washington, DC. National Academy Press; Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. Washington, DC: National Academy Press; and Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium and Zinc*. Washington, DC: National Academy Press. AI, Adequate Intake; ND, not determinable due to lack of data of adverse effects in infants; UL, upper limit.

complementary food becomes varied. Any evaluation of dietary intake of vitamin K should use the recently updated vitamin K values for raw and cooked foods available from the USDA National Nutrient Database for Standard Reference, Release 17 (http://www.nal.usda.gov/fnic/foodcomp/Data/SR17/wtrant/wt_rank.html). Although carotenoids are present in human milk, a factor to calculate their bioconversion to vitamin A is not known so their contribution to vitamin A was not included.

For vitamin D, it was determined that a dietary (or supplement) intake of 100 IU would likely prevent rickets but not maintain normal circulating concentrations of 25-hydroxyvitamin D. Thus, assuming that most infants obtain minimal or no vitamin D via exposure to sunlight, an AI of 200 IU (5 μg) was established. This amount of vitamin D was also recommended for infants 7–12 months of age assuming most infants could maintain normal vitamin D status with this intake. The AI for vitamin D set by the Institute of Medicine was adopted by the American Academy of Pediatrics (AAP). The AAP recommended a minimum intake of 200 IU vitamin D per day for all infants beginning during the first 2 months of life in recognition of the risk in vitamin D-deficiency rickets in the United States, especially among infants who are breast fed for a number of months without vitamin D supplementation. In Canada, a recent recommendation by Health Canada (2004) was for all breast fed infants to receive 10 μg (400 IU)/day of vitamin D from birth until their diet included equal amounts of vitamin D from other food sources. Considerations for this recommendation (rather than 5 μg (200 IU)/day of the DRI) included lack of sun exposure owing to Canada's northern geographic latitude, current practices related to protection from the sun, and an increasing prevalence of vitamin D deficiency rickets in infants.

Water-soluble B vitamins, folate, choline, and vitamin C The AIs for infants age 0–6 months for most water-soluble vitamins were based on the content of human milk (**Table 7**). This approach may be problematic for water-soluble B vitamins, in which milk content is dependent on maternal intake of vitamins. An example of clinical relevance is the vegan mother

Table 7 DRI for water-soluble vitamins[a]

Nutrient	*0–6 months*	*7–12 months*
Vitamin C		
AI (mg/day)	40	50
UL	ND	ND
Thiamin		
AI (mg/day)	0.2	0.3
UL	ND	ND
Riboflavin		
AI (mg/day)	0.3	0.4
UL	ND	ND
Niacin		
AI (mg/day)	2	4
UL	ND	ND
Vitamin B_6		
AI (mg/day)	0.1	0.3
UL	ND	ND
Folate		
AI (μg/day)	65	80
UL	ND	ND
Vitamin B_{12}		
AI (μg/day)	0.4	0.5
UL	ND	ND
Pantothenic acid		
AI (mg/day)	1.7	1.8
UL	ND	ND
Biotin		
AI (μg/day)	5	6
UL	ND	ND
Choline		
AI (mg/day)	125	150
UL	ND	ND

[a]From the Institute of Medicine (1998) *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin, and Choline*. Washington, DC: National Academy Press; and Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. Washington, DC: National Academy Press. AI, Adequate Intake; UL, upper limit.

who may have subclinical vitamin B_{12} deficiency and produce B_{12}-deficient milk. For vitamin C, the effect of maternal supplementation on milk content remains uncertain, but available reports do not indicate that excessive amounts of vitamin C are secreted in milk, even in mothers taking supplements of 1000 mg or more. For age 7–12 months, the AI for thiamin, riboflavin, niacin, folate, pantothenic acid, and choline was derived by extrapolation down from values for older children or adults due to a lack of information of dietary intake of these nutrients from solid foods. Tolerable ULs for infants were not established for any of the water-soluble vitamins.

Water and electrolytes Optimal water intake in infants is more critical than at any other period of life. Not only do infants have higher total body water content per body mass than children or adults but also they have a higher water turnover rate, a less well-developed sweating mechanism, and little ability to indicate when they are thirsty. The AI for water intake of infants age 0–6 months is 0.7 l/day and is based on the water content of human milk. Assuming infants are breast-fed on demand, infants will drink to meet thirst needs; thus, even in hot and humid climates supplemental water should not be required. The AI for water intake of 0.8 l/day set for 7–12 months is based on the sum of the water content of human milk, complementary foods, and beverages, the latter obtained from reported food intakes from surveys in the United States.

For sodium and potassium, the AI of 0.12 g/day and 0.4 g/day, respectively, is based on the human milk model. For 7–12 months, the AI for sodium is 0.37 g/day and for potassium is 0.7 g/day based on the sum of observed intakes from human milk and complementary foods. No ULs were established for infants due to lack of data on adverse effects of these nutrients on infant health. However, particularly because the renal excretory capacity of young infants may not be able to handle excess amounts of ingested electrolytes, the DRI report notes that intake of sodium, chloride, and potassium should be limited to human milk (or infant formula) and solid foods appropriate for age.

Assessment of Growth as an Indication of Adequate Nutrition

Assessment of growth in weight, length and head circumference is an internationally accepted measure of health and nutritional status of infants, albeit not an indicator that is nutrient specific. The interpretation of growth measures requires comparison with reference data from normal populations of infants that have been complied into growth charts with centiles indicated. The growth charts from the Center for Disease Control in the United States as revised in 2000 (http://www.cdc.gov/nchs/about/major/nhanes/growthcharts/charts.htm), were adopted for use in the USA and Canada as well as by the World Health Organization (WHO) for use internationally. The growth charts as shown in **Figures 1–6** may be downloaded from the CDC website and copied. Weight gain of exclusively breast fed infants is more rapid than formula-fed infants in the first 2 to 3 months but they weigh less from 6 to 12 months. The longitudinal data that form the basis of the growth charts for infants from birth to 36 months growth represent a mix of both breast and formula fed infants from the American population. A Working Group of WHO has undertaken a project to develop growth charts specifically for exclusively or predominantly breast fed infants but these are not yet available.

Research Needs

The paucity of sound evidence on which to provide a substantial basis for estimating the nutrient requirements for infants is highlighted at the end of each chapter in the DRI reports. Since infants and children are not just "little adults," the DRI values must be carefully defined for the specific stages of growth and development and with consideration for nutritional programming that occurs in early life in response to dietary exposures as our knowledge of this area becomes more complete.

Practical Aspects of Meeting the Nutrient Needs of Infants

Adequate amounts of breast milk meet the nutrient needs of most infants for the first 6 months of life. However, there is not universal agreement on the optimal duration of exclusive breast feeding and the precise timing or the order of introduction of complementary foods. Internationally, recommendations from most health agencies state that the ideal feeding of infants is exclusive breast feeding for the first 6 months of life with appropriate introduction of foods from 6 months onward including partial breast feeding through 2 years of age or beyond. When assessing intakes of infants fed marketed formulas, it must be kept in mind that intakes of most nutrients will exceed the new DRI values for AI given that these are based on the composition of human milk. In many cases, the greater concentration of nutrients in infant formula is appropriate due

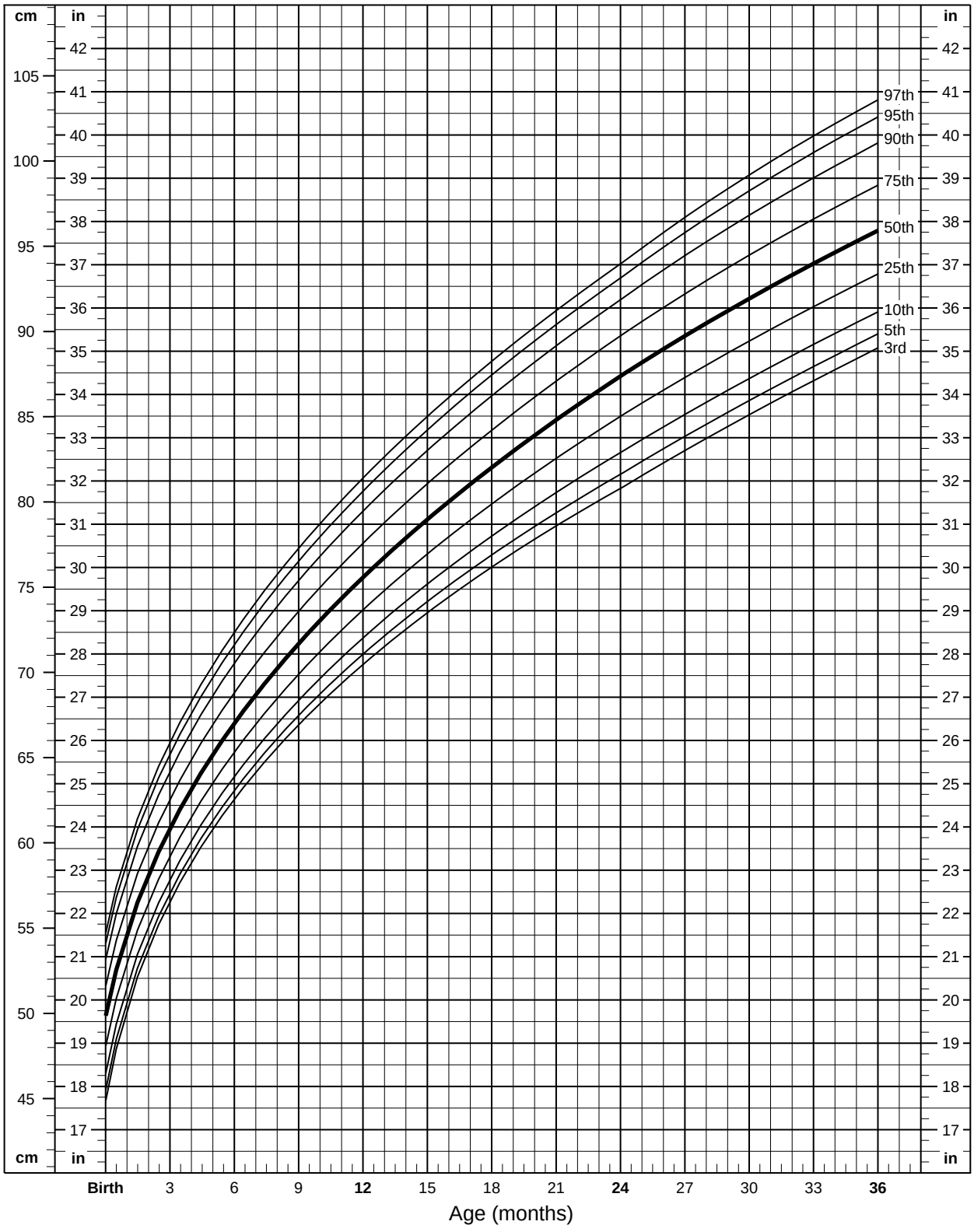


Figure 1 Length-for-age percentiles: Boys, birth to 36 months. CDC Growth Charts: United States from Centers for Disease Control and Prevention, National Center for Health Statistics. Source: Developed by the National Center for Chronic Disease Prevention and Health Promotion (2000).

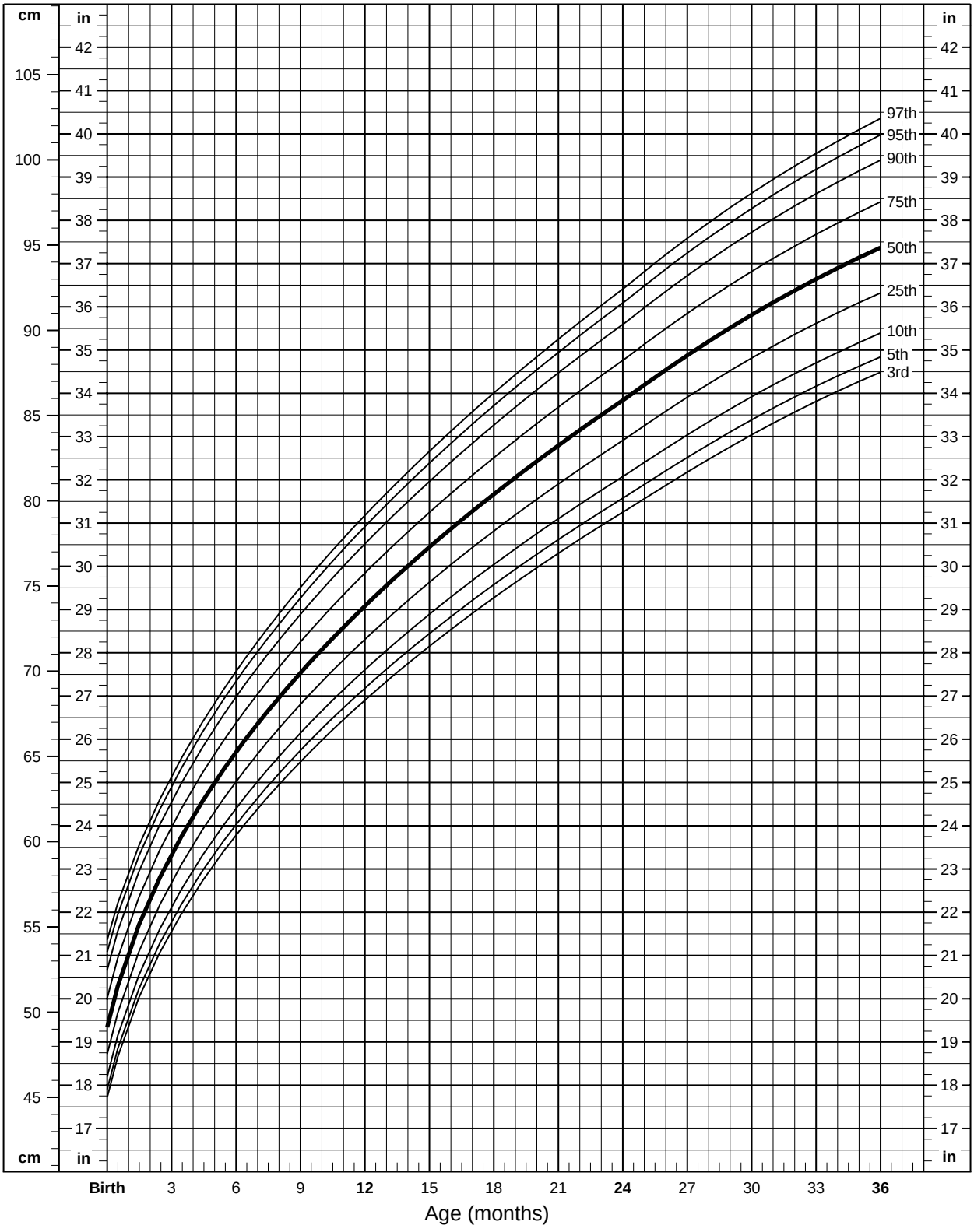


Figure 2 Length-for-age percentiles: Girls, birth to 36 months. CDC Growth Charts: United States from Centers for Disease Control and Prevention, National Center for Health Statistics. Source: Developed by the National Center for Chronic Disease Prevention and Health Promotion (2000).

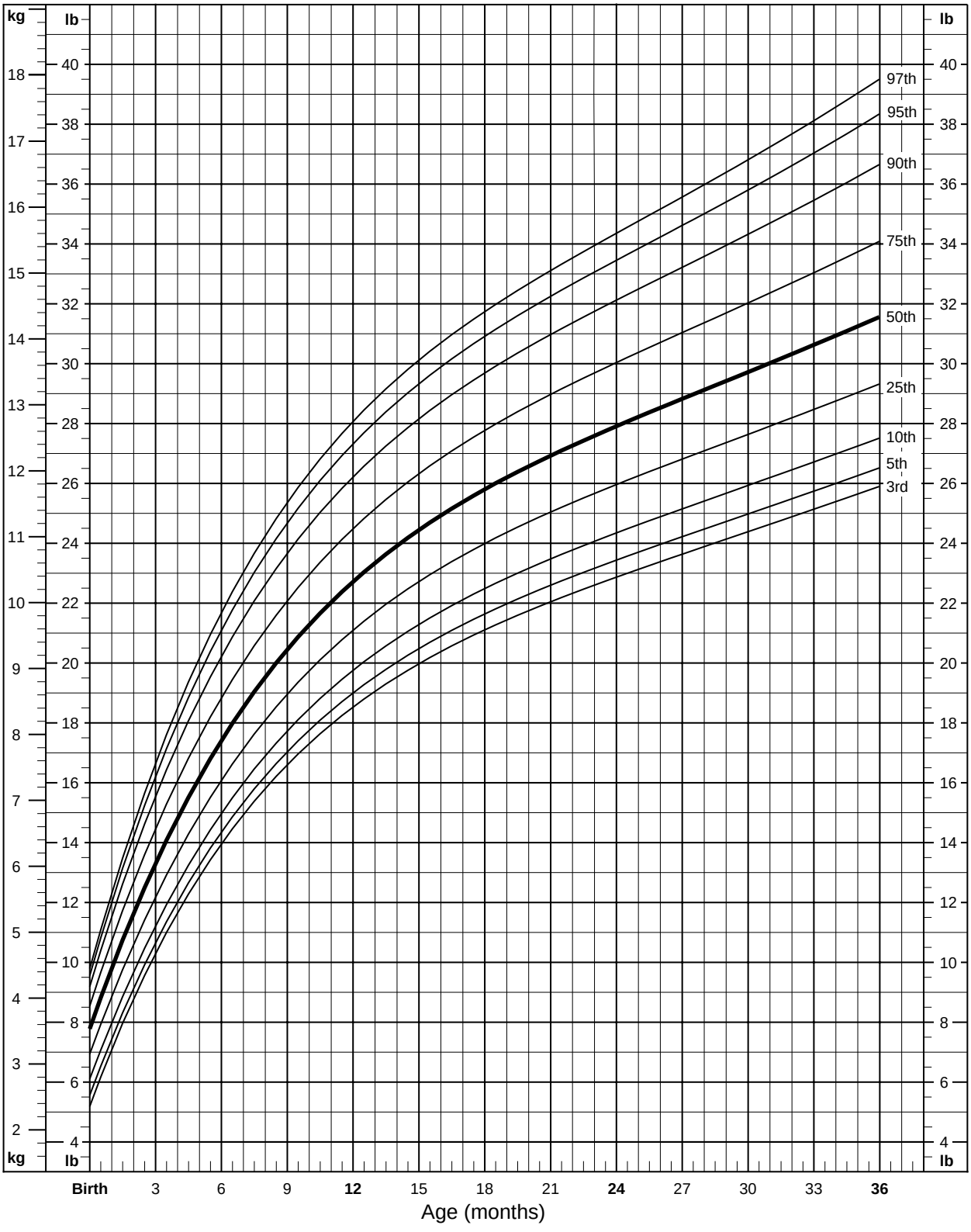


Figure 3 Weight-for-age percentiles: Boys, birth to 36 months. CDC Growth Charts: United States from Centers for Disease Control and Prevention, National Center for Health Statistics. Source: Developed by the National Center for Chronic Disease Prevention and Health Promotion (2000).

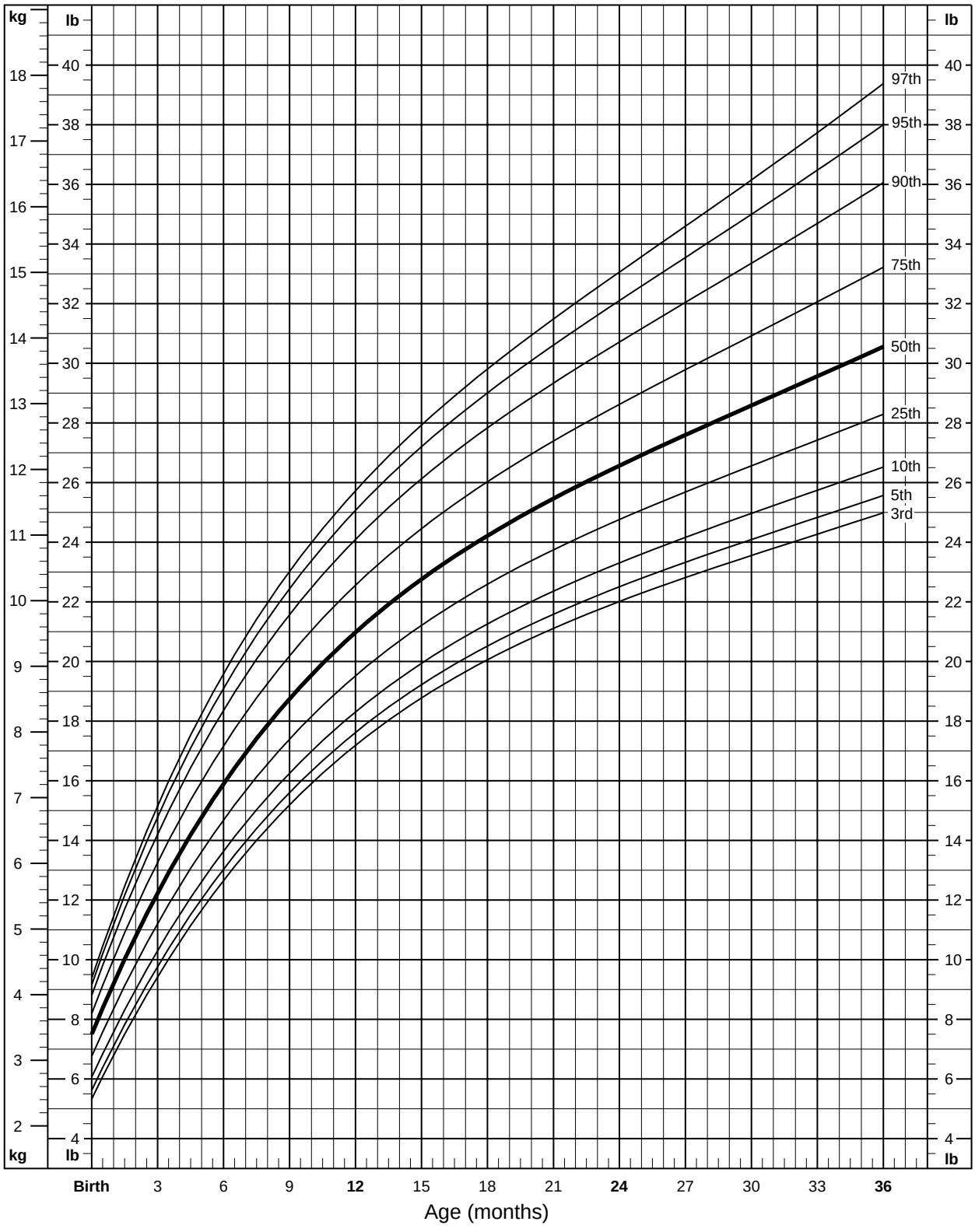


Figure 4 Weight-for-age percentiles: Girls, birth to 36 months. CDC Growth Charts: United States from Centers for Disease Control and Prevention, National Center for Health Statistics. Source: Developed by the National Center for Chronic Disease Prevention and Health Promotion (2000).

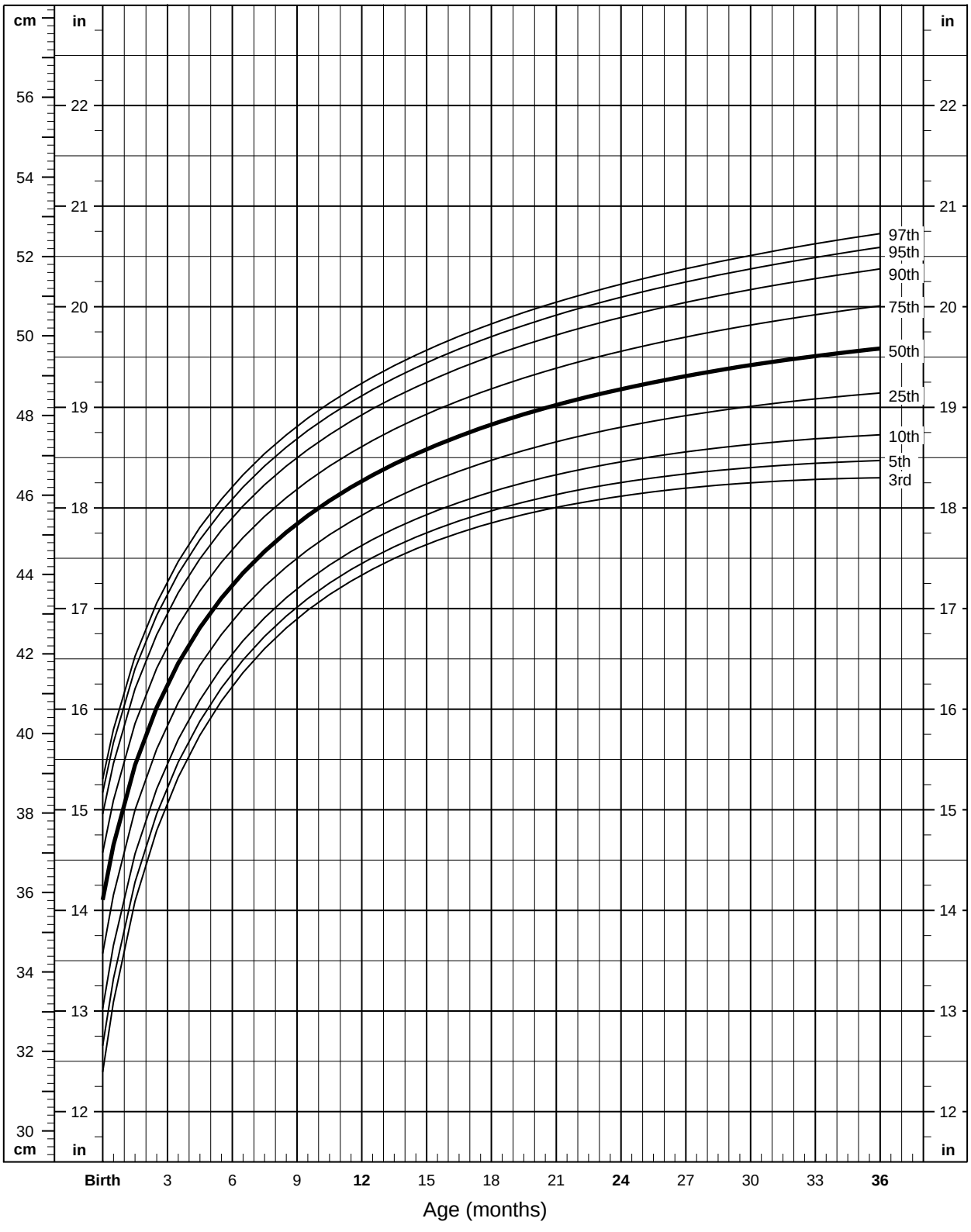


Figure 5 Head circumference-for-age percentiles: Boys, birth to 36 months. CDC Growth Charts: United States from Centers for Disease Control and Prevention, National Center for Health Statistics. Source: Developed by the National Center for Chronic Disease Prevention and Health Promotion (2000).

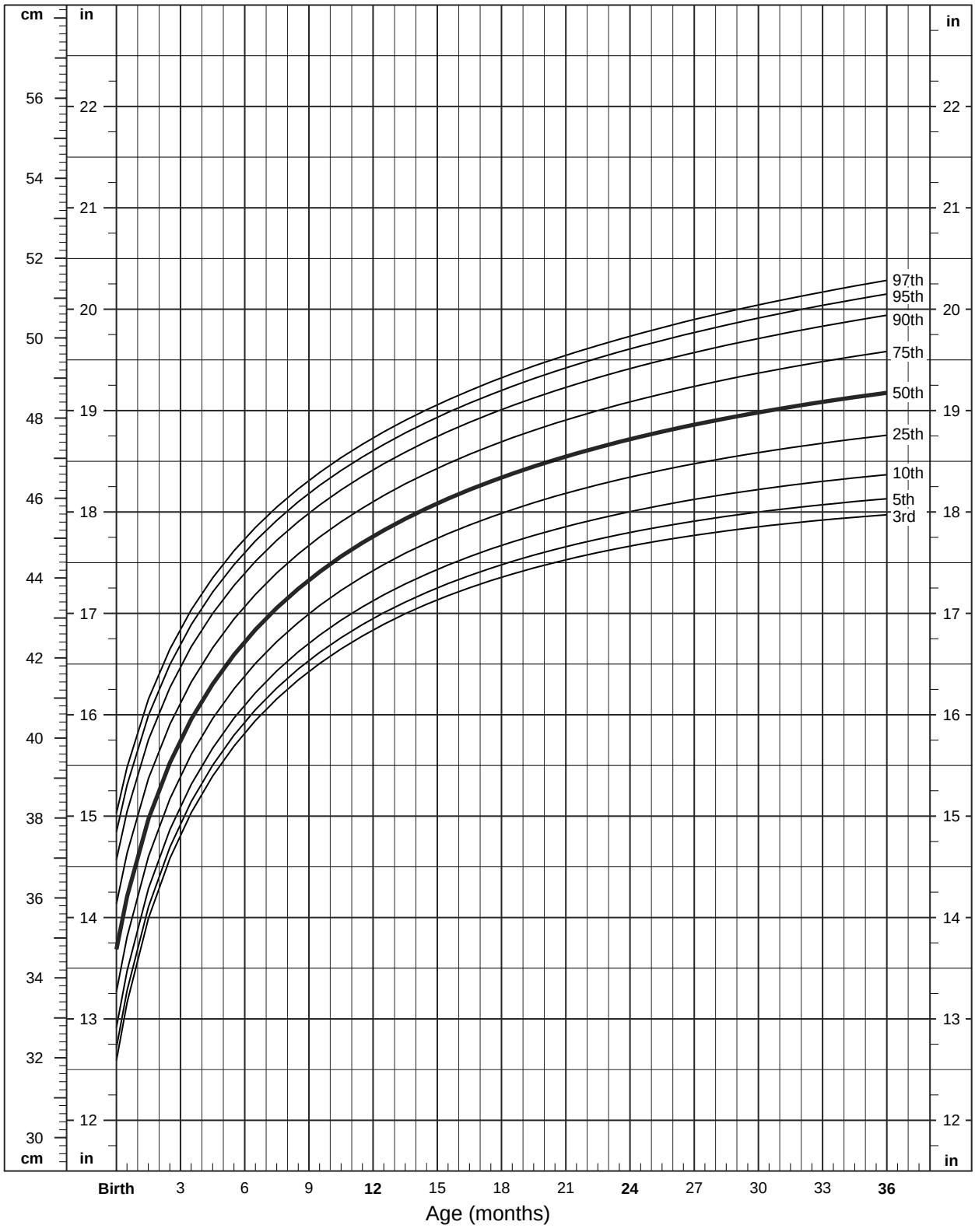


Figure 6 Head circumference-for-age percentiles: Girls, birth to 36 months. CDC Growth Charts: United States from Centers for Disease Control and Prevention, National Center for Health Statistics. Source: Developed by the National Center for Chronic Disease Prevention and Health Promotion (2000).

to lower digestibility or bioavailability of nutrients from cow milk- or soy-based protein in formulas compared to human milk.

The introduction of complementary foods, especially solids and eventually finger foods, is important for infants to develop normal oral and motor skills related to eating and to attain adequate intakes of nutrients that may be low in breast milk (e.g., protein or iron). In a report by the March of Dimes, three common inappropriate complementary feeding practices were delineated: (i) introducing foods too early or too late, (ii) introducing foods of low nutrient density, and (iii) feeding contaminated foods. It is noted in the report that early introduction of foods may reduce the intake of breast milk due to limited gastric capacity of very young infants or precipitate an allergic reaction in infants with a family history of food allergy or atopy. By delaying introduction of foods beyond 6 months, there is increasing risk of deficiencies of nutrients known to be relatively low in breast milk and yet essential to support rapid growth of infants, such as iron and zinc. The choice of first foods is important so that they contain adequate energy and micronutrients to meet the needs of infants. For example, reduced-fat cow milk (skim 2%, and 1% fat) should not be fed to infants before 2 years of age. Excessive amounts of fruit juices or 'empty calorie' fast foods should not be fed to infants. To achieve adequate intakes of micronutrients such as iron, choice of nutrient-fortified foods (e.g., iron-fortified infant cereal or other weaning food) may be required in areas where natural sources of micronutrients are not available. Finally, both solid and liquid foods offered to infants need to be free of contamination since the transmission of infections through food is thought to be a primary cause of diarrhea in young infants, particularly in developing countries.

The March of Dimes report (2002) outlined three key recommendations for ensuring optimal nutrition of term-born infants through breast-feeding and complementary feeding practices. The rationale for each recommendation and suggestions for implementation strategies on a global basis are provided in the report. The three key recommendations are as follows:

> Recommendation 1: Promote and support exclusive breast feeding for 6 months, with the introduction of complementary foods and continued breast feeding thereafter—up to 2 years of age or longer as mutually desired by the mother and infant.
>
> Recommendation 2: Promote and support programs to ensure that pregnant women and breast feeding mothers receive adequate nutrient intakes.
>
> Recommendation 3: Promote the appropriate introduction of safe, nutritionally adequate, and developmentally appropriate complementary foods.

The recommendations from the March of Dimes report are universally applicable and will ensure that infants attain nutrient intakes that match the nutrient requirements as set out in dietary standards such as the DRI reports.

See also: **Amino Acids**: Chemistry and Classification; Metabolism. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Breast Feeding**. **Calcium**. **Carbohydrates**: Requirements and Dietary Importance. **Complementary Feeding**. **Electrolytes**: Water–Electrolyte Balance. **Fats and Oils**. **Fatty Acids**: Omega-3 Polyunsaturated. **Folic Acid**. **Infants**: Nutritional Requirements. **Iron**. **Lactation**: Dietary Requirements. **Magnesium**. **Phosphorus**. **Protein**: Requirements and Role in Diet. **Vitamin A**: Biochemistry and Physiological Role. **Vitamin B_6**. **Vitamin D**: Rickets and Osteomalacia. **Vitamin E**: Metabolism and Requirements. **Vitamin K**. **Zinc**: Physiology.

Further Reading

Atkinson SA (2004) Nutritional requirements for fetal & neonatal bone health & development. In: Holick M and Dawson-Hughes B (eds.) *Nutrition & Bone Health*, pp. 157–172. Clifton, NJ: Humana Press.

Atkinson SA and Zlotkin S (1997) Recognizing deficiencies and excesses of zinc, copper and other trace elements. In: Tsang R, Zlotkin S, Nichols B, and Hansen J (eds.), *Nutrition during Infancy: Birth to Two Years*, pp. 635–641. Cincinnati OH: Digital Education Publishing, Inc.

Institute of Medicine (1997) *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D and Fluoride*. Washington, DC: National Academy Press. Available at http://www.nap.edu/catalog/5776.html.

Institute of Medicine (1998) *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin, and Choline*. Washington, DC: National Acadamy Press. Available at http://www.nap.edu/catalog/6015.html.

Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. Washington, DC: National Acedemy Press. Available at http://www.nap.edu/catalog/9810.html.

Institute of Medicine (2001) *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium and Zinc*. Washington, DC: National Acedemy Press. Available at http://www.nap.edu/catalog/10026.html.

Institute of Medicine (2002) *Dietary Reference Intakes for Energy, Carbohydrates, Fiber, Fat, Protein, and Amino Acids (Macronutrients)*. Washington, DC: National Academy Press. Available at http://www.nap.edu/catalog/10490.html.

Institute of Medicine (2004) *Dietary Reference Intakes for Water, Potassium, Sodium, Chloride and Sulfate*. Washington, DC: National Acadamy Press. Available at http://www.nap.edu/catalog.

Jain A, Concato J, and Levanthal J (2002) How good is the evidence linking breastfeeding and intelligence? *Pediatrics* **109**: 1044–1053.

Lawrence M, Gartner MD, Frank R, Greer MD and the Section on Breastfeeding and Committee on Nutrition (2003) Prevention of rickets and vitamin D deficiency: new guidelines for vitamin D intake. *Pediatrics* **111**(4): 908–910.

March of Dimes, Task Force on Nutrition and Optimal Human Development (2002) *Nutrition Today Matters Tomorrow*. White Plains, NY: Education Services, March of Dimes.

Raiten DJ, Talbot JM, and Waters JH (eds.) (1998) Executive summary for the report Assessment of nutrient requirements for infant formulas. *Journal of Nutrition* **11**(supplement).

Feeding Problems

R M Katz, Johns Hopkins University School of Medicine and Mount Washington Pediatric Hospital, Baltimore, MD, USA
L Schuberth, Kennedy Krieger Institute, Baltimore, MD, USA
C S Gulotta, Johns Hopkins University and Kennedy Krieger Institute, Baltimore, MD, USA

Feeding is the process by which growing children accept and digest food in amounts adequate to meet their nutritional needs. What seems at first glance to be a simple intuitive act is actually a complex process requiring successful caregiver interaction, adequate oral motor skills, and intact gastrointestinal motility and absorption. The term 'feeding disorder' is applied to situations in which young children are unable or unwilling to eat enough to maintain their nutritional needs. The *Diagnostic and Statistical Manual of Mental Disorders*, a compendium of diagnoses and the related criteria, more specifically defines pediatric feeding disorders as "persistent failure to eat adequately as reflected in significant failure to gain weight or significant weight loss over at least one month." Feeding disorders are surprisingly common in children, and it has been reported that 25–35% of normal children will have mild feeding disorders and up to 70% of premature infants will have more severe feeding problems. Clinical manifestations include food refusal/selectivity, gagging, vomiting, swallowing difficulty, poor weight gain, or failure to thrive. These can be grouped into medical, oral motor, and behavioral categories, although many children have overlapping problems.

Certain groups may be at a higher risk for feeding difficulties. For example, children with food allergy may have accompanying gastroesophageal reflux and motility disorders, which then result in food refusal. A variety of medical conditions, such as cardiopulmonary, genetic, and metabolic disorders, can lead to poor appetite and slow weight gain. Oral motor and/or swallowing problems are commonly seen in children with congenital and acquired neurologic conditions such as cerebral palsy, structural abnormalities, or traumatic brain injury. Premature and medically fragile infants may miss sensitive periods of oral motor development resulting in delayed acquisition of feeding skills. This early interruption of feeding skills can lead to serious feeding disorders and food refusal due to lack of experience and impaired oral sensitivity.

Lastly, behavioral difficulties such as food refusal or selectivity are not always isolated problems. More often, they develop when medical illness adversely affects feeding patterns and caregiver interactions. If a child is failing to thrive, the most immediate solution to address the lack of weight gain and growth is to start nasogastric or gastrostomy tube feeding. However, this supplemental feeding often results in a decrease in oral intake, which ultimately impacts on hunger, experience, and endurance. Medical issues (i.e., reflux, cleft palate, etc.) that occur very early in infancy can be the initial cause for food refusal. Consequently, for the majority of children with a feeding disorder, an early avoidance pattern is established. The parent–child interaction usually exacerbates this pattern. For example, because of severe reflux the child learns to associate eating with pain. Consequently, when the parent tries to feed the child, he or she will often encounter severe refusal behavior, which leads most parents to terminate the meal prematurely. At this point, the child not only has associated food with pain but also has learned that by having severe food refusal the meal will be terminated. Even when the reflux is medically managed, the child will still have the learned history of pain associated with eating, and the child will also have the new history of having refusal behaviors to escape the meal.

Normal Development of Feeding and Swallowing

In order to understand feeding and swallowing disorders, one must recognize that there are dynamically changing developmental skills and social abilities in the growing child. Progression through the normal stages of feeding (**Table 1**) requires the attainment of physical abilities such as postural stability, oral motor coordination, and sensory awareness. In addition, factors such as emerging cognitive skills and

Table 1 Normal infant feeding

Age	*Stage*
Birth–12 months	Suck/swallow liquids (breast or bottle)
4–6 months	Pureed solids by spoon (cereal, fruits, vegetables, and meats)
8–9 months	Cup drinking liquids
	Ground or junior foods by spoon
	Finger feeding soft dissolvables
10–12 months	Soft table food
24 months	Self-feed with utensils

socialization play an important role in an effective caregiver–child feeding interaction.

The Swallowing Process

Understanding the mechanisms involved in eating can also be useful in understanding why a child is refusing to eat. The swallowing process is usually divided into three phases: oral, pharyngeal, and esophageal. In the newborn and young infant, all phases are driven reflexively by typical rooting and sucking behavior. As children age, the oral phase of chewing and managing food comes under more voluntary control, requiring cortical integration of sensory/motor input to coordinate the complex patterns of jaw, tongue, and oral movements. Factors such as smell, taste, and emotion become increasingly important. Once the process of chewing is completed, the tongue and soft palate propel the bolus toward the pharynx, initiating the pharyngeal phase of swallowing. As food progresses through the pharynx, a complex sequence of movements allows the safe passage of food around the airway into the esophagus. Closure of the mouth and nasal/laryngeal passages prevents aspiration while elevation and anterior displacement of the larynx opens the upper esophageal sphincter. This automatically generates a pressure gradient, which propels the bolus toward the esophageal opening. Once the food progresses to the esophageal phase, the subsequent movements are almost entirely automatic and no longer subject to cortical control. After passing the lower esophageal sphincter, food normally enters the stomach, beginning the gastrointestinal and absorptive phase of feeding. Food is emptied from the stomach based on the volume, nutrient composition, and caloric density of the meal.

Classification of Feeding Disorders in Children

A single underlying cause for why children refuse to eat enough to sustain normal growth is rarely evident, and therefore this problem presents a significant diagnostic and therapeutic challenge to clinicians and parents. Given the complexity of this challenge, numerous attempts at classifying feeding disorders have been made based on the apparent etiology, physical condition, or associated behaviors. Because most feeding disorders are the result of multiple factors (i.e., physical, motivational, skill, and parent/child relationships), a more functional classification has been developed that allows differentiation of patient types by symptoms rather than an arbitrary disease-based diagnostic approach (**Table 2**).

Table 2 Common feeding disorder symptoms

Food refusal—partial/total
Liquid dependent
Enteral tube dependent
Food selectivity
Texture
Type

Children with food refusal who require any kind of enteral tube feed would be categorized as 'food refusal—enteral tube dependent,' whereas a child who drinks more than 80% of his or her calorie requirement would be considered 'food refusal—liquid dependent.' Another feeding problem category is 'food selectivity—type.' In this category, children would eliminate 75% of the four basic food groups. Typically, a child with this categorization would have the skill to eat but would choose to only eat one or two different foods and restrict all other foods. The child may or may not be able to sustain normal growth with this kind of diet. 'Food selective—texture' describes a child who does not eat an age-appropriate texture of food due to lack of skill or oversensitivity to a particular food texture—for example, a 5-year-old child who only eats pureed foods when he or she should be able to handle regular textured food. Again, the child may or may not sustain normal growth.

Assessment

An appropriate assessment of a child's feeding disorder is a critical first step in initiating treatment. The management of complicated feeding disorders usually requires a multidisciplinary team devoted to establishing diagnosis, assessment of need, and developing a thorough treatment plan. This team may include a variety of pediatric specialists, including physicians (e.g., general pediatricians, developmental pediatricians, pediatric gastroenterologists, allergists, and otolaryngologists), nurse practitioners, nutritionists,

occupational therapists, speech therapists, psychologists, and social workers. The assembled team must begin its approach to diagnosis and therapy with complete history taking by all interested parties. This includes a careful prenatal, birth, and neonatal history. In addition to determining the nutritional and medical status of the child, an appropriate psychological and developmental pediatric evaluation must be performed.

Physicians

An important goal of the physician history taking is to assess for any comorbid conditions that would require treatment prior to the implementation of a therapeutic treatment program for the food refusal (**Table 3**). As part of the initial evaluation, an observation of a feeding session between the child and primary caregiver will often provide insight into the feeding problem, especially from an oral motor/sensory and behavioral perspective. Clinical signs of oral motor dysfunction, length of meals, and nature of the caregiver–child interaction are all noted. Observation of the muscle tone, posture, and positioning as well as special seating systems and feeding devices is routine because this can provide insight into the child's overall neurological functioning. Physical examination of the child includes a general survey examination for the determination of any underlying medical disorders that may preclude safe feeding. This includes evaluation of tongue and jaw movement, dentition, airway sounds, speech, and oral cavity assessment. Additionally, a complete physical examination including cardiac, pulmonary, and abdominal exams is mandatory.

Table 3 Medical conditions associated with pediatric feeding disorders

Disorders of the oral and pharyngeal phases of swallowing
Anatomic lesions
– Cleft lip and/or palate
– Pierre–Robin sequence
– Choanal atresia
– Laryngeal clefts
– Macroglossia
– CHARGE association
Acquired structural abnormalities
– Dental caries
– Tonsillar hypertrophy
– Viral/inflammatory stomatitis
– Retropharyngeal mass
– Candida stomatitis
Cardiopulmonary effects
– Chronic lung disease
– Complex congenital heart disease
– Reactive airway disease
– Tachypnea
Neuromuscular disorders
– Familial dysautonomia
– Cerebral palsy
– Pseudo-bulbar palsy
– Bulbar atresia or palsy
– Cranial nerve anomalies
– Muscular dystrophic disorders
– Arnold–Chiari malformation
– Myelomeningocele
– Intracranial mass lesions

Disorders of the esophageal phase of swallowing
Anatomic lesions
– Esophageal atresia
– Cricopharyngeal achalasia
– Tracheoesophageal fistula
– Esophageal mass
– Esophageal stricture
– Esophageal web
– Esophageal rings
– Vascular rings/aberrant vessels
– Foreign bodies

Disorders of the lumen
– Peptic esophagitis
– Candida esophagitis
– Viral esophagitis
– "Pill" esophagitis
– Inflammatory bowel disease
– Behcet syndrome
Motility disorders
– Achalasia
– Diffuse esophageal spasm
– Chronic pseudo-obstruction
– Systemic lupus erythematosis
– Polymyositis
Genetic disorders
– Prader–Willi syndrome
– Trisomy 21
– Cornelia de Lange syndrome
– Velocardiofacial syndrome
– Rett syndrome
Metabolic disorders
– Urea cycle abnormalities
– Hereditary fructose intolerance
– Hypothyroidism
Miscellaneous
– Gastroesophageal reflux
– Constipation
– Gas-bloat syndrome
– Dumping syndrome
– Food allergies
– Sensory loss (visual/auditory impairment)

Diagnostic Testing

Diagnostic evaluations may be warranted to better assess swallowing and anatomy (**Table 4**). The modified barium swallow study (MBS) is the procedure of choice to assess oral, pharyngeal, and upper esophageal phases of swallowing. Seat positioning, food texture, and rate and amount of food presented can be manipulated during the performance of the MBS

Table 4 Diagnostic evaluation for patients with feeding disorders

Detailed history and physical examination
Upper gastrointestinal contrast radiography
– Esophogram
– Small bowel follow-through
Videofluoroscopic swallow study
Gastric emptying study
pH monitoring
Esophagogastroduodenoscopy with biopsies
Antroduodenal manometry
Fiberoptic endoscopic evaluation of swallowing
CBC
Comprehensive metabolic panel
Thyroid function
RAST analysis for food allergies
Skin test for food allergies
Plasma amino acids
Urine organic acids
Karyotype

to determine the safest and most efficient method of feeding. Clinical evaluation prior to the MBS is essential so that appropriate food textures and liquid consistencies are available at the time of the study. Changes in head and neck position, such as chin tuck, should be tried before the actual study is performed to better correlate clinical and radiologic findings.

Additionally, a standard upper gastrointestinal contrast series utilizing barium is required for assessment of anatomy of the gastrointestinal tract. Children with repetitive vomiting or abdominal pain require endoscopic evaluation, and many will also need colonoscopy to rule out the possibility of underlying inflammatory bowel disease. Some children will need cranial imaging, such as computed tomography or magnetic resonance imaging, to search for evidence of intracranial mass lesions, hydrocephaly, or posterior fossa anomalies such as the Chiari malformation. Fiberoptic endoscopic evaluation of swallowing (FEES) allows for direct visualization of the hypopharynx and larynx during swallowing by use of a flexible laryngoscope. This will allow evaluation of the valleculae and pyriform sinuses as well as the assessment of anatomy during swallowing and potential aspiration problems. This procedure, however, does not provide information on the oral phase of swallowing. FEES may also be combined with sensory testing to induce a laryngeal adductor response. Lastly, increasingly important is the need for allergy evaluation, including consultation by an allergist. Appropriate skin testing may be necessary as well as appropriate RAST testing to search for response to food allergy.

Feeding Specialist

The generic term 'feeding specialist' refers to the team member whose responsibilities include assessing and treating oral motor and swallowing dysfunction and performing MBS studies when warranted. This may be either an occupational therapist or a speech/language pathologist depending on the training and local facility. Occupational therapists can also evaluate fine motor, sensorimotor, and visual motor function as well as positioning and the need for adaptive equipment. Speech/language pathologists can evaluate and make recommendations for communication skills when necessary.

Nutritionists

Nutritionists dedicated to pediatric care are also essential in the diagnostic team functioning. The role of the nutritionist in the assessment of current nutritional status, anticipated growth, and recommended caloric intake that would be age and diagnosis appropriate is essential.

Psychologists/Behavioral Therapists

Behavioral therapists help to provide detailed observation and an analysis of variables that may be contributing to food refusal behaviors. An integral part of the therapist's approach is performing an in-depth assessment of the child's behavior patterns with regard to eating. The goal of an assessment is to help the therapist identify what behaviors have been shaped in a child with regard to eating patterns and to help identify rewards preferred by the child to help reinforce or shape new eating patterns. The behavioral therapist then designs a treatment plan oriented toward shaping new child behaviors, and ultimately the therapist teaches the parents how to implement these strategies in the home environment.

Social Workers

Because the medical issues, behavioral needs, and family psychodynamics play a central role in the development of abnormal feeding patterns, a clinical social work evaluation is necessary for assessment and treatment of underlying familial interactions and support systems. These assessments and planning help to ensure continued success once the child has returned to the home environment.

Treatment of Feeding Disorders

The goal of all therapy is directed toward allowing parents to safely feed their children in a developmentally appropriate manner. The physician in the

treatment team must ensure that all appropriate diagnostic studies have been performed to determine if an underlying medical condition has predisposed a child to developing an unusual feeding pattern. This includes appropriate utilization of consultants and diagnostic modalities (**Table 4**). Once these studies have been performed, the physician must coordinate all the resources and direct care so that feeding therapy may proceed with minimal risk to the patient—keeping the child safe from aspiration and other complications.

The initial, and perhaps most important, part of any therapeutic approach to introducing or increasing oral food intake is to establish the safety of eating as well as the types and textures of food the child can consume most efficiently. Approaches to therapy are often described as nutritive or nonnutritive. Nonnutritive oral stimulation is performed to decrease hypersensitivity, facilitate management of secretions, establish or retrain the swallowing mechanism, maintain coordination of breathing and swallowing, and develop oral movement for sound production and communication.

Objectives for a nutritive approach include increasing oral intake, advancing food texture, transitioning to utensil use, and improvement of self-feeding. Oral motor techniques to improve muscle tone and postural control as a foundation for feeding and swallowing are largely based on a neurodevelopmental framework. The use of adaptive seating systems is a key component to feeding a child with physical disabilities that require external devices to provide head, neck, and trunk support. Attention must be paid to how positioning affects the feeding process because a change in head and neck posture and oral motor structures may affect oral motor control.

Once airway safety, positioning, and sensitivity have been controlled, a variety of treatment approaches have been suggested for children with pediatric feeding disorders. These range from individual child psychotherapy to interactional therapy between child and caregiver. However, the most widely employed treatments for feeding disorders are behavioral interventions usually included within an interdisciplinary team approach that also addresses physiology, oral motor functioning, parent–child interactions, and community or social support.

Behavioral interventions for pediatric feeding disorders are the most common modality of therapy and are often a mixture of antecedent and consequence-based treatment packages. Antecedent interventions include the establishment of a systematic feeding routine (i.e., the same time and place to eat), reducing or increasing the level of texture of food (i.e., puree vs chopped fine), and presenting a preferred food along with a nonpreferred food. Consequence-based treatments include rewarding appropriate eating behavior and/or ignoring (i.e., escape extinction) or punishing food refusal behavior. Thus, if a child accepts a bite, he or she is rewarded with attention or an arbitrary reinforcer, such as a toy or music. If the child engages in food refusal behavior, such as batting at the spoon or turning his or her head away from the food, the consequence is to ignore or extinguish the food refusal behavior and continue to present the bite to the child until it is accepted. If the child continues to refuse by expelling the food, this refusal behavior is ignored/extinguished by re-presenting the expelled bite of food to the child. In some cases, a child refuses food by holding the bite of food in his or her mouth. This form of food refusal behavior can also be ignored or extinguished by moving or redistributing the food from between the child's cheek and teeth onto the tongue, where it is more likely to be swallowed. Finally, training the parents in the use of the various feeding techniques is critical in maintaining long-term treatment gains. Skill-based parent training involving step-by-step criteria-based training has been shown to be superior to didactic methodology. Parent training, including instruction, discussion, handouts, role-playing, feedback, and the practice of techniques with a trained clinician, can result in increased parent treatment integrity.

Conclusion

Despite the increased awareness of feeding disorders in young children, there remain many challenges in implementing the specialized treatment necessary for these children. Foremost among these challenges is the financial burden associated with diagnosis and therapy. Children who cannot or will not eat require a systematic diagnostic and therapeutic approach by a team of dedicated professionals. The goal of safe oral feeding is attainable in most children when those involved in the care of children understand the complexity of eating and the associated medical and psychological conditions that comprise a feeding disorder. Helping these children to eat will allow independence from artificial sources of nutrition, such as gastrostomy feeds and parenteral nutrition, and ultimately reduce the total cost of health care for these children.

See also: **Dietetics**. **Food Allergies**: Etiology; Diagnosis and Management. **Infants**: Nutritional Requirements.

Further Reading

American Psychiatric Association (1994) *Diagnostic and Statistical Manual*, 4th edn. Washington, DC: American Psychiatric Association.

Babbitt RL, Hoch TA, Coe DA *et al.* (1994) Behavioral assessment and treatment of pediatric feeding disorders. *Journal of Developmental and Behavioral Pediatrics* **15**: 278–291.

Burklow K, Phelps A, Schultz J, McConnell K, and Rudolph C (1998) Classifying complex pediatric feeding problems. *Journal of Pediatric Gastroenterology and Nutrition* **27**: 143–147.

Dellert S *et al.* (1993) Feeding resistance and gastroesophageal reflux in infancy. *Journal of Pediatric Gastroenterology and Nutrition* **17**: 66–71.

Kerwin ME (1999) Empirically supported treatments in pediatric psychology: Severe feeding problems. *Journal of Pediatric Psychology* **24**(3): 193–214.

Logemann J (1983) *Evaluation and Treatment of Swallowing Disorders*. San Diego: College Hill Press.

Manikam R and Perman J (2000) Pediatric feeding disorders. *Journal of Clinical Gastroenterology* **30**(1): 34–36.

Morris SE and Klein MD (2000) *Pre Feeding Skills. A Comprehensive Resource for Mealtime Development*. Tucson, AZ: Therapy Skill Builders.

Munk DD and Repp AC (1994) Behavioral assessment of feeding problems of individuals with severe disabilities. *Journal of Applied Behavior Analysis* **27**: 241–250.

Piazza CC, Fisher WW, Brown KA *et al.* (2003) Functional analysis of inappropriate mealtime behaviors. *Journal of Applied Behavior Analysis* **37**: 187–204.

Rommel N, DeMeyer AM, Feenstra L, and Veereman-Wauters G (2003) The complexity of feeding problems in 700 infants and young children presenting to a tertiary care institution. *Journal of Pediatric Gastroenterology and Nutrition* **37**: 75–84.

Rudolph C and Link D (2002) Feeding disorders in infants and children. *Pediatric Clinics of North America* **49**: 97–112.

Sevin BM, Gulotta CS, Sierp BJ, Rosica LA, and Miller LJ (2002) Analysis of a response class hierarchy of food refusal behavior. *Journal of Applied Behavior Analysis* **35**(1): 65–68.

Shore BA and Piazza CC (1992) Pediatric feeding disorders. In: Konarski EA and Favell JE (eds.) *Manual for the Assessment and Treatment of the Behavior Disorders of People with Mental Retardation*. Morgantown NC: Western Carolina Center Foundation.

Tuchman D (1994) *Physiology of the Swallowing Apparatus*. San Diego: Singular Publishing.

Wolf LS and Glass R (1992) *Feeding and Swallowing Disorders in Infancy—Assessment and Management*. Tucson, AZ: Therapy Skill Builders.

INFECTION

Contents

Nutritional Interactions

H Ghattas, London School of Hygiene and Tropical Medicine, London, UK

Introduction

Immunological competence (the ability of the immune system to mount a response in the presence of a pathogen) and nutritional status are major determinants of morbidity and mortality, particularly in children of the world's least developed countries. Communicable and nutritional deficiency diseases are often grouped together in mortality statistics, as many infection-related deaths occur in individuals who are also malnourished, making it difficult to disentangle infectious causes from malnutrition-related causes of death. It is estimated that infectious and nutritional deficiency diseases are responsible for 32% of global mortality and up to 59% of deaths in the world's poorest countries (World Health Organization 2004).

The relationship between malnutrition and infection has often been described as synergistic, and is the result of multifaceted interactions between nutritional intake, nutritional status, immunity, and vulnerability to infections. However, the role of nutrition in host resistance to infection is such that both nutrient deficiencies and excesses can increase susceptibility to infection.

The Cycle of Malnutrition and Infection

Malnutrition and infection interact in a cycle of adverse events (**Figure 1**) whereby malnutrition impairs immunocompetence by affecting both nonimmunological defense mechanisms (such as epithelial membrane integrity) and immunological defenses (e.g., cytokine activity, neutrophil function, T-cell maturation) thereby increasing host

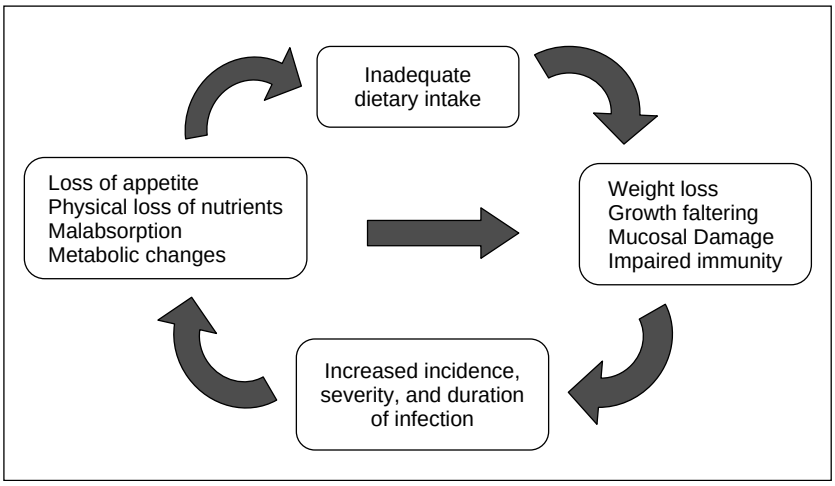


Figure 1 The cycle of malnutrition and infection. (Reproduced from Tomkins A and Watson F (1989) *Malnutrition and Infection: A Review. Nutrition Policy Discussion Paper No.5 (ACC/SCN State of the Art Series).* Geneva: United Nations.)

susceptibility to infection. Conversely, infection can affect energy requirements and appetite, and can lead to weight loss in adults and growth faltering in children. This occurs through a simultaneous increase in energy requirements during the acute phase response of an infection, anorexia (primarily mediated by interleukin (IL)-1 released by infected macrophages), physical loss of nutrients from the intestine, and malabsorption. The resulting deterioration in nutritional status is associated with additional mucosal damage, which can in turn further prolong and increase the severity of the infection as well as leaving the individual susceptible to further pathogenic invasion, thus bridging the vicious cyclical relationship between malnutrition, impaired immunity, and infection.

The Effect of Infection on Nutritional Status

Infection triggers several processes that lead to the deterioration of nutritional status (**Figure 2**). The innate response to an acute infection induces a catabolic response that increases basal metabolism (therefore increasing energy expenditure), places individuals in negative nitrogen balance (as a result of amino acid mobilization from peripheral muscle for gluconeogenesis), and leads to loss of body weight. This primarily occurs in the febrile stage of an infection during which the increase in body temperature is accompanied by an increase in basal energy requirements. This energy is required to fuel the increased rates of enzymatic reactions that occur when body temperature is elevated and to provide energy for the synthesis of proteins involved in the response to infection, e.g., acute phase proteins and immunoglobulins. The latter explains why energy metabolism can also be increased in subclinical nonfebrile infections.

The acute phase reaction is mainly driven by cytokines produced by infected leucocytes. Interleukin-1 (IL-1) is the primary mediator of the acute phase response and stimulates endocrine changes that lead to amino acid mobilization as well as the initiation of anorexia (loss of appetite). This can be compounded by physical discomfort associated with eating or swallowing that can occur in certain infections. For example, dehydration due to diarrhea can lead to mouth dryness, and opportunistic oral infections may occur following acute infections. Nutritional intake can be further reduced as a result of the cultural practice of withdrawal of food from individuals with signs of infection (such as fever or diarrhea).

Amino acid mobilization during the acute phase response to infection is also accompanied by redistribution of other nutrients among tissues as well as vitamin (e.g., retinol, folate, riboflavin, ascorbic acid) and mineral (e.g., potassium, zinc, copper) losses from the body. These changes reflect a shift in the transport of nutrients by nutrient transport proteins, the synthesis of which is reduced in response to infection, in order to prioritize the synthesis of acute phase proteins by the liver. Consequently, plasma nutrient concentrations fall due to reduced circulating levels of nutrient transport proteins.

Reductions in circulating levels of iron also occur through the sequestration of iron by the reticuloendothelial system as well as the release of the iron-binding protein lactoferrin by neutrophils, increased storage of iron as ferritin in the liver and spleen, and reduced intestinal iron absorption. This is

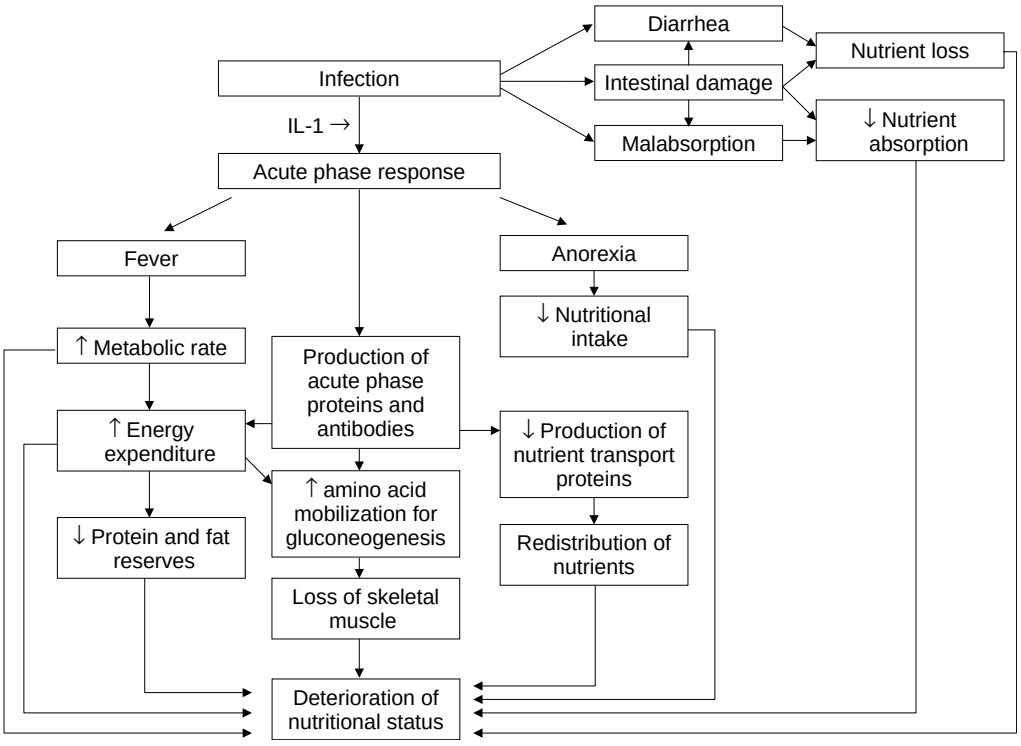


Figure 2 The effect of infection on nutritional status.

a protective mechanism that deprives microorganisms of the iron required for microbial growth and replication, therefore restricting further spread of infection.

Serum retinol levels are also reduced in a range of infections including acute respiratory infections, gastroenteritis, measles, malaria, pneumonia, and hookworm infection. Retinol depletion in measles infection has been shown to be closely related to the severity of infection.

Direct nutrient losses also occur in infection through diarrhea and nutrient malabsorption. Various infections of the gastrointestinal (GI) tract can cause diarrhea including viral, bacterial, protozoan, and helminthic infections, although non-GI infections such as malaria may also precipitate diarrheal episodes. GI infections can damage the gastrointestinal epithelium leading to flattening of microvilli, this decreases the absorptive surface area of the intestine resulting in malabsorption and electrolyte imbalance.

Table 1 lists examples of major infections, how they affect nutritional status, and the ways in which these infections may be modulated by certain nutrients.

The Effect of Nutrition on Immunity and Infection

In nutritionally compromised individuals with infection, a persistent catabolic state prevails, impairing the capacity for recovery from infection, as several recovery processes are dependent on active protein synthesis, for example, tissue repair, replacement of structural and functional proteins, and synthesis of immunoproteins. Additionally, the coexistence of malnutrition and infection increases susceptibility to secondary infection.

Severe Malnutrition and Reductive Adaptation

The acute phase response to an infection is muted in severe protein-energy malnutrition in part of a process referred to as 'reductive adaptation,' whereby the structure and function of cells or tissues cannot be maintained due to the limited supply of energy resulting from decreased nutritional intake. Protein synthesis from amino acids is highly energy-dependent and proteins have a wide variety of structural and functional roles in the body, including the cytokines that initiate the acute phase reaction (IL-1, Il-6, tumor necrosis factor (TNF-α)). Severely malnourished individuals are therefore immunocompromised, in that they cannot produce an adequate immune response to infection. The generalized responses to infection such as fever and increased pulse, as well as localized responses such as inflammation and delayed cutaneous hypersensitivity, may also not manifest. Silent infections must therefore always be suspected and treated in severe malnutrition.

Table 1 Major infections: their effect on nutritional status and ways in which these infections may be modulated by nutrition

Nutritional modulation	*Infection*	*Nutritional effects*
• Zinc: ↓ incidence and morbidity • Ascorbic acid: may have protective effect (conflicting results)	Acute Respiratory infections (ARI)	• Anorexia • Dysphagia
• Vitamin A: ↓ incidence of diarrhea • Breast feeding: ↓incidence of diarrhea • Malnutrition → ↓ epithelial integrity and ↑ diarrhea • Zinc: ↓ duration and mortality	Diarrheal diseases	• Nutrient losses • Intestinal damage and malabsorption • Dehydration • Electrolyte imbalance • Nutrient deficiencies (e.g., vitamin A)
• Vitamin A: ↓ morbidity and mortality in HIV + children • ↓ plasma selenium associated with ↑ HIV severity • Ascorbic acid: may ↓ HIV viral load • Zinc supplementation → conflicting results in HIV • Vitamin A: in lactation may ↑ MTCT[a]	Human immunodeficiency virus (HIV)	• Anorexia • Wasting syndrome • When treated with HAART[b] can lead to metabolic changes and lipodystrophy
	Intestinal parasites	• Nutrient losses • Malabsorption • Anorexia • Anemia (in hookworm and *Trichuris*) • Impaired growth and weight loss
• Vitamin A: ↓ malaria-anemia • Iron: ↓ malaria-anemia but may ↑ morbidity	Malaria	• Anemia • ↑ Protein metabolism • Anorexia • Fever → ↑ energy needs • ↓ Plasma retinol • Malaria in pregnancy → low-birth-weight baby
• Vitamin A: ↑ measles specific Ab in response to measles vaccination • Vitamin A: ↓ morbidity	Measles	• Anorexia • Buccal mucosal lesions → ↓ intake • ↑ Catabolism → growth faltering and weight loss • ↓ Plasma retinol
	Tuberculosis (TB)	• ↑ Energy metabolism • ↑ Protein breakdown • Anorexia • Anemia

[a]MTCT: mother to child transmission.
[b]HAART: highly active anti-retroviral therapy.

Malnutrition and the Breakdown of Defenses

The skin and mucous membranes provide the first layer of physical and chemical defense against an invading pathogen, and the mucosae are the major sites of entry of infectious agents into the host. The structural integrity of these barriers is compromised in malnutrition through the reduced production of mature epithelial cells, and the decreased secretion of mucin, gastric acid, lysozyme, and secretory immunoglobulins. This results in reduced gastric acidity and intestinal villous atrophy, which facilitate pathogen entry into the host. The host gut-associated lymphoid tissue (GALT) is the principal site of stimulation of mucosal immune responses and the gastrointestinal epithelium functions in the transport of nutrients as well as in immunological surveillance. Reduced epithelial integrity therefore impairs mucosal immune function and can further exacerbate nutritional status.

Lymphoid organs and cell-mediated immunity are also affected by malnutrition. In the context of malnutrition-related immunodeficiency, changes in both thymic morphology and function have been observed. These include thymic involution, thymus atrophy, circulation of immature lymphocytes, and increased thymocyte apoptosis. Similar changes occur in the spleen and lymph nodes. Where innate immunity is an organism's first level of defense in response to an infective agent, adaptive immunity and hence thymic involvement may be considered to be less crucial in the short term. The case has

therefore been made that when faced with a stress such as malnutrition, adaptive immunity becomes less of a priority and is shut down in order to prioritize other more critical organ functions. It has been suggested that critical periods in lymphoid organ development, such as the fetal and neonatal phases of development, may be particularly susceptible to malnutrition-induced changes, which may be irreversible and have long-term consequences on host resistance to immunity.

Cell-mediated immunity depends on thymus-derived T lymphocytes, which may be reduced in both number (lymphopenia) and function (impaired maturation) in malnutrition. This may result from reduced production of thymocytes by the thymus or from impaired T-cell differentiation and proliferation. Lymphocyte response to mitogens is undermined in both protein-energy and specific micronutrient deficiencies.

Moreover, neutrophil activity and bactericidal capacity are decreased in undernutrition. The neutrophil respiratory burst (which involves the production of toxic metabolites including hydrogen peroxide, superoxide anion, and nitric oxide that cause direct damage to bacteria) is impaired.

Complement proteins are also reduced in malnutrition. However, B-lymphocyte function and humoral immunity appear unaffected and normal antibody response to infectious agents is seen (except antibody responses that are highly dependent on T-cell help). However, secretory immunoglobulin (Ig) A responses may be decreased in malnutrition, possibly due to reduced secretion of IgA from damaged or atrophied mucosal surfaces.

Low Birth Weight

Low birth-weight babies are born with low nutrient reserves, an immature immune system, and small sized lymphoid organs. Studies have also linked low birth weight to reductions in T-cell counts, altered proportions of lymphocyte subsets, and reduced *in vitro* lymphocyte proliferation in infants and children. Additionally, mucosal surfaces (of the gastrointestinal and respiratory tract) are underdeveloped, thereby weakening the first layer of defense against an invading pathogen. Low birth-weight infants have increased neonatal mortality, mainly from diarrhea, pneumonia, and measles, and there is evidence that they remain at an increased risk of infection past the first year of life.

Breast feeding and Immunity to Infection

Human milk is the first form of nutrition for a neonate. Mammary glands are part of the integrated mucosal immune system and produce antibodies against mucosal pathogens that the mother is exposed to and which the infant is most likely to encounter. Breast milk contains several factors that protect against infections in the breast-fed infant either through passive immunity or by activating the infant's immune system. These include secretory IgA and IgM antibodies specific to maternal pathogenic encounters, short-chain fatty acids (SCFA), which can inhibit bacterial growth, block bacterial toxins and activate eosinophils, bactericidal lactoferrin, lysozymes, and mucins, as well as lymphocytes (both T cells and B cells), which may transfer primed immunity to the infant.

Additionally, cytokines and other growth factors in human milk contribute to the activation of the lactating infant's immune system, rendering breast-fed infants less susceptible to diarrheal diseases, respiratory infections, otitis media, and other infections and may impart long-term protection against diarrhea. Breast feeding also reduces mortality from diarrhea and respiratory infections. However, human immunodeficiency virus (HIV) infection (and other viral infections) can be transmitted from a virus-positive mother to her child through breast milk, and breast-feeding is responsible for a significant proportion of childhood HIV infection.

Key Nutrients Involved in Host Resistance to Infection

Ascorbic acid (vitamin C) Ascorbic acid is rapidly mobilized and utilized in infection and high levels of ascorbic acid are found in leucocytes. Studies in humans and animals have found a reduced T-cell response, delayed cutaneous hypersensitivity, and reduced epithelial integrity in vitamin C deficiency. Vitamin C supplementation is associated with increased lymphocyte proliferation in response to mitogen, increased phagocytosis by neutrophils, and decreased serum lipid peroxides. A role for vitamin C has been suggested in the treatment of autoimmune diseases as well as in delaying the progression of HIV to AIDS; however, further research is required to confirm such a role. The effectiveness of ascorbic acid in preventing and reducing the duration of acute respiratory infection also remains controversial. Claims that high intakes of vitamin C can prevent the common cold have not been corroborated, although there is evidence of a decrease in duration and alleviation of symptoms of the common cold.

Iron The effects of iron on infection are bipolar, with both deficiency and excess leading to increased

susceptibility to infection. Iron deficiency has been found to impair cell-mediated immunity (CMI), neutrophil function, natural killer (NK) cell activity, and bactericidal activity of macrophages, and to delay the development of CMI. The assessment of iron status in infection is however complicated by the redistribution of iron-binding proteins in the inflammatory response, making it difficult to interpret studies that find associations between measures of iron deficiency and infection.

Excess iron has immunosuppressive effects and can promote bacterial growth. Iron overload (hemochromatosis) decreases the phagocytic capacity of macrophages, alters lymphocyte subset distribution, and increases incidence of infection. Free iron acts as a catalyst in the production of reactive oxygen species (ROS) thereby increasing lipid peroxidation and cell membrane damage (of both host and microbial cells).

Invading organisms compete with the host for available iron necessary for cell function and proliferation. Pathogen replication and virulence is increased when iron supplements are administered to individuals with both iron deficiency and infection, leading to increased morbidity and mortality from infection (there is accumulating evidence for such an effect in malaria, tuberculosis, and HIV). In malaria, iron supplementation reduces malaria anemia but some studies have found increased incidence and severity of the disease, although this may be related to mode and dosage of iron administration.

The detrimental effect of iron supplementation in infection remains controversial; however, the variation in the response to supplementation may be explained by polymorphisms in genes that affect iron metabolism. For example, studies have shown poorer prognosis in HIV-infected patients with the haptoglobin (Hp) 2-2 phenotype (characterized by high serum iron, transferrin saturation and ferritin levels).

Retinol (vitamin A) Vitamin A is essential for the maintenance of mucosal surfaces and plays a role in cytokine regulation. Vitamin A supplementation has been reported to reduce child mortality from diarrheal diseases and HIV/AIDS, and to decrease the prevalence, severity, and duration of diarrheal episodes (particularly in nonbreastfed infants). Vitamin A is also involved in increasing levels of long-term measles-specific antibodies in response to measles vaccination and in reducing measles-related morbidity.

However, vitamin A does not necessarily play a beneficial role in all infections and supplementation may not always achieve the desired outcome; for example, the role of vitamin A supplementation in reducing malaria morbidity and increasing resistance remains contentious. Despite evidence for reductions in malaria-anemia, supplementation studies in respiratory infections have found no beneficial effect and a recent study found increased mother to child transmission of HIV when HIV-infected mothers were supplemented with vitamin A.

Selenium The role of selenium in resistance to infection mainly derives from its antioxidant function, but an increasing number of studies have shown that selenium also functions in both cell-mediated immunity and humoral immunity. Selenium containing proteins (selenoproteins) are the major modulators of the effects of selenium on immunity. Of these, the selenium-dependent glutathione peroxidases have an antioxidant function, and play a regulatory role in the synthesis pathways of both anti-inflammatory and proinflammatory eicosanoids.

Selenium deficiency leads to impaired antioxidant capacity, resulting in cell damage as well as a diminished respiratory-burst reaction (a microbicidal reaction in neutrophils and monocytes/macrophages). Low serum selenium levels are associated with severity of HIV, and selenium supplementation may be beneficial in HIV infection where it has been shown to reduce oxidative stress and to modulate cytokine production. Keshan disease, a juvenile cardiomyopathy with a viral etiology, is precipitated by selenium deficiency.

Selenium appears to be able to modulate pathogens as well as the immune response, as demonstrated by the increased virulence of Coxsackie virus and influenza A in response to selenium deficiency.

Zinc Zinc acts as a cofactor for many enzymes and plays a role in cellular DNA synthesis, RNA transcription, cell division, and activation. Zinc modulates both humoral and cell-mediated immunity and zinc deficiency is marked by lymphoid organ atrophy, lymphopenia, decreased T-helper 1 cell function, impaired B-cell function, and reduced phagocytic capacity.

Zinc supplementation has been found to improve infectious morbidity in individuals with sickle cell disease. In malnourished children, zinc supplementation improves epithelial integrity, decreases the duration of diarrheal episodes and mortality from diarrhea, decreases the incidence of respiratory tract infection respiratory morbidity, and improves T-cell-mediated immunity. Maternal zinc deprivation results in small thymus and spleen size in the neonate.

Results of zinc supplementation studies in HIV-positive individuals are conflicting with both benefits and adverse effects reported. The role of zinc in

reducing the severity and duration of the common cold is also still contested. High doses of zinc however have been associated with increased mortality in severe malnutrition and sepsis indicating a potential detrimental effect associated with pharmacological (versus physiological) doses of zinc.

Multiple micronutrients Much of the research in the field of nutrition–infection interactions has focused on the effect of single nutrient deficiencies on immunity, and how single nutrient supplementation may modulate infectious outcomes. However, micronutrient deficiencies often occur concurrently and significant interactions exist between different micronutrients, whereby large supplementation doses of one nutrient may inhibit uptake of another. Recent studies have shown beneficial effects of a multiple micronutrient supplement including vitamins B-complex, C, E, and folate on lymphocyte counts and birth outcomes of pregnant HIV-positive women. Few other studies have investigated the effects of multiple micronutrient supplementation on infectious outcomes, although these are warranted in populations where multiple micronutrient deficiencies coexist with infection.

The effects of nutrition on immune function are mainly modulated through the action of nutrients as essential cofactors and substrates in biosynthetic pathways or as antioxidants; with cell-mediated immunity and mucosal immunity being the most cited immunological outcomes affected by nutrition. **Table 2** lists the key nutrients involved in host defense against infection and their major effects on infectious outcomes.

Table 2 Role of key nutrients in host defense against infection

Nutrient	*Effect on infection*
Ascorbic acid	• Prevents oxidative damage in infection • Role in reducing ARI is controversial • May improve HIV outcome
Iron	• ↓ Malaria-anemia • Supplementation may ↑ morbidity from malaria, TB, and HIV
Retinol (vitamin A)	• ↓ Child mortality from diarrhea and AIDS • ↓ Prevalence of diarrhea • ↑ Measles-specific antibodies, ↓ measles morbidity • Benefits in malaria controversial • ↑ Mother-to-child transmission of HIV
Selenium	• Deficiency can precipitate Keshan disease • Deficiency ↑ virulence of Coxsackie virus and influenza A • Deficiency may ↑ severity of HIV
Zinc	• ↑ Epithelial integrity • ↓ Diarrheal duration and mortality • ↓ Incidence of and morbidity from respiratory tract infection • High doses associated with ↑ mortality in severe malnutrition and sepsis

Confounding Factors in the Nutrition-Infection Relationship

Infection and malnutrition are interdependent indices; however, the nature of the interaction between nutrition and immunity is complex and is confounded by factors that affect both nutritional status and immunocompetence (**Figure 3**).

Despite the clinical evidence for increased frequency and severity of infections in malnourished individuals, the major factor confounding the relationship between nutrition and immune status is that the environmental conditions of poverty simultaneously lead to both malnutrition and increased exposure to infections. The increase in morbidity and mortality from infectious diseases in malnourished children is often attributed to mucosal and epithelial damage caused by malnutrition, whereas this damage may be caused by the infection itself, making it difficult to confirm a causal path in the malnutrition–infection paradigm.

Furthermore, studies investigating the effects of single nutrient deficiency on infection are often confounded by coexisting nutritional deficiencies, and although nutrition can modulate immunocompetence, susceptibility to secondary opportunistic infection can be independent of nutritional status.

Additionally, the assessment of nutritional status is complicated by the presence of infection, as is the diagnosis of infection in malnourished individuals. During the acute phase response to infection, the plasma concentrations of many micronutrients are altered making it difficult to assess nutritional status, whereas the metabolic and physiological changes (reductive adaptation) that occur in severe malnutrition impair immune responses, making it difficult to diagnose infection.

Conclusions

Owing to the complex nature of the interaction between nutritional status and host-susceptibility to infection, establishing a better understanding of the underlying mechanisms behind the nutrition–infection relationship is crucial to the formulation of intervention strategies to reduce morbidity and mortality from communicable and nutritional deficiency diseases in developing countries.

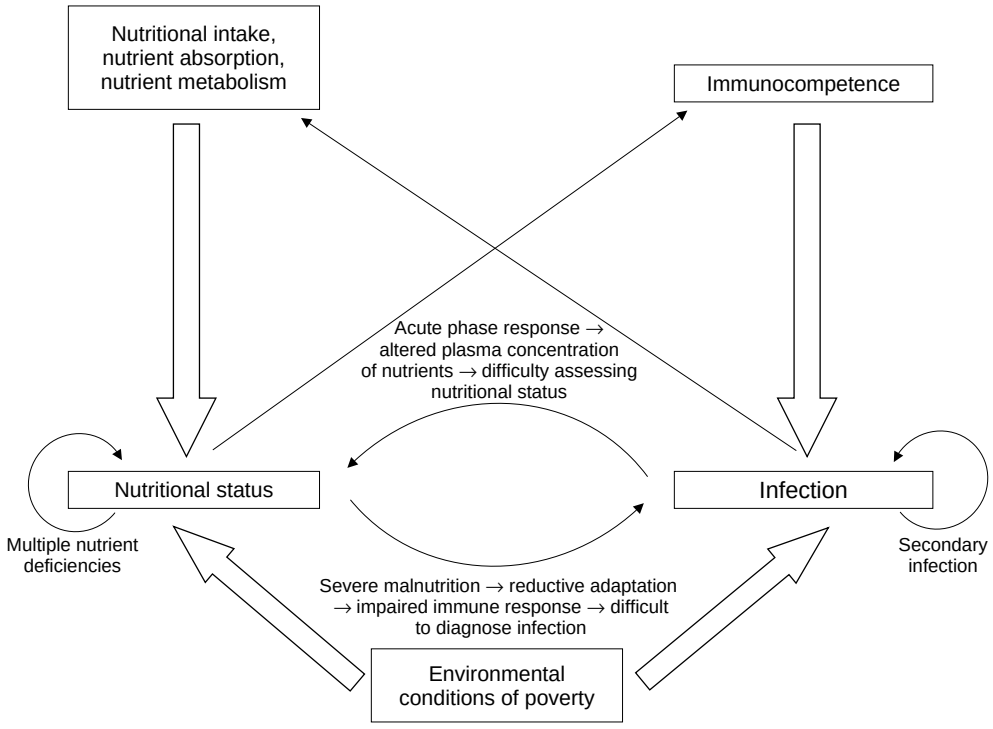


Figure 3 Confounding factors in the nutrition–infection paradigm.

See also: **Breast Feeding**. **Diarrheal Diseases**. **Fatty Acids**: Omega-3 Polyunsaturated. **Immunity**: Physiological Aspects; Effects of Iron and Zinc. **Infection**: Nutritional Management in Adults. **Lung Diseases**. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **Parasitism**. **Tuberculosis**: Nutrition and Susceptibility; Nutritional Management.

Further Reading

Calder PC, Field CJ, and Gill HS (2002) *Nutrition and Immune Function*. Oxon: CAB International.

Dreyfuss ML and Fawzi WW (2002) Micronutrients and vertical transmission of HIV-1. *American Journal of Clinical Nutrition* **75**(6): 959–970.

Doherty CP, Weaver LT, and Prentice AM (2002) Micronutrient supplementation and infection: a double-edged sword? *Journal of Pediatric Gastroenterology and Nutrition* **34**(4): 346–352.

Field CJ, Johnson IR, and Schley PD (2002) Nutrients and their role in host resistance to infection. *Journal of Leukocyte Biology* **71**(1): 16–32.

Gershwin ME, Nestel P, and Keen CL (2004) *Handbook of Nutrition and Immunity*. New Jersey: Humana Press.

Harbige LS (1996) Nutrition and immunity with emphasis on infection and autoimmune disease. Nutrition and Health **10**: 285–312.

Powanda MC and Beisel WR (2003) Metabolic effects of infection on protein and energy status. *Journal of Nutrition* **133**(1): 322S–327S.

Scrimshaw NS, Taylor CF, and Gordon JF (1968) *Interactions of Nutrition and Infection*. Monograph Series 57. Geneva: World Health Organization.

Scrimshaw NS and SanGiovanni JP (1997) Synergism of nutrition, infection and immunity: an overview. *American Journal of Clinical Nutrition* **66**: 464S–477S.

Suskins MS and Tontisirin K (2001) *Nutrition, Immunity and Infection in Infants and Children*. Philadelphia: Lippincott Williams and Wilkins.

Tomkins A and Watson F (1989) *Malnutrition and Infection: A review. Nutrition Policy Discussion Paper No.5 (ACC/SCN State of the Art Series)*. Geneva: United Nations.

Nutritional Management in Adults

J A Tayek, Harbor–UCLA Medical Center, Torrance, CA, USA

Metabolic and Nutritional Changes in Patients with Infection

An increased blood glucose concentration is the most common abnormality in the infected hospitalized patient. This section discusses the metabolic abnormalities in glucose, protein, and fat metabolism as well as abnormalities in specific nutrients in this population. Specific nutritional treatment plans are presented. In addition, the host response to injury and why

patients may not be able to become anabolic with conventional nutritional support are discussed. The acute phase response typifies the host's response to infection. Mechanisms to blunt the catabolic state are important because the extent of muscle wasting and weight loss is inversely correlated with long-term survival. The potential uses of conventional nutritional support and newer nutritional adjunctive techniques utilized for patients are discussed.

Glucose Utilization in Injury and Infection

In nearly all studies of glucose metabolism in patients with infection, injury, or cancer, there is a significant reduction in glucose utilization. This occurs even when the insulin concentrations are in the physiological range. This effect is not overcome even with administration of supraphysiological insulin concentrations. In sepsis, the insulin resistance associated with injury is due to defective insulin-mediated activation of the glycogen storage pathway. By approximately 7 h after the onset of injury, there is a reduction in glucose utilization via the nonoxidative pathway. This injury response persists until the source of injury, infection, or tumor is removed.

Hepatic Glucose Metabolism

During infection, the liver increases glucose production to defend against hypoglycemia. In fact, the increase in hepatic glucose production is the major reason why patients with infection have an elevated blood glucose concentration. For example, patients with active malaria can have an increase in fasting glucose concentration due to an increase in gluconeogenesis and overall glucose production. Approximately 75% of cancer patients, like patients with infection, also have an elevated rate of glucose production. Cancer patients also have a mild form of injury; approximately 75% have an elevated rate of hepatic glucose production. In 18 studies, hepatic glucose production for normals ranges between 1.6 and 3.0 mg/kg/min, with an average of 2.1 mg/kg/min. Glucose production for cancer patients without weight loss ranges from 1.7 to 5.1 mg/kg/min, with a mean of 2.75 mg/kg/min. This is a 30% increase in the fasting rate of hepatic glucose production. For cancer patients with weight loss, glucose production ranges from 2.3 to 3.3 mg/kg/min, with a mean of 2.96 mg/kg/min. This represents a 41% increase in the rate of hepatic glucose production. Not all cancer types have an elevation in hepatic glucose production. For example, head and neck cancer patients may not have an elevation in fasting hepatic glucose production, but it is commonly elevated in lung cancer patients, probably because they have an increased injury response. In cancer patients, the etiology for the elevated rate of fasting hepatic glucose production is not known. Early studies tested whether excessive growth hormone (GH) release in cancer patients might be responsible. However, there was no direct correlation between GH secretion pattern and hepatic glucose production. Furthermore, the administration of GH to cancer patients for a 3-day period failed to increase the rate of glucose production. Koea and Shaw suggested that the rate is related to the bulk of the tumor, and others have suggested it is related to cytokines or other factors. Earlier studies on normal volunteers demonstrated that the loss of the first-phase insulin response causes a delay in the normal inhibition of glucose production. Although the latter effect may explain postprandial hyperglycemia, it is an unlikely explanation for fasting hepatic glucose production.

Gluconeogenesis is elevated in head and neck cancer patients and also in lung cancer patients. Gluconeogenesis accounts for approximately 50% of the overall glucose production after an overnight fast. It was demonstrated that glucose carbon recycling was elevated in five of seven published studies. Glucose carbon recycling is an indicator of increased gluconeogenesis. The ability to measure gluconeogenesis was not possible in humans until recently, when a method using [U-^{13}C] glucose and isotopomer analysis was developed. The Cori cycle is increased in cancer patients and has been estimated to account for 300 kcal of energy loss per day. In 70% of published studies, cancer patients have a significant elevation in the rate of gluconeogenesis compared to normal weight-matched controls. Gluconeogenesis was directly related to the morning blood cortisol concentration in both the normal volunteers ($r = 0.913$, $p < 0.01$) and the cancer patients ($r = 0.595$, $p < 0.05$). In the septic host, the increase in glucose production is likely due to an elevation of multiple counterregulatory hormones (cortisol, GH, catecholamines, and glucagon) and cytokines (interleukin-1 (IL-1), tumor necrosis factor-α (TNF-α), etc.).

It is important to note that unlike diabetic patients with an elevated blood glucose concentration, cancer patients with an elevated glucose production rate frequently have a normal blood glucose concentration. Fasting glucose concentrations may be 110–120 mg/dl, which may be overlooked as a subtle indicator of an elevated glucose production rate. The increased rate may contribute to an increased energy cost. Data indicate that the resting energy expenditure is elevated in lung cancer patients and those with other types of cancer compared to weight-matched controls. As expected, energy expenditure is increased in most critically ill patients a few days after admission. However, the precise measurement of energy expenditure is difficult in this setting. Early

in the course of critically ill patients, one should focus on excellent blood glucose control. A total caloric intake of 20–25 kcal/kg/day should be provided to the nonthermal injured patient. Protein intake should be 1.5 g/kg body weight/day.

Unlike the normal fasting blood glucose that is seen in cancer patients, patients with injury or infection most commonly have an increase in blood glucose. This has been associated with a large increase in hospital mortality (**Table 1**). Hyperglycemia as a marker of intensive care unit (ICU) mortality may be greater in surgical patient compared to medical ICU patients. In a prospective randomized clinical trial in which intravenous insulin was provided to surgical patients, preventing the increase in blood glucose associated with injury and infection, there was significantly reduced mortality.

Protein Metabolism

Sepsis is associated with an increase in skeletal muscle catabolism and a reduction in the rate of skeletal protein synthesis. Both contribute to a large loss of lean body mass during injury and infection. Skeletal protein breakdown occurs more in the fast-twitch or white muscle fibers than in the red fibers. In addition to sepsis, injury and cancer are also associated with muscle wasting and malnutrition. The etiology is multifactorial, including poor dietary intake, insulin resistance, elevated resting energy expenditure, and other unknown factors. Muscle wasting is due to a combination of increased skeletal muscle protein catabolism and reduced skeletal muscle protein synthesis. For example, in an experimental model of cancer cachexia, protein synthesis was reduced in rats with several tumor types and it occurred at small tumor burdens. In humans with renal cell cancer, the rate of muscle protein synthesis was reduced. In this cancer host, the loss of skeletal muscle appears to be due in part to a reduced protein synthesis and in part to a normal rate of protein catabolism. This can occur even in the face of an adequate dietary intake.

Whole body protein metabolism can be measured in many ways. The most common isotope is that of the essential amino acid leucine. In the majority of studies on cancer, injury, and infected patients, the rate of plasma amino acid appearance or turnover is elevated. This rate of plasma appearance is a reflection of multiple sites of protein metabolism. The most important are the skeletal muscle, liver, and gastrointestinal (GI) mucosa. Other sites also play an important role, as does the tumor. Studies have demonstrated that the rate of plasma amino acid appearance is related to the bulk of the tumor mass. Measurements of protein metabolism in tumor tissue have demonstrated that the tissue has a very high fractional protein synthesis rate of 50–90% per day. This is similar to that of the liver, and it contrasts with a rate of 1–3% for the skeletal muscle. However, since the body is composed mostly of skeletal muscle, its overall contribution to whole body amino acid metabolism is large and it contributes to a significant proportion of plasma amino acid appearance rates. Data suggest that the increase in the protein catabolism in humans is via the effect of cytokines (IL-1, IL-6, and TNF) and the glucocorticoids, which are known to stimulate the ubiquitin–proteasome pathway of skeletal muscle protein catabolism. Earlier work demonstrates that TNF administration reduces skeletal muscle amino acid content by 20%, but it has no effect on skeletal muscle protein synthesis. The loss of amino acids without stimulation of protein synthesis suggests that TNF stimulates protein catabolism via a loss of amino acids from inside the skeletal muscle. This effect of TNF wanes after 6 h since animals studied at 60 h have a 30% increase in the rate of protein synthesis and a normal skeletal muscle amino acid content. The increased rate of protein synthesis probably reflects the recovery of the depleted amino acid pool due to

Table 1 Mean blood glucose concentrations, hospital mortality

Patients	*Controls*		*IV insulin*		*Reference*
	Glucose (mg/dl)	*Mortality (%)*	*Glucose (mg/dl)*	*Mortality (%)*	
1600 mixed ICU	152	20.9	131	14.8*	Krinsley (2004)
1548 C-T surgery	153	10.9	103	7.2*	Van den Berghe *et al.* (2001)
139 DM with acute MI	162	26.1[a]	153	18.6*	Malmeberg (1995)
620 DM with acute MI	162	43.9[b]	148	33.3*	Malmeberg (1999)
3554 DM with C-T surgery	213	5.3	177	2.4*	Furnary (2003)
Mean ± SEM	168 ± 11	21.4 ± 6.7	142 ± 12	15.3 ± 5.3	

[a]One-year mortality.
[b]Three-year mortality.
*$p < 0.01$ vs mortality at baseline.

earlier administration of TNF. The increased intracellular concentration of amino acids in the skeletal muscle may stimulate synthesis. The direct effect of TNF 60 h after a single administration is not likely since it has a short half-life. Chronic administration results in a reduction in whole body protein synthesis and a net loss of skeletal muscle protein but an increase in liver protein synthesis. An increase in the thyroid hormone triiodothyronine also plays an important role in promoting protein breakdown in both the ubiquitin–proteasome pathway and the lysomal pathway. However, under most conditions, patients with malignancy have either a normal or a reduced triiodothyronine concentration. Similar processes are responsible for the loss of protein seen in infection.

Data suggest that humans make and break down approximately 300 g of protein per day, which is exchanged and reused. This is meditated by the flow of amino acids into and out of cells. Since the amino acid pool is small (only 60 g), the turnover is large. An average person ingests approximately 70 g of protein per day and loses approximately 70 g per day in the form of nitrogen. The cellular proteins, including muscle and extracellular proteins, are approximately 10 400 g. These proteins are broken down and reused at various rates. The key to a small intake of amino acids in the diet is the reutilization of amino acids locally inside the cell and the maintenance of the plasma amino acid pool. The amino acid pool is only 0.6% of the whole body amino acid content, but it plays a vital role in the maintenance of protein synthesis.

Cancer patients who have an elevated plasma amino acid appearance rate survive and those with a normal rate have a worse survival. In one study, stage D colorectal cancer patients who were able to sustain an increased whole body protein metabolism over a 3-month period, as measured by amino acid kinetics, survived and those who had a normal or reduced rate died. Although fasting plasma glucose concentrations were greater in the survivors (100 ± 2 vs 92 ± 3 mg/dl), there was no difference in glucose production rate, age, and body weight. Carcinoembryonic Antigen (CEA) concentrations were higher in the patients who died, which suggests that they had a larger tumor burden. There may be subgroups of patients who are able to mount an acute phase response, which may improve survival. It is not known why some patients mount an increased amino acid appearance rate with cancer, and further research is needed to confirm that it may predict survival. Historically, an elevated plasma amino acid appearance rate was believed to represent protein wasting, but recent data suggest that an elevated rate of whole body protein metabolism may not reflect a maladaptive processes but rather a healthy response to the tumor. An adequate acute phase response to tumor may reflect a greater fight against cancer. The absence of a response may be unfortunate, as data from patients with colorectal carcinoma suggest. Unfortunately, there are no similar data from infected patients for this comparison.

Lipid Metabolism

Energy in the body is stored mainly in body fat, which is depleted during the wasting process. This process is normally increased during fasting without tumor or injury. When the patient has a tumor, there is a metabolic response to the injury that also promotes lipid mobilization. Several authors have implicated a lipid mobilization factor as being responsible for this process, which is believed to occur in both infection and cancer. Data suggest that this factor may also be responsible for the depletion of liver glycogen in cancer cachexia. This factor(s) increases lipolysis and plasma triglyceride concentrations. The former effect may be due to an increase in the hormone-sensitive lipase and the latter effect due to inhibition of lipoprotein lipase activity. However, the exact factor(s) that is responsible for these effects is not known.

Cancer patients with weight loss have an increase in whole body lipid turnover measured by radioactively labeled fatty acids. However, when weight loss is prevented, there is no increase in the rate of lipolysis. Similarly, the rates of lipid oxidation are normal in cancer patients compared to weight-matched controls. In more severe injury, as seen in sepsis, the rate of lipolysis is increased.

Hormonal Response to Injury, Infection, and Cancer

Infection, cancer, or any injury to the body result in an increase in counterregulatory hormones as well as insulin concentration. As a result of cancer, sepsis, or injury, many patients develop the syndrome of insulin resistance even though they had no history of diabetes prior to cancer. In cancer patients, when the overall injury is smaller, many studies have failed to demonstrate an elevation in counterregulatory hormones. Mild elevations in cortisol concentrations may contribute to the protein catabolism and increased gluconeogenesis. When serum insulin is measured with a sensitive assay, cancer patients demonstrate a small but significant elevation in serum insulin concentration. This is consistent with the observation that these patients have insulin resistance. Cancer patients, like diabetics, have a reduced glucose utilization and loss of the first-phase insulin

response, and many have an increased fasting hepatic glucose production rate. As mentioned previously, underweight cancer patients frequently have increased fatty acid oxidation and plasma fatty acid appearance rates. Triglyceride hydrolysis involves much more than fat oxidation, so albumin-bound fatty acids are used partially for energy but many are utilized for reesterification or substrate cycling back to triglyceride.

The rise in serum cortisol as the host's response to the tumor is one of many factors that are responsible for the development of insulin resistance. Insulin resistance is easy to diagnosis because the patient's fasting glucose will be elevated. An elevated fasting glucose level of approximately 110 mg/dl is a good marker of insulin resistance. This is not likely seen in mild injury alone unless the patient has a predisposition to the development of diabetes mellitus. Although insulin resistance is present, the presence of frank diabetes (blood glucose level >126 mg/dl or >7 mm) is not common in cancer or mild injury. It is more common in patients with severe infection or injury. Although most of the counterregulatory hormones are usually normal, serum cortisol and/or glucagon can be mildly elevated. Newer glucagon assays measure the normal value as 35–45 ng/ml, so a significant increase in injury can be detected, which was difficult to do with the older Unger assay. Recent data from pancreatic cancer patients have shown elevated glucagon concentrations, which may be contributing to the development of diabetes. Earlier work found that GH secretion was increased in cancer patients by 24-h analysis and by random sampling. However, after careful study, the increase in GH does not appear to have a major influence on hepatic glucose metabolism. Although there may be a small effect on glycogen breakdown, the major effect is likely via inhibition of glucose utilization in the skeletal muscle.

The sick euthyroid state, in which total triiodothyronine (T_3) concentrations are reduced in severely injured and infected patients, is common. This is likely a normal response to conserve energy in the injured person as the body's ability to convert the stored form of a thyroid hormone (thyroxine (T_4)) into the active form of thyroid hormone, T_3, becomes impaired. T_4 is converted to an inactive thyroid hormone known as reverse-T_3 hormone (rT_3). This event may have evolved as a necessary energy-saving response during a severe injury or illness to reduce the known contribution of T_3 to resting energy expenditure. The low T_3 syndrome is an adaptive way to reduce the normal day-to-day effect of T_3 on resting energy expenditure. This process can occur in the aggressive cancers, for which the patient's response is similar to that of an injury response.

In septic and injured patients, all counterregulatory hormones are routinely elevated, contributing to an increase in protein catabolism, glucose production, gluconeogenesis, and glycogen breakdown and a major reduction in glucose utilization and anabolism.

Acute Phase Response

The development of injury, infection, or cancer cachexia elicits an acute phase response. This is one of the most basic responses of the body to defend itself against injury. Phylogenetically, this response could be considered the most primitive response of the body. This stereotypical response is similar for injury from an accident, burn, infection, foreign objects, and, in some cases, from a tumor. Unfortunately, this response does not occur for most tumors, but it is seen when the malignancy presents with infection, such as in lung cancer, or in other more aggressive malignancy, such as seen in leukemia. The host develops a response that includes reductions in serum iron and zinc levels, increased serum copper and ceruloplasmin levels, alterations in amino acid distribution and metabolism, an increase in acute phase globulin synthesis, and gluconeogenesis. Although not common, fever can occur, and a negative nitrogen balance results. The tumor can elicit a sequence of events that include changes in cytokine levels as well as several classical hormone levels. For example, a malignant process in the lung will attract monocytes that will be transformed into macrophages at the tissue site of tumor. These macrophages will secrete proteins known as cytokines and other peptides that can attract other white blood cells and initiate an inflammatory response common to many types of injury. Cytokines include TNF-α and IL-1 to IL-20. TNF and other cytokines circulate to the liver, inhibit albumin syntheses, and stimulate the synthesis of acute phase proteins. Acute phase proteins include C-reactive protein, which promotes phagocytosis, modulates the cellular immune response, and inhibits the migration of white blood cells into the tissues; α_1-antichymotrypsin, which minimized tissue damage due to phagocytosis and reduces intravascular coagulation; and α_2-macroglobulin, which forms complexes with proteases and removes then from circulation, maintains antibody production, and promotes granulopoiesis and other acute phase proteins. Unfortunately, the majority of tumors do not elicit a large acute phase response. This limited response may result in a decreased inflammatory and tumorcidal effect.

Urine Urea Nitrogen Loss as a Marker of Catabolism

As part of the host response to injury, infection, or tumor, patients frequently lose protein in the urine in the form of nitrogen. For example, 16 g of urea nitrogen in the urine per day represents a 1-lb loss of lean body mass, such as muscle tissue. In some aggressive cancers, urea nitrogen loss can be as high as 24 g per day. The loss of 1 g of urinary urea nitrogen is equal to 6.25 g of dry protein. A total of 6.25 g of dry protein is equal to approximately 1 oz. of lean body mass. A loss of 16 g of urinary urea is equal to the loss of 1 lb of skeletal muscle or lean body mass per day. Specific areas of lean body mass loss that may result in a functional impairment of the respiratory muscles include the diaphragm, heart muscle, and GI mucosa. The loss of lean body mass in these areas can contribute to the development of respiratory failure, heart failure, and diarrhea, respectively. The rapid development of malnutrition can occur in patients with infection due to large losses of lean body mass per day.

Vitamin Deficiencies

Reduced serum concentrations of several vitamins, including vitamins C and E, have been reported in patients with sepsis. In one study, the administration of additional vitamin E and C resulted in a significant reduction in 28-day mortality (67.5 vs 45.7). Clearly, cancer patients with a poor intake can have deficiencies of many vitamins. For example, cancer patients have been noted to have significant reductions in plasma levels of many of vitamins, especially folate, vitamin A, and vitamin C.

Vitamin C and Vitamin A Patients with a premalignant lesion called leukoplakia also have reductions in plasma levels of retinol (vitamin A), β-carotene, and vitamin C. A study of healthly elderly demonstrated that approximately 20% had a reduced vitamin C level (<0.5 mg/dl) and 10% had a reduced serum vitamin A level (<33 μg/dl). The replacement of multiple vitamins and minerals with 80 mg of vitamin C and 15 000 IU of vitamin A per day for 1 year resulted in a significant reduction in the number of days associated with an infection-related illnesses (48 ± 7 to 23 ± 5 days per year). The multiple vitamin and mineral supplement improved the lymphocyte response to phytohemagglutin and the natural killer cell activity. In another study, the administration of a multivitamin for 1 year demonstrated a 41% reduction in infectious illnesses. In addition, there was a 63% reduction in infection-related absenteeism compared to that of placebo-treated individuals. The administration a MVI to pregnant HIV mothers also reduced HIV progression and mortality (24.7 vs 31.1% mortality, $p < 0.05$).

Vitamin deficiency states are difficult to diagnose. Plasma levels of vitamins are not the best way to assess deficiency. Vitamin C decreases during injury. Although plasma vitamin C concentrations reflect whole body stores, the measurement of plasma vitamin A (retinol) is not the best marker of an actual deficiency state. Liver vitamin A measurements may be a better marker. Patients who die of cancer and subsequent infections have an 18% incidence of moderate liver deficiency of vitamin A at autopsy. Serum vitamin A (retinol) levels are low in up to 92% of patients with serious infections. This depletion of liver stores of vitamin A may be due to excessive loss of retinol in the urine in patients with sepsis. In contrast to what is noted in patients with cancer or serious infections, trauma patients who die within 7 days of hospitalization only have a 2% incidence of severe liver vitamin A deficiency. Vitamin A can be provided by supplementation dietary intake, parenteral intake, or intramuscular vitamin A administration. In addition to the changes in folate, vitamin A, and vitamin C mentioned previously, excessive losses of several vitamins have been observed in patients receiving medications that interfere with normal utilization or elimination (**Table 2**).

Table 2 Drug-induced nutrient deficiencies

Drug	*Nutrient(s) affected*
Steroids	Vitamin A, potassium
Phenothiazines	Vitamin B_2
Tricyclic antidepressants	Vitamin B_2
Hydralazine	Vitamin B_6
Isoniazid	Vitamin B_6, niacin
Penicillamine	Vitamin B_6
Ammonium chloride	Vitamin C
Aspirin	Vitamin C
Phenobarbital and phenytoin	Vitamin C, vitamin D
Tetracycline	Vitamin C
Coumadin	Vitamin K
Estrogen and progesterone compounds	Folic acid, vitamin B_6
Aminoglycoside	Magnesium, zinc
Platinum	Magnesium, zinc
Diphenylhydantoin	Niacin
Antacid	Phosphorus, phosphates
Diuretics	Sodium, potassium, magnesium, zinc
Laxatives	Sodium, potassium, magnesium
Cholestyramine	Triglycerides, fat-soluble vitamins

Mineral Deficiencies

Multiple elevated cytokines are likely responsible for the commonly observed reduction in serum mineral concentrations. This is known as part of the cytokine-mediated inflammatory response. In addition, in patients with injury, infection, or cancer, the reduced mineral content may also occur secondary to poor oral intake, increased requirements, and excessive urinary and stool losses.

Magnesium Total body stores are 2028 g of magnesium. Communications with several experts on magnesium and current work on the antiarrhythmic actions of magnesium suggest a that the commonly used normal values for serum magnesium levels should be increased from 1.7–2.3 mg/dl to 2.0–2.6 mg/dl. Large losses can occur in conditions such as diarrhea, in which the stool may have up to 12 meq of magnesium per liter and the urine may have up to 25 meq per day. Large urinary losses can occur in cancer patients given aminoglycides, diuretics, and ketoconazole. Furthermore, large losses can occur in some of the intestinal fluids (**Table 3**) in cancer and other operative patients who develop GI fistulas.

Zinc Total body stores are only 2 or 3 g of zinc. Zinc concentration in the blood decreases as an early response to cytokines. This is commonly seen in many different types of injury as well as in cancer patients. There are minor tissue stores of zinc in skin, bone, and intestine. Zinc is redistributed to liver, bone marrow, thymus, and the site of injury or inflammation. This redistribution is mediated by IL-1 and the other cytokines secreted from macrophages. In hospitalized cancer patients, a reduced serum zinc concentration ($<70\,\mu g/dl$) is not uncommon. The administration of approximately 50 mg of zinc per day is associated with a normalization of the zinc level after 3 weeks of feeding. Fifteen percent of healthy elderly have been found to have reduced serum zinc levels ($<67\,\mu g/dl$). The replacement of a multivitamin with 14 mg of zinc per day for 1 year resulted in a significant reduction in the number of days associated with infection-related illnesses (48 ± 7 to 23 ± 5 days per year). This vitamin and mineral supplementation improved the lymphocyte response to phytohemagglutin and the natural killer cell activity. There was no change in the placebo-treated group. Zinc supplementation in hospitalized patients may help with normal immune response for minor infection and wound healing. Zinc is needed for cell mitosis and cell proliferation. It has also been demonstrated to improve wound healing in patients provided 600 mg of zinc sulfate (136 mg of elemental zinc) orally per day who had a serum zinc level on admission of less than 100 μg/dl. In this double-blind study, the healing rate increased more than twofold in those randomized to receive zinc supplementation. In addition, large losses of zinc can occur via intestinal losses (**Table 3**). It is important to note that intestinal fluids can contain up to 17 mg of zinc per liter, so the replacement rate of zinc should take into account the abnormal sources of zinc loss as well as the routine nutritional requirements.

Copper Total body stores are very small at 60–80 mg. Serum copper status is normal or increased compared to that of serum zinc, and cytokines are also believed to be responsible for these changes. The benefits of or rational for these increased concentrations are not known.

Iron Total body stores are 3.5–4.5 g of iron. An increase in cytokines also contributes to the observed decrease in serum iron concentration. This is a mediated response to cancer, injury, or infection. The exact mechanism is not known, but iron is stored in Kupffer cells of the liver until the injury wanes. This is probably a beneficial effect

Table 3 Electrolyte contents of body fluids

Body fluid	*Electrolyte and mineral concentration (meq/l)*					
	Sodium	*Potassium*	*Chloride*	*Bicarbonate*	*Magnesium*	*Zinc (mg)*
Bile	145	5	100	15–60	1–2	—
Colonic fluids	50	30–70	15–40	30	6–12	17
Diarrheal fluids	50	35	40	45	1–13	17
Duodenum	130	5–10	90	10	1–2	12
Ileal fluids	140	10–20	100	20–30	6–12	17
Pancreatatic juice	140	5	75	70–115	0.5	—
Saliva	10	20–30	15	50	0.6	—
Stomach fluids	100	10	120	0	0.9	—
Urine	60–120	30–70	60–120	—	5	0.1–0.5
Urine post Lasix	15× normal	2× normal	—	—	20× normal	—

since many microbes use iron as a source of energy. Iron administration should be restricted in patients who have a serious infection because it has been shown to cause harm with fungal, parasitic, malarial, or other types of low-grade or quiescent infections.

Summary

Vitamins and minerals act as cofactors for essential processes in health and in illness. The requirements for the healthy person have been well established and are published as the recommended daily requirements (**Tables 4** and **5**). The exact needs for the infected, injured, or cancer patient are not well documented and evaluations are in progress. Reduced levels of vitamin C, vitamin A, copper, manganese, and zinc have been observed, and all of these are associated with poor wound healing. Wound dehiscence is eight times more common with decreased vitamin C levels. This is probably due to the fact that vitamin C enhances capillary formation and decreases capillary fragility, is a necessary component of complement, and is key to the hydroxylation of proline and lysine in collagen synthesis. Vitamin A enhances collagen synthesis and crosslinking of new collagen, enhances epithelialization, and antagonizes the inhibitory effects of glucocorticoids on cell membranes. Manganese is a cofactor in the glycosylation of hydroxylysine in procollagen. Copper acts a cofactor in the polymerization of the collagen molecule and as a cofactor in the formation of collagen crosslinks.

Table 4 Adult daily vitamin nutritional requirements (RDA, 1989)

Nutrient	*Oral*	*Intravenous*	*Special requirements (diagnosis)*
Vitamin A	3300 IU/day	3300 IU/day (1 mg)	5000+ IU/day (serious infections)
Vitamin B (Biotin)	100 μg/day	60 μg/day	
Vitamin B (Folic acid)	0.2 mg/day	0.4 mg/day	5 mg/day (ICU patients/thrombocytopenia)
Vitamin B (Niacin)	20 mg/day	40 mg/day	
Vitamin B_1 (Thiamin)	1.5 mg/day	3 mg/day	50 mg/day (alcoholics/Wernike–Korsakoff)
Vitamin B_2 (Riboflavin)	1.8 mg/day	3.6 mg/day	
Vitamin B_6 (Pyridoxine)	2 mg/day	4 mg/day	
Vitamin B_{12}	2 μg/day	5 μg/day	
Vitamin C	60 mg/day	100 mg/day	
Vitamin D	400 IU/day	200 IU/day (5 μg)	
Vitamin E	10 mg/day	10 mg	
Vitamin K	80 μg/day	[a]	
Pantothenic acid	7 mg/day	15 mg/day	

[a]Vitamin K is routinely given as 10 mg SQ on admission and then every Monday.

Table 5 Daily nutritional requirements

Nutrient	*Adult daily nutritional requirements*		
	Oral	*IV*	*Special requirements (diagnosis)*
Macronutrients			
Protein	1.5–2.0 g/kg	1.5–2.0 g/kg	2–3 g/kg (thermal injury)
Glucose	20–25 kcal/kg	20–25 kcal/kg	3000 kcal goal in alcoholic liver disease patients
Lipid	4% of kcals	4% of kcals	Can administer up to 60% of calories to prevent hyperglycemia
Micronutrients			
Sodium	60–150 meq	60–150 meq	
Potassium	40–80 meq	40–80 meq	
Chloride	40–100 meq	40–100 meq	
Acetate	10–40 meq	10–40 meq	
Phosphorus	10–60 mmol	10–60 mmol	
Calcium	5–20 meq	5–20 meq	100 meq or more severe hypocalcemia and hungry bone syndrome
Magnesium	10–20 meq	10–20 meq	50–100 meq (cardiac arrthymias, diarrhea)
Zinc	3 mg	2.5–4 mg	10–100 mg (diarrhea, fistula, wounds)
Copper	1.5–3 mg	1–1.5 mg	
Chromium	50–200 μg	10–15 μg	40 μg (diarrhea, gastrointestinal losses)
Molybdenum	75–250 μg	100–200 μg	
Manganese	2–5 mg	150–800 μg	
Selenium	40–120 μg	40–120 μg	120–200 μg (thermal injury, wounds)

Nutritional Assessment and Predictors of Hospital Outcome

Markers of Nutritional Assessment

Conventional nutritional assessment in injured, infected, or cancer patients is of clinical value. Body weight and history of weight loss is one of the best indicators of survival in patients with infection or cancer. In addition, serum albumin concentration upon admission is probably one of the best predictors of hospital survival (**Table 2**). Serum albumin is commonly used as an indicator of nutritional status. Its level provides the clinician with an index of visceral and somatic protein stores for most medical illnesses. A level less than 3.0 is considered malnutrition and may also be called hypoalbuminemic malnutrition or protein malnutrition. Exceptions to this include the isolated starved state such as anorexia nervosa, severe edema, and the rare case of congenital analbuminemia. Serum albumin has a 21-day half-life, and this can reflect processes that have been ongoing for a few weeks. The benefit of serum albumin is that it is also an inverse acute phase reactant. The further it declines, the more severe the injury response on top of the severity of malnutrition at the time of the injury or cancer.

Predictors of Clinical Outcome

The best marker of injury is serum albumin concentration. It is an excellent predictor of survival in patients with cancer and other types of illnesses (**Table 6**). More than 20 studies have shown that a serum albumin level below normal can be used to predict disease outcomes in many groups of patients. One of the first studies in this area was a Veterans Administration study in which 30-day mortality rates were evaluated for a total of 2060 consecutive medical and surgical admissions. Investigators found that 24.7% of the patient population had a low albumin level defined as 3.4 g/dl or lower. The 30-day mortality rate for hypoalbuminemia patients was 24.6% compared to 1.7% for patients with a normal albumin level. These investigators demonstrated an excellent correlation between serum albumin levels and 30-day mortality rates. A 1-g decrease in serum albumin levels (3.5 to 2.5 g/dl) translated into a 33% increase in mortality. Patients with an average albumin level of 1.8 g/dl had a mortality rate of 65%. It is interesting to note that of 15 hypoalbuminemia patients in this study who were provided with total parenteral nutrition, only 1 died (7% mortality).

Protein malnutrition is associated with a greater risk for infection, especially fungal infections. In one

Table 6 Serum albumin and mortality

Patient population	*Mortality*				
	With normal albumin		*With low albumin (%)*	*Increased risk (-fold)*	*Albumin cutoff level (g/dl)*
	n	*%*			
VA hospital	2060	1.7	24.7	14.7	3.5
Medical and surgical patients	500	1.3	7.9	6.1	3.5
Hodgkins	586	1.0	10	10.0	3.5
Lung CA	59	49	85	1.7	3.4
VA hospital	152	3.3	25.8	7.8	3.5
Surgical patients	243	4.7	23	4.9	3.5
Malnutrition	92	8.0	40	5.0	3.5
Surgery (colorectal)	83	3.0	28	9.3	3.5
ETOH hepatitis	352	2.0	19.8	9.9	3.5
Pneumonia	38	0	100	—	3.0
Cirrhosis	139	32	52	1.6	2.9
ICU patients	55	10	76	7.6	3.0
Cardiovascular disease	7735	0.0	2.0	—	4.0
Trauma	34	15.4	28.6	1.9	3.5
Sepsis	199	0.7	15.9	22.7	2.9
Pneumonia	456	2.1	8.3	4.0	3.5
Multiple meyloma	23	25	50	2.0	3.0
CABG/cardiac valve surgery	5156	0.2	0.9	5.7	2.5
Preoperative (VA hospital)	54 215	2.0	10.3	5.1	3.5
Beth Israel Hospital	15 511	4.0	14.0	3.5	3.4
Hemodialysis	13 473	8.0	16.6	2.1	4.0
Average ± SEM	4275	7.8 ± 2.8	31 ± 7	15 ± 2-fold risk	3.4 ± 0.1
Total No. of patients	101 178				

study, the most important risk factor for the development of candidemia was malnutrition. A reduced serum albumin level is an independent risk factor for nosocomial infections. The greater the protein malnutrition, the greater the risk for nosocomial infections.

In summary, serum albumin concentrations provide the clinician with a tool to help predict recovery or mortality. Albumin levels should be monitored at regular intervals (every 3 or 4 days) for hospitalized patients who are ill and at risk for malnutrition. Once hypoalbuminemia is documented, it is not an ideal indicator of nutritional rehabilitation since it returns to normal slowly (21-day half-life) and lags behind other indices of nutritional status, such as transferrin (7-day half-life), prealbumin (1-day half-life), or retinol binding protein (4-h half-life). Albumin replacement does not reverse the metabolic process that the hypoalbuminemia state represents. The reduced level of protein reserves in the patient and the severity of the metabolic injury or cancer are the two most important determinants of serum albumin level.

Nutritional Diagnoses Commonly Seen in Hospitalized Patients

The diagnosis of malnutrition is made by taking a good history and obtaining a physical exam. It is important to ask the patient if he or she has been able to maintain his or her appetite and body weight during the past several months. A history of a recent hospitalization is also important to note due to the common development of protein malnutrition during hospital stay. The physical exam involves inspection of the muscle mass, especially noting a loss of temporals muscle, 'squaring off' of the deltoid muscle, and loss of the thigh muscles. Obtaining a measured body weight should be standard on all admissions, and this weight should be followed on a daily basis.

Up to 50% of hospitalized surgical and medical patients have either hypoalbuminemia or marasmic-type malnutrition. Hypoalbuminemia or protein malnutrition can be diagnosed with the measurement of reduced albumin, transferrin, prealbumin, or retinol binding protein levels. Albumin levels are most commonly used to make the diagnosis. Marasmic malnutrition is the diagnosis for anyone who has lost 20% of usual body weight during the preceding 3–6 months or who is less than 90% of ideal body weight. Of these two types of malnutrition, protein malnutrition is most common. The presence of hypoalbuminemia malnutrition in one study was associated with a 4-fold increase in dying and a 2.5-fold increased risk of developing a nosocomial infection and sepsis. As indicated in **Table 6**, a low serum albumin level predicts a significant increase in mortality across many diseases.

Loss of Lean Body Mass

The use of body weight as an index of muscle mass in the cancer patient is very difficult due to the possible fluid shifts that occur in the extracellular compartment. Body weight can be divided into three compartments: extracellular mass, lean body mass, and fat mass. Extracellular mass is known to increase in malnutrition and as a result of hypoalbuminemia. An increase in extracellular fluid occurs more commonly in the malnourished patient. A large portion of the fluid shift noted in cancer patients is due to a reduction in the plasma colloid oncotic pressure. Lean body mass is the mixture of skeletal muscle, plasma proteins, skin, skeleton, and visceral organs. The skin and skeleton account for 50% of the lean body mass. Currently, there are no convenient markers to determine the loss of nitrogen from either skin or skeleton. The plasma proteins account for only 2% of the lean body mass, but albumin measurement can reflect the overall status of the lean body mass. The viscera accounts for 12% of the lean body mass, and decreases in some of visceral sizes (gut atrophy and cardiac atrophy) are noted in cancer patients. Unfortunately, there is no convenient marker of loss of lean body mass that originates from the visceral organs. On the other hand, urine creatinine is a marker of skeletal muscle mass. The skeletal muscle accounts for 35% of the lean body mass, and it provides the major storage area for amino acids needed during illness. The standard way to assess the size of the skeletal mass is to determine the creatinine height index by collecting 24-h urine and comparing the value to normal values of creatinine excretion for age, sex, and height. A simplified way is to collect 24-h urine and divide the total amount by the ideal body weight based on the patient's height. The normal value for an adult male is 23 mg/kg of ideal body weight, and that for a female is 18 mg/kg. A value of 10% less than normal would be consistent with a 10% loss in the muscle mass for unit height. A value of 20% less than the lower range of normal would classify patients as having mild muscle loss. A 20–40% loss would classify them as having a moderate loss, and a 40% or greater reduction in the creatinine per weight would document severe muscle loss. The most accurate estimate is to obtain urine creatinine over a 3-day period and to repeat at intervals to

document the loss of muscle mass over an extended period of time. Dietary creatine and creatinine intake has only a minor influence (<20%) on urinary creatinine in the normal eating individual. Changes in dietary intake may influence the accuracy of the collection, but repeating the values over 3 days will help average variations in dietary intake. Impairment of renal function reduces the normal creatinine excretion and excludes the creatinine height index as a marker of muscle mass.

Elevated Resting Energy Expenditure

Resting energy expenditure (REE) is directly linked to the size of the lean body mass. REE is difficult to determine accurately in volunteers since the method of indirect calorimetry has variations when the same individual is restudied. Several studies have demonstrated an elevated rate of energy expenditure when compared to controls of similar weight. The use of D_2O^{18} has helped in the estimate of energy expenditure and will improve our understanding of energy expenditure in the future.

Nutritional Feeding of the Patient: Enteral versus Parenteral

Vitamins and Minerals

The standard oral and intravenous vitamin intake and what is currently being given at Harbor-UCLA Medical Center and UCLA Medical Center are listed in **Table 4**. Also included are the few exceptions to the routine intravenous amounts for both **Tables 4** and **5**. The mineral and trace element requirements are listed in **Table 5**. These vitamin, mineral, and trace mineral recommendations are for hospitalized cancer patients and noncancer patients who are hospitalized. They should not have oliguric renal failure or cholastatic liver disease. In acute oliguric renal failure, vitamins A and D should be reduced or eliminated from the enteral or parenteral solutions. Potassium, phosphorus, magnesium, zinc, and selenium should be reduced or eliminated. Iron and chromium are known to accumulate in renal failure and should be removed from the parenteral or enteral formulations. In cholastatic liver disease, the trace elements copper and manganese are excreted via the biliary tree in the bile and should be reduced or eliminated to prevent toxicity. In comparison, large amounts of electrolytes and minerals can be lost in gastrointestinal fluids and in urine (**Table 3**). It is essential to replace the estimated amounts lost on a daily basis in the parenteral nutrition.

Enteral versus Parenteral Feeding

In all situations, if the gut is functional, then it should be used as the route of calorie administration. Gut atrophy predisposes bacterial and fungal colonization and subsequent invasion associated with bacteremia. Sepsis due to microbial or toxin translocation into the portal system is a frequent source of fever evaluations that do not indicate an obvious source of infection. Utilization of the GI track can reduce the incidence of bacterial translocation.

Enteral Products

Enteral nutrition is best taken by mouth if the patient can ingest the required amount. If the patient cannot, then either supplements or full tube feeding is the method of choice. Protein in the peptide form is better absorbed than the free amino acid form due to specific transporters in the small intestines for amino acids, dipeptides, and tripeptides. Feeding tube placement is best in the small bowel up to the ligament of Treitz. This can be obtained best by the direct use of fluoroscopy or may be obtained by the passage of the feeding tube into the small bowel by a corkscrew technique, in which the distal tip of the feeding tube is bent at an approximately 30° angle with the wire stilet in place. Upon placement into the stomach, the tube is rotated so that the tip may pass via the pylorus into the small bowel. The infusion of enteral products into the small bowel will reduce the incidence of aspiration because the infusion is below the pylorus. Intubated patients have a low risk for aspiration due to the endotracheal cuff, so placement of a feeding tube into the small bowel is less essential.

Supplementation of enteral products with higher than standard amounts of the amino acid arginine has been done to enhance immune function. Published data on its beneficial effect in surgical patients have demonstrated some benefit; however, data from nonsurgical patients suggest harm. Immunonutrition should not be given to patients with severe infection, especially patients with pneumonia.

Branched-chain amino acid-enriched enteral products are available and have been shown to improve mental function and mortality in patients with hepatic encephalopathy. Albumin synthesis is also stimulated by branched-chain-enriched amino acid solutions. However, additional branched-chain amino acids did not improved morbidity or mortality in trauma or septic patients randomized to receive branched-chain-enriched amino acids compared to conventional feeding.

Glutamine-enriched enteral formulas are very common. There are many enteral products used in hospitalized patients and for home enteral nutritional support. These can be found at several enteral nutrition pharmaceutical Web sites.

The choice of lipid composition in enteral products is a field that is rapidly evolving, and this is an important decision to be made by the clinician depending on the type of disease being treated. The use of omega-3-enriched fatty acids in the enteral product (fish oil-enriched) has been associated with an ability to modify the inflammatory response that may be related to the increased arachondic acid metabolism and a decrease in the omega-6 pathway fatty acid metabolism. Unfortunately, most commercially available enteral products that have omega-3 fatty acids also have other additives, such as arginine, glutamine, and nucleotides, so that the benefits attributed to the use of an omega-3-enriched fatty aid enteral diet await future clinical studies.

Energy Intake for Patients with Malnutrition

The diagnosis of protein malnutrition can be made when the serum albumin level is less than 2.8 g/dl. Many of these patients have a 20% weight loss during the preceding 3 months, or they have a reduced ideal body weight (<90% for height). Patients at high risk for the development of malnutrition are those who are unlikely to ingest a minimum of 1500 kcal by day 5.

There are currently only three studies that support the importance of energy intake in malnourished patients. Elderly hospitalized patients who consume less than 50% of their estimated maintenance caloric requirement have an 8-fold increase in hospital mortality (11.8 vs 1.5%). This suggests that an intake of less than 1000 kcal may not be helpful. In a prospective study providing approximately 400 additional calories as 'sip feeds,' reduced mortality was seen in severely malnourished (body mass index <5th percentile), medically ill elderly patients. In this study, patients were randomized to receive 120 ml of enteral supplements provided by the registered nurse three times per day or provided no additional sip feeds. Patients who received the sip feeds had a significantly better energy intake (1409 kcal) than nonsupplemented patients (1090 kcal), and they had an increased overall weight gain compared with a loss in the controls. Patients in the severely undernourished group who received intervention had a significant reduction in mortality compared to controls (15 vs 35%, $p < 0.05$). The less malnourished or normals did not demonstrate the same benefit. In the third study, patients with less than 25% of recommended calorie intake (<600 kcal) had a 3.7-fold increased rate of nosocomial bloodstream infections. Candida and coagulase-negative *Staphylococcus* accounted for 63% of the nosocomial infections, with candida accounting for 29%.

See also: **Anemia**: Iron-Deficiency Anemia. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements; Deficiency States. **Cancer**: Epidemiology of Gastrointestinal Cancers Other Than Colorectal Cancers; Epidemiology of Lung Cancer. **Carbohydrates**: Regulation of Metabolism. **Cholesterol**: Sources, Absorption, Function and Metabolism. **Copper**. **Cytokines**. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Fatty Acids**: Metabolism. **Glucose**: Chemistry and Dietary Sources; Metabolism and Maintenance of Blood Glucose Level; Glucose Tolerance. **Iodine**: Deficiency Disorders. **Iron**. **Lipids**: Chemistry and Classification. **Magnesium**. **Malnutrition**: Secondary, Diagnosis and Management. **Nutritional Assessment**: Anthropometry; Biochemical Indices; Clinical Examination. **Nutritional Support**: Adults, Enteral; Adults, Parenteral. **Protein**: Deficiency. **Vitamin A**: Physiology; Deficiency and Interventions. **Zinc**: Deficiency in Developing Countries, Intervention Studies.

Further Reading

Barringer TA, Kirk JK, Santaniello AC, Foley KL, and Michielutte R (2003) Effect of multivitamin and mineral supplement in infection and quality of life. *Annals of Internal Medicine* **138**: 365–371.

Carson GL (2004) Insulin resistance in human sepsis: Implications for nutritional and metabolic care of the critically ill surgical patient. *Annals of the Royal College of Surgeons* **86**: 75–81.

Christiansen C, Tolf P, Jorgensen HS, Andersen SK, and Tonnesen E (2004) Hyperglycemia and mortality in critically ill patients. *Intensive Care Medicine* **30**: 1685–1688.

Fawzi WW, Msamanga GI, Spiegelman D *et al.* (2004) A randomized trial of multivitamin supplements and HIV disease progression and mortality. *New England Journal of Medicine* **35**: 23–32.

Koea J and Shaw JFH (1992) The effect of tumor bulk on the metabolic response to cancer. *Annals of Surgery* **215**: 282–288.

Plank LD and Hill GL (2003) Energy balance in critically illness. *Proceedings of the Nutrition Society* **62**: 545–552.

Ramaswamy G, Rao VR, Kumaraswamx SV, and Anantha N (1996) Serum vitamin status in oral leucoplakia: A preliminary study. *European Journal of Cancer* **328**(2): 120–122.

Rubinson L, Diette GB, Song X, Grower RG, and Krishnan JA (2004) Low calorie intake is associated with nosocomial bloodstream infections in patients in the medical intensive care unit. *Critical Care Medicine* **32**: 350–357.

Scjmeoder SM, Veyres P, Pivot X *et al.* (2004) Malnutrition is an independent factor associated with nosocomial infections. *British Journal of Nutrition* **92**: 105–111.

Tayek JA (1992) A review of cancer cachexia and abnormal glucose metabolism in humans with cancer. *Journal of the American College of Nutrition* **11**: 445–456.

Tayek JA and Brasel JA (1995) Failure of anabolism in malnourished cancer patients receiving growth hormone. *Journal of Clinical Endocrinology & Metabolism* **80**: 2082–2087.

Tayek JA and Katz J (1996) Glucose production, recycling, and gluconeogenesis in normals and diabetics; Mass isotopomer U-13C glucose study. *American Journal of Physiology* **270**: E709–E717.

Tayek JA and Katz J (1997) Glucose production, recycling, Cori cycle and gluconeogenesis in humans with and without cancer: Relationship to serum cortisol concentration. *American Journal of Physiology* **272**: E476–E484.

Van den Berghe G, Wouters P, Weekers F *et al.* (2001) Intensive insulin therapy in critically ill patients. *New England Journal of Medicine* **345**: 1359–1367.

Ziegler TR (1992) Clinical and metabolic efficacy of glutamine-supplemented parenteral nutrition after bone marrow transplantation. *Annals of Internal Medicine* **116**: 821.

Intestine *see* **Small Intestine**: Structure and Function; Disorders. **Microbiota of the Intestine**: Probiotics; Prebiotics

IODINE

Contents

Physiology, Dietary Sources and Requirements

R Houston, Emory University, Atlanta, GA, USA

> A relevant thing, though small, is of the highest importance
>
> MK Gandhi

Iodine is classified as a nonmetallic solid in the halogen family of the Periodic Table of the elements and therefore is related to fluorine, chlorine, and bromine. The halogen family lies between the oxygen family and the rare gases. Iodine sublimates at room temperature to form a violet gas; its name is derived from the Greek *iodes*, meaning 'violet-colored.' Iodine was discovered by Bernard Courtois in Paris in 1811, the second halogen (after chlorine) to be discovered. It took nearly 100 years to understand its critical importance in human physiology. In 1896, Baumann determined the association of iodine with the thyroid gland, and in 1914 Kendall, with revisions by Harrington in 1926, described the hormone complexes synthesized by the thyroid gland using iodine that are so integral to human growth and development.

As the biochemistry of iodine and the thyroid was being established, the scarcity of the element in the natural environment became evident and the link between deficiency and human disease was revealed. Enlargement of the thyroid, or goitre, is seen in ancient stone carvings and Renaissance paintings, but it was not until years later that the link with lack of iodine was firmly established. Even with this knowledge, many years passed before preventive measures were established. From 1910 to 1920 in Switzerland and the USA work was done on the use of salt fortified with iodine to eliminate iodine deficiency, with classic work being done by Dr David Marine in Michigan. Recently the linkage of iodine deficiency with intellectual impairment has brought iodine into the international spotlight.

Recent work has demonstrated that the halogens, including iodine, are involved through the halo-peroxidases in enzymatic activity and production of numerous active metabolites in the human body. While the importance of iodine for the

thyroid has been known for some time, recent research on halogen compounds in living organisms suggests additional more complex roles including antibiotic and anticancer activity. Yet it is the critical importance of iodine in the formation of the thyroid hormones thyroxine (T_4) and triiodothyronine (T_3) that makes any discussion of this element and human physiology of necessity bound up with a review of thyroid function.

Existence of Iodine in the Natural Environment

The marine hydrosphere has high concentrations of halogens, with iodine being the least common and chlorine the most. Halogens, including iodine, are concentrated by various species of marine organisms such as macroalgae and certain seaweeds. Release from these organisms makes a major contribution to the atmospheric concentration of the halogens. Iodine is present as the least abundant halogen in the Earth's crust. It is likely that in primordial times the concentration in surface soils was higher, but today the iodine content of soils varies and most has been leached out in areas of high rainfall or by previous glaciation. Environmental degradation caused by massive deforestation and soil erosion is accelerating this process. This variability in soil and water iodine concentration is quite marked, with some valleys in China having relatively high iodine concentrations in water, and other parts of China with negligible amounts in soil and water. **Table 1** shows the relative abundance of various halogens in the natural environment, while **Figure 1** illustrates the cycle of iodine in nature.

Table 1 Relative abundance of halogens in the natural environment

Element	*Abundance in oceans (ppm)*	*Abundance in Earth's crust (ppm)*	*Abundance in human body (mol)*
Fluorine	1.3	625	0.13
Chlorine	19 400	130	2.7
Bromine	67	2.5	0.0033
Iodine	0.06	0.05	0.00013

Commercial production of iodine occurs almost exclusively in Japan and Chile, with iodine extracted from concentrated salt brine from underground wells, seaweed, or from Chilean saltpetre deposits.

Absorption, Transport, and Storage

Iodine is usually ingested as an iodide or iodate compound and is rapidly absorbed in the intestine. Iodine entering the circulation is actively trapped by the thyroid gland. This remarkable capacity to concentrate iodine is a reflection of the fact that the most critical physiological role for iodine is the normal functioning of the thyroid gland. Circulating iodide enters the capillaries within the thyroid and is rapidly transported into follicular cells and on into the lumen of the follicle. This active transport is likely to be based on cotransport of sodium and iodine, allowing

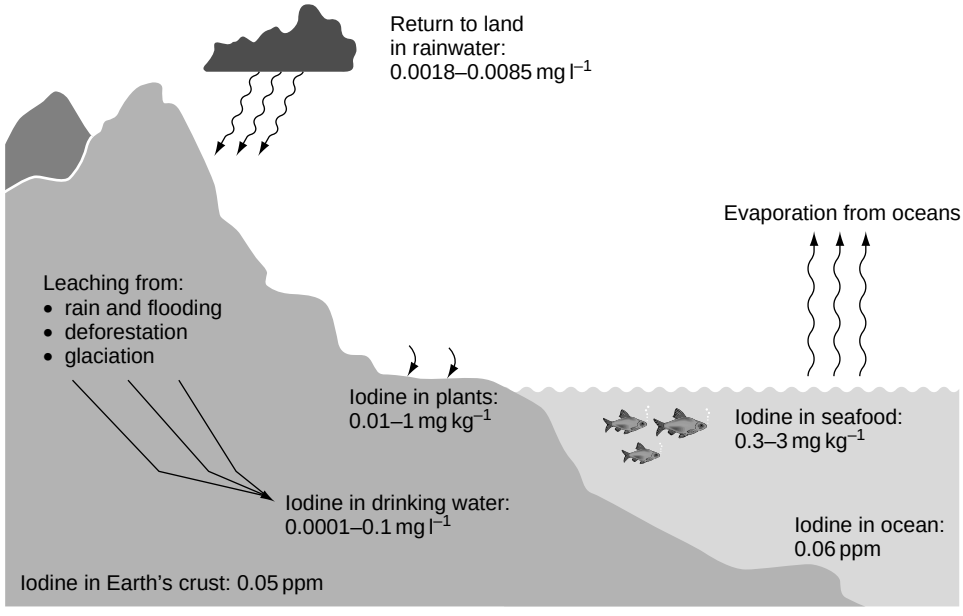


Figure 1 Cycle of iodine in nature.

iodine to move against its electrochemical gradient. Several anions, such as thiocyanate, perchlorate, and pertechnetate, inhibit this active transport. There is evidence that the active transport clearly demonstrated in the thyroid gland is also true for extrathyroidal tissues, including the salivary glands, mammary glands, and gastric mucosa.

In addition to trapping iodine, follicular cells also synthesize the glycoprotein, thyroglobulin (Tg), from carbohydrates and amino acids (including tyrosine) obtained from the circulation. Thyroglobulin moves into the lumen of the follicle where it becomes available for hormone production. Thyroid peroxidase (TPO), a membrane-bound hem-containing glycoprotein, catalyzes the oxidation of the iodide to its active form, I_2, and the binding of this active form to the tyrosine in thyroglobulin to form mono- or diiodotyrosine (MIT or DIT). These in turn combine to form the thyroid hormones triiodothyronine (T_3) and thyroxine (T_4). Thyroglobulin is very concentrated in the follicles through a process of compaction, making the concentration of iodine in the thyroid gland very high. Only a very small proportion of the iodine remains as inorganic iodide, although even for this unbound iodide the concentration in the thyroid remains much greater than that in the circulation. This remarkable ability of the thyroid to concentrate and store iodine allows the gland to be very rapidly responsive to metabolic needs for thyroid hormones. **Figure 2** shows the structures of the molecules tyrosine and thyroxine.

Formation of thyroid hormones is not restricted to humans. Marine algae have an 'iodine pump' that facilitates concentration; invertebrates and all vertebrates demonstrate similar mechanisms to concentrate iodine and form iodotyrosines of various types. Although the function of these hormones in invertebrates is not clear, in vertebrates these iodine-containing substances are important for a variety of functions, such as metamorphosis in amphibians, spawning changes in fish, and general translation of genetic messages for protein synthesis.

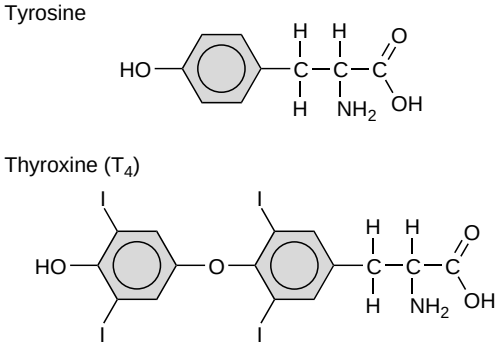


Figure 2 Structures of tyrosine and thyroxine (T_4).

Metabolism and Excretion

Once iodine is 'captured' by the thyroid and thyroid hormones formed in the lumen of the follicles, stimulation of the gland causes release of the hormones into the circulation for uptake by peripheral tissues. Both production and release of the hormones are regulated in two ways. Stimulation is hormonally controlled by the hypothalamus of the brain through thyroid releasing hormone (TRH) which stimulates the pituitary gland to secrete thyroid stimulating hormone (TSH), which in turn stimulates the thyroid to release T_3 and T_4. In addition to the regulation of thyroid hormones by TSH, iodine itself plays a major role in autoregulation. The rate of uptake of iodine into the follicle, the ratio of T_3 to T_4, and the release of these into the circulation, among other things, are affected by the concentration of iodine in the gland. Thus, an increase in iodine intake causes a decrease in organification of iodine in the follicles and does not necessarily result in a corresponding increase in hormone release. Recent research suggests that this autoregulation is not entirely independent of TSH activity and that several other factors may contribute. However, regardless of the mechanism, these regulatory mechanisms allow for stability in hormone secretion in spite of wide variations in iodine intake.

When stimulated to release thyroid hormones, thyroglobulin is degraded through the activity of lysosomes and T_3 and T_4 are released and rapidly enter the circulation. Iodide freed in this reaction is for the most part recycled and the iodinated tyrosine reused for hormone production. Nearly all of the released hormones are rapidly bound to transport hormones, with 70% bound to thyroxine binding globulin (TBG). Other proteins, such as transthyretin (TTR), albumin, and lipoproteins, bind most of the remainder; with significant differences in the strengths of the affinity for the hormones, these proteins transport the hormones to different sites.

This remarkable ability of the thyroid to actively trap and store the iodine required creates a relatively steady state, with daily intake used to ensure full stores. T_4, with a longer half-life, serves as a reservoir for conversion to the more active hormone, T_3, with a much shorter half-life of 1 day. Target organs for thyroid hormone activity all play a role in the complex interplay between conversion of T_4 to T_3 deiodination, and metabolism of various other proteins involved with thyroid function. The liver,

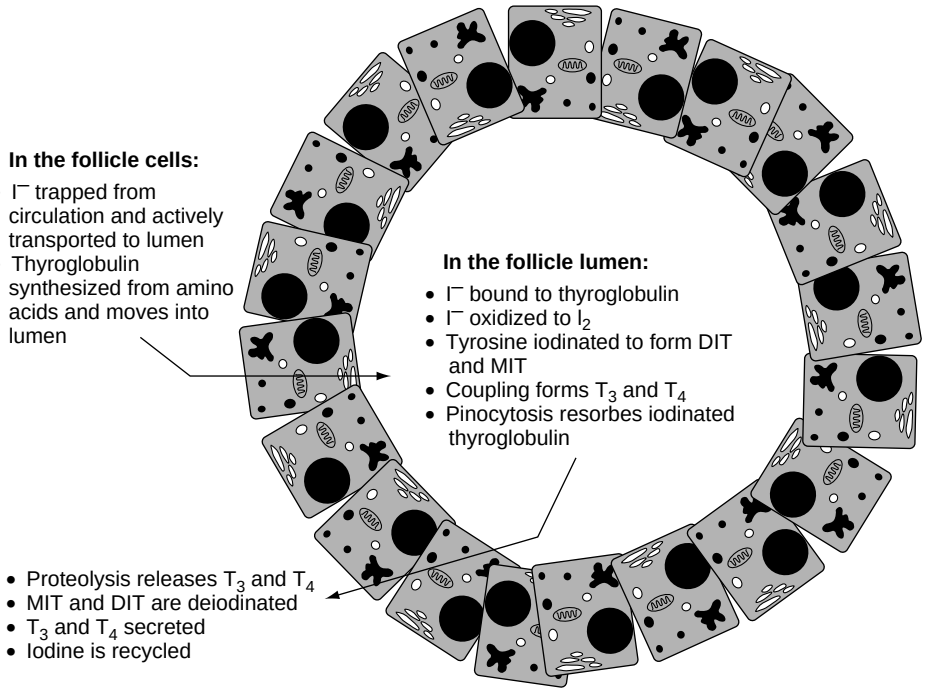


Figure 3 Thyroid follicle (courtesy of Kiely Houston).

which is estimated to contain 30% of the extrathyroidal T_4, is responsible, through the activity of the liver cell enzyme, deiodinase, for ensuring adequate supply of T_3 to peripheral tissues and degradation of metabolic by-products. The kidney demonstrates a strong ability to take up the iodothyronines. Iodine is ultimately excreted in the urine, with average daily excretion rates of approximately 100 μg per day. This accounts for the vast majority of iodine excretion, with negligible amounts excreted in feces. **Figure 3** illustrates a thyroid follicle and summarizes iodine transport.

Metabolic Functions

Separating the role of iodine from the complex and pervasive function of the thyroid gland is difficult since iodine is a critical component of the hormones that mediate these functions, and whatever other roles iodine may have are poorly understood. Thyroid hormones affect a wide range of physiological functions, from liver and kidney to heart and brain. Earlier work supported a role for thyroid hormones in affecting the energy generating capacity of cells through biochemical changes in mitochondria. More recent work has shown, however, that these hormones act on specific genetic receptors in cell nuclei, and perhaps through other extranuclear mechanisms. The nuclear receptors belong to a large family of receptors that bind other extranuclear molecules including vitamins A and D and steroids. Through this interaction, along with a number of other proteins, thyroid hormones modify genetic expression. A great deal of research currently focuses on these thyroid hormone receptors, and the effect primarily of T_3 on the physiological function of the target organ through genetic transcription. These receptors are present in pituitary, liver, heart, kidney, and brain cells.

In the pituitary gland, thyroid hormones, along with many cofactors, regulate the synthesis and secretion of growth hormone by increasing gene transcription. Similarly, as part of the feedback loop for hormone regulation and release, thyroid hormones affect transcription of TSH in the pituitary. In cardiac and skeletal muscle, thyroid hormones affect production of the muscle tissue myosin in a variety of ways, depending on the stage of life and specific muscle tissue affected. In addition, the hormones affect muscle contraction through genetic alteration of calcium uptake within the cell. Carbohydrate metabolism and formation of certain fats (lipogenesis) are affected through hormone-induced changes in gene transcription in liver cells.

In the adult brain, receptors have been identified, but the specific genes affected by thyroid hormones have not yet been located. However, in the developing brain of the fetus and neonate, the effects of thyroid hormones are significant even though the

Table 2 Estimated iodine concentrations in selected organs

Total body	*Thyroid gland*	*Brain*	*Liver*	*Blood*
15–20 mg	8–12 mg (for a 15–25 g gland)	0.02 $\mu g\,g^{-1}$ (wet weight)	0.2 $\mu g\,g^{-1}$ (wet weight)	0.08–0.60 $\mu g\,dl^{-1}$ (plasma inorganic iodide)

exact mechanisms are still not fully understood. The effects of thyroid hormones on brain development are suggested by failure in development of the nerve elements, failure of differentiation of cerebellar cells, and reduced development of other brain cells, in hypothyroid states. It is this early effect that has recently elevated the status of iodine from an element whose deficiency caused goitre to one whose deficiency is the leading cause of mental impairment worldwide.

In addition to these nuclear mechanisms, several alternative pathways have been suggested, some based on earlier historical studies. The thermogenic effects of thyroid hormones were originally felt to be a direct action on mitochondria, though this has recently been questioned. Thyroid hormones stimulate glucose transport, and again though originally attributed to a direct action on the plasma membrane, recent evidence suggests a genetic mechanism. There may also be a direct effect of thyroid hormones on brain enzymatic activity.

The overall effect of these cellular and systemic actions is to stimulate respiratory and other enzyme synthesis, which results in increased oxygen consumption and resultant increased basal metabolic rate. This affects heart rate, respiratory rate, mobilization of carbohydrates, cholesterol metabolism, and a wide variety of other physiological activities. In addition, thyroid hormones stimulate growth and development and, as noted earlier, are critical for the normal proliferation, growth, and development of brain cells. **Table 2** shows the estimated iodine concentration in selected organs.

Iodine Deficiency and Excess

Iodine Deficiency

Iodine deficiency is the most common cause of preventable mental retardation in the world. This fact, along with the recognition that iodine deficiency is not limited to remote rural populations, has stimulated agencies and governments to mobilize resources to eliminate this problem. This global effort, focusing primarily on iodization of salt for human and animal consumption, is slowly succeeding in eliminating a hidden set of disorders that have plagued mankind for centuries.

Unlike many nutritional deficiencies that are more directly related to socioeconomic status, insufficient intake of iodine is a geographical disease, related to lack of iodine in the environment. Iodine originally present in soil was subjected to leaching by snow and rain, and while a portion of the iodine in the oceans evaporates and is returned to the soil in rainwater, this amount is small. Thus, many areas have insufficient iodine in the environment, and this is reflected in plants grown in that environment. The diets in many developing countries are limited in variability and contain few processed foods. This places large populations at risk of iodine deficiency. The World Health Organization (WHO) estimates that at least 1572 million people are at risk in 118 countries, with 43 million affected by 'some degree of mental impairment.'

In the most simplistic physiological model, inadequate intake of iodine results in a reduction in thyroid hormone production, which stimulates increased TSH production. TSH acts directly on thyroid cells, and without the ability to increase hormone production, the gland becomes hyperplastic. In addition, iodine trapping becomes more efficient, as demonstrated by increased radioactive iodine uptake in deficient individuals. However, this simplistic model is complicated by complex adaptive mechanisms which vary depending on the age of the individual affected. In adults with mild deficiency, reduced intake causes a decrease in extrathyroidal iodine and reduced clearance, demonstrated by decreased urinary iodine excretion, but iodine concentration in the gland may remain within normal limits. With further reduction in intake, this adaptive mechanism is overwhelmed, and the iodine content of the thyroid decreases with alterations in iodination of thyroglobulin, in the ratio of DIT to MIT, and reduction in efficient thyroid hormone production. The ability to adapt appears to decrease with decreasing age, and in children the iodine pool in the thyroid is smaller, and the dynamics of iodine metabolism and peripheral use more rapid. In neonates, the effects of iodine deficiency are more directly reflected in increased TSH. Diminished thyroid iodine content and increased turnover make neonates the most vulnerable to the effects of iodine deficiency and decreased hormone production, even with mild deficiency.

A number of other factors influence iodine balance. Active transport of iodide is competitively inhibited by several compounds, including complex ions such as perchlorate, and by thiocyanate, a metabolic product of several foods. Other compounds, such as propylthiouracil, affect coupling reactions and iodination, doing so regardless of iodine intake, e.g., without blocking iodide transport. Several pharmaceuticals affect peripheral hormone action. Dietary goitrogens, as these compounds have been called, include cassava, lima beans, sweet potatoes, cabbage, and broccoli; these contain cyanide compounds that are detoxified to thiocyanate, which may inhibit iodide transport. Cabbage and turnips, and other plants of the genus *Brassica*, also contain thionamide compounds which block iodination. Certain industrial waste products, such as resorcinol from coal processing, contain phenols that cause irreversible inhibition of TPO and block iodination. In some countries the staple diet includes such goitrogens, and iodine deficiency may be exacerbated, as has been well documented for cassava. While this may be a significant problem in some geographical areas, in most instances adequate dietary iodine can reverse the goitrogenic effect.

The most important clinical effect of deficiency relates to the fact that thyroid hormone is required for the normal development of the brain in both humans and other animals. Numerous studies have demonstrated reduced psychomotor skills and intellectual development in the presence of iodine deficiency, and most experts now believe that there is a continuum of deficits, from mild impairment in IQ to severe mental retardation. Studies in China demonstrated shifts in IQ point distributions in rural communities that were deficient, suggesting an impact of deficiency of 10–15 IQ points. In Europe, where mild deficiency still exists, studies have demonstrated decreased psychomotor, perceptual integrative motor ability as well as lower verbal IQ scores in schoolchildren. Studies in Iran showed similar findings. A recent meta-analysis of 18 studies demonstrated a strong relationship, with an overall 13.5 IQ point difference between deficient and non deficient populations. These findings, coupled with the high prevalence of deficiency in many countries, have major implications for development.

The most severe effect of iodine deficiency is cretinism, which is rare in areas of mildly endemic deficiency but may have reached 5–10% or more in areas with severe deficiency. There are general classifications of cretinism, the symptoms of which frequently overlap. Neurological cretinism presents as extreme mental retardation, deaf-mutism, and impaired motor function including spastic gait. Myxoedematous cretinism presents as disturbances of growth and development including short stature, coarse facial features, retarded sexual development, mental retardation, and other signs of hypothyroidism. It appears likely that severe deficiency resulting in decreased maternal T_4 may be responsible for the impaired neurological development of the fetus occurring early in pregnancy. The effect of deficiency on the fetus after 20 weeks' gestation may result in hyperstimulation of the developing fetal thyroid, with the extreme manifestation being thyroid failure causing myxoedematous cretinism. Other factors may affect thyroid hormone metabolism. Selenium deficiency, when present with iodine deficiency, may alter the clinical manifestations. Selenium deficiency decreases the activity of the enzyme, glutathione peroxidase (GPX), which, along with thyroid hormone synthesis, reduces hydrogen peroxide (H_2O_2). Combined with iodine deficiency and reduced hormone synthesis, it has been speculated that selenium deficiency may contribute to accumulation of H_2O_2 which may in turn lead to cell damage and contribute to thyroid failure. Selenium is also essential for the deiodinase enzyme activity affecting thyroid hormone catabolism, and deficiency may actually increase serum thyroxine. The balance between these two effects is still not fully understood. The study of cretinism has been critical to the evolution of our understanding of the critical role of iodine for normal mental development.

Iodine deficiency has a number of other effects, including development of goitre, clinical or subclinical hypothyroidism, decreased fertility rates, increased stillbirth and spontaneous abortion rates, and increased perinatal and infant mortality. This spectrum of clinical effects, collectively called 'iodine deficiency disorders,' underlines the importance of iodine in human health.

The most effective method to eliminate iodine deficiency in populations is through iodization of salt. The most classic success of salt iodization was demonstrated in Switzerland. Salt is universally consumed, and in most countries the amount consumed is relatively constant between 5 and 10 g per person per day. Iodine is usually added as iodide or iodate (which is more stable) to achieve 25–50 ppm iodine at consumption. This provides about 150–250 μg of iodine per person per day.

The challenge for national iodine deficiency elimination programs is to mobilize the various sectors that must be involved in a sustainable national program, including education, industry, health, and the political arena. There must be an appropriate regulatory environment, effective demand creation,

adequate production to make iodized salt available, and quality assurance of both the product and all program elements to ensure that the program is sustained forever. Success in these efforts has the potential to have a greater impact on development than any public health program to date.

Iodine Excess

Iodine is used in many medications, food preservatives, and antiseptics with minimal adverse effects on populations. Pure iodine crystals are toxic, and ingestion can cause severe stomach irritation. Iodine is allergenic, and acute reactions to radiographic contrast media are not rare. Yet because of the thyroid's unique ability to regulate the body's iodine pool, quite a wide range in intake is tolerated without serious effects, particularly when the exposure is of limited duration.

When ingestion of iodine is in excess of the daily requirement of approximately 150 μg per day, changes in thyroid hormones can occur. A variety of clinical problems can occur, and these differ depending on the dose, the presence of thyroid disease, and whether the individual has been deficient in the past. In iodine-replete individuals without thyroid disease, goitre can result, and rarely, hypothyroidism, although the latter is more common in individuals with other illnesses such as lung disease or cystic fibrosis. The relationship of iodine excess to other diseases such as Hashimoto's thyroiditis remains controversial. In the US iodine levels were quite high from 1960 to 1980, with estimates for adult males as high as 827 μg per day. There was no immediate evidence of an impact on thyroid disease, although longer term longitudinal data are lacking. Effects usually remain subtle and transient, even with ingestion of up to 1500–4500 μg per day.

In the presence of thyroid disease, and in areas with endemic iodine deficiency, suddenly raising daily iodine intake may precipitate hyperthyroidism, and this has been the subject of some concern as salt iodization efforts proceed with fledgling quality assurance. This effect is felt to be related in part to autonomous nodules in the gland that synthesize and release excess thyroid hormone. The exact prevalence of iodine-induced hyperthyroidism in deficient areas is not clear. Many countries initiating salt iodization programs have reported increases in the incidence of toxic nodular goitre and iodine-induced thyrotoxicosis, usually in older people. While this may be a significant clinical problem, the risk is estimated to be between 0.01 and 0.06% and must be considered in the light of the benefit from correction of deficiency.

Assessment of Iodine Status

A standard set of indicators of iodine status has been established by the WHO in response to the need to determine prevalence in countries with endemic deficiency. These indicators reflect iodine status as mediated through the response of the thyroid gland to fluctuations in iodine intake. There are several additional indicators that are used to assess thyroid function, such as T_4 and T_3, but these are less accurate in reflecting iodine status since conversion of T_4 to T_3 and cellular uptake is so responsive to peripheral need.

Urinary iodine reflects iodine sufficiency, and output decreases with diminished intake. Since this indicator reflects the amount of iodine per unit volume of urine, its accuracy is impaired by variable fluid intake and factors affecting the concentration of the urine. Therefore, as a measure of iodine status in an individual, it is less accurate than as a measure of iodine status of a population. Median urinary iodine values are used extensively to assess population prevalence of iodine deficiency.

Thyroid size, either estimated by palpation or using ultrasound volume determination, reflects iodine status since deficiency results in thyroid enlargement, or goitre. Due to the relative ease of palpation, that measure has been a traditional standard to assess populations for iodine deficiency and has been particularly useful in schoolchildren. In adults, where long-standing thyroid enlargement from iodine deficiency may be minimally responsive to corrected iodine intake, palpation may be misleading and could overestimate the current level of iodine sufficiency. In children, palpation becomes increasingly difficult and significantly less accurate when deficiency is mild. Ultrasound volume determination provides a more accurate estimate of thyroid size. For any measure of thyroid size, other factors besides iodine deficiency can cause enlargement, including iodine excess, carcinoma, and infection. In areas of the world where deficiency is a problem, the prevalence of these other diseases compared with goitre from iodine deficiency is negligible.

TSH is produced in response to decreased iodine intake and diminished thyroid hormone production and is used as a measure of iodine status. TSH is best measured in neonates—in the developed world for surveillance against congenital hypothyroidism, and in endemic countries to estimate the magnitude of iodine deficiency. Neonatal TSH has been a useful advocacy tool to demonstrate to policy makers that iodine deficiency is not limited to rural remote populations but affects children born in big city

Table 3 WHO criteria for iodine deficiency as a public health problem in populations

Indicator	*Population assessed*	*Mild deficiency (%)*	*Severe deficiency (%)*
Goitre by palpation	Schoolchildren	5–19.9	≥30
Thyroid volume by ultrasound (>97th percentile)	Schoolchildren	5–19.9	≥30
Median urinary iodine ($\mu g\,l^{-1}$)	Schoolchildren	50–99	<20
TSH (>5 $mU\,l^{-1}$ whole blood)	Neonates	3–19.9	≥40

hospitals. However, with the complexity of the interactions between TSH and other hormones, TSH has not been shown to be as useful in older children or adults in estimating prevalence of iodine deficiency. Also, use of iodine containing antiseptics affects TSH distributions in neonates.

Uptake of radioactive iodine isotopes can be used to scan the gland, and determine the affinity of the gland to introduced iodine, and is a measure of deficiency. The most common isotope used is ^{123}I because of its relatively short 13-h half-life and γ photon emission. Uptake is increased in iodine deficiency. Isotopes can also be used to examine the organification of iodine in the formation of thyroid hormones. This is an impractical method for surveying populations. **Table 3** provides the WHO criteria for defining iodine deficiency as a public health problem.

Requirements and Dietary Sources

The daily requirement for iodine in humans has been estimated based on daily losses, iodine balance, and turnover, with most studies ranging from 40 to 200 μg per day, depending on age and metabolic needs, as shown in **Table 4**.

Natural sources of iodine include seafood, seaweeds, and smaller amounts from crops grown on soil with sufficient iodine, or from meat where livestock has grazed on such soil. The contribution of the latter two is small, and in most countries other sources are required. Iodine added to salt, as noted above, is the primary source for many populations. **Table 5** shows sample iodine content for various sources.

Table 4 Recommended dietary intake

Age	*WHO recommended intake (μg per day)*	*US RDA 1989 (μg per day)*
0–6 months	40	40
6–12 months	50	60 (at age 1 year)
1–10 years	70–120	60–120
11 years–adult	120–150	150
Pregnancy	175	175
Lactation	200	200

In the US and Britain, as well as in other developed countries, most dietary iodine comes from food processing. Intake can vary, as illustrated in **Table 6**. Iodophors used as antiseptics in the dairy and baking industries provide residual iodine in milk and processed foods. In addition, iodine is present in several vitamin and pharmaceutical preparations.

Iodine as a trace element in low concentrations in most environments plays a critical role in the normal growth and development of many species. In humans, iodine is critical for brain development and correction of global deficiencies is an unparalleled opportunity to improve the well-being of our global community.

Table 6 Iodine intake from average US and British diets

Country	*Milk (μg per day)*	*Grains (μg per day)*	*Meat, fish, and poultry (μg per day)*
US	534	152	103
Britain	92	31	36

Table 5 Sample iodine content for various sources

Water	*Cabbage*	*Eggs*	*Seafood*	*Sugar*	*Iodized salt*
0.1–2 $\mu g\,l^{-1}$ in endemic area	0–0.95 $\mu g\,g^{-1}$	4–10 $\mu g\ egg^{-1}$	300–3000 $\mu g\,kg^{-1}$	<1 $\mu g\,kg^{-1}$ in refined sugar	20–50 ppm (at household level, depending on climate, and currently subject to review)
2–15 $\mu g\,l^{-1}$ in nonendemic area				30 $\mu g\,kg^{-1}$ in unrefined brown sugar	

See also: **Fruits and Vegetables**. **Iodine**: Deficiency Disorders. **Legumes**.

Further Reading

Braverman LE and Utiger RD (eds.) (1996) *Werner and Ingbar's The Thyroid, A Fundamental and Clinical Text*. Philadelphia: Lippincott-Raven.

Burgi H, Supersaxo Z, and Selz B (1990) Iodine deficiency diseases in Switzerland one hundred years after Theodor Kocher's survey: A historical review with some new goitre prevalence data. *Acta Endocrinologica (Copenhagen)* **123**: 577–590.

Gaitan E (1990) Goitrogens in food and water. *Annual Review of Nutrition* **10**: 21–39.

Hall R and Kobberling J (1985) *Thyroid Disorders Associated with Iodine Deficiency and Excess*. New York: Raven Press.

Hetzel BS (1994) Iodine deficiency and fetal brain damage. *New England Journal of Medicine* **331**(26): 1770–1771.

Hetzel BS (1989) In *The Story of Iodine Deficiency: An International Challenge in Nutrition*. Oxford: Oxford University Press.

Hetzel BS and Pandav CS (eds.) (1994) *SOS for a Billion—The Conquest of Iodine Deficiency Disorders*. Delhi: Oxford University Press.

Mertz W (1986) *Trace Elements in Human and Animal Nutrition*, 5th edn. New York: Academic Press.

Patai S and Rappoport Z (eds.) (1995) *The Chemistry of Halides, Pseudo-halides and Azides,* Supplement D2: part 2. New York: John Wiley & Sons.

Stanbury JB (ed.) (1994) *The Damaged Brain of Iodine Deficiency*. New York: Cognizant Communication Corporation, The Franklin Institute.

Sullivan KM, Houston RM, Gorstein J, and Cervinskas J (1995) *Monitoring Universal Salt Iodization Programmes* Ottowa: UNICEF, MI, ICCIDD, WHO publication.

Thorpe-Beeston JG and Nicolaides KH (1996) *Maternal and Fetal Thyroid Function in Pregnancy* New York: The Parthenon Publishing Group.

Todd CH, Allain T, Gomo ZAR *et al.* (1995) Increase in thyrotoxicosis associated with iodine supplements in Zimbabwe. *Lancet* **346**: 1563–1564.

Troncone L, Shapiro B, Satta MA, and Monaco F (1994) *Thyroid Diseases: Basic Science Pathology, Clinical and Laboratory Diagnosis*. Boca Raton: CRC Press.

WHO, UNICEF, and ICCIDD (1994) *Indicators for Assessing Iodine Deficiency Disorders and Their Control through Salt Iodization*, (limited publication). Geneva: WHO, UNICEF, ICCID.

Wilson JD and Foster DW (eds.) (1992) *Williams Textbook of Endocrinology*. Philadelphia: WB Saunders.

Deficiency Disorders

B S Hetzel, Women's and Children's Hospital, North Adelaide, SA, Australia

Iodine deficiency is discussed as a risk factor for the growth and development of up to 2.2 million people living in iodine-deficient environments in 130 countries throughout the world. The effects of iodine deficiency on growth and development, called the iodine deficiency disorders (IDD), comprise goiter (enlarged thyroid gland), stillbirths and miscarriages, neonatal and juvenile thyroid deficiency, dwarfism, mental defects, deaf mutism, and spastic weakness and paralysis, as well as lesser degrees of loss of physical and mental function.

Iodine deficiency is now accepted by the World Health Organization as the most common preventable cause of brain damage in the world today.

Since 1990, a major international health program to eliminate iodine deficiency has developed that uses iodized salt. The progress of this program and the continuing challenge are discussed as a great opportunity for the elimination of a noninfectious disease, which is quantitatively a greater scourge than the infectious diseases of smallpox and polio.

History

The first records of goiter and cretinism date back to ancient civilizations, the Chinese and Hindu cultures and then to Greece and Rome. In the Middle Ages, goitrous cretins appeared in the pictorial art, often as angels or demons. The first detailed descriptions of these subjects occurred in the Renaissance. The paintings of the madonnas in Italy so commonly showed goiter that the condition must have been regarded as virtually normal. In the seventeenth and eighteenth centuries, scientific studies multiplied and the first recorded mention of the word 'cretin' appeared in Diderot's *Encyclopédie* in 1754. The nineteenth century marked the beginning of serious attempts to control the problem; however, not until the latter half of the twentieth century was the necessary knowledge for effective prevention acquired.

Mass prophylaxis of goiter with iodized salt was first introduced in Switzerland and in Michigan in the United States. In Switzerland, the widespread occurrence of a severe form of mental deficiency and deaf mutism (endemic cretinism) was a heavy charge on public funds. However, following the introduction of iodized salt, goiter incidence declined rapidly and cretins were no longer born. Goiter also disappeared from army recruits.

A further major development was the administration of injections of iodized oil to correct iodine deficiency in Papua New Guinea for people living in inaccessible mountain villages. These long-lasting injections corrected iodine deficiency and prevented goiter for 3–5 years, depending on the dosage.

Subsequently, the prevention of cretinism and stillbirths was demonstrated by the administration of iodized oil before pregnancy in a controlled trial

in the Highlands of Papua New Guinea. This proved the causal role of iodine deficiency.

To further establish the relation between iodine deficiency and fetal brain development, an animal model was developed in the pregnant sheep given an iodine-deficient diet. Subsequently, similar models were developed in the primate marmoset monkey and in the rat.

Studies with animal models confirmed the effect of iodine deficiency on fetal brain development (as already indicated by the results of the field trial with iodized oil in Papua New Guinea). The combination of the controlled human trials and the results of the studies in animal models clearly indicated that prevention was possible by correction of the iodine deficiency before pregnancy.

This work led Hetzel to propose the concept of the IDD resulting from all the effects of iodine deficiency on growth and development, particularly brain development, in an exposed population that can be prevented by correction of the iodine deficiency. Iodine deficiency is now recognized by the World Health Organization (WHO) as the most common form of preventable mental defect.

Although the major prevalence of iodine deficiency is in developing countries, the problem continues to be very significant in many European countries (France, Italy, Germany, Greece, Poland, Romania, Spain, and Turkey) because of the threat to brain development in the fetus and young infant.

Ecology of Iodine Deficiency

There is a cycle of iodine in nature. Most of the iodine resides in the ocean. It was present during the primordial development of the earth, but large amounts were leached from the surface soil by glaciation, snow, or rain and were carried by wind, rivers, and floods into the sea. Iodine occurs in the deeper layers of the soil and is found in oil well and natural gas effluents, which are now a major source for the production of iodine.

The better known areas that are leached are the mountainous areas of the world. The most severely deficient soils are those of the European Alps, the Himalayas, the Andes, and the vast mountains of China. However, iodine deficiency is likely to occur to some extent in all elevated regions subject to glaciation and higher rainfall, with runoff into rivers. It has become clear that iodine deficiency also occurs in flooded river valleys, such as the Ganges in India, the Mekong in Vietnam, and the great river valleys of China.

Iodine occurs in soil and the sea as iodide. Iodide ions are oxidized by sunlight to elemental iodine, which is volatile so that every year approximately 400,000 tons of iodine escapes from the surface of the sea. The concentration of iodide in the seawater is approximately 50–60 μg/l, and in the air it is approximately 0.7 μg/m^3. The iodine in the atmosphere is returned to the soil by rain, which has a concentration of 1.8–8.5 μg/l. In this way, the cycle is completed (**Figure 1**).

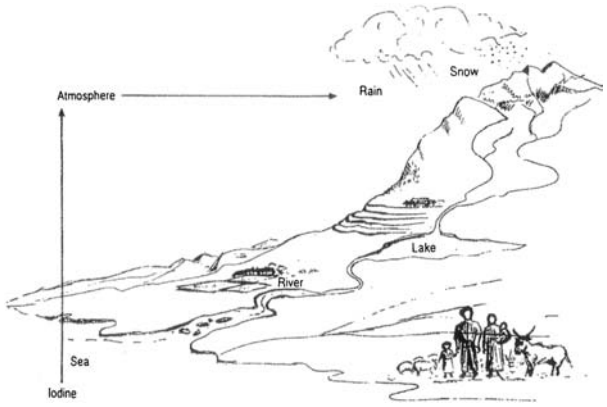


Figure 1 The iodine cycle in nature. The atmosphere absorbs iodine from the sea, which then returns through rain and snow to mountainous regions. It is then carried by rivers to the lower hills and plains, eventually returning to the sea. High rainfall, snow, and flooding increase the loss of soil iodine, which has often been already denuded by past glaciation. This causes the low iodine content of food for man and animals. (Reproduced from Hetzel BS (1989) *The Story of Iodine Deficiency: An international Challenge in Nutrition.* Oxford: Oxford University Press.)

However, the return of iodine is slow and the amount is small compared to the original loss of iodine, and subsequent repeated flooding ensures the continuity of iodine deficiency in the soil. Hence, no natural correction can take place and iodine deficiency persists in the soil indefinitely. All crops grown in these soils will be iodine deficient. The iodine content of plants grown in iodine-deficient soils may be as low as 10 μg/kg compared to 1 mg/kg dry weight in plants in a non-iodine-deficient soil.

As a result, human and animal populations that are totally dependent on food grown in such soil become iodine deficient. This accounts for the occurrence of severe iodine deficiency in vast populations in Asia that live within systems of subsistence agriculture in flooded river valleys (India, Bangladesh, Burma, Vietnam, and China).

Iodine Deficiency Disorders

The effects of iodine deficiency on the growth and development of a population that can be prevented by correction of iodine deficiency, denoted by the term IDD, are evident at all stages, including

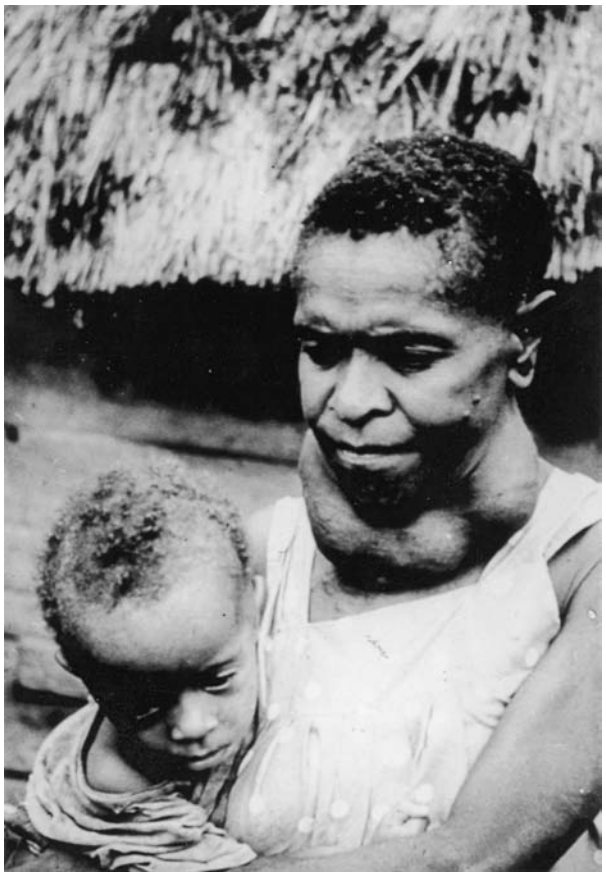


Figure 2 A mother and child from a New Guinea village who are severely iodine deficient. The mother has a large goiter and the child is also affected. The larger the goiter, the more likely it is that she will have a cretin child. This can be prevented by eliminating the iodine deficiency before the onset of pregnancy. (Reproduced from Hetzel BS and Pandav CS (eds.) (1996) *SOS for a Billion: The Conquest of Iodine Deficiency Disorders*, 2nd edn. Oxford: Oxford University Press.)

particularly the fetus, the neonate, and in infancy, which are periods of rapid brain growth. The term goiter has been used for many years to describe the enlarged thyroid gland caused by iodine deficiency (**Figure 2**). Goiter is indeed the obvious and familiar feature of iodine deficiency, but knowledge of the effects of iodine deficiency on brain development has greatly expanded in the past 30 years so that the term IDD was introduced to refer to all the effects of iodine deficiency on growth and development, particularly brain development, in a population that can be prevented by correction of the deficiency (**Table 1**).

The following sections discuss in detail the IDD at various stages of life: the fetus, the neonate, the child and adolescent, and the adult (**Table 1**).

The Fetus

Iodine deficiency of the fetus is the result of iodine deficiency in the mother (**Figure 2**). The condition is associated with a greater incidence of stillbirths, abortions, and congenital abnormalities, which can be prevented by iodization.

Another major effect of fetal iodine deficiency is the condition of endemic cretinism, which is quite distinct from the condition of sporadic cretinism or congenital hypothyroidism due to a small or absent thyroid gland.

Endemic cretinism-associated with an iodine intake of less than 25 μg per day, in contrast to a normal intake of 100–150 μg per day, has been widely prevalent, affecting up to 10% of populations living in severely iodine-deficient regions in India, Indonesia, and China. In its most common form, it is characterized by mental deficiency, deaf mutism, and spastic diplegia (**Figure 3**). This form of cretinism is referred to as the nervous or neurological type, in contrast to the less common hypothyroid or myxedematous type characterized by hypothyroidism with dwarfism (**Figure 4**).

In addition to Asia, cretinism also occurs in Africa, (Zaire, now the Republic of the Congo), South America in the Andean region (Ecuador, Peru, Bolivia, and Argentina), and the more remote areas of Europe. In all these areas, with the exception of the Congo, neurological features are predominant. In the Congo, the hypothyroid form is more common, probably due to the high intake of the root vegetable cassava, which contains substances inhibiting the function of the thyroid gland.

However, there is considerable variation in the clinical manifestations of neurological cretinism,

Table 1 Spectrum of Iodine Deficiency Disorders

Fetus	Abortions
	Stillbirths
	Congenital anomalies
	Neurological cretinism
	Mental deficiency, deaf mutism, spastic diplegia, squint
	Hypothyroid cretinism
	Mental deficiency, dwarfism, hypothyroidism
	Psychomotor defects
Neonate	Increased perinatal mortality
	Neonatal hypothyroidism
	Retarded mental and physical development
Child and adolescent	Increased infant mortality
	Retarded mental and physical development
Adult	Goiter with its complications
	Iodine-induced hyperthyroidism
All ages	Goiter
	Hypothyroidism
	Impaired mental function
	Increased susceptibility to nuclear radiation

Reproduced with permission from Oxford University Press and the World Health Organization, WHO/UNICEF/ICCIDD (2001).

Figure 3 A mother with her four sons, three of whom (ages 31, 29, and 28 years) are cretins born before iodized salt was introduced, and the fourth is normal (age 14 years), born after iodized salt became available in Chengde, China. (Reproduced from Hetzel BS and Pandav CS (eds.) (1996) *SOS for a Billion: The Conquest of Iodine Deficiency Disorders*, 2nd edn. Oxford: Oxford University Press.)

Figure 4 A hypothyroid cretin from Sinjiang, China, who is also deaf mute. This condition is completely preventable. (Right) The barefoot doctor of her village. Both are approximately 35 years old. (Reproduced from Hetzel BS (1989) *The Story of Iodine Deficiency: An international Challenge in Nutrition.* Oxford: Oxford University Press.)

which include isolated deaf mutism and mental defect of varying degrees. In China, the term cretinoid is used to describe these individuals, who may number 5–10 times those with overt cretinism.

The Neonate

Apart from the question of mortality, the importance of the state of thyroid function in the neonate relates to the fact that at birth the brain of the human infant has only reached approximately one-third of its full size and continues to grow rapidly until the end of the second year. The thyroid hormone, dependent on an adequate supply of iodine, is essential for normal brain development, as has been confirmed by animal studies.

Data on iodine nutrition and neonatal thyroid function in Europe confirm the continuing presence of severe iodine deficiency. This affects neonatal thyroid function and hence represents a threat to early brain development. These data have raised great concern about iodine deficiency, which is also heightened by awareness of the hazard of nuclear radiation with carcinogenic effects following the Chernobyl disaster in the former Soviet Union (**Table 1**).

These observations of neonatal hypothyroidism indicate a much greater risk of mental defects in iodine-deficient populations than is indicated by the presence of cretinism. Apart from the developing world, there has been a continuing major problem in many European countries, such as Italy, Germany, France, and Greece, and Romania, Bulgaria, and Albania still have very severe iodine deficiency with overt cretinism.

The Child

Iodine deficiency in children is characteristically associated with goiter. The goiter rate increases with age and reaches a maximum at adolescence. Girls have a higher prevalence than boys. Goiter rates in schoolchildren over the years provide a useful indication of the presence of iodine deficiency in a community.

In a review of 18 studies, a comparison was made between IQ scores in iodine-deficient children and carefully selected control groups. The iodine-deficient group had a mean IQ that was 13.5 points lower than that of the non-iodine-deficient control group. Detailed individual studies demonstrating these defects in Italian and Spanish schoolchildren as well as those from Africa, China, Indonesia, and Papua New Guinea have been published. There is a serious problem in Europe as well as in many developing countries.

The Adult

Long-standing large goiter may require surgery to reduce pressure in the neck. Long-standing goiter may also be associated with iodine-induced hyperthyroidism (IIH) due to an increase in iodine intake. IIH is associated with nervousness, sweating, and tremor, with loss of weight due to excessive levels of circulating thyroid hormone. This condition no longer occurs following correction of iodine deficiency and therefore is within the spectrum of IDD.

In northern India, a high degree of apathy has been noted in whole populations living in iodine-deficient areas. This may even affect domestic animals such as dogs. It is apparent that reduced mental function is widely prevalent in iodine-deficient communities, with effects on their capacity for initiative and decision making. This is due to the effect of hypothyroidism on brain function. This condition can be readily reversed by correction of the iodine deficiency, unlike the effects on the fetus and in infancy, so that villages can come to life.

Thus, iodine deficiency is a major block to the human and social development of communities living in an iodine-deficient environment. Correction of the iodine deficiency is indicated as a major contribution to economic development. An increase in physical and mental energy leads to improved work output, improved learning by children, and improved quality of life. Improved livestock productivity (chickens, cattle, and sheep) is also a major economic benefit.

Magnitude of the Problem

The number of cases of IDD throughout the world was estimated by WHO in 1990 to be 1.6 billion, including more than 200 million cases with goiter and more than 20 million cases with some degree of brain damage due to the effects of iodine deficiency in pregnancy. Recent estimates of the population at risk have been increased to 2.2 billion, with the recognition that even mild iodine deficiency in the mother has effects on the fetus. There are now estimated to be 130 IDD-affected countries, including the most populous: Bangladesh, Brazil, China, India, Indonesia, and Nigeria. Therefore, there is a global scourge of great magnitude, which provides one of the major challenges in international health today.

Correction of Iodine Deficiency

Iodized Salt

Since the successful introduction of iodized salt in Switzerland and the United States in the 1920s, successful programs have been reported from a number of countries, including those in Central and South America (e.g., Guatemala and Colombia) and Finland and Taiwan. However, there has been great difficulty in sustaining these programs in Central and South America mainly due to political instability. Following the breakup of the Soviet Union, iodine deficiency recurred in the Central Asian republics.

The difficulties in the production and quality maintenance of iodized salt for the millions who are iodine deficient, especially in Asia, were vividly demonstrated in India, where there was a breakdown in supply. These difficulties led to the adoption of universal salt iodization (USI) for India and subsequently for many other countries. This policy includes legislation to provide for compulsory iodization of all salt for human and animal consumption, and this legislation makes it illegal for noniodized salt to be available for human or animal consumption.

In Asia, the cost of iodized salt production and distribution is on the order of 3–5 cents per person per year. This must be considered cheap in relation to the social benefits that have already been described.

However, there is still the problem of the iodine in the salt actually reaching the iodine-deficient subject. There may be a problem with distribution or preservation of the iodine content: It may be left uncovered or exposed to heat. Thus, it should be added after cooking to reduce the loss of iodine.

Potassium iodate is the preferred vehicle compared to potassium iodide because of its greater stability in the tropical environment. A dose of 20–40 mg iodine as potassium iodate per kilo is recommended to cover losses to ensure an adequate household level. This assumes a salt intake of 10 g per day; if the level is below this, then an appropriate correction can readily be made by increasing the concentration of potassium iodate.

Iodized Oil

Iodized oil by injection or by mouth is singularly appropriate for isolated communities characteristic of mountainous endemic goiter areas. The striking regression of goiter following iodized oil administration, with improved well-being from correction of hypothyroidism, ensures general acceptance of the measure (**Figure 5**).

Iodized oil is more expensive than iodized salt but is used especially for severe iodine deficiency in remote areas. It provides instant correction of the deficiency and the consequent prevention of brain damage.

Figure 5 Subsidence of goiter in a New Guinea woman 3 months after the injection of iodized oil. This is accompanied by a feeling of well-being due to a rise in the level of the thyroid hormone in the blood. This makes the injections very popular. (Reproduced from Hetzel BS (1989) *The Story of Iodine Deficiency: An international Challenge in Nutrition.* Oxford: Oxford University Press.)

In a suitable area, the oil (1 ml contains 480 mg iodine) should be administered to all females up to the age of 40 years and all males up to the age of 20 years. A dose of 480 mg will provide coverage for 1 year by mouth and for 2 years by injection.

Iodized Milk

This is particularly important for infants receiving formula milk as an alternative to breast-feeding. An increase in levels from 5 to 10 μg/dl has been recommended for full-term infants and 20 μg/dl for premature infants. However, breast-fed infants will be iodine deficient if the mother is iodine deficient.

Iodized milk has been available in the United States, the United Kingdom and Northern Europe, Australia, and New Zealand as a result of the addition of iodophors as disinfectants by the dairy industry. This has been a major factor in the elimination of iodine deficiency in these countries. However, in most countries of Southern Europe and Eastern Europe, this has not occurred and the risk of iodine deficiency continues. Recently, the use of iodophors has been phased out, with a substantial decrease in the level of urine iodine excretion. Recurrence of iodine deficiency has been confirmed in Australia and New Zealand.

The Role of the United Nations

In 1990 the United Nations Sub-Committee on Nutrition recognized IDD as a major international public health problem and adopted a global plan for the elimination of IDD by the year 2000 proposed by the International Council for Control of Iodine Deficiency Disorders (ICCIDD) working in close collaboration with UNICEF and WHO.

The ICCIDD, founded in 1986, is an independent multidisciplinary expert group of more than 700 professionals in public health, medical, and nutritional science, technologists, and planners from more than 90 countries.

In 1990, the World Health Assembly and the World Summit for Children both accepted the goal of elimination of IDD as a public health problem by the year 2000. These major meetings included government representatives, including heads of state at the World Summit for Children, from 71 countries, and an additional 88 countries signed the plan of action for elimination of IDD as well as other major problems in nutrition and health.

Since 1989, a series of joint WHO/UNICEF/ICCIDD regional meetings have been held to assist countries with their national programs for the elimination of IDD. The impact of these meetings has been that governments now better realize the importance of iodine deficiency to the future potential of their people.

A dramatic example is provided by the government of the People's Republic of China. As is well-known, China has a one child per family policy, which means that an avoidable hazard such as iodine deficiency should be eliminated. In China, iodine deficiency is a threat to 40% of the population due to the highly mountainous terrain and flooded river valleys—in excess of 400 million people at risk. In recognition of this massive threat to the Chinese people, in 1993 the government held a national advocacy meeting in the Great Hall of the People sponsored by the Chinese Premier, Li Peng. The commitment of the government to the elimination of iodine deficiency was emphasized by Vice Premier Zhu Rongyi to the assembly of provincial delegations led by the provincial governors and the representatives of international agencies.

In 1998, an international workshop was held in Beijing by the Ministry of Health of China with the ICCIDD. Dramatic progress was reported, as indicated by a reduction in mean goiter rate (from 20 to 10%) with normal urine iodine levels. Severe iodine deficiency has persisted in Tibet due to difficulty in the implementation of salt iodization. In other provinces, excess iodine intake was noted in 10% of the population. The need for continuation of monitoring with urine iodine was emphasized at the meeting. Tibet is now receiving special assistance with a program supported by WHO, UNICEF, and the Australian Aid Program (AusAID).

Elimination of Iodine Deficiency Disorders at the Country Level

It is now recognized that an effective national program for the elimination of IDD requires a multisectoral approach as shown in **Figure 6**, which provides a model in the form of a wheel.

This wheel model represents the continuous feedback process involved in the national IDD control (elimination) program. All actors in the program need to understand the whole social process. The wheel must keep turning to maintain an effective program.

The wheel model also shows the social process involved in a national IDD control program. The successful achievement of this process requires the establishment of a national IDD control commission, with full political and legislative authority to carry out the program.

The program consists of the following components:

1. Assessment of the situation requires baseline IDD prevalence surveys, including measurement of urinary iodine levels and an analysis of the salt economy.
2. Dissemination of findings implies communication to health professionals and the public so that there is complete understanding of the IDD problem and the potential benefits of elimination of the most common preventable cause of brain damage.
3. Development of a plan of action includes the establishment of an intersectoral task force on IDD and the formulation of a strategy document on achieving the elimination of IDD.

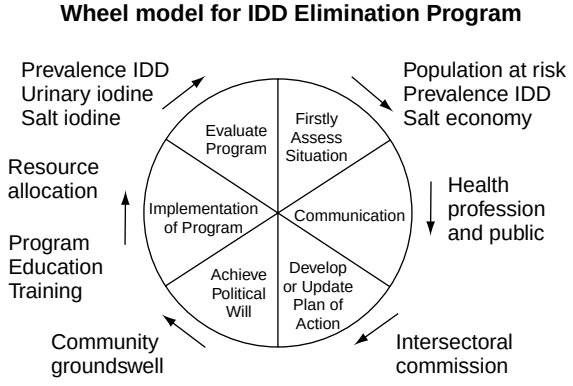


Figure 6 Wheel model for the iodine deficiency disorders (IDD) elimination program. The model shows the social process involved in a national IDD control program. The successful achievement of this process requires the establishment of a national IDD control commission, with full political and legislative authority to carry out the program. (Reproduced from Hetzel BS (1989) *The Story of Iodine Deficiency: An international Challenge in Nutrition.* Oxford: Oxford University Press.)

4. Achieving political will requires intensive education and lobbying of politicians and other opinion leaders.
5. Implementation requires the complete involvement of the salt industry. Special measures, such as negotiations for monitoring and quality control of imported iodized salt, will be required. It will also be necessary to ensure that iodized salt delivery systems reach all affected populations, including the neediest. In addition, the establishment of cooperatives for small producers, or restructuring to larger units of production, may be needed. Implementation will require training in management, salt technology, laboratory methods, and communication at all levels.

   In addition, a community education campaign is required to educate all age groups about the effects of iodine deficiency, with particular emphasis on the brain.
6. Monitoring and evaluation require the establishment of an efficient system for the collection of relevant scientific data on salt iodine content and urinary iodine levels. This includes suitable laboratory facilities.

Striking progress with USI has occurred, as indicated by the WHO/UNICEF/ICCIDD report to the 1999 World Health Assembly. Data show that of 5 billion people living in countries with IDD, 68% now have access to iodized salt. Of the 130 IDD-affected countries, it was reported that 105 (81%) had an intersectoral coordinating body and 98 (75%) had legislation in place.

Criteria for tracking progress toward the goal of elimination of IDD have been agreed on by ICCIDD, WHO, and UNICEF. These include salt iodine (90% effectively iodized) and urine iodine in the normal range (median excretion, 100–200 μg/l).

The major challenge is not only the achievement but also the sustainability of effective salt iodization. In the past, a number of countries have achieved effective salt iodization, but in the absence of monitoring the program lapsed with recurrence of IDD. To this end, ICCIDD, WHO, and UNICEF offer help to governments with partnership evaluation to assess progress toward the goal and also provide help to overcome any bottlenecks obstructing progress.

The Global Partnership

Since 1990, a remarkable informal global partnership has come together composed of the people and countries with an IDD problem, international agencies (particularly UNICEF, WHO, and

ICCIDD), bilateral aid agencies (Australia (AusAID) and Canada (CIDA)), the salt industry (including the private sector), and Kiwanis International. Kiwanis International is a world service club with 500,000 members throughout the world that has achieved a fundraising target of $75 million toward the elimination of IDD through UNICEF.

This partnership exists to support countries and governments in their elimination of IDD.

A more recent development is the establishment of the Global Network for the Sustainable Elimination of Iodine Deficiency, in collaboration with the salt industry.

The achievement of the global elimination of iodine deficiency will be a great triumph in international health in the field of noninfectious disease, ranking with the eradication of the infectious diseases smallpox and polio.

However, the goal of elimination is a continuing challenge. Sustained political will at both the people and the government level is necessary to bring the benefits to the many millions who suffer the effects of iodine deficiency.

Nomenclature

Endemic Occurrence of a disease confined to a community

Endemic Cretinism A state resulting from the loss of function of the maternal thyroid gland due to iodine deficiency during pregnancy characterised by mental defect, deaf-mutism and spastic paralysis in its fully developed form

Goiter An enlarged thyroid gland most commonly due to iodine deficiency in the diet

Hypothyroidism The result of a lowered level of circulating thyroid hormone causing loss of mental and physical energy

Hyperthyroidism The result of excessive circulating thyroid hormone with nervousness, sweating, tremor, with a rapid heart rate and loss of weight

ICCIDD International Council for Control of Iodine Deficiency Disorders-an international non-government organization made up of a network of 700 health professionals from more than 90 countries available to assist IDD elimination programs in affected countries

IDD Iodine Deficiency Disorders referring to all the effects of iodine deficiency in a population that can be prevented by correction of the iodine deficiency

IIH Iodine Induced Hyperthyroidism-due to increase in iodine intake following long standing iodine deficiency. The condition is transient and no longer occurs following correction of iodine deficiency

Iodization The general term covering fortification programs using various agents (iodide, iodate) or various vehicles (salt, oil, bread and water)

Iodized Oil Iodine in poppy seed oil-lipiodol is extensively used in radiology as a radio-contrast medium to demonstrate holes (cavities) in the lung. Available both by injection (lipiodol) and by mouth (oriodol) for the instant correction of iodine deficiency

Iodized Salt Salt to which potassium iodate or potassium iodide has been added at a recommended level of 20–40 milligrams of iodine per kilogram of salt

Kiwanis International A World Service Group including more than 10,000 clubs and over 500,000 members based in the USA

Thyroid size Measured by ultrasound-a much more sensitive and reproducible measurement than is possible by palpation of the thyroid

Thyroxine Thyroid Hormone (T_4) an amino acid which includes four iodine atoms

Triiodothyronine A more rapidly active thyroid hormone (T_3) which includes 3 iodine atoms on the amino acid molecule

UNICEF United Nations Children's Fund

USI Universal Salt Iodization-iodization of all salt for human and animal consumption which requires legislation and has been adopted by a number of countries

WHO World Health Organization-the expert group on health within the UN System

See also: **Food Fortification**: Developing Countries. **Iodine**: Physiology, Dietary Sources and Requirements. **Supplementation**: Role of Micronutrient Supplementation. **World Health Organization**.

Further Reading

Buttfield IH and Hetzel BS (1967) Endemic goiter in Eastern New Guinea with special reference to the use of iodized oil in prophylaxis and treatment. *Bulletin of the World Health Organization* **36**: 243–262.

Delange F, Dunn JT, and Glinoer D (eds.) (1993) *Iodine Deficiency in Europe: A Continuing Concern*, NATO ASI Series A: Life Sciences vol. 241. New York: Plenum.

Hetzel BS (1983) Iodine deficiency disorders (IDD) and their eradication. *Lancet* **2**: 1126–1129.

Hetzel BS (1989) *The Story of Iodine Deficiency: An International Challenge in Nutrition*. Oxford: Oxford University Press.

Hetzel BS and Pandav CS (eds.) (1996) *SOS for a Billion: The Conquest of Iodine Deficiency Disorders*, 2nd edn. Oxford: Oxford University Press.

Hetzel BS, Pandav CS, Dunn JT, Ling J, and Delange F (2004) *The Global Program for the Elimination of Brain Damage Due to Iodine Deficiency*. Oxford: Oxford University Press.

Ma T, Lu T, Tan U *et al.* (1982) The present status of endemic goiter and endemic cretinism in China. *Food and Nutrition Bulletin* **4**: 13–19.

Pharoah POD, Buttfield IH, and Hetzel BS (1971) Neurological damage to the fetus resulting from severe iodine deficiency during pregnancy. *Lancet* **1**: 308–310.

Stanbury JB (ed.) (1994) *The Damaged Brain of Iodine Deficiency*. New York: Cognizant Communication Corporation.

Stanbury JB and Hetzel BS (eds.) (1980) *Endemic Goiter and Endemic Cretinism*. New York: John Wiley.

World Health Organization (1990) *Report to the 43rd World Health Assembly*. Geneva: World Health Organization.
World Health Organization (1996) *Recommended Iodine Levels in Salt and Guidelines for Monitoring Their Adequacy and Effectiveness*, WHO/NUT/96.13. Geneva: WHO/UNICEF/ICCIDD.
World Health Organization (1999) *Progress towards the Elimination of Iodine Deficiency Disorders (IDD)*, WHO/NHD/99.4. Geneva: World Health Organization.
WHO/UNICEF/ICCIDD (2001) Assessment of Iodine Deficiency Disorders and their Elimination: A guide for Program Managers WHO/NHD/01.1.

IRON

J R Hunt, USDA-ARS Grand Forks Human Nutrition Research Center, Grand Forks, ND, USA

Iron, the Earth's most abundant metal and fourth most common element, is also the essential nutrient that is most commonly deficient in human diets. At the beginning of the 21st century, the World Health Organization recognizes iron deficiency as one of the 10 greatest global health risks, ranked according to the number of lost healthy life years. Iron deficiency impairs reproductive performance, cognitive development, and work capacity. Effectively resolving this problem with preventative nutritional strategies remains an unmet challenge.

Iron Chemistry and Physiology

Body Content, Forms, and Function

Iron, the 26th element of the periodic table, has a molecular weight of 55.85. Two common aqueous oxidation states, ferrous (Fe^{2+}) and ferric (Fe^{3+}), enable iron to participate in oxidation/reduction reactions that are essential to energy metabolism by accepting or donating electrons. However, this property also enables free iron to catalyze oxidative reactions, resulting in reactive and damaging free radicals. Accordingly, body iron must be chemically bound to facilitate appropriate physiological function, transport, and storage, with minimal opportunity for free ionic iron to catalyze harmful oxidative reactions.

Most of the body's iron functions in heme protein complexes that transport oxygen as hemoglobin and myoglobin. Approximately two-thirds of the body iron is in hemoglobin, a 68,000 MW structure containing four subunits of heme, a protoporphyrin ring with iron in the center (**Figure 1**), and four polypeptide chains (two chains each of α- and β-globin). For transport by hemoglobin, oxygen bonds directly to the iron atom, stabilized in a Fe^{2+} oxidation state surrounded by the protoporphyrin ring and histidine residues. Hemoglobin iron easily binds and releases oxygen, circulating in blood erythrocytes. Myoglobin, consisting of a single heme molecule and globin, enables oxygen transfer from erythrocytes to cellular mitochondria in muscle cytoplasm.

Smaller quantities of iron in the heme form function in mitrochondrial cytochromes involved with electron transfer, oxygen utilization, and the production of ATP. A small fraction of body iron functions in heme-containing hydrogen peroxidases such as catalase that protect against excessive hydrogen peroxide accumulation by catalyzing its conversion to hydrogen and oxygen.

Iron also functions in non-heme proteins that contain an iron–sulfur complex, a cubical arrangement of four iron and four sulfur atoms. This is the principal form of iron in mitochondria, functioning in enzymes of energy metabolism such as aconitase, NADH dehydrogenase, and succinate dehydrogenase. In both mitochondria and cytosol, aconitase

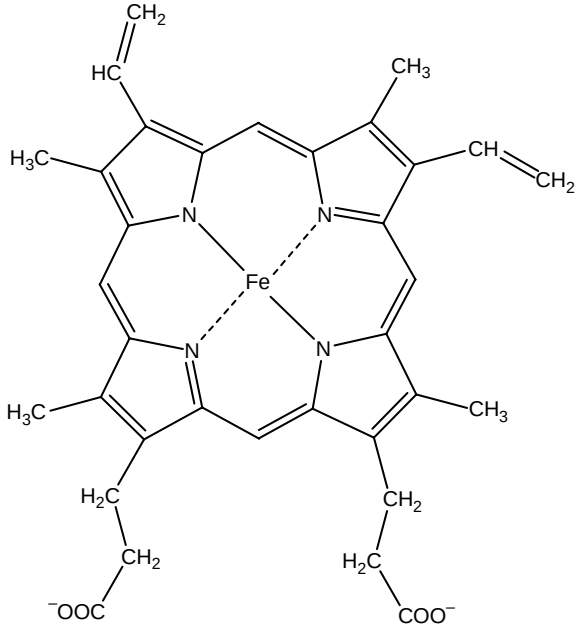


Figure 1 Heme (ferroprotoporphyrin 9).

is sensitive to iron concentrations. When iron is abundant, the aconitase enzyme assumes the full iron–sulfur cubic structure that is associated with carbohydrate metabolism. However, when iron concentrations are reduced, the protein loses aconitase activity and functions as an iron binding protein (IRP). IRPs interact with iron response elements (IREs) of the mRNA to regulate the synthesis of proteins involved with iron transport, storage, and use, in response to changes in cellular iron concentrations.

Absorption, Excretion, Transport, and Storage

Absorption Both heme and non-heme (inorganic) iron are absorbed in an inverse proportion to body iron stores (indicated by serum ferritin; **Figure 2**). Heme iron is absorbed more efficiently than the non-heme form. Non-heme iron absorption can vary from 0.1 to >35% and that of heme iron from 20 to 50%, depending on body iron status (stores, erythropoiesis, and hypoxia) and dietary bioavailability. These ranges indicate greater control of non-heme compared to heme iron absorption. When iron stores are high, absorption of non-heme iron can be minimized more completely, and when iron stores are low, non-heme iron is absorbed nearly as efficiently as heme iron. Because there is considerably more non-heme iron in the diet (~85–100%), this form accounts for most of the physiological control of iron absorption in relation to iron needs.

The upper portion of the duodenum, with its low pH luminal conditions, is the primary site for both heme and non-heme iron absorption (**Figure 3**). Non-heme iron absorption is better understood

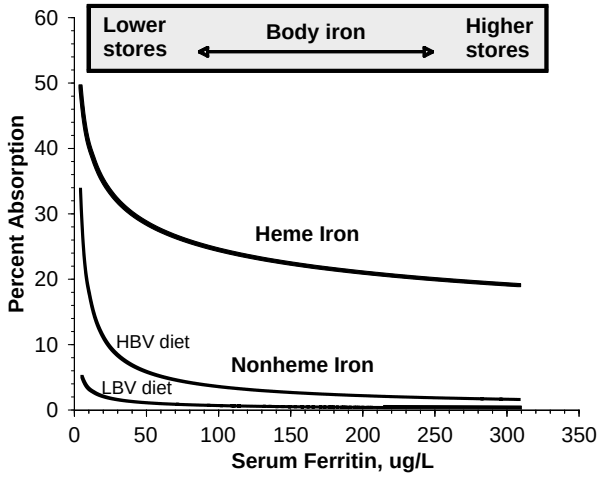


Figure 2 Heme and non-heme iron absorption as influenced by body iron stores and dietary bioavailability. HBV and LBV indicate high and low dietary bioavailability, respectively.

than heme iron absorption, and only receptors for mucosal uptake of non-heme iron have been identified. The globin proteins of hemoglobin are proteolytically digested in the intestinal lumen, producing peptide remnants that may enhance the absorption of the heme molecule by preventing heme polymerization. The heme molecule is absorbed as an intact porphyrin structure, possibly involving endocytosis. In the mucosal cell, heme iron is split into ferrous iron and bilirubin by heme oxygenase, adding to a common pool of cellular iron for transport into plasma or intracellular storage and exfoliation.

Non-heme iron is best absorbed if presented to the intestinal villi as soluble ions (preferably reduced, ferrous ions) or as low-affinity, low-molecular-weight iron ligands. Stomach acid facilitates these conditions. Ascorbic acid concurrently ingested with iron helps to maintain the iron in a soluble, reduced, low-molecular ligand form in the intestinal lumen. Mucin, an intraluminal protein, has been proposed to bind iron and facilitate duodenal uptake.

Proteins involved in mucosal uptake and transfer of non-heme iron as well as possible regulatory molecules have been identified (**Figures 3 and 4**). These include duodenyl cytochrome *b* (Dcytb), which converts ferric to ferrous iron at the apical mucosal surface. A divalent metal transporter (DMT-1) transfers ferrous iron into the mucosal cell. Mutations in DMT-1 impair iron absorption and produce microcytic anemia in rodents. Ferrous iron has the highest affinity for DMT-1, but it will also transport other divalent ions, such as manganese, lead, cadmium, zinc, and copper. This may contribute to competitive inhibition observed in the absorption of these metals. Ferric iron is transported into the mucosal cell by mobilferrin, followed by ferroreduction with the protein paraferritin. Iron transported into the enterocyte may be further transported to the body at the basolateral membrane, completing absorption, or may be held and returned to the intestinal lumen with cellular desquamation. Ireg-1, or ferroportin, is involved in efflux of iron from the mucosal cell at the basolateral membrane. A mutation in Ireg-1 results in an uncommon form of hemochromatosis, an iron storage disorder. The mRNA for both DMT-1 and Ireg-1 contain an IRE, enabling regulation of mRNA translation by intracellular iron concentrations. Dcytb, DMT-1, and Ireg-1 are all upregulated in iron deficiency. Intestinal transfer of iron to the circulation also involves hephaestin, an intestinal ferroxidase with a protein sequence similar to that of ceruloplasmin (a copper-containing ferroxidase in serum). A defective hephaestin gene in mice results in anemia and

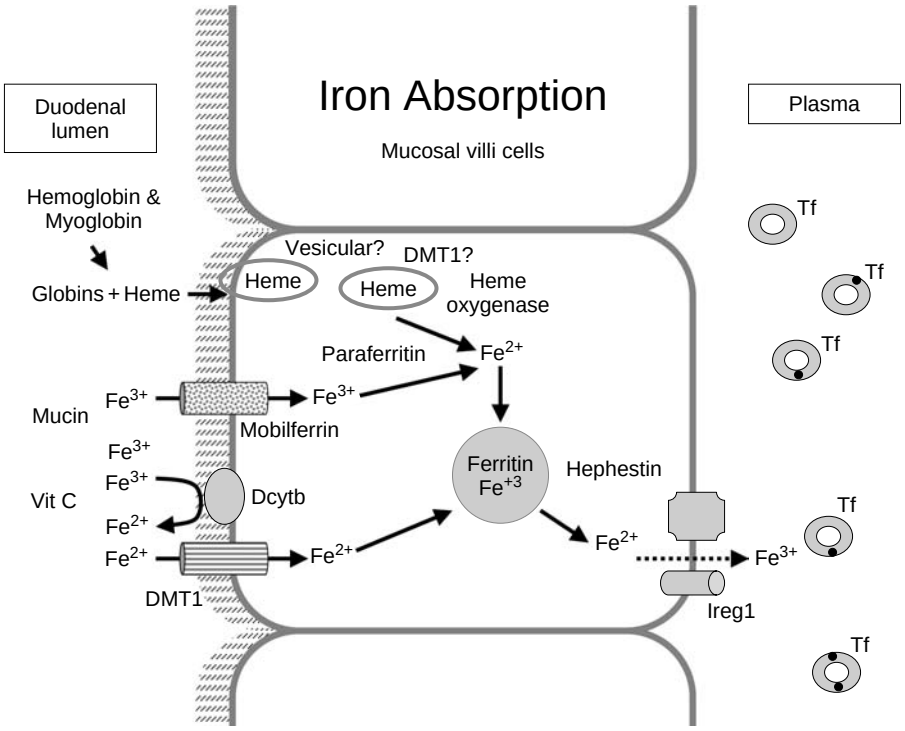


Figure 3 Absorption of iron in the intestinal mucosa.

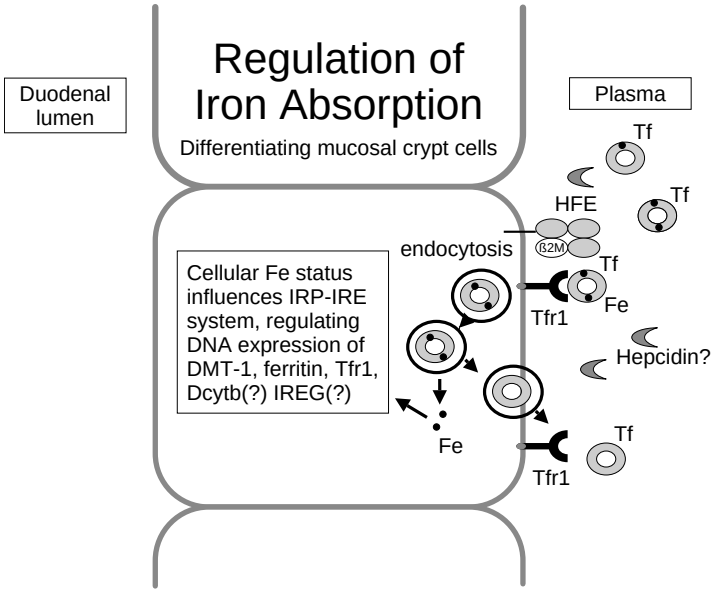


Figure 4 Regulation of iron absorption in the mucosal crypt cells before differentiation and development into actively absorbing intestinal villi cells.

accumulation of iron in intestinal cells. However, unlike Dcytb, DMT-1, or Ireg-1, hephaestin is not preferentially expressed in the duodenum, the main site of iron absorption.

Iron absorption is responsive to recent iron intake, iron stores, erythropoiesis, hypoxia, pregnancy, and inflammation. A newly identified peptide, hepcidin, may be related to several of these stimuli of regulatory control. Hepcidin is an antimicrobial peptide found in human blood and urine that apparently serves as a signal for limiting iron absorption. Control of absorption also likely involves the HFE protein located in the basolateral membrane of intestinal crypt cells. A specific point mutation in

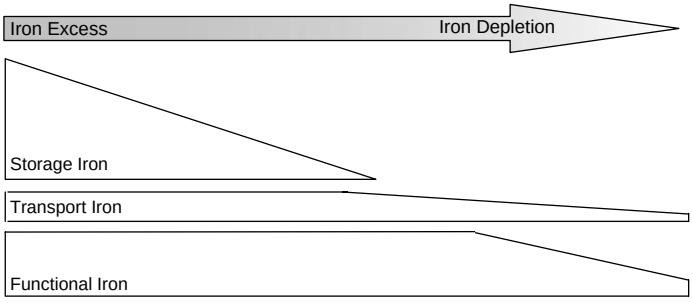
the HFE gene is associated with the most common form of hemochromatosis, a disorder involving excessive iron absorption and accumulation. The HFE protein interacts with β_2-microglobin and transferrin receptor, apparently influencing iron uptake from serum transferrin, the primary protein involved in serum iron transport (**Figure 4**). Knowledge of the control of iron absorption is growing rapidly.

Transport Transferrin transports essentially all of the 3 or 4 mg of iron in blood serum, including dietary iron absorbed from the duodenum as well as iron from macrophages after the degradation of hemoglobin. Each transferrin molecule binds two iron atoms; the transferrin in serum is normally approximately one-third saturated with iron. The amount of iron that can be bound by transferrin is measured as the total iron binding capacity (TIBC). In iron deficiency, serum iron is reduced, and TIBC is elevated; expressing serum iron as a fraction of the TIBC defines the transferrin saturation, which is reduced in iron deficiency. As iron deficiency develops, these measures of iron transport signal iron deficiency before the functional pool of circulating hemoglobin is reduced (**Figure 5**).

Membrane transferrin receptors enable the cellular uptake of iron. Transferrin receptors complex with transferrin, the complex is internalized by endocytosis, and the iron is released to the cell from transferrin upon vesicular acidification (**Figure 4**). Transferrin receptors are abundant in erythrocyte precursors, placenta, and liver, and the number of receptors changes inversely with cellular iron status. Serum transferrin receptors are a soluble, truncated form of the cellular receptors, present in proportion to the cellular receptors, which serve as a clinical indicator of cellular iron status that is useful in distinguishing between iron deficiency and other causes of anemia.

Other proteins involved in iron transport include lactoferrin, which is structurally similar to transferrin and occurs in body fluids such as milk and semen. Haptoglobin and hemopexin proteins clear hemoglobin and heme, respectively, from circulation as they are released from senescent red blood cells.

Storage Iron is primarily stored in liver, spleen, and bone marrow in the form of ferritin or hemosiderin. Ferritin is a water-soluble protein complex of 24 polypeptide subunits in a spherical cluster with a hollow center that contains up to 25% iron by weight, or 4000 atoms of iron per molecule.



	Excessive Body Iron	Adequate Iron	Low Iron Stores	Iron Deficiency	Iron Deficiency Anemia
Serum Ferritin	↑	---	↓	↓↓	↓↓
Transferrin Receptor (sTfR)	---	---	---	↑	↑↑
Index: Log (sTfR/ferritin)	↑	---	↓	↓↓	↓↓↓
Transferrin Saturation	↑↑	---	---	↓	↓
Serum Iron	↑↑	---	---	↓	↓
Total Iron Binding Capacity	---	---	---	↑	↑
Erythrocyte Protoporphyrin	---	---	---	↑	↑↑
Hemoglobin, Hematocrit	---	---	---	---	↓

Figure 5 Clinical indicators of body iron status.

Hemosiderin is a water-insoluble complex, immunologically similar to ferritin, containing up to 35% iron. Ferritin and hemosiderin each account for approximately half of the storage iron in liver.

Excretion The approximately 1 mg of iron lost daily by men and postmenopausal women represents mainly obligatory fecal losses from exfoliated mucosal cells, bile, and extravasated red cells, with minor additional amounts in desquamated skin cells and sweat. Urine contains minimal amounts of iron.

Adolescent girls and premenopausal women excrete considerable amounts of iron through menstruation. The menstrual losses of individual women vary considerably; half of women lose less than 14 mg of iron per menstrual period, but the distribution is highly skewed, and 5% lose 50 mg or more. Iron deficiencies among women in prosperous countries are commonly attributable to these high iron excretion rates rather that to differences in dietary intakes.

Body iron balance The body contains 2–4 g of total iron, or approximately 50 mg/kg in men and 40 mg/kg in women. Red blood cells contain approximately two-thirds of body iron and have an average life span of 120 days; consequently, approximately 20 mg of iron daily is efficiently recycled from senescent to newly formed erythrocytes through the reticuloendothelial system.

In contrast to other nutrients, controlled through both absorption and excretion, body iron balance is controlled almost exclusively by absorption. Approximately 10–20 mg iron is consumed daily from food. Average absorption and excretion of iron for adult men or postmenopausal women is approximately 1 mg daily. Menstruation can more than double iron losses in women of child-bearing age, increasing their requirement for absorbed iron. Iron balance is also challenged by the growth demands of pregnancy and early childhood.

Clinical Assessment of Iron Status

With adequate iron status, there is sufficient iron to meet all of iron's functional roles and a small reserve of storage iron that can be mobilized when needed (**Figure 5**). Excessive body iron, stored in liver and bone marrow, is marked by elevated serum ferritin and also serum iron and transferrin saturation. Ferritin in plasma corresponds well with body iron stores, but its use as an indicator is limited under inflammatory conditions. Iron deficiency occurs when iron stores are depleted and the iron transported for physiological function is reduced. Iron deficiency without anemia is commonly detected from abnormal values for two out of three blood indices, usually serum ferritin, transferrin saturation, and free erythrocyte protoporphyrin (**Figure 5**). As iron deficiency becomes more severe, iron deficiency anemia results, with small, pale erythocytes and reduced blood hemoglobin and hematocrit. Measurement of hemoglobin in reticulocytes, or immature red blood cells, is a possible new tool to assess developing anemias. The ratio of serum transferrin receptor to serum ferritin provides a single, sensitive indicator of iron status across the full range of body iron status, except under conditions of inflammatory stress.

Iron Nutrition

Iron Deficiency

Iron deficiency is the most common of nutrient deficiencies, affecting as many as two-thirds of all children and women of child-bearing age worldwide. Iron deficiency severe enough to cause anemia affects 20–25% of infants and as many as 40% of women and 25% of men. Iron deficiency occurs in industrially developed as well as developing populations. In the United States, 9–11% of toddlers, adolescent girls, and women of child-bearing age have iron deficiency, and 2–4% have iron deficiency anemia. The prevalence of iron deficiency is approximately doubled in US black and Hispanic women.

Consequences of Iron Deficiency

Iron deficiency adversely affects pregnancy, impairs early childhood development and cognitive function, and reduces the ability to do physical work. These serious problems are almost exclusively associated with iron deficiency severe enough to cause anemia; however, small reductions in exercise capacity, detectable in a laboratory setting, are also detectable in women with low iron stores and no anemia.

Physical work capacity Iron deficiency anemia adversely affects physical work capacity, reflecting the element's key role in oxygen and energy utilization. Maximal oxygen consumption during exercise is reduced, in association with decreased muscle myoglobin and other iron-containing enzymes. Iron supplementation has improved productivity among Guatemalan sugar and coffee plantation workers, Indian tea pickers, and Indonesian road construction workers and rubber tappers. Iron supplementation programs are clearly cost-effective in addition to providing a positive impact on human health and well-being.

Cognitive development In infants, iron deficiency anemia has been associated with reduced mental and motor test scores and behavioral changes such as being more hesitant and wary. This impaired mental and motor functioning appears to persist after treatment with iron, emphasizing the need for early detection and treatment and preferably prevention of iron deficiency during early development.

Reproduction Iron deficiency anemia has been associated with greater perinatal maternal and infant mortality, premature birth, and low birth weight. Iron supplementation during pregnancy has not been completely effective in preventing maternal anemia, leading to suggestions for promoting adequate iron stores in all women of child-bearing age prior to conception.

Other Iron deficiency increases the susceptibility to lead poisoning. It may also impair resistance to infection and regulation of body temperature. Iron deficiency has been associated with the eating of non-food material (pica) or ice (pagophagia). Clinical signs may include spoon-shaped fingernails and abnormalities of the mucosa of the mouth and gastrointestinal tract.

Recommended Dietary Intakes

The US and Canadian recommended iron intakes are intended to meet the requirements of 97.5% of the healthy population, replacing excreted iron and maintaining essential iron functions with a minimal supply of body iron stores. They also assume a relatively high bioavailability of the dietary iron. The recommended 8 mg daily for adult men and postmenopausal women can easily be met with varied Western-style diets. More careful food choices are needed to obtain the 18 mg recommended to meet requirements for 97.5% of adult menstruating women. This higher recommendation reflects the high menstrual iron losses of some women; the median iron requirement is 8.1 mg for menstruating women.

During pregnancy, dietary iron recommendations are increased to 27 mg daily, based on the iron content of the fetus and placenta (approximately 320 mg) as well as the expanded blood volume associated with a healthy pregnancy. Meeting this recommendation generally requires iron supplementation. Supplementation with 30–60 mg daily is commonly recommended. Lactation has minimal impact on maternal iron balance and recommendations.

The high iron requirements of early growth put infants and toddlers at risk of iron deficiency. Breast-feeding is recommended for the first year of life. Although iron in breast milk is relatively low (0.35 mg/l, or 0.27 mg daily), it is well absorbed, possibly because of lactoferrin. Breast milk alone is assumed to be adequate for the first 6 months of infancy, with the addition of iron-rich foods in the next 6 months. When prepared formula is used, iron fortification of the formula is recommended.

Dietary recommendations at other ages reflect the increased needs of active growth periods, such as adolescence. Western dietary recommendations have been based on mixed diets with a relatively high bioavailability of iron and may need to be increased twofold or more for low meat, plant-based diets with greater phytic acid content (see Bioavailability).

Other factors that may increase dietary requirements include achlorhydria, which decreases iron absorption, hookworm or other parasites that increase gastrointestinal blood loss, or intrauterine contraceptive devices that may increase menstrual losses by 30–50%. In contrast, hormonal contraceptives reduce iron requirements by reducing menstrual losses by approximately 50%.

Dietary Iron

Food Sources

Typical Western diets contain approximately 6 mg iron per 1000 kcal. Men and women consume approximately 16–18 and 12–14 mg daily, respectively. In the United States, 24% of dietary iron is supplied by breads, pasta, and bakery products. An additional 21% comes from (mostly fortified) cereal products. Other abundant dietary sources are red meats (9% from beef), poultry, legumes, and lentils. In countries such as the United States, fortification practices increase the influence of grain and cereal products as sources of iron. In countries without fortification to at least replace the iron lost during milling, the refinement of grain products considerably reduces dietary iron content. The populations of developing countries that eat little meat and do not include legumes or lentils as a dietary staple are at increased risk of inadequate iron intake.

Bioavailability

In underdeveloped countries, diets may be inadequate in both iron content and bioavailability (the amount that is absorbed and utilized by the body). However, the bioavailability of iron can be more important than the iron content in determining the amount of iron absorbed from food. Diets with similar total iron contents can differ 8- to 10-fold

in the amount of absorbable iron. Dietary iron bioavailability is high from refined Western diets containing meat, poultry, and fish and abundant sources of ascorbic acid with low consumption of phytic acid from whole grains and legumes and limited drinking of coffee and tea with meals. On average, men absorb 1 mg daily from such diets, and women, with their lower iron stores, absorb approximately 2 mg. Individuals may absorb considerably more or less, depending on their body iron stores (**Figure 2**).

Despite the considerable differences in dietary iron bioavailability observed with absorption measurements, dietary changes are slow to influence biochemical indices of iron status. However, people following vegetarian diets for years have lower iron stores than their omnivorous counterparts, and consumption of red meat is often a predictor of iron status in epidemiological studies.

Heme iron Approximately 10%, or 1 or 2 mg, of the iron in a mixed, Western diet is in the well-absorbed heme form. Heme iron accounts for approximately 40% of the iron in meat, poultry, or fish flesh. There is little to no heme iron in organ meats, dairy products, or foods of plant origin. Heme iron is absorbed as an intact porphyrin structure. Heme iron absorption is enhanced by meat, poultry, or fish and is reduced by calcium consumed concurrently, but it is not influenced by the other enhancers and inhibitors of non-heme iron absorption.

Non-heme iron Non-heme iron accounts for 85–100% of dietary iron. In contrast to heme iron, the absorption of non-heme iron is substantially influenced by dietary enhancers and inhibitors consumed concurrently. These factors appear to affect the solubility of a single exchangeable pool of non-heme iron absorbed from the intestinal digestate.

Absorption of non-heme iron is enhanced by ascorbic acid, which reduces ferric iron to ferrous iron, resulting in a soluble iron–ascorbic acid complex. Enhanced absorption has been demonstrated with synthetic as well as several food sources of ascorbic acid. The enhancement increases logarithmically with the dose, approximately doubling absorption with as little as 25 mg of ascorbic acid and increasing absorption by nearly 10-fold with 1000 mg of ascorbic acid.

Non-heme iron absorption is also enhanced by concurrently consuming meat, poultry, or fish. Despite intensive study, the factor responsible for this enhancement by animal flesh has not been identified and may involve the general matrix of low-molecular-weight peptides released during digestion.

Non-heme iron absorption is reduced by phytic acid (inositol hexaphosphate), present in legumes, rice, and grains, that binds iron and makes it insoluble. Both phytate and iron are concentrated in the aleurone layer and germ of grains, and they are reduced with milling, which increases the bioavailability of the remaining iron. An additional unidentified factor in soy beans, independent of the phytic acid, also impairs iron absorption. Polyphenols in grains, fruits, and vegetables, and including the tannins in tea and coffee, also inhibit non-heme iron absorption. Ascorbic acid consumed concurrently can partially reduce the inhibition of non-heme iron absorption by both phytic acid and polyphenols. Calcium in supplemental quantities inhibits both heme and non-heme iron absorption from foods. Supplemental zinc also inhibits non-heme iron absorption.

Supplementation and Fortification

The serious international problem of iron deficiency has been met with poor success by supplementation and fortification efforts. Both approaches suffer from difficulties in delivery and acceptance. Supplements that readily ionize into the ferrous form, such as ferrous sulfate, ferrous fumerate, or ferrous gluconate, are highly bioavailable but may cause gastrointestinal discomfort. Iron injections are poorly tolerated and can result in serious infections. Because daily supplementation reduces the physiological efficiency of iron absorption, routine weekly iron supplementation with 60 mg iron has been suggested in developing countries for women of childbearing age, beginning in adolescence. Menstruating women in more prosperous countries are advised to obtain assessment from a health professional before taking iron supplements in excess of 20 mg daily.

Fortification of staple foods with 3–10 mg iron daily, depending on the needs of the population, is a long-term preventative strategy. In the United States, bread and cereal products are routinely fortified with 20 mg iron per pound (460 g) of flour, and additional fortification at the option of food suppliers is common. However, fortification is difficult when food processing is decentralized, as is common in poor populations. Food fortification carries the additional challenge that the chemical forms of iron most bioavailable also tend to be the most reactive with the food fortified, resulting in adverse changes in flavor, color, and shelf life. Promising approaches include the fortification of food sauces with iron chemically bound with amino acids or with EDTA (sodium iron ethylenediaminetetraacetic acid), which are well absorbed even in the presence

of phytic acid. Elemental iron powders, commonly referred to as carbonyl, electrolytic, and reduced forms of iron, are relatively inert in foods and inexpensive, but their bioavailability may be 30–80% less than iron from ferrous sulfate, depending on the dissolution in the gastrointestinal tract. Ferric orthophosphate and ferric pyrophosphate do not adversely affect foods but are poorly bioavailable; however, efforts are under way to enhance their bioavailability by reducing the particle size and encapsulating the particles with various lipids or carbohydrates to prevent agglomeration.

Excessive Intakes

An extensive biological control system limits the occurrence of free ionic iron that can readily participate in toxic, free radical-producing reactions. Large quantities of ingested iron are acutely toxic, and accidental ingestion of medicinal iron preparations is a leading cause of poisoning deaths in young children. Iron supplementation is also associated with gastrointestinal irritation. Iron supplements adversely affect absorption of zinc. Iron absorption is well controlled, but iron overload can result from excessive parenteral iron administration or blood transfusions. Dietary iron overload, possibly exacerbated by genetic factors, occurs in sub-Saharan tribes that consume a high-iron traditional beer prepared and stored in iron containers. Genetic factors can substantially influence body iron retention, as indicated by hemochromatosis, a relatively frequent iron storage disorder of northern European descendants characterized by excessive iron absorption and leading to life-threatening iron damage of organs in adulthood. The possible association of high iron stores with increased risk of diseases related to oxidative stress, including cardiovascular disease, diabetes, and cancer, is an area of epidemiological investigation.

See also: **Adolescents**: Nutritional Requirements. **Anemia**: Iron-Deficiency Anemia. **Bioavailability**. **Breast Feeding**. **Food Fortification**: Developed Countries; Developing Countries. **Pregnancy**: Nutrient Requirements. **Supplementation**: Dietary Supplements; Role of Micronutrient Supplementation; Developing Countries; Developed Countries.

Further Reading

Brugnara C (2003) Iron deficiency and erythropoiesis: New diagnostic approaches. *Clinical Chemistry* **49**: 1573–1578.

Centers for Disease Control and Prevention (1998) Recommendations to prevent and control iron deficiency in the United States. *MMWR Recommendations and Reports* **47**: 1–29.

Eisenstein RS and Ross KL (2003) Novel roles for iron regulatory proteins in the adaptive response to iron deficiency. *Journal of Nutrition* **133**: 1510S–1516S.

Institute of Medicine, Food and Nutrition Board (2001) *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc*. Washington, DC: National Academy Press.

Knutson M and Wessling-Resnick M (2003) Iron metabolism in the reticuloendothelial system. *Critical Reviews in Biochemistry and Molecular Biology* **38**: 61–88.

Mielczarek EV and McGrayne SB (2000) *Iron, Nature's Universal Element: Why People Need Iron and Animals Make Magnets*. New Brunswick, NJ: Rutgers University Press.

Miret S, Simpson RJ, and McKie AT (2003) Physiology and molecular biology of dietary iron absorption. *Annual Review of Nutrition*.

Ischemic Heart Disease *see* **Coronary Heart Disease**: Lipid Theory

K

Keshan Disease *see* **Selenium**

KETOSIS

D H Williamson†, Radcliffe Infirmary, Oxford, UK

This article is a revision of the previous edition article by D H Williamson, pp. 1160–1167, © 1999, Elsevier Ltd.

The two ketone bodies, acetoacetate ($CH_3COCH_2COO^-$) and D-3-hydroxybutyrate ($CH_3CHOHCH_2COO^-$), are the only freely soluble lipids in the circulation.

The name ketone bodies originates from the German *Ketonkörper* (literally, ketones excreted from the body) and refers to their discovery in the urine of diabetic patients in the latter half of the nineteenth century. In reality, the term is a misnomer because 3-hydroxybutyrate is not a ketone. It arose because the reagent originally used reacted positively with ketones in diabetic urine. Acetone (CH_3COCH_3), the product of the spontaneous decarboxylation of acetoacetate, is also a ketone and is present in blood and urine when the plasma concentration of acetoacetate is elevated. It is excreted via the kidneys and lungs and is responsible for the sweet smell on the breath in ketotic states.

The association of ketone bodies with the pathology of diabetes resulted in the view that they were toxic waste products. It is only in the past 30 years that this view has been convincingly reversed. Two factors led to this change, namely the development of an enzymatic method for the determination of acetoacetate and 3-hydroxybutyrate, which in turn allowed the dramatic finding of Cahill and colleagues in 1967 that adult human brain removed appreciable amounts of ketone bodies from the circulation in prolonged starvation.

†Deceased.

The aim in this contribution is to review (a) the formation of ketone bodies in physiological and pathological situations, and (b) the function of ketone bodies as physiological substrates and signals.

Formation of Ketone Bodies

It is well established that in humans and other mammals the only organ that contributes significant amounts of ketone bodies to the blood is the liver; this organ, unlike peripheral tissues, is unable to utilize ketone bodies to any appreciable extent. More recently it has been found that during the suckling period (high-fat diet) the intestine also has the capacity (about 10% of the liver) to produce ketone bodies. Whether ketone bodies are used *in situ* or are transported via the portal blood to supplement the existing hyperketonemia is an open question.

The main blood-borne substrates for the synthesis of ketone bodies (ketogenesis) are the nonesterified fatty acids; others of lesser importance are the branched-chain amino acids, leucine and isoleucine. In addition, acetate (sources: intestinal fermentation, in vinegar or an oxidation product of ethanol) is a ketogenic substrate.

Long-chain fatty acids contained in dietary lipids do not enter the portal blood directly but are esterified in the intestinal cells, packaged with proteins and phospholipids to form chylomicrons (large lipoproteins), and transported via the lymphatic system to the thoracic duct where they enter the blood. In contrast, the short- and medium-chain fatty acids (below C_{14}) contained in dairy products or in clinical medium-chain triacylglycerol preparations are directly absorbed as the respective fatty acids and are transported to the liver via the portal blood (**Figure 1**). The long-chain

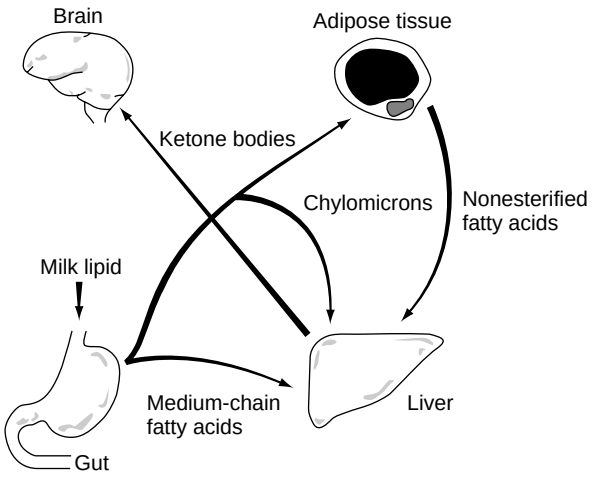


Figure 1 Intertissue fluxes of substrates in the suckling neonate. Thickness of line denotes rate of flux.

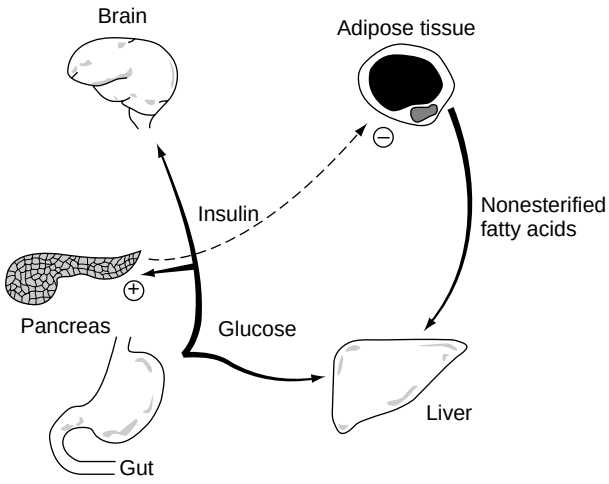


Figure 2 Intertissue fluxes of substrates in the fed state. Thickness of line denotes rate of flux.

fatty acids in the plasma are bound to albumin and are released from adipose tissue triacylglycerol stores by the process of lipolysis.

Extrahepatic Regulation

A key factor in the regulation of ketogenesis is the availability of nonesterified long-chain fatty acids to the liver, which in turn is controlled by their release from adipose tissue. The enzyme responsible for the initiation of the hydrolysis of stored triacylglycerols to fatty acids is hormone-sensitive lipase. As its name implies, this enzyme is exquisitely sensitive to hormones: adrenaline (in the plasma) and noradrenaline (released from sympathetic nerve endings) are activators, whereas insulin inhibits the activity. In small mammals glucagon is also an activator of the enzyme, but this does not seem to be the case in the human.

Insulin has an additional effect on the net release of long-chain fatty acids from adipose tissue in that it stimulates their reesterification to triacylglycerols. Thus after a high-carbohydrate meal, when insulin secretion and its concentration in the plasma is high, the release of fatty acids from adipose tissue is suppressed and their concentration in the plasma is low (**Figure 2**). In contrast, during stress, when adrenaline and noradrenaline are elevated, the release of fatty acids is increased and their plasma concentration is high.

In experimental animals increased plasma ketone body concentrations (hyperketonemia) can inhibit adipose tissue lipolysis (a) indirectly by increasing the secretion of insulin or (b) by a direct effect on the tissue (**Figure 3**). This can be viewed as a feedback mechanism for controlling the rate of ketogenesis via fatty acid supply to the liver, but whether this is important in the human is not known. In contrast, the supply of short- and medium-chain fatty acids to the liver is mainly dependent on the dietary intake and on the proportion that escapes further metabolism in the intestinal tract; there is no known involvement of hormones in the process.

Intrahepatic Regulation

There are situations (e.g., stress) where the supply of fatty acids to the liver may be increased, but there is no necessity to increase the availability of ketone bodies to the peripheral tissues. Consequently, there is a requirement that the rate of hepatic ketogenesis should be controlled independently of the supply of fatty acids. However, it must be stressed that without an increase in the supply of fatty acids the rate of ketogenesis cannot increase.

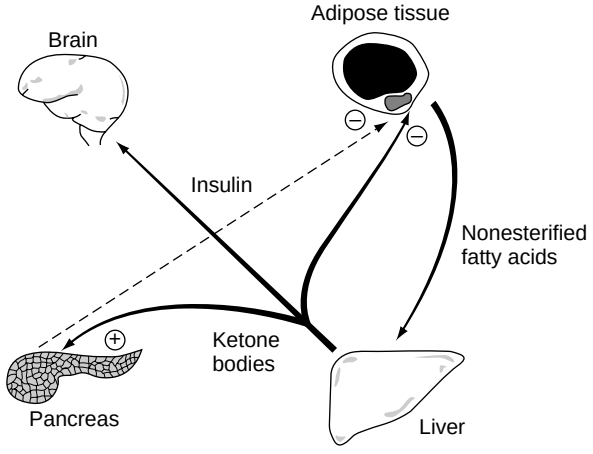


Figure 3 Role of ketone bodies as feedback regulators.

Much of the current interest is concerned with how the intrahepatic metabolism of fatty acids (**Figure 4**) is regulated. Long-chain fatty acids entering the liver have three main fates:

1. They can be re-esterified to phospholipids and triacylglycerols and then be secreted as very low-density lipoproteins (VLDL).
2. They can be oxidized via the mitochondrial β-oxidation complex to acetyl-CoA. The latter can combine with another molecule of acetyl-CoA in the reaction catalysed by acetoacetyl-CoA thiolase and then enter the hydroxymethylglutaryl-CoA pathway to form acetoacetate.
3. The acetyl-CoA derived from the fatty acids can be completely oxidized in the tricarboxylate cycle.

The short- and medium-chain fatty acids cannot be re-esterified to any appreciable extent in mammalian liver and therefore they are either metabolized to ketone bodies or completely oxidized. In addition, unlike the long-chain fatty acids, they are transported directly into the mitochondrial matrix without the need to be converted first to the corresponding acyl-CoA derivatives.

The role of malonyl CoA The entry of free long-chain fatty acids into the hepatocyte is via a specific carrier on the plasma membrane. Once inside the cytosol the long-chain fatty acids are bound to binding proteins, converted to the acyl-CoA derivatives, and then can either be esterified or enter

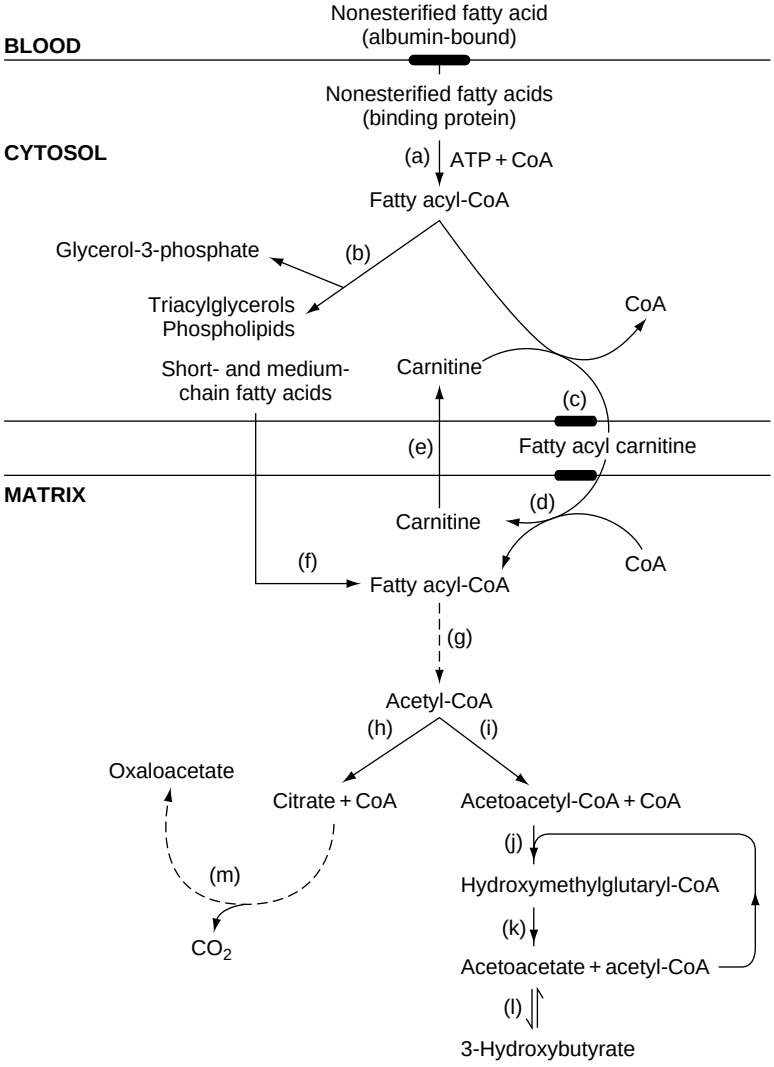


Figure 4 Pathway of fatty acid catabolism in liver. Enzymes involved: (a) long-chain fatty acyl-CoA synthetase; (b) glycerol-3-phosphate acyl-CoA transferase; (c) CAT I; (d) CAT II; (e) carnitine exchange; (f) short- and medium-chain fatty acyl-CoA synthetase; (g) fatty acid oxidation complex; (h) citrate synthase; (i) acetoacetyl-CoA thiolase; (j) hydroxymethylglutaryl-CoA synthase; (k) hydroxymethylglutaryl-CoA lyase; (l) hydroxybutyrate dehydrogenase; (m) tricarboxylate cycle.

the mitochondria via a complex transport system, the carnitine–acyl-CoA transferase (CAT) system. This consists of two proteins: CAT I located on the outer mitochondrial membrane and CAT II on the inner mitochondrial membrane (**Figure 5**). The overall action of the two enzymes results in the transfer of a long-chain fatty acyl-CoA to the mitochondrial matrix and the return of free carnitine to the cytosol via an exchange mechanism. Although carnitine is not consumed in the reaction, the available concentration can be critical. In nutritional carnitine deficiency there is impairment of long-chain fatty acid oxidation and ketogenesis.

The activity of CAT I is the key to the intrahepatic regulation of fatty acid metabolism in most situations. Its activity increases in ketogenic situations. More importantly, CAT I is inhibited by malonyl-CoA and the sensitivity of CAT I to this inhibitor changes in various pathophysiological situations such as fasting or diabetes.

As malonyl-CoA is a key intermediate in the synthesis of fatty acids (lipogenesis) from products (pyruvate and lactate) of glucose metabolism, this interaction provides a regulatory link between lipid and carbohydrate metabolism (**Figure 5**). Thus on high-carbohydrate diets, when the rate of hepatic lipogenesis, and consequently the cytosolic concentration of malonyl-CoA, is high, the activity of CAT I will be inhibited and fatty acids will be diverted to esterified products and secretion as VLDL rather than oxidation and conversion to ketone bodies. Conversely, on high-fat diets or in starvation, when lipogenesis is inhibited, malonyl-CoA concentration is low and CAT I is active. The sensitivity of CAT I to malonyl-CoA generally correlates with the prevailing concentration of the latter.

The short- and medium-chain fatty acids do not utilize the CAT I and II system to enter the mitochondrial matrix and therefore their oxidation is not greatly influenced by the prevailing 'carbohydrate status' (amount of glycogen, direction of carbohydrate flux, glycolysis, or gluconeogenesis) of the liver (**Figure 5**).

Insulin can rapidly depress the rate of ketogenesis *in vitro*. This effect is thought to result mainly from its stimulatory action on a key enzyme of lipogenesis, acetyl-CoA carboxylase, which in turn increases the concentration of malonyl-CoA. Glucagon and the catecholamines have the opposite effect. Thus hormonal effects can be exerted at both the extrahepatic (lipolysis) and intrahepatic (modulation of lipogenesis) levels.

Intramitochondrial regulation Once the fatty acyl-CoA molecule is attached to the mitochondrial β-oxidation complex there appears to be little regulation exerted until release of the acetyl-CoA fragments. As indicated above, the acetyl-CoA can enter the tricarboxylate cycle and be oxidized to CO_2 or can be converted to ketone bodies via the hydroxymethylglutaryl-CoA pathway.

It appears that in most experimental situations the complete oxidation of fatty acids proceeds at a low, but relatively similar, rate and it is the activity of the

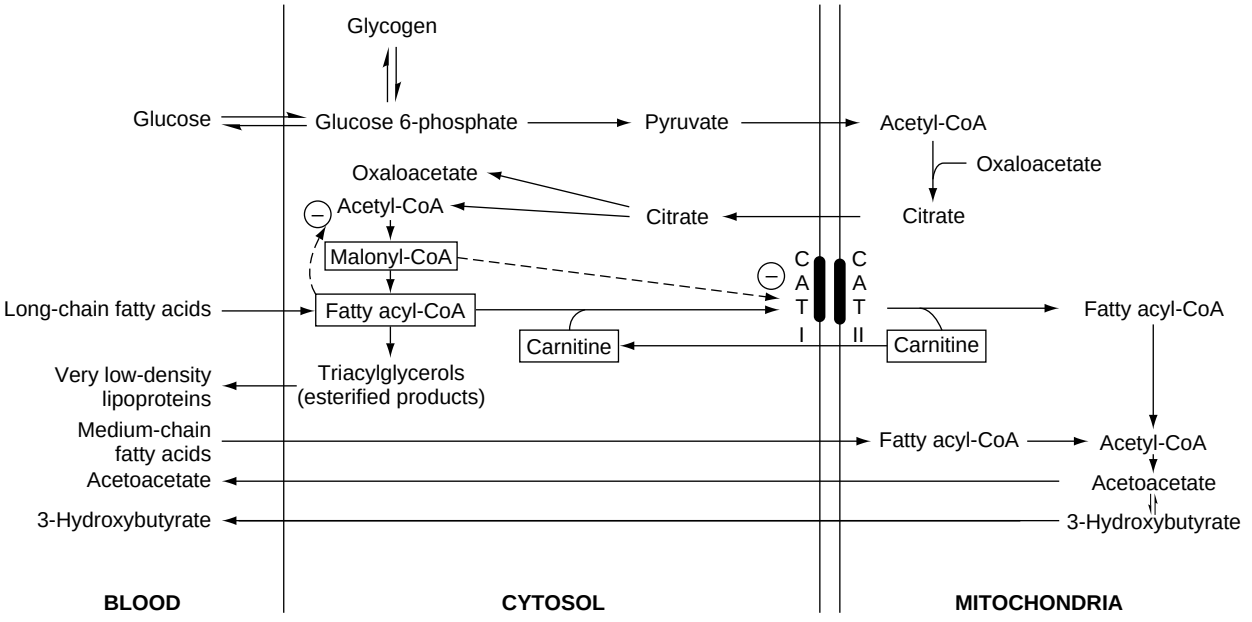


Figure 5 Interrelationship between hepatic carbohydrate metabolism, lipogenesis, and ketogenesis. Circled minus signs indicate inhibition by the metabolite.

hydroxymethylglutaryl-CoA pathway that shows larger changes. This has led to the view that the pathway might be regulated by mechanisms other than substrate supply.

Studies on the expression of 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA) synthase have shown that both the mRNA coding for the protein and the amount of protein increase during the onset of ketogenic states (fasting, diabetes) and that these changes are rapidly reversed (refeeding, insulin treatment). However, the finding that rates of ketogenesis from medium-chain fatty acids (CAT I and II) do not alter greatly with change in physiological state, if the rate of fatty acid supply is held constant, would seem to rule out appreciable regulation within the hydroxymethylglutaryl-CoA pathway. Indeed, current thinking suggests that the activity of CAT I is the primary intrahepatic site for the regulation of fatty acid oxidation and ketogenesis. If there is another important site, particularly during situations associated with the reversal of ketogenesis, it is likely to be proximal to the step catalysed by this protein (e.g., the supply of fatty acids to the liver). Thus *in vivo* there is little doubt that the primary step that controls ketogenic flux is the rate of long-chain fatty acid release from adipose tissue.

Function of Ketone Bodies

The major role of ketone bodies is to supply an alternative oxidizable substrate to glucose for the brain in situations where the availability of the latter is impaired (e.g., starvation). In addition, ketone bodies can act as precursors for the acetyl-CoA required in neural lipid synthesis (myelin). Other mammalian tissues, including heart, skeletal muscle, kidney, and lactating mammary gland, can utilize ketone bodies but, in contrast to glucose utilization, no energy can be obtained in the absence of oxygen. In these tissues metabolism of ketone bodies results in the inhibition of glucose utilization and inhibition of the oxidation of pyruvate. The net result is a sparing of carbohydrate for the brain and the strictly glycolytic tissues (erythrocytes, retina).

Pathways of Ketone Body Utilization

Mitochondrial pathway The major site of ketone body utilization in peripheral tissues is the mitochondria (**Figure 6**). Although transporters for ketone bodies have been described on the plasma and inner mitochondrial membranes of some tissues, these do not appear to limit the flux. The initiating enzyme for acetoacetate metabolism is 3-oxoacid-CoA transferase:

$$\text{Acetoacetate} + \text{succinyl-CoA} \rightleftharpoons \text{acetoacetyl-CoA} + \text{succinate}$$

The resulting acetoacetyl-CoA is cleaved to two molecules of acetyl-CoA by acetoacetyl-CoA thiolase; they are then oxidized in the tricarboxylate cycle.

3-Hydroxybutyrate is converted to acetoacetate by 3-hydroxybutyrate dehydrogenase:

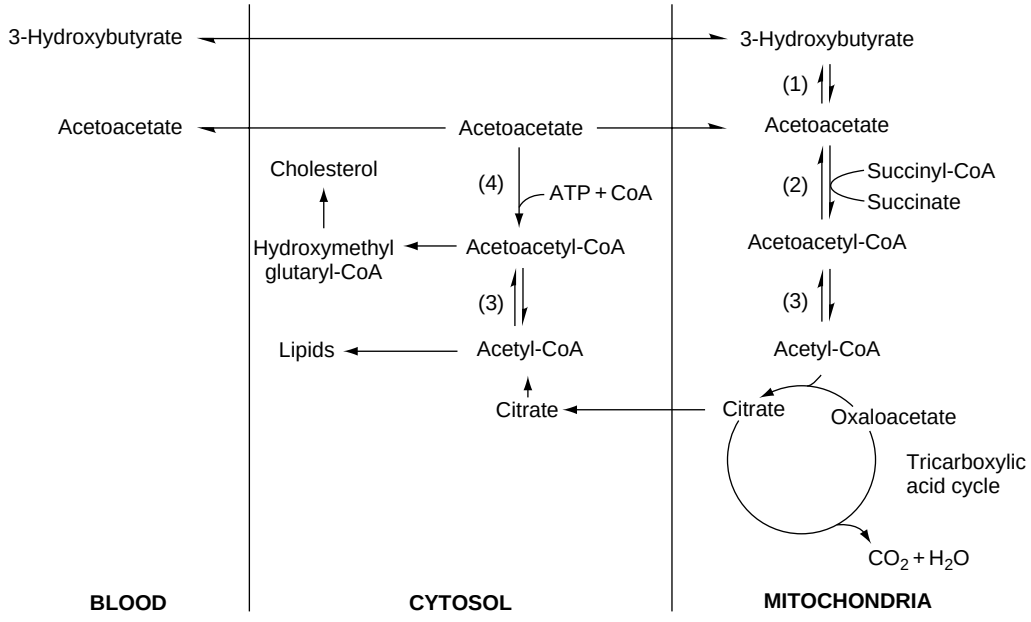


Figure 6 Pathways of ketone body utilization in peripheral tissues. (1) Hydroxybutyrate dehydrogenase, (2) 3-oxoacid-CoA transferase; (3) acetoacetyl-CoA thiolase; (4) acetoacetyl-CoA synthetase.

$$\text{3-Hydroxybutyrate} + NAD^+ \rightleftharpoons \text{acetoacetate} + NADH + H^+$$

The ready reversibility of the three enzymes of the mitochondrial pathway (**Figure 6**) means that if the overall system is near equilibrium within the cell *in vivo*, the utilization of the ketone bodies will be dependent on their respective concentrations and on the rate of removal of the products. Thus acetoacetate utilization will be promoted when mitochondrial acetyl-CoA is decreased, whereas an increase in the latter will have the opposite effect. Similarly, oxidation of hydroxybutyrate will increase if the concentrations of $NADH_2$ and acetoacetate fall. Unlike the hepatic hydroxymethylglutaryl-CoA pathway for ketogenesis, which is essentially irreversible, the free reversibility of this pathway in peripheral tissues can be viewed as means of buffering the mitochondrial acetyl-CoA pool and hence energy production. Some of the acetyl-CoA can be transported to the cytosol in the form of citrate to act as a precursor for lipogenesis (**Figure 6**).

Cytosolic pathway The cytosol of tissues where active lipogenesis occurs (adipose tissue, developing brain, lactating mammary gland, and liver) contains an enzyme, acetoacetyl-CoA synthetase, which converts acetoacetate to acetoacetyl-CoA (**Figure 6**):

$$\text{Acetoacetate} + ATP + CoA \rightarrow \text{acetoacetyl-CoA} + AMP + \text{pyrophosphate}$$

Its activity is at least an order of magnitude lower than that of the mitochondrial 3-oxoacid-CoA transferase, whereas its affinity for acetoacetate is appreciably higher. The presence of acetoacetyl-CoA thiolase in the cytosol allows the conversion of acetoacetate to acetyl-CoA and then to lipids without the involvement of the mitochondria.

Brain cytosol also contains 3-hydroxy-3-methylglutaryl-CoA synthase, allowing acetoacetate to act as a direct precursor for sterol synthesis. Evidence from *in vivo* experiments with ^{14}C-labelled acetoacetate has confirmed the existence of this pathway in developing brain and liver. The cytosolic route for acetoacetate utilization can be seen as a mechanism for directing this substrate to lipid or sterol synthesis rather than to oxidation.

Ketosis

The concentration of ketone bodies in the blood at any time represents a balance between the rate of hepatic ketogenesis and the rate of utilization by peripheral tissues. It is generally assumed that an increase in ketogenesis leads to a rise in blood ketone bodies, which in turn results in their increased utilization. In rare situations, such as congenital absence of key enzymes involved in ketone body utilization (e.g., 3-oxoacid-CoA transferase) or inhibition of these enzymes by pharmacological agents, blood ketone bodies may increase without any concomitant increase in ketogenesis.

The concentration of ketone bodies in the blood is exquisitely sensitive to changes in pathophysiological state. It is therefore useful to define *normoketonemia* in mammals as a concentration of total ketone bodies in blood below $0.2\,mmol\,l^{-1}$, *hyperketonemia* as above this level, and *ketoacidosis* (ketosis; by analogy to the definition of lactic acidosis) as above $7\,mmol\,l^{-1}$. In adult mammals there are small but characteristic diurnal variations in ketone body concentrations. Larger increases in concentration occur in man in response to change in pathophysiological state (**Table 1**). The concentrations span a 200-fold range and it is this which underlines the important role of ketone bodies as substrates and signals.

Physiological Ketosis

Physiological hyperketonemia is found in the suckling neonate (high-fat diet of the milk; **Figure 1**), postexercise (depletion of hepatic glycogen reserves), and after prolonged fasting (more than 24 h; **Figure 7**). All these situations have in common a low hepatic carbohydrate status (depletion of glycogen and/or activation of gluconeogenesis) and therefore from a physiological standpoint one would expect an increased rate of ketogenesis. Comparison of the factors which can influence ketogenesis in suckling and fasting (**Table 2**) shows the expected broad agreement.

Table 1 Range of blood ketone body concentrations in humans

Situation	*Ketone body concentration (mmol l^{-1})*
Fed normal diet	about 0.1
Fed high-fat diet	up to 3
Fasted: 12–24 h	up to 0.3
Fasted: 48–72 h	2.0–3.0
Postexercise	up to 2
Late pregnancy	up to 1
Late pregnancy: fasted 48 h	4.0–6.0
Neonate: 0–1 days	0.2–0.5
Neonate: 5–10 days	0.7–1.0
Hypoglycemia	1.0–5.0
Untreated diabetes mellitus	up to 25

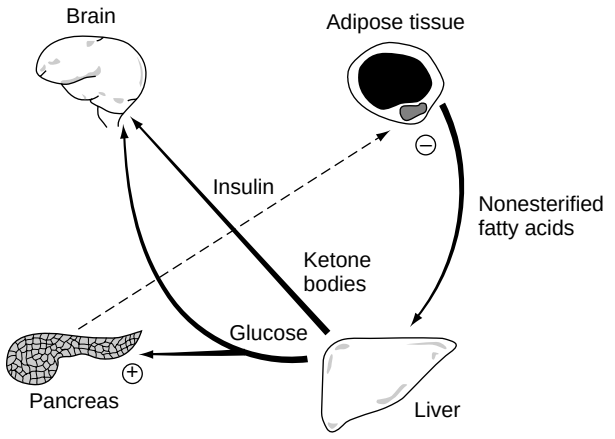


Figure 7 Intertissue fluxes of substrates in the starved state. Thickness of line denotes rate of flux.

More detailed information on the hierarchy of the regulatory factors during onset and reversal of ketogenesis has been obtained for the fasting state by measurements at short time intervals. The first event after withdrawal of food is a lowering of plasma insulin accompanied by an increase in plasma fatty acids (stimulation of lipolysis). However, for an appreciable period (8–10 h) there is no increase in blood ketone bodies or in the *in vitro* rates of hepatic ketogenesis (measured with saturating fatty acid concentrations). The major increment in ketogenic rate occurs at the nadir of the hepatic malonyl-CoA concentrations and when the sensitivity of CAT I to malonyl-CoA is starting to increase rapidly. This long time lag before a change in sensitivity of the protein to malonyl-CoA inhibition is thought to be due to the time required to bring about alterations to the lipid environment of the outer mitochondrial membrane.

Confirmation of this view is that on refeeding, when insulin rapidly increases and plasma fatty acids decrease with a parallel decrease in blood ketone bodies, there is again a time lag before malonyl-CoA concentrations rise and a longer one before sensitivity returns. In physiological and nutritional terms this delay of return to the normal fed settings of intrahepatic regulation makes excellent sense. It is only when the refeeding consists primarily of large amounts of carbohydrate that the starved liver needs to inhibit the activity of CAT I to prevent the oxidation of newly synthesized fatty acids. If the meal consists mainly of lipid with little carbohydrate the activity of CAT I needs to remain high to allow oxidation of the excess fatty acids. Thus the liver must sense a prolonged increase in plasma insulin before the high activity of CAT I is suppressed.

Table 2 Comparison of factors influencing ketogenesis in suckling and fasted states

Factor	*Suckling*	*Fasted*
Plasma nonesterified fatty acids	Increased	Increased
Plasma insulin	Decreased	Decreased
Plasma glucagon	Increased	Increased
Hepatic carnitine	Increased	Increased
Hepatic lipogenesis	Decreased	Decreased
Hepatic malonyl-CoA	Decreased	Decreased
Hepatic CAT I activity	Increased	Increased
Sensitivity to malonyl-CoA	Decreased	Decreased

Pathological ketosis

The major example of pathological ketosis is of course insulin-dependent or type 1 diabetes. Essentially the changes in this condition are similar to those that occur during fasting, but they are more pronounced. Insulin is absent or very low in the plasma and therefore there is no antagonistic action to restrain the opposing hormones, adrenaline, noradrenaline, and glucagon. Consequently, lipolysis in adipose tissue is greatly stimulated and plasma fatty acids increase to high levels.

The lack of insulin and the large flux of fatty acids to the liver means that lipognesis is inhibited at the level of acetyl-CoA carboxylase and there is the expected decrease in malonyl-CoA concentration. In addition, the sensitivity of CAT I to inhibition by malonyl-CoA is considerably decreased. The level of expression of hepatic CAT I and II proteins also increases several-fold in diabetes. Thus the liver is in the ideal mode for producing excessive amounts of ketone bodies.

It has been suggested that diversion of oxaloacetate to hepatic glucose synthesis (which is also increased in insulin deficiency) may also play a role in the increased rate of ketogenesis by diverting acetyl-CoA from the tricarboxylate cycle. However, present evidence suggests that this makes a minor contribution. Although the excessive output of ketone bodies by the liver undoubtedly makes the major contribution to their high levels in the blood, it is likely that there is also a degree of underutilization by peripheral tissues. The net result is ketoacidosis and excretion of large amounts of energy as ketone bodies in the urine.

A rare, but intriguing, example of pathological ketosis (ketone bodies up to $10\,mmol\,l^{-1}$) is the inborn error of hepatic glycogen synthase deficiency (**Figure 8**). Here glycogen is virtually absent from the liver so that after short-term fasting (5–10 h) the glucose falls to hypoglycemic levels, plasma insulin is decreased, plasma fatty acids increase, and

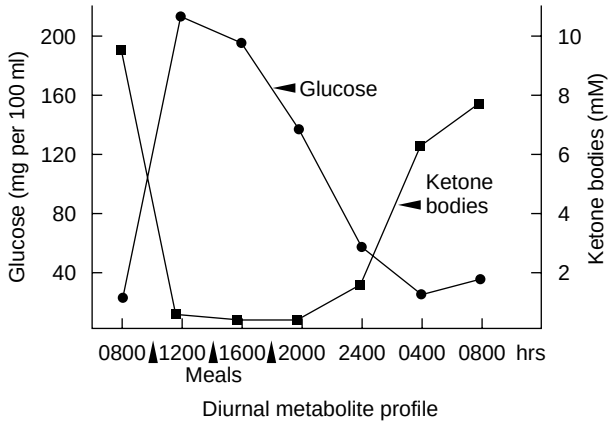


Figure 8 Diurnal blood metabolite profile of a child with glycogen synthetase deficiency. Values taken from Aynsley-Green A, Williamson DH and Gitzelmann R (1977) *Archives of Disease in Childhood* **52**: 573–579. (With permission from BMJ Publishing Group.)

ketogenesis is switched on. On consuming a meal the pattern is reversed until the blood glucose falls again. This case illustrates the importance of hepatic glycogen (and its mobilization) in the smooth transition of substrate supply from the fed to the fasted state. Treatment in this case was to recommend the consumption of more frequent high-carbohydrate snacks. It is of interest that this particular child suffered no ill effects from the daily exposure to high concentrations of ketone bodies, underlining their role as normal substrates for the brain when available.

Metabolic Acidosis

The great disadvantage of ketone bodies is that both acetoacetate and hydroxybutyrate are relatively strong acids. When they increase to high concentration there is the expected decrease in the blood pH, the plasma hydrogen carbonate concentration, and the partial pressure of carbon dioxide in blood and body fluids. The symptoms of acidosis include malaise, weakness, anorexia, and vomiting and these may eventually lead to coma. Treatment of diabetic ketoacidosis is to give insulin as soon as possible, usually as a continuous intravenous infusion. This rapidly decreases the raised plasma fatty acids and more slowly lowers the blood glucose and ketone bodies. Prolonged starvation, where the blood ketone bodies may reach 8–10 $\text{mmol}\,\text{l}^{-1}$, does not usually cause a serious disturbance of the acid–base balance. Loss of ketone bodies via the urine occurs but is not excessive. The nonenzymic decarboxylation of acetoacetate to acetone and carbon dioxide can be seen as a primitive mechanism for removing the potential acidotic effects of ketone bodies. The fact that acetone can be converted to glucose by the liver at low rates is an extra bonus.

The other common form of metabolic acidosis is lactic acidosis. This can arise because of infection, tissue hypoxia (anaerobic glycolysis), can be drug induced (ethanol, hypoglycemic biguanides), or can arise because of a congenital defect (pyruvate dehydrogenase or pyruvate carboxylase deficiency). In addition to the acidosis caused by lactic acid or ketone bodies there is a group of organic acidurias (some 25–30 different types) in which an inborn error results in the accumulation of an organic acid in the blood and urine. However, frank acidosis is not always associated with these conditions. The key investigation is chromatographic identification of the organic acid.

See also: **Adipose Tissue**. **Carbohydrates**: Regulation of Metabolism. **Cholesterol**: Sources, Absorption, Function and Metabolism. **Fatty Acids**: Metabolism. **Lactation**: Physiology. **Starvation and Fasting**.

Further Reading

Bach AC, Ingenbleek Y, and Frey A (1996) The usefulness of dietary medium-chain triglycerides in body weight control: fact or fancy? *Journal of Lipid Research* **37**: 708–726.

Girard JR, Ferré P, Pégorier JP, and Duée PH (1992) Adaptations of glucose and fatty acid metabolism during perinatal period and suckling–weanling transition. *Physiological Reviews* **72**: 507–562.

Krebs HA, Woods HF, and Alberti KGMM (1975) Hyperlactataemia and lactic acidosis. *Essays in Medical Biochemistry* **1**: 81–103.

McGarry JD and Foster DW (1980) Regulation of hepatic fatty acid oxidation and ketone body production. *Annual Review of Biochemistry* **49**: 395–420.

Nehlig A and de Vasconcelos AP (1993) Glucose and ketone body utilization by the brain of neonatal rats. *Progress in Neurobiology* **40**: 163–221.

Owen OE, Morgan AP, Kemp HG *et al.* (1967) Brain metabolism during fasting. *Journal of Clinical Investigation* **46**: 1589–1595.

Page MA and Williamson DH (1971) Enzymes of ketone body utilisation in human brain. *Lancet* **2**: 66–68.

Porter R and Lawrenson G (eds.) (1982) Metabolic acidosis. *Ciba Foundation Symposium*, **87**. London: Pitman.

Robinson AM and Williamson DH (1980) Physiological roles of ketone bodies as substrates and signals in mammalian tissues. *Physiological Reviews* **60**: 143–187.

Williamson DH (1982) The production and utilization of ketone bodies in the neonate. In: CT Jones (ed.) *The Biochemical Development of the Fetus*, pp. 621–650. Amsterdam: Elsevier Biomedical.

Williamson DH (1987) Brain substrates and the effects of nutrition. *Proceedings of Nutrition Society* **46**: 81–87.

Zammit VA (1996) Role of insulin in hepatic fatty acid partitioning: Emerging concepts. *Biochemical Journal* **314**: 1–14.

L

LACTATION

Contents
Physiology
Dietary Requirements

Physiology

J L McManaman and M C Neville, University of Colorado, Denver, CO, USA

Lactation is a uniquely mammalian physiological process in which the caloric and nutrient reserves of the mother are transformed into a complex fluid capable of supporting the nutritional demands of newborns for sustained periods. Milk, the product of lactation, is a mixture of solutes whose composition reflects the activities of distinct secretion and transport processes of the mammary gland and mirrors the differing nutritional requirements of mammalian neonates. In humans, this fluid is capable of providing the full-term infant with all the nutrients required for the first 4–6 months of life as well as offering significant protection against infectious disease. Although artificial formulas are widely utilized for human infant nutrition in developed countries, many components of human milk, including critical growth factors, long-chain polyunsaturated fatty acids, antiinfectious oligosaccharides and glycoconjugates, and the protein lactoferrin, are not duplicated in formula. Although it is likely that such substances are beneficial even to healthy infants in well-protected environments, they are particularly important to infants living in conditions of inadequate sanitation, as well as to preterm infants and infants with feeding problems. Despite the obvious importance of milk to neonatal nutrition and the selective advantage of lactation in mammalian evolution, the physiological mechanisms underlying milk secretion and utilization are not well understood and the molecular mechanisms involved in the production of individual milk components are still poorly characterized. In this article, the functional anatomy of the mammary gland is described, followed by a brief description of human milk composition and a review of the transport mechanisms involved in the secretion of individual milk components. We then summarize the functional differentiation of the mammary gland and the initiation of lactation—a process that involves a series of carefully programmed functional changes that transform a prepared, but nonsecretory, gland into a fully functioning organ during the first week postpartum in humans.

Functional Anatomy of Lactation

The lactating mammary gland consists of an arborizing ductal network that extends from the nipple and terminates in grape-like lobular clusters of alveoli forming the lobuloalveolar unit, which is the site of milk secretion. A stylized diagram of these structures is shown in **Figure 1**. Alveoli are composed of a single layer of polarized secretory epithelial cells that possess specialized features indicative of highly developed biosynthetic and secretory capacities, including numerous mitochondria, an extensive rough endoplasmic reticulum network, and a well-developed Golgi apparatus. Secretory components including lipid droplets and casein containing secretory vesicles are found juxtaposed to the apical membrane of these cells. The epithelial cells are connected to each other through a junctional complex composed of adherens and tight-junctional elements that function to inhibit transfer of extracellular substances between the vascular system and milk compartments during lactation (**Figure 2**). The basal portion of alveolar epithelial cells is surrounded by a meshwork of myoepithelial cell processes that contract to bring about milk

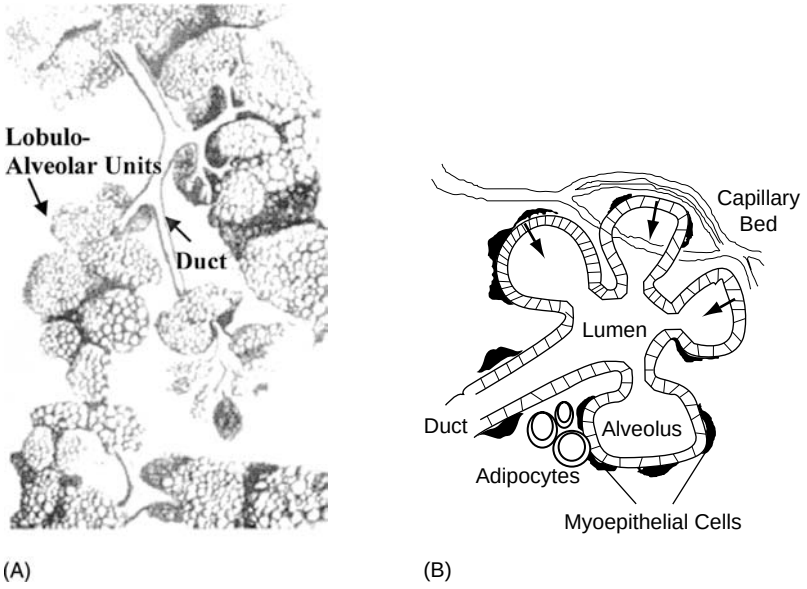


Figure 1 (A) Camera lucida drawing of a section of the breast of a woman who died 2 days after last suckling her infant. The drawing clearly shows collecting ducts and the grape-like lobuloalveolar units, which are engorged with milk. (From Dabelow A (1941) *Morphology Journal* **85**: 361–416.) (B) Cross-sectional diagram showing the relationship of the lobuloalveolar unit composed of milk secreting alveoli and ducts to the other cellular compartments of the mammary gland. Arrows indicate milk secretion by the alveolar epithelial cells into the lumen.

ejection and by a connective tissue stroma that supports and separates the lobules. The stromal component also contains lymphatics and becomes extensively vascularized during lactation to sustain the biosynthetic demands of alveolar epithelial cells. In nonpregnant, nonlactating animals the stroma contains a large adipose component.

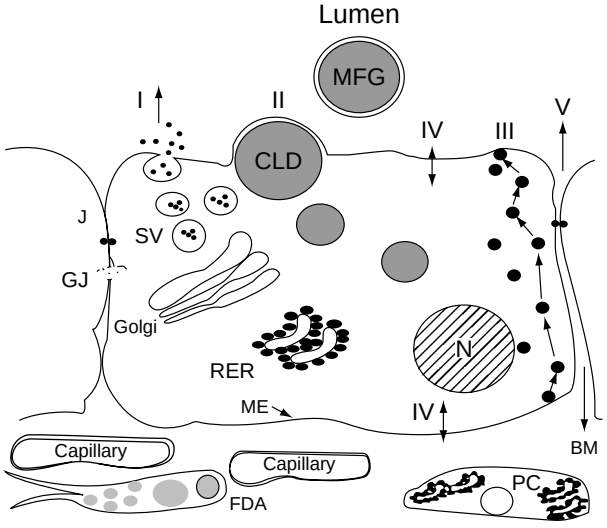


Figure 2 Diagram of a mammary epithelial cell showing pathways for milk secretion described in the text. SV, secretory vesicle; RER, rough endoplasmic reticulum; BM, basement membrane; N, nucleus; PC, plasma cell; FDA, fat-depleted adipocyte; J, junctional complex containing the tight and adherens junctions; GJ, gap junction; ME, myoepithelial cell; CLD, cytoplasmic lipid droplet; MFG, milk fat globule. (Redrawn from Neville MC, Allen JC and Watters C (1983) The mechanisms of milk secretion. In: Neville MC and Neifert MR (eds.) *Lactation: Physiology, Nutrition and Breast-Feeding*, p. 50. New York: Plenum Press.)

The nipple, which is the termination point of the mammary ductal network, is innervated by the fourth intercostal nerve. Afferent sensory stimuli from suckling are transmitted to the spinal cord and the brain, resulting in release of prolactin and oxytocin from the pituitary. Prolactin, secreted from the anterior pituitary, acts directly on alveolar epithelial cells to foster synthesis and secretion of milk components. Oxytocin, secreted from the posterior pituitary, stimulates contraction of the myoepithelial cells that surround the alveoli and ducts. This process, called the 'letdown reflex,' forces the milk from the alveoli through ductules into ducts draining several clusters of alveoli. In the human, the small ducts converge into 15–25 main ducts that drain sectors of the gland and open directly on the nipple. The secretory product is stored in the alveolar space until myoepithelial cell contractions force it through the ducts toward the nipple, where it is available to the suckling infant.

Milk Composition

The major macronutrients in milk are lactose (a disaccharide unique to milk); lipids; proteins, including casein, α-lactalbumin, lactoferrin, secretory immunoglobulin A (sIgA), and many others present at much lower concentrations; and minerals

such as sodium, chloride, calcium, and magnesium. Minor nutrients in milk are enzymes, vitamins, trace elements, and growth factors. The lipid content of milk varies considerably between species. In human and cow's milk, the fat accounts for approximately 4% of milk volume, whereas in whales and seals it can account for as much as 60% of milk volume. Milk fat is primarily composed of triglycerides, a major source of neonatal calories, but it also contains cholesterol and phospholipids, essential for early neonatal development. Casein micelles form a separate phase that can be pelleted by high-speed centrifugation or acidification. These micelles have a high calcium and phosphate content. The aqueous fraction of milk, often called whey, is a true solution that contains all the milk sugar as well as the major milk proteins lactoferrin, α-lactalbumin, and sIgA and nonprotein nitrogen compounds (mostly urea); the monovalent ions sodium, potassium, and chloride; citrate; calcium; free phosphate; and most of the water-soluble minor components of milk.

The casein fraction from cow's milk, usually obtained by rennin precipitation, is used in cheese making, whereas the whey has a multiplicity of uses, most notably as the base for infant formula. Urea and other nonprotein nitrogen components of milk are a source of nitrogen for amino acid and protein synthesis. Isotope utilization studies indicate that on average 10–20% of urea nitrogen is converted into protein by breast-fed infants. Significantly higher utilization rates, however, have been measured in children recovering from infection, suggesting that alterations in urea nitrogen utilization may be a homeostatic response. Human and bovine milk differ primarily in their concentrations of lactose, mono- and divalent ions, and casein levels and the existence of antiinfectious agents in human milk (**Table 1**). These differences are related to the specific needs of these species. Human milk, for example, possesses higher concentrations of lactose and lower divalent ion concentrations than cow's milk. The high lactose concentration provides a large amount of 'free water,' via osmotic regulation, that serves as a reserve for temperature regulation via sweating in human infants. Human milk also contains a number of agents that protect against gastrointestinal and respiratory infections, including oligosaccharides that interact specifically with pathogen receptors, lactoferrin and sIgA. Bovine milk, on the other hand, contains high concentrations of casein, which provides protein and associated calcium and phosphate needed to support rapid growth of young calves.

Table 1 Comparison of the macronutrient contents of human and bovine milk

Component	*Human milk*	*Bovine milk*
Carbohydrates (g/dl)[a]		
Lactose	7.3	4.0
Oligosaccharides	1.2	0.1
Proteins (g/dl)[a]		
Caseins	0.2	2.6
α-Lactalbumin	0.2	0.2
Lactoferrin	0.2	Trace
Secretory IgA	0.2	Trace
β-Lactoglobulin	0	0.5
Nonprotein nitrogen (NPN) (g/l)		
Total NPN	0.42[b]	0.29[c]
Urea	0.16[b]	0.14[c]
Milk lipids (%)[a]		
Triglycerides	4.0	4.0
Phospholipids	0.04	0.04
Minerals and other ionic constituents (mM)[a]		
Sodium	5.0	15
Potassium	15.0	43
Chloride	15.0	24
Calcium	7.5	30
Magnesium	1.4	5
Phosphate	1.8	11
Bicarbonate	6.0	5

[a]Data from Neville MC (1998) Physiology of lactation. *Clinical Perinatology* **26**: 251.
[b]Data from Atkinson SA and Lonnerdal B (1995) In: Jensen RG (ed.) *Handbook of Milk Composition*. San Diego: Academic Press.
[c]Data from Alston-Mills B (1995) In: Jensen RG (ed.) *Handbook of Milk Composition*. San Diego: Academic Press.

Synthesis and Secretion of Milk Components

Solutes enter milk through five general pathways (**Figure 2**). Endogenously generated substances, including the major milk proteins, oligosaccharides, and nutrients such as lactose, citrate, phosphate, and calcium, are secreted through an exocytotic pathway (pathway I). Lipids and lipid-associated proteins are secreted by a process that is unique to mammary epithelial cells (pathway II). The transcytosis pathway (pathway III) transports a wide range of macromolecular substances derived from serum or stromal cells, including serum proteins such as immunoglobulins, albumin, and transferrin; endocrine hormones such as insulin, prolactin, and insulin-like growth factor-1; and stromal-derived agents such as IgA, cytokines, and lipoprotein lipase. In addition, various membrane transport pathways (pathway IV) exist for the transfer of ions and small molecules, such as glucose, amino acids, and water,

across basal and apical plasma membranes. Finally, there is a paracellular pathway (pathway V) that provides a direct route for entry of serum and interstitial substances into milk. This pathway, however, closes during the first few days of lactation in the human. Transport through these pathways is affected by the functional state of the mammary gland and regulated by direct and indirect actions of hormones and growth factors. The general cellular and physiological properties of these pathways are summarized next.

Exocytotic Pathway (I)

Like exocytotic secretion mechanisms found in other cells, proteins, oligosaccharides, and nutrients such as lactose and citrate are packaged into secretory vesicles within the Golgi that are then transported to the apical region of the cell, where they fuse with the apical plasma membrane, discharging their contents into the extracellular space. A unique feature of this pathway in the mammary gland is the presence of high concentrations of lactose, phosphate, citrate, and calcium within the vesicles. Lactose is synthesized in the Golgi from UDP-galactose and glucose, which have entered from the cytoplasm using specific transporters, by the enzyme β-galactosidase, with α-lactalbumin acting as a cofactor. The high concentration of lactose present in the Golgi during lactation osmotically stimulates the influx of water that contributes to the fluidity of milk. Casein micelle formation begins in the terminal Golgi with condensation, and simultaneous phosphorylation, of casein molecules. Addition of calcium, possibly in the secretory vesicle, leads to maturation of casein micelles into particles sufficiently dense to be seen in the electron microscope. This complex thus delivers an efficient package of protein, calcium, and phosphate that provides the nutrients necessary for bone growth, among other things. Calcium enters the cytoplasm from the plasma by a poorly defined transport process. Cytoplasmic calcium is then transported into secretory vesicles by an ATP-dependent Ca^{2+} pump localized on Golgi and secretory membranes. The phosphate in secretory vesicles is derived from the hydrolysis of UDP-galactose during the synthesis of lactose. Citrate is generated endogenously within the cytoplasm of alveolar epithelial cells and transported into the Golgi lumen by an undefined process.

Lipid Secretion Pathway (II)

Estimates of the quantity of milk lipid secretion during lactation in humans and rodents indicate that in many species the lactating mammary gland may be one of the most lipogenic organs in the body. In a fully lactating woman secreting 800 ml/day of milk containing 4% fat, the mammary gland synthesizes approximately 32 g of triglyceride daily or approximately 6 g, 10% of the weight of the woman, in a typical 6-month lactation. The fatty acids for triglyceride synthesis are synthesized from glucose or derived from the plasma lipids by the action of lipoprotein lipase. Once available in the mammary alveolar cells, fatty acids are either bound to a fatty acid binding protein or activated by combination with coenzyme A (CoA) and then bound to an acyl-CoA binding protein. Activated fatty acids are joined with glycerol-3-phosphate by transacylases located in the endoplasmic reticulum to form triglycerides, which enter the cytoplasm as protein-coated structures called cytoplasmic lipid droplets. These structures are translocated to the apical membrane, where they are enveloped by a novel budding process that leads to their release as membrane-bound lipid droplets known as milk fat globules.

The fatty acid composition of milk triglycerides reflects differences in maternal diet. Medium-chain (C_{8-14}) fatty acids are synthesized only in the mammary gland using glucose (or acetate in ruminants) as substrate, whereas long-chain fatty acids are derived from the plasma. Nigerian women who have high-carbohydrate, low-fat diets have significantly more medium-chain fatty acids in their milk than Western women who consume a high-fat diet (**Table 2**).

Transcytosis Pathway (III)

Transport of proteins and other macromolecules by transcytotic pathways involves endocytic uptake of substances at the basal membrane, formation and maturation of endosomes, and sorting to lysosomes for degradation or to the apical recycling compartment for exocytosis at the apical membrane. The best studied molecule in this regard is immunoglobulin A (IgA). IgA is synthesized by plasma cells in the interstitial spaces of the mammary gland or elsewhere in the body and binds to receptors on the basal surface of the mammary alveolar cell; the entire IgA–receptor complex is endocytosed and transferred to the apical membrane, where the extracellular portion of the receptor is cleaved and secreted together with the IgA. It is thought that many other proteins, hormones, and growth factors that find their way into milk from the plasma are secreted by a similar mechanism.

Transmembrane Pathway (IV)

Transport processes for sodium, potassium, and chloride exist on the basal and apical plasma

Table 2 Major fatty acids of human and bovine milk (wt%)

Fatty acid	*Human milk*		*Bovine milk*
	Western diet	*Nigerian diet*	
Saturated fatty acids			
Medium and intermediate chain (formed in mammary gland)			
8:0, octanoic acid	0.46		1.3
10:0, decanoic acid	1.03	0.54	2.7
12:0, lauric acid	4.40	8.34	3.0
14:0, myristic acid	6.27	9.57	10.6
Long chain			
16:0, palmitic acid	22.0	23.35	28.2
18:0, stearic acid	8.06	10.15	12.6
Monounsaturated fatty acids			
16:1 n-7 (*cis*), palmitoleic acid	3.29	0.91	1.6
18:1 n-9 (*cis*), oleic acid	31.3	18.52	21.4
18:1 n-9 (*trans*), oleic acid	2.67	0.86	1.7
Polyunsaturated fatty acids (PUFA) (essential fatty acids)			
18:2 n-6, linoleic acid	10.76	11.06	2.9
18:3 n-3, linolenic acid	0.81	1.41	0.3
Long-chain PUFA (n-6)			
18:3 n-6, γ-linolenic acid	0.16	0.12	2.9
20:2 n-6,	0.34	0.26	0.03
20:3 n-6, dihomo-γ-linolenic acid	0.26	0.49	0.1
20:4 n-6, arachadonic acid	0.36	0.82	0.2
Long-chain PUFA (n-3)			
20:5 n-3, eicosapentenoic acid	0.04	0.48	0.08
22:5 n-3	0.17	0.39	
22:6 n-3, docashexenoic acid	0.22	0.93	0.09

Data from Jensen RG (1995) *Handbook of Milk Composition*. San Diego: Academic Press.

membranes of alveolar epithelial cells. Uptake mechanisms for calcium, phosphate, and iodide, however, are thought to be limited to the basal membrane. The mammary epithelial cells possess a GLUT1 glucose transporter and a sodium-dependent glucose transporter. The GLUT1 transporter is thought to mediate glucose transport at the basal and Golgi membranes, but it does not contribute to glucose transport at the apical membrane. Both sodium-dependent and sodium-independent amino acid transport mechanisms analogous to those found in other organs are located in the basolateral component of the mammary epithelium. It is unclear if apical membranes have similar transport mechanisms for amino acids, and it is unknown how amino acids get into milk.

Paracellular Transport Pathway (V)

Pathway V (**Figure 2**) involves passage of substances between epithelial cells rather than through them, and for this reason it is designated the paracellular pathway. During full lactation the passage of even low-molecular-weight substances between alveolar cells is impeded by the gasket-like tight junction structures that join the epithelial cells tightly, one to another. During pregnancy, with mastitis and after involution the tight junctions become leaky and allow components of the interstitial space, such as sodium and potassium, to pass unimpeded into the milk, a fact that is sometimes useful in diagnosing breast-feeding problems.

Regulation of Milk Synthesis, Secretion, and Ejection

Milk volume production is a primary indicator of lactational function; the most precise methods for measuring the volume of milk produced involve weighing infants before and after each feed for 24 h or longer or using an isotope dilution technique with stable isotopes. Clinically, the amount of milk that can be expressed with a breast pump or the change in infant weight after a single feed can be used as a rough index. The volume of milk secreted by women exclusively breast-feeding a single infant at 6 months postpartum is remarkably constant at approximately 800 ml/day in populations throughout the world. Mothers of twins, and occasionally even triplets, are able to produce volumes of milk sufficient for complete nutrition of their multiple infants,

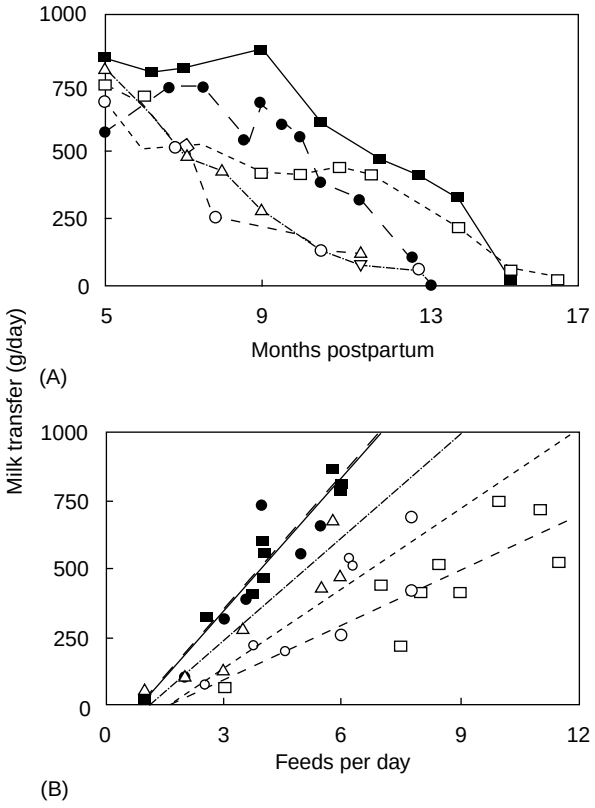


Figure 3 Changes in milk volume during weaning and in response to increased feeding frequency. (A) Milk volume transfer as a function of time postpartum. (B) Relation between feeding frequency (feeds/day) and the milk volume. Data are from five breast-feeding dyads; each symbol represents an individual dyad. (Reproduced with permission from Neville MC *et al.* (1991) *American Journal of Clinical Nutrition* **54**: 81–92.)

and studies of wet nurses indicate that at least some women are capable of producing up to 3.5 l of milk per day. On the other hand, if infants are supplemented with foods other than breast milk, milk secretion is proportionately reduced. This point is illustrated in **Figure 3**, which shows that milk volumes gradually decline during weaning and increase as feeding frequency increases. These observations illustrate the important principle that the volume of milk secretion in lactating women is regulated by infant demand. If milk cannot be removed from the breast, local mechanisms bring about an inhibition of milk secretion and downregulation of milk synthetic machinery. With partial removal of milk on a consistent basis, these local factors adjust milk secretion to a new steady-state level. If milk removal ceases for extended periods, involution sets in and the gland loses its competency to secrete milk.

Hormonal Control of Milk Synthesis and Secretion

In most species, the presence of high levels of plasma prolactin appears to be essential for lactation. In rats, the ergot alkaloid bromocriptine (an inhibitor of prolactin release from the pituitary) inhibits lactation, and in women it inhibits the onset of lactation when given in appropriate doses. How prolactin influences lactation is not known in any detail. However, it appears to promote mammary epithelial cell survival. In addition, it is an osmoregulator in some species of fish, birds, and amphibians and may function to maintain solute transport in the mammary gland.

Local Control of Synthesis and Secretion

Two local mechanisms have been postulated to regulate milk volume production. In one, it is thought that buildup of a specific inhibitory substance occurs in milk as it accumulates in the lumen of the mammary gland. However, the identity of this factor, called feedback inhibitor of lactation, has not been defined. In the second, it is thought that a stretch response of alveoli regulates milk production. Understanding this regulation may be very important in helping women to increase their milk supply, particularly in the postpartum period; therefore, further research is needed.

Regulation of Milk Ejection

When the infant is suckled, afferent impulses from sensory stimulation of nerve terminals in the areolus travel to the central nervous system, where they promote the release of oxytocin from the posterior pituitary. This neuroendocrine reflex can be conditioned, and in the woman oxytocin release is often associated with such stimuli as the sight or sound, or even the thought, of the infant. The oxytocin is carried through the bloodstream to the mammary gland, where it interacts with specific receptors on myoepithelial cells, initiating their contraction and expelling milk from the alveoli into the ducts and subareolar sinuses. The passage of milk through the ducts is facilitated by longitudinally arranged myoepithelial cell processes whose contraction shortens and widens the ducts, allowing free flow of milk to the nipple. Milk is removed from the nipple not so much by suction as by the stripping motion of the tongue against the hard palate. This motion carries milk through the teat into the baby's mouth. The letdown response is decreased by psychological stress or pain, which interfere with oxytocin release. Oxytocin also appears to be involved in regulating maternal behavior in laboratory animals and may play a similar role in humans.

Initiation of Lactation

Pregnancy transforms the mammary gland from a simple ductal tree into a highly efficient exocrine organ with expansive lobuloalveolar structures. This

transformation is hormonally regulated and involves changes in the cellular composition of the mammary gland and alterations in the structural, cellular, and biochemical properties of alveolar cells that are critical to development of efficient solute transport and secretory functions. Alveolar epithelial cells begin to differentiate into secretory cells at midpregnancy in most species. The differentiation process occurs heterogeneously and has been divided into initiation and activation phases based on differences in the composition of mammary secretions, gene expression, and structural and functional properties of alveolar cells. Alveolar cells become capable of limited secretion of some milk components during the initiation phase, which in humans is detected by measurement of increased concentrations of lactose and α-lactalbumin in the plasma. Copious milk secretion, however, is induced during the secretory activation phase (sometimes called lactogenesis II) that occurs in response to the decrease in serum progesterone levels. In rodents and ruminants, this decrease is closely associated with parturition; in humans it occurs after parturition.

Changes in Milk Composition during Secretory Activation

Secretory activation is reflected in dramatic modifications of the solute composition of milk and increased secretory volume, which in turn reflect the maturation of secretory mechanisms and transport pathways during this period. In women, there are three temporally distinct changes in milk composition at the onset of lactation. The earliest is a decrease in sodium and chloride concentrations and an increase in the lactose concentration of milk (**Figure 4**). These modifications occur immediately after delivery and are largely complete by 72 h postpartum. They precede increases in milk volume by at least 24 h and can be explained by closure of the tight junctions that block the paracellular pathway. Blocking this pathway prevents lactose, made by the epithelial cells, from passing from the lumen of the alveolus to the plasma, and it prevents sodium and chloride from directly entering the lumen from the interstitial space. These changes result in decreased concentrations of sodium and chloride and increased concentrations of lactose in the mammary secretion. The increased lactose concentration is reflective of decreased water entering the lumina as monovalent ion secretion decreases rather than an increase in the lactose secretion rate.

Secondarily, there are transient increases in the rates of secretion of sIgA and lactoferrin into milk of women soon after delivery. The concentrations of these two important protective proteins remain high, comprising

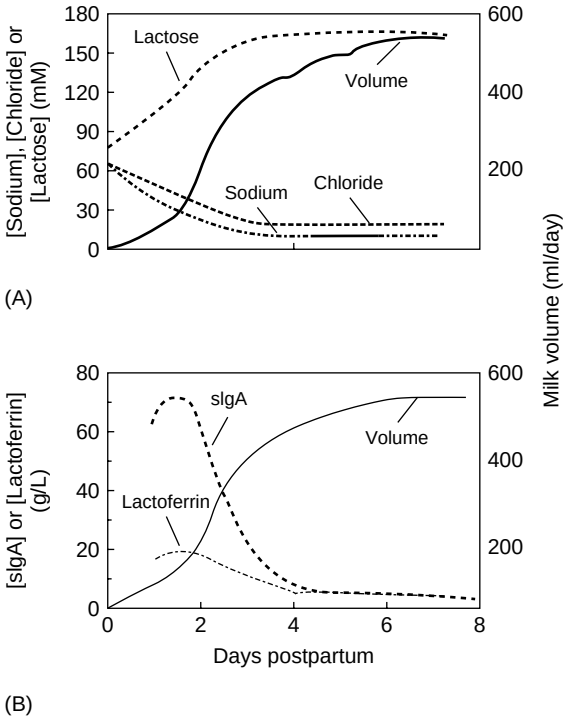


Figure 4 Changes in milk composition and volume in women during secretory activation and early lactation. (Reproduced with permission from McManaman JL and Neville MC (2003) *Advanced Drug Delivery Reviews* **55**: 630–641.)

as much as 10% of milk, for the first 48 h after birth. The concentration of each protein decreases rapidly after day 2, both from dilution as milk volume secretion increases and from actual decreases in their rates of secretion, particularly of immunoglobulins. Although both these proteins are found at high concentrations in colostrum, they are likely to be secreted by different mechanisms; lactoferrin, an endogenous protein of alveolar cells, is secreted by the exocytotic pathway (pathway I), whereas sIgA, a plasma-derived protein, is secreted by receptor-mediated transcytosis (pathway III). In addition, the peak secretion rate of lactoferrin occurs at the same time as that of lactose and the major milk proteins, whereas sIgA secretion peaks 1 day earlier, indicating the possibility that the exocytotic and transcytosis pathways are regulated differently during early lactation.

The third phase occurs approximately 36 h postpartum and is associated with massive and concerted increases in milk volume and the rates of synthesis and/or secretion of almost all the components of mature milk, including, but not limited to, lactose, protein (mainly casein), lipid, calcium, sodium, magnesium, potassium, citrate, glucose, and free phosphate. Considering that the secretion of these substances involves the actions of several distinct transport pathways and biosynthetic

processes, such tightly synchronized increases imply the presence of a common activation switch for coordinating their activities.

Hormonal Control of Secretory Activation

The decrease in progesterone around parturition is generally agreed to be required for the onset of milk secretion. In humans, it is known that removal of the placenta, the source of progesterone, is necessary for the initiation of milk secretion. In swine, timing of the increase in milk lactose correlates closely with timing of the decrease in plasma progesterone at parturition. Exogenous progesterone prevents lactose and lipid synthesis in mammary glands of pregnant rats and sheep after removal of their ovaries, the source of progesterone in these species. Progesterone also suppresses β-casein expression in the rat mammary gland during pregnancy and the decrease in progesterone levels is linked to increased β-casein synthesis at parturition. Receptors for progesterone are not detected in lactating mammary tissues, which explains why progesterone does not inhibit established lactation. It is likely that the decline in progesterone is insufficient to activate secretion and that the actions of other hormones, including prolactin and glucocorticoids, are necessary to complete this process. In all *in vitro* mammary systems, insulin and corticoids, in addition to prolactin, are necessary to maintain synthesis of milk components. Further more, cortisol replacement is required for maintenance of milk production in adrenalectomized animals. An early notion that a surge of glucocorticoids is the initiator of lactation is likely incorrect since the increase in cortisol seen in unanesthetized women associated with the stress of labor is complete by the time milk volume begins to increase to any extent. Because secretory activation proceeds at parturition in severely diabetic rats, a role for insulin in lactogenesis as opposed to metabolic adjustments during lactation seems improbable. In summary, the most reasonable interpretation of the data from both animal and human studies is that the hormonal trigger for lactogenesis is a decrease in progesterone in the presence of maintained prolactin. Since postpartum prolactin levels are similar in both breast-feeding and non-breast-feeding women, the basic process appears to be initiated whether or not breast-feeding occurs. The caveat, of course, is that the mammary epithelium must be sufficiently prepared by the hormones of pregnancy to respond with milk synthesis.

Delays in Secretory Activation

A delay in the onset of milk secretion is a problem for the initiation of breast-feeding in a significant number of parturient women. A number of pathological conditions may delay secretory activation in women, including cesarean section, diabetes, obesity, and stress during parturition. The role of cesarean section is controversial, but if there is one it is likely to have only a modest effect. However, poorly controlled diabetes, stress from delivery, or obesity are associated with significant decreases in early milk production. Because each of these conditions is related to higher blood glucose, hyperglycemia may be an underlying factor in the delay in lactation. However, once it is established, diabetics do not have a problem in maintaining lactation. Thus, compensatory factors may override initiation defects to ensure infant nutrition in these disorders.

See also: **Breast Feeding**. **Fatty Acids**: Omega-3 Polyunsaturated; Omega-6 Polyunsaturated. **Lactation**: Dietary Requirements. **Lipids**: Chemistry and Classification. **Pregnancy**: Energy Requirements and Metabolic Adaptations.

Dietary Requirements

N M F Trugo and C M Donangelo, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Introduction

Milk secretion imposes a considerable nutritional demand on lactating women. The challenge to the maternal organism to sustain milk production and nutrient composition while maintaining an adequate nutritional status is high and must be met by increased dietary intake of energy and nutrients. Otherwise, maternal nutrient depletion may occur due to excessive mobilization of maternal stores. Owing to the major gaps in the knowledge on maternal nutrient requirements and the impact of lactation on maternal nutrient status, and the quantitative and qualitative importance of milk production for the incremental nutrient requirements, the recommended nutrient intakes for lactating women are based mainly on the volume of milk secreted and its nutritional content. The high nutritional demands for milk production result in recommended intakes of most nutrients for lactation that are higher (10%–90%) than in nonreproductive stages.

The dietary recommendations for lactating women considered in this article are those of the FAO/WHO Reports on fats (1994) and micronutrients (2001), and the Dietary Reference Intakes (DRIs) of the Institute of Medicine (US) on micronutrients (1997, 1998, 2000, 2001) and macronutrients (2002). The rationale for recommended nutrient intakes, and the nutrient requirements and dietary recommendations for energy, fat, protein, calcium, zinc, folate, and vitamin A are specifically addressed.

Rationale for Recommended Nutrient Intakes

Recommendations on dietary nutrient intakes for lactating women by different scientific authorities are typically based on the estimated total amount of each nutrient secreted daily into breast milk, taking into account, where known, the efficiency of milk synthesis and the bioavailability of the nutrient in the maternal diet. This estimate for each nutrient is then added to the recommended nutrient intake for non-pregnant, non-lactating women.

The onset of lactation after parturition is brought about by the major hormonal changes that occur in this period. During the first 2–7 days post-partum a thick yellow fluid (colostrum) is secreted. With the progress of lactation, the volume of milk secreted increases and its nutrient composition changes, with an increase, decrease or no change in concentration, depending on the nutrient. After about 21 days the milk secreted is considered mature milk. The volume of breast milk secreted daily increases rapidly in the first post-partum days, being ~500 ml on day 5, ~650 ml at 1 month, and ~750 ml at 3 months, remaining relatively stable during full lactation but decreasing during weaning. In industrialized countries, the average volume of breast milk produced is 750–800 ml day^{-1} in the first 4–5 months post-partum and decreases to 600 ml day^{-1} during 6–12 months after delivery. In this period, the volume of milk produced may be even lower and more variable, depending on the weaning practices adopted.

The FAO/WHO and DRI committees considered 750 and 780 ml, respectively, as the average milk volume produced during full lactation and the basis for recommendations. For most nutrients, average concentration in mature milk multiplied by the average milk volume was used to estimate the total amount of nutrient secreted daily into breast milk. A correction factor was then applied to account for the nutrient bioavailability in the maternal diet and, where known, for the anabolic cost of milk synthesis, and the final value was added to the recommended intake of nonpregnant, nonlactating women. The stage of lactation was considered to be a factor for some nutrients and, where applicable, separate values were given according to the period of time post-partum.

The volume of milk secreted during lactation is not influenced by maternal nutritional status, unless maternal undernutrition is severe. The composition of breast milk for most nutrients is adequate to support infant growth and development in a wide range of maternal nutritional status. However, maternal diet and nutritional status do have an influence on the concentration of some micronutrients such as vitamin A, thiamin, riboflavin, vitamins B_6 and B_{12}, iodine, and selenium. Also, the fatty acid composition of breast milk can be affected by maternal diet.

An important step taken by the DRI committees when setting recommendations was taking maternal age into account, thus giving separate values for adolescent ($\leq$18 years) and adult (19–50 years) lactating women. For some nutrients, adolescent lactating women may have greater requirements than adult women because they are still growing and they need to cover their own nutrient demands. Recommendation of intakes during lactation of calcium, phosphorus, magnesium, iron, and zinc are higher for adolescent than for adult women.

In general, there is considerable uncertainty in establishing dietary nutrient recommendations for lactation due to high intra- and interindividual variability in breast milk volume output and in several specific nutrient concentrations in breast milk, and to temporal changes in milk volume and nutrient concentrations during the lactation period. The composition of breast milk is affected by several factors depending on the nutrient, such as stage of lactation, changes during nursing, diurnal rhythm, maternal diet, gestational age at birth, and parity. Moreover, the total amount of nutrients secreted into breast milk depends on the extent and duration of breast feeding. In addition, physiological adaptation to the increased nutrient lactation demands such as increased nutrient absorption and conservation, and use of maternal nutrient stores, which are quite specific for each nutrient and not easily quantified, contributes to the degree of uncertainty. Maternal age and maternal nutritional status during pregnancy and lactation may influence the homeostatic adaptations during lactation such as the efficiency of nutrient absorption and the degree of mobilization of maternal nutrient stores. These factors are not well known and are difficult to quantify.

Requirements and Dietary Recommendations

Macronutrients

Energy The dietary energy intake recommended for healthy adults of normal weight (body mass index between 18.5 and 25 $kg\,m^{-2}$) is the energy required to maintain energy balance, considering gender, age, weight, height, and level of physical activity. The energy requirements of lactating women include the additional energy that is necessary for milk production. The stage and extent of breastfeeding affect the incremental energy requirements for lactation.

The energy density of human milk is mainly determined by its fat content, which represents 50–60% of the total energy in mature milk and is the most variable energy-yielding component. Protein and lactose contribute to approximately 5% and 38% of energy, respectively. The mean energy density of representative 24-h pooled mature milk samples from well-nourished women ranges from 0.64 to 0.74 $kcal\,g^{-1}$ (2.7–3.1 $kJ\,g^{-1}$).

The estimated energy requirements (EER) for lactating women by the DRI committee are based mainly on studies done in the 1990s, using the doubly labeled water method. The main findings in women who were fully breastfeeding their infants up to 6 months of age were: total energy expenditure of 2109–2580 $kcal\,day^{-1}$ (8860–10840 $kJ\,day^{-1}$) or 35.8–41.0 $kcal\,kg^{-1}\,day^{-1}$ (150–172 $kJ\,kg^{-1}\,day^{-1}$), milk energy output of 483–538 $kcal\,day^{-1}$ (2030–2260 $kJ\,day^{-1}$), and energy mobilization from tissue stores of 72–287 $kcal\,day^{-1}$(300–1200 $kJ\,day^{-1}$). It was concluded that the energy requirements of lactating, well-nourished women are met primarily from the diet and partially by mobilization of tissue stores, without evidence for adaptations in basal metabolism and physical activities. The EER for lactating adult women during the first 6 months of lactation is calculated as the sum of the EER obtained from the equation for adult nonlactating women (using current age, weight, and physical activity level), and the milk energy output (500 $kcal\,day^{-1}$ or 2100 $kJ\,day^{-1}$), subtracting the energy derived from tissue mobilization during lactational weight loss (170 $kcal\,day^{-1}$ or 714 $kJ\,day^{-1}$). The committee considered a milk production rate of 0.78 $l\,day^{-1}$ from birth through 6 months of age, with a milk energy density of 0.67 $kcal\,g^{-1}$ (2.8 $kJ\,day^{-1}$), and an average maternal weight loss of 0.8 $kg\,month^{-1}$. For the second 6 months of lactation, the incremental EER is calculated considering a milk energy output of 400 $kcal\,day^{-1}$ or 1680 $kJ\,day^{-1}$ (milk production rate of 0.6 $l\,day^{-1}$) and no maternal weight loss. The EER for lactating adolescents (14–18 years) is calculated in the same manner as for adult lactating women, but using the appropriate equation to estimate the EER of nonlactating adolescents.

The acceptable macronutrient distribution ranges as percentage of total dietary energy for lactating women are the same as for the general adult population: 10–35% protein, 20–35% fat, and 45–65% carbohydrates. Natural simple sugars, such as those present in fruit, and complex carbohydrates (polysaccharides), such as in cereals (rice, wheat), cereal products (flour, pasta) and starchy roots, should be the preferred sources of carbohydrates in the diet. Added sugars, usually sucrose, should not be higher than 25% of dietary energy. Many of the energy-yielding carbohydrate food sources are also sources of dietary fiber, which are beneficial in reducing the risk of coronary heart disease, ameliorating constipation, and other ways. A total fiber intake of 29 $g\,day^{-1}$ is recommended for lactating women. Whole grain cereals, nuts, legumes, and fruit are good fiber and energy sources, and are also nutrient-rich foods. Restriction of energy intake during lactation to values below 1800 kcal (7500 kJ) per day may lead to low intakes of several micronutrients such as calcium, magnesium, zinc, folate, vitamin B_6, and vitamin A.

Fat Total fat content in human milk is affected by several factors, including stage of lactation, moment of feeding, and parity, but maternal intake of energy, fat, and fatty acids and maternal status have little influence, except when there is a long-term or severe maternal undernutrition. Milk fat content is highly variable, being on average 35–40 $g\,l^{-1}$ in mature milk from well-nourished women delivering at term gestation. The content of individual fatty acids in milk is also highly variable, especially for the long-chain polyunsaturated fatty acids (LCPUFA; mainly C_{20} and C_{22}), and more dependent on maternal diet than total fat. Fatty acid intake and relative contribution of carbohydrate and fat to the total energy intake, as well as maternal body stores and endogenous synthesis, influence the fatty acid composition of human milk. In well-nourished mothers, the polyunsaturated essential fatty acids (EFA) linoleic acid (18:2*n*-6) and α-linolenic acid (18:3*n*-3) represent approximately 11 and 1% (wt/wt), respectively, of the total fatty acids in milk. LCPUFA of the *n*-6 and *n*-3 series account for 1.2 and 0.6%, respectively.

The adequate transfer of polyunsaturated fatty acids from maternal circulation to milk and the

maternal synthesis of LCPUFA, especially arachidonic acid (20:4*n*-6), dihomo-γ-linolenic acid (20:3*n*-6), eicosapentenoic acid (EPA, 20:5*n*-3), and docosahexenoic acid (DHA, 22:6*n*-3), from their respective EFA precursors, are important for infant growth, neurodevelopment, and visual function. These polyunsaturated fatty acids are structural components of all cell membrane phospholipids. Arachidonic acid and DHA are the two quantitatively most important LCPUFA in the brain and retina, and the LCPUFA with 20 carbon atoms are precursors for the synthesis of eicosanoids, a group of signaling molecules. The major part of the polyunsaturated fatty acids in human milk (70–85% in women on omnivorous diet) is derived from maternal body stores, which reflects long-term intake, and not from direct dietary transfer.

The metabolic fate of individual fatty acids depends on dietary energy and on energy balance. Therefore, the intake and requirements for fat, EFA, and LCPUFA are usually expressed as a percentage of the total energy in the diet (en %), rather than total intake (g). The fat intake recommended for lactating women is in the range of 20–35 en %, which is the same range as recommended for the adult population. Concerning the fatty acid intake, FAO/WHO recommends an additional maternal intake of 1–2 en % as EFA (3–4 g day^{-1}) during the first 3 months of lactation, and up to 4 en % (about 5 g day^{-1}) thereafter due to depletion of maternal fat stores. Based on the median linoleic and α-linolenic acid intakes of lactating women in the US, the DRI committee recommends an intake of 5–10 en % (average 13 g day^{-1}) of *n*-6 (as linoleic acid) and of 0.6–1.2 en % (average 1.3 g day^{-1}) of *n*-3 (as α-linolenic acid) polyunsaturated fatty acids throughout lactation, with a 10% contribution of LCPUFA of the *n*-6 and *n*-3 series to these ranges. The ratio of *n*-6:*n*-3 unsaturated fatty acids in the diet is important because these fatty acids are desaturated and elongated using the same series of enzymes. Increased intakes of linoleic acid result in decreased conversion of α-linolenic acid to EPA and DHA, whereas the conversion of linoleic acid to arachidonic acid is inhibited by EPA and DHA, and also by arachidonic acid, α-linolenic acid, and linoleic acid itself. The *n*-6:*n*-3 ratio recommended for adults by both DRI and FAO/WHO committees is 5:1 to 10:1. Vegetable oils are the main sources of *n*-6 fatty acids in the diet and also of *n*-3 fatty acids, although in lower amounts. Fish such as herring, mackerel, and salmon are good sources of *n*-3 fatty acids.

The intake of *trans* fatty acids (*trans* isomers of oleic and linoleic acid) present in hydrogenated food fats and oils, deep-fried foods, and meats are of special concern in lactating women when their intake is excessively high or when EFA intake is low during pregnancy and lactation. An inverse correlation of arachidonic acid and DHA with *trans* fatty acids in plasma lipids has been reported in infants, suggesting impairment in LCPUFA synthesis and metabolism.

Protein The average protein content in colostrum is 15–20 g l^{-1} decreasing to approximately 8–10 g l^{-1} in mature human milk during the first 6 months of lactation. The protein concentration in human milk is not affected by diet, body composition, or maternal undernutrition.

The recommended dietary allowance (RDA) of protein for adolescent and adult lactating women by the DRI committee is 1.1 g per kg of body weight per day. This corresponds to an increment of 25 g day^{-1} of protein intake above the RDA for nonlactating women, and it is the same as for pregnant women. Recent data have shown that protein intakes of 1 g kg^{-1} day^{-1} are able to maintain good milk production, and promote conservation of maternal skeletal muscle apparently by downregulating protein metabolism. The recommended range of percentage of energy from dietary protein is the same as for the general adult population (10–35%).

The factorial approach was used to estimate the protein RDA for lactation, assuming that the maintenance protein requirement of the lactating women is not different from that of the nonlactating women, and that the additional protein and/or amino acid requirements are proportional to milk production. The additional protein requirement for lactation is defined as the output of total protein and nonprotein nitrogen (converted in protein by multiplying by 6.25) in milk. Nonprotein nitrogen represents 20–25% of total milk nitrogen, mainly as urea. It is taken into account because it is assumed that the nitrogen needed to cover the total nitrogen loss in milk should be derived from dietary protein. The total nitrogen output in milk is converted to total protein output (approximately 10 g day^{-1}) and divided by the incremental efficiency of nitrogen utilization (0.47), which is assumed to be the same in adult and adolescent lactating women. The additional estimated average requirement due to milk production is 21.2 g day^{-1}. After correction by the coefficient of variation and rounding off, the RDA for lactation amounts to +25 g day^{-1}, which corresponds to +0.46 g protein kg^{-1} day^{-1} (based on a reference woman of 57 kg), above the RDA for nonlactating women.

Recommendations for individual indispensable amino acids for lactation by the DRI committee assume that the incremental needs correspond to the

amino acids secreted in milk, since there are no specific data on the amino acid requirements in lactating women. Therefore, the RDA of amino acids for lactation are calculated by adding the average amounts of amino acids in human milk in the first 6 months of lactation (expressed as milligrams per kilogram per day) to the respective RDA for the nonlactating women. Recommendations of indispensable amino acids for the lactating women are 36% (histidine) to 80% (tryptophan) higher than those for nonlactating women. The intake of good-quality protein such as in eggs, milk, meat, and fish provide the requirements for all indispensable amino acids. Individuals who restrict their diets to plant proteins (cereals, legumes, nuts, starchy roots, vegetables, and fruits) may be at risk of not getting adequate amounts of certain indispensable amino acids. However, adequate complementary mixtures of plant proteins, with increased digestibility through processing and preparation, can provide high-quality protein.

Micronutrients

Daily requirements for several micronutrients (riboflavin, vitamin B_{12}, vitamin C, vitamin A, vitamin E, copper, iodine, manganese, selenium, and zinc) are higher during lactation than during pregnancy, indicating that lactation is a very demanding process. The only micronutrient needed in lower amounts during lactation is iron, due to the small amount of iron secreted into breast milk and to the usual amenorrhea of nursing women. However, iron requirements may be high post-partum for women who need to replace major blood losses during delivery.

The recommended intakes for micronutrients during lactation established by FAO/WHO and DRI committees are summarized in **Table 1**. The

Table 1 Daily recommended micronutrient intakes for adult lactating women

Nutrient	*FAO/WHO*[a]		*IOM*[b]	
	Recommended value	*Per cent change*[c]	*Recommended value*	*Per cent change*[c]
Vitamin A (μg RAE day^{-1})	–	–	1300	↑ 86%
Vitamin A (μg RE day^{-1})	850	↑ 70%	–	–
Vitamin D (μg day^{-1})	5	No change	5	No change
Vitamin E (mg α-TE day^{-1})	7.5	No change	19	↑ 27%
Vitamin K (μg day^{-1})	55	No change	90	No change
Thiamin (mg day^{-1})	1.5	↑ 36%	1.4	↑ 27%
Riboflavin (mg day^{-1})	1.6	↑ 45%	1.6	↑ 45%
Niacin (mg NE day^{-1})	17	↑ 21%	17	↑ 21%
Vitamin B_6 (mg day^{-1})	2.0	↑ 54%	2.0	↑ 54%
Pantothenate (mg day^{-1})	7.0	↑ 40%	7.0	↑ 40%
Biotin (μg day^{-1})	35	↑ 17%	35	↑ 17%
Folate (μg DFE day^{-1})	500	↑ 25%	500	↑ 25%
Vitamin B_{12} (μg day^{-1})	2.8	↑ 17%	2.8	↑ 17%
Vitamin C (mg day^{-1})	70	↑ 55%	120	↑ 60%
Calcium (mg day^{-1})	1000	No change	1000	No change
Iodine (μg day^{-1})	200	↑ 82%	290	↑ 93%
Iron (mg day^{-1})	15[d]	↓ 49%	9	↓ 50%
Zinc (mg day^{-1})	9.5[e]	↑ 94%	12	↑ 50%
	8.8[f]	↑ 80%		
Magnesium (mg day^{-1})	270	↑ 23%	310	No change
Selenium (μg day^{-1})	35	↑ 35%	70	↑ 27%
Chromium (μg day^{-1})	–	–	45	↑ 80%
Copper (μg day^{-1})	–	–	1300	↑ 44%
Fluoride (mg day^{-1})	–	–	3	No change
Manganese (mg day^{-1})	–	–	2.6	↑ 44%
Molybdenum (μg day^{-1})	–	–	50	↑ 11%
Phosphorus (mg day^{-1})	–	–	700	No change

[a]FAO/WHO (2001) *Human Vitamin and Mineral Requirements*. Report of a Joint FAO/WHO Expert Consultation. Rome: Food and Agriculture Organization.
[b]Institute of Medicine (IOM) (2001) *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc*. Washington, DC: National Academy Press.
[c]Changes from recommendations for nonpregnant nonlactating women: ↑, per cent increase; ↓, per cent decrease.
[d]Considering 10% bioavailability.
[e]0–3 months post-partum, considering moderate bioavailability.
[f]4–6 months post-partum, considering moderate bioavailability.
RAE, retinol activity equivalent; α-TE, alpha-tocopherol equivalent; NE, niacin equivalent; DFE, dietary folate equivalent.

percentages of change from the recommendations for nonpregnant nonlactating women are also shown. In order to meet these intakes, lactating women should be guided to consume daily a large variety of foods rich in micronutrients, since food diversification contributes to improve the intake of limiting nutrients. Micronutrients most commonly at risk of inadequate intakes by lactating women are calcium, zinc, folate, and vitamin A.

Calcium It is estimated that lactating women secrete an average of 200 mg of calcium per day into mature breast milk although this amount is variable among women, usually ranging from 150 to 300 mg day^{-1}. The maternal diet does not affect the milk calcium concentration except when maternal calcium intake is very low (<300 mg day^{-1}). The primary source of calcium for milk production appears to be the increased mobilization of calcium from maternal bone due to the increased bone resorption that occurs during lactation favored by the low estrogen concentration. This results in a net loss of maternal bone mass during lactation that is regained after weaning upon return of ovarian function. The decreased urinary calcium excretion during lactation also contributes to the calcium economy for milk secretion. The efficiency of intestinal calcium absorption is not increased during lactation and, therefore, does not contribute to the extra calcium needed for milk production.

Several studies have shown that the adaptive changes in calcium homeostasis during lactation are independent of maternal calcium intake. It was demonstrated that the loss of bone mass during lactation was not affected by calcium supplementation (1000 mg day^{-1}) of nursing women with habitual dietary calcium intakes of 300 mg day^{-1}, 800 mg day^{-1}, and 1200 mg day^{-1}. Since the loss of maternal bone calcium that occurs during lactation is not prevented by increased dietary calcium, and the calcium lost appears to be regained after weaning, the recommended intake of calcium of lactating women is the same as for nonpregnant nonlactating women of the same age, being 1000 mg day^{-1} and 1300 mg day^{-1} for adult and adolescent women, respectively. Even if not increased during lactation, the recommended calcium intake may be difficult to obtain by women with low habitual intake of dairy products. Therefore, lactating women should be guided to consume dairy products such as milk, yogurt, and cheese, and other calcium-rich foods such as fish with edible bones, broccoli, and kale.

Lactating adolescents are a group of special concern regarding calcium intake due to the already high calcium requirements of nonpregnant nonlactating adolescents. These young women are still increasing their own bone density besides the increased calcium requirement to support lactation. Studies are needed to investigate if these women are able to regain bone after weaning to the same level as when they were nonpregnant nonlactating and if they would benefit from increased calcium intake.

Zinc Zinc concentrations in human milk are highest in colostrum, decrease rapidly during the first 3 months post-partum, and more gradually at later stages of lactation. Typical milk zinc concentrations are 4 mg l^{-1} at 2 weeks, 3 mg l^{-1} at 4 weeks, 2 mg l^{-1} at 8 weeks and 1.2 mg l^{-1} at 24 weeks. These concentrations are not influenced by either maternal dietary intake or zinc supplementation at least in well-nourished women. Less is known about the effect of low maternal zinc intakes on milk zinc concentrations, but the available data indicate that concentrations in developing countries may be lower than those in developed countries at comparable times post-partum.

Average losses of zinc via the mammary gland range from 2.2 mg day^{-1} during the first month post-partum to 1 mg day^{-1} at 6 months. The average estimate of daily output of zinc in milk during the first 3 months of lactation is 1.6 mg day^{-1}, which would theoretically double the minimum endogenous zinc losses in lactating women compared to those of nonlactating nonpregnant women. However, maternal homeostatic mechanisms such as enhanced zinc absorption and reduced urinary zinc excretion contribute to compensate for the secretion of zinc into human milk, independent of maternal zinc intake. Intestinal conservation of endogenous fecal zinc appears to contribute to zinc homeostasis during lactation at low zinc intakes (<8 mg day^{-1}). Involution of the uterus, decreased maternal blood volume and increased resorption of trabecular bone in the postpartum period also contribute to mobilizable zinc pools to compensate for the increased needs. These sources appear to provide up to 0.5 mg day^{-1} of zinc during the first 3 months of lactation. Taking all these adaptation mechanisms into account, the average estimate of increased requirement for absorbed zinc during the first 6 months of lactation is 1.35 mg day^{-1}. Therefore, dietary zinc requirements during lactation are substantially increased compared to nonpregnant nonlactating women, both in adults and adolescents.

Bioavailability is an important factor in setting dietary zinc recommendations since the efficiency of dietary zinc utilization may vary up to fivefold depending on the overall composition of the diet, particularly the balance between promoters (animal

protein) and antagonists (phytic acid and possibly calcium, iron, and copper) of zinc absorption.

Dietary zinc recommendations during lactation are set at $12\,mg\,day^{-1}$ for adult lactating women consuming a mixed diet, but recommended intake may be as high as $19\,mg\,day^{-1}$ for nursing women with habitual diets of low zinc bioavailability, such as those based mainly on unrefined cereals and legume seeds, with high phytate:zinc ratio (>15), and low in animal protein. This high-zinc intake may be difficult to obtain using plant-based diets. Therefore, nursing women in developing countries and strict vegetarian women worldwide may be at risk of inadequate zinc status during lactation. Red meat, poultry, eggs, and seafood provide highly available zinc, and their consumption should be encouraged in lactating women.

Folate Concentration of folate in breast milk increases during the lactation period, with lower values for colostrum ($10–40\,\mu g\,l^{-1}$) than for mature milk ($79–133\,\mu g\,l^{-1}$). These concentrations are several-fold higher than in maternal plasma, independent of maternal folate status, suggesting that the mammary gland actively transports and regulates the secretion of this vitamin into milk. Folate concentration in breast milk is maintained with the concomitant depletion of maternal folate when maternal dietary intake is low. Maternal supplementation during lactation has little effect on milk folate but it benefits maternal folate status. Folate deficiency has been implicated in disorders such as neural tube defects, low infant birth weight, aborption, cervical dysplasia, atherosclerosis, and colon cancer.

Dietary folate requirements during lactation are based on the average milk folate concentration of $85\,\mu g\,l^{-1}$ and assume a 50% dietary absorption factor from a mixed diet, to account for dietary bioavailability. The average extra amount of dietary folate needed to cover the lactation needs is thus estimated as $133\,\mu g\,day^{-1}$, an increase of about 40% of the nonpregnant nonlactating average folate requirements. Dietary folate recommendations during lactation are set at $500\,\mu g\,day^{-1}$, as dietary folate equivalents (DFEs). A DFE is defined as 1 μg of food folate, or 0.6 μg of folic acid from fortified food or as a supplement taken with meals, or 0.5 μg of folic acid as a supplement taken on an empty stomach. Thus, in order to meet lactation requirements, much less of this vitamin is needed when given as pure folic acid than as natural food folate.

Present recommendations are very difficult to meet by dietary means and most nursing women worldwide appear to have a much lower dietary folate intake and, therefore, to be at risk for folate deficiency. Although folate is found in a variety of foods, such as fresh green vegetables, oranges, legumes and nuts, it is present in relatively small amounts, and several servings per day of these foods are needed to meet recommended intake. Moreover, considerable losses of folate occur during food harvesting, storage, and cooking. Fortification of cereal grains with folate has become mandatory or encouraged in many countries in order to reduce the risk of folate deficiency.

Vitamin A Vitamin A is present in human milk, primarily as retinyl esters (95%) and free retinol. Vitamin A activity is also provided as carotenoid precursors, mainly as beta-carotene, which accounts for up to 30% of total carotenoids in breast milk. Concentration of vitamin A in human milk is high in early lactation ($600–2000\,\mu g\,l^{-1}$) and declines thereafter ($200–1100\,\mu g\,l^{-1}$), being responsive to maternal intake, particularly in nursing women with poor vitamin A status. These women are at risk of providing insufficient amounts of vitamin A to their infants, as is often the case in developing countries.

Dietary recommendations of vitamin A during lactation are based on replacement of the amount of vitamin A secreted into breast milk during the first 6 months of lactation, while preserving maternal vitamin A stores. Because the bioconversion of carotenoids in human milk and in infants is still unknown, the contribution of maternal carotenoids in breast milk to meeting the vitamin A lactation requirement cannot yet be established.

Based on the average vitamin A milk concentration of $485\,\mu g\,l^{-1}$, an extra intake of 400 μg of retinol activity equivalents (RAE) per day is recommended for lactating women, which represents an increase of over 70% of recommended intakes for nonpregnant nonlactating adolescent and adult women. RAE is defined as 1 μg all-*trans*-retinol, 12 μg beta-carotene, and 24 μg alpha-carotene or beta-cryptoxanthin. The amounts of carotenoids equivalent to 1 RAE are double the equivalent to 1 RE (retinol equivalents). The new equivalency value (RAE) is based on recent studies demonstrating that bioconversion of carotenoids to vitamin A is 50% less than previously thought.

The vitamin A intake recommended for lactating women can be obtained as the preformed vitamin from foods of animal origin (primarily milk products, eggs, and liver) and as the carotenoid precursors by regular consumption of green leafy vegetables and ripe, colored fruits. However, meeting the recommended intake by consumption of plant foods alone, as is the case in many developing countries, may be difficult.

See also: **Adolescents**: Nutritional Requirements. **Breast Feeding**. **Calcium**. **Dietary Guidelines, International Perspectives**. **Energy**: Requirements. **Fatty Acids**: Metabolism. **Folic Acid**. **Lactation**: Physiology. **Nutrient Requirements, International Perspectives**. **Protein**: Requirements and Role in Diet. **Vitamin A**: Biochemistry and Physiological Role. **Zinc**: Physiology.

Further Reading

Allen LH (2001) Pregnancy and lactation. In: Bowman A and Russell RM (eds.) *Present Knowledge in Nutrition*, 8th edn., pp. 403–415. Washington, DC: International Life Science Institute Press.

Donangelo CM and Trugo NMF (2003) Lactation/human milk: composition and nutritional value. In: Caballero B, Trugo LC, and Finglas PM (eds.) *Encyclopedia of Food Sciences and Nutrition*, 2nd edn., vol. 6, pp. 3449–3458. London: Academic Press.

FAO/WHO (1994) *Fats and Oils in Human Nutrition*. Report of a Joint FAO/WHO Expert Consultation. Rome: Food and Agriculture Organization.

FAO/WHO (2001) *Human Vitamin and Mineral Requirements*. Report of a Joint FAO/WHO Expert Consultation. Rome: Food and Agriculture Organization.

Haskell MJ and Brown KH (1999) Maternal vitamin A nutriture and the vitamin A content of human milk. *Journal of Mammary Gland Biology and Neopllasia* **4**: 243–257.

Institute of Medicine (IOM) (1997) *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride*. Washington, DC: National Academy Press.

Institute of Medicine (IOM) (1998) *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Panthothenic Acid, Biotin, and Choline*. Washington, DC: National Academy Press.

Institute of Medicine (IOM) (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. Washington, DC: National Academy Press.

Institute of Medicine (IOM) (2001) *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc*. Washington, DC: National Academy Press.

Institute of Medicine (IOM) (2002) *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein and Amino Acids*. Washington, DC: National Academy Press.

Krebs NF (1998) Zinc supplementation during lactation. *American Journal of Clinical Nutrition* **68**(suppl): 509S–512S.

O'Connor DL (1994) Folate status during pregnancy and lactation. In: Allen L, King J, and Lonnerdal B (eds.) *Nutrient Regulation during Pregnancy, Lactation, and Infant Growth*. pp. 157–172. New York: Plenum Press.

Prentice A (2000) Calcium in pregnancy and lactation. *Annual Review of Nutrition* **20**: 249–272.

Prentice AM, Goldberg GR, and Prentice A (1994) Body mass index and lactation performance. *European Journal of Clinical Nutrition* **48**(suppl 3): S78–S86; discussion S86–S89.

LACTOSE INTOLERANCE

D M Paige, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Lactose maldigestion and intolerance result from an inability to digest varying amounts of the milk sugar lactose. This is a result of an inadequate amount of the genetically regulated milk sugar enzyme lactase. The most common reason for lactose maldigestion is a decline of lactase activity with increasing age. Lactose maldigestion may also occur secondary to intestinal tract infection and diarrhea. A rare form of alactasia, an absence of the milk sugar enzyme, can occur at birth. The symptoms associated with lactose maldigestion are a result of the incomplete hydrolysis, or splitting, of the disaccharide lactose into its absorbable monosaccharide components, glucose and galactose. Lactose maldigestion may result in abdominal bloating and/or pain, flatulence, loose stools, and diarrhea, singly or in combination. The symptoms associated with lactose maldigestion result in lactose intolerance. The most common form of lactose maldigestion and lactose intolerance, as observed in the majority of the world's adult population, is due to genetically determined low lactase levels. Lactase deficiency due to genetic nonpersistence is reported in approximately 70% of the world's adult population.

The prevalence is lowest in people of Northern European descent (15%) and highest in many Asian populations (near 100%). The prevalence of lactase deficiency in individuals of African descent is approximately 70–80%. Similar levels are reported for Latinos and those of Eastern European and South American ancestry. Not all individuals with a reduced level of the enzyme lactase exhibit symptoms with the ingestion of dietary lactose. The presence or absence of symptoms varies with the amount and type of food consumed, intestinal transit time, and level of residual intestinal lactase. Individuals with low lactase levels may tolerate a

moderate intake of lactose. Lactase deficiency can generally be identified by a breath hydrogen test measuring the level of undigested lactose reaching the colon. Bacterial fermentation of the undigested lactose is responsible for the volume of breath hydrogen production. A lactose tolerance test measuring blood sugar rise has also been used. Individuals experiencing discomfort with lactose ingestion can elect to consume commercially hydrolyzed milk that is readily available, milk substitutes, or alternative food sources equally rich in calcium.

Historical and Geographic Perspective

The first herd animal kept by humans, sheep, seems to have been domesticated in approximately 10 000 BC. Herd animals were primarily used for meat and perhaps certain other purposes. The historical record suggests that herd animals during this period were not milked. Evidence that humans milked domesticated animals dates to approximately 4000 to 3000 BC in northern Africa and southwest Asia. Following that time, dairying spread across Eurasia and into sub-Saharan Africa. However, dairying was not adopted by all groups in Asia and Africa who had suitable herd animals. Even as late as 1500 AD, the beginning of the great European overseas expansion, there were sizable areas occupied by nonmilking groups. In Africa, the zone of nonmilking centered on the Congo Basin but extended beyond to cover approximately one-third of the continent. In Asia, the zone of nonmilking covered the bulk of the eastern and southeastern portions of the continent, including Thailand, Vietnam, China, and Korea, as well as the islands to the east. Moreover, dairying remained unknown in the Pacific region and in the Americas in pre-European times. In those days, the nonmilking peoples of Asia, Africa, and the Americas consumed mother's milk as infants but normally ingested no milk after weaning. Animal milk was not part of their diet.

It was striking that adults of all groups whose origins lay in the traditional zone of nonmilking were predominantly maldigesters, usually 70–100% of the individuals tested. Also striking was the fact that the peoples with low prevalences of lactose maldigestion (northwest Europeans and certain East African pastoral groups) came from a long tradition of consuming milk, much of it in lactose-rich forms. This suggested the geographic or culture–historical hypothesis. By that hypothesis, in the hunting and gathering stage, human groups everywhere were like most other land mammals in their patterns of lactase activity. That is, in the normal individual lactase activity would drop at weaning to low levels, which prevailed throughout life. With the beginning of dairying, however, significant changes occurred in the diets of many human groups. In some of these, moreover, there may have been a selective advantage for those aberrant individuals who experienced high levels of intestinal lactase throughout life. That advantage would have occurred only in certain situations: Where milk was a specially critical part of the diet, where the group was under dietary stress, and where people did not process all their milk into low-lactose products such as aged cheese. Under these conditions, most likely to occur among pastoral groups, such aberrant individuals would drink more milk, would benefit more nutritionally as a result, and would enjoy increased prospects of survival, well-being, and bearing progeny and supporting them. In a classical Mendelian way, the condition of high intestinal lactase activity throughout life, or lactase persistence, would come to be typical of such a group.

Lactase Nonpersistence

In its pure form, lactose cannot be transported across the mucosa of the small intestine. To be absorbed, it must be hydrolyzed by lactase to free glucose and galactose. These two simple sugars are rapidly and completely absorbed in the normal small intestine. The rate of lactase synthesis is high from birth until ages 3–5 years. Between ages 5 and 14 years, many people undergo a genetically programmed reduction in lactase synthesis that results in a lactase activity level only 5–10% of that of infancy. This reduction, known as lactase nonpersistence or primary lactase deficiency, is not related to the continued intake of milk or lactose. As noted, less than one-third of the world's adult population is genetically predisposed to maintaining a high degree of lactase activity or lactase persistence throughout adulthood.

Lactase persistence in the human population is inherited as a dominant genetic trait. It has been observed that lactose intolerance is 'ancient and globally distributed,' predating the appearance of a persistent lactase variant that was naturally selected in dairying regions. Hollox *et al.* report, "the continued adult production of lactase results from the persistent expression of the protein lactose–phlorizin hydrolase which is encoded by the lactase gene (LCT) on chromosome 2." Swallow notes, "the distribution of different lactase phenotypes in human populations is highly variable and is controlled by a polymorphic element *cis*- acting to the lactase gene. A putative causal nucleotide change

has been identified and occurs on the background of a very extended haplotype that is frequent in Northern Europeans, where lactase persistence is frequent."

Lactose Digestion and Gastrointestinal Function

Lactose is hydrolyzed at the intestinal jejunal brush border by the enzyme lactase into its absorbable monosaccharides glucose and galactose. Lactase activity is robust during infancy and, as is the case in humans and most mammals, declines after weaning. Accordingly, the general pattern of lactase nonpersistence is a continuous decline in genetically programmed populations. A shifting pattern of lactose digestion and gastrointestinal function is a result of lactase nonpersistence. The pattern can be described and monitored during three distinct clinical phases.

First, there is a decreasing ability to digest the large lactose load consumed during the screening test. It is important to recognize that this is not an all-or-none phenomenon but rather a slowly progressive decline in available lactase activity, and that this decline, as noted previously, can be influenced by transit time, the vehicle in which the lactose is consumed, and/or the intake of additional foods along with lactose.

Next, with the continued decline of lactase activity, a point is reached when available lactase activity is no longer sufficient to hydrolyze more modest levels of lactose. Therefore, the consumption of a glass of milk or another product containing the equivalent level of lactose will result in incomplete hydrolysis of the lactose consumed. The individuals so tested frequently do not recognize signs or symptoms associated with the incomplete digestion of lactose.

Finally, with the continued decline of lactase activity with increasing age, individuals become symptomatic as a result of the undigested lactose. The decline in available lactase activity reaches a recognizable clinical threshold with increasing age (**Figure 1**).

Initially, many reports treated the population studied as a single unit and paid incomplete attention to age-specific considerations. Distinctions between secondary lactose malabsorption due to short-term intestinal injury and primary lactose malabsorption that has a genetic basis were not always made. This introduced additional confounding variables. Differences in an individual's capacity to hydrolyze and tolerate a lactose challenge dose compared to his or her ability to utilize lesser amounts of lactose found in usually consumed amounts of milk created additional areas of confusion.

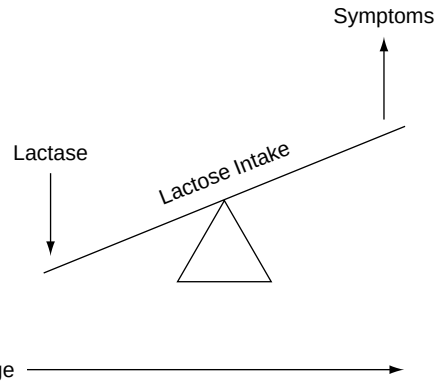


Figure 1 Symptoms associated with lactose maldigestion result from the decline in lactase levels with age and increase with the amount of lactose consumed.

When attention is paid to the many factors associated with lactose digestion from infancy to old age, it is possible to place many of the seeming contradictions into perspective. What may have appeared to be incongruities in reported data appear to merge into a relatively predictable pattern of lactose digestion.

Lactose maldigestion and intolerance are influenced by age, infection, size of the lactose bolus, gastric emptying time, intestinal transit time, individual sensitivities, eating habits, genetics, environment, food ideologies, and cultural patterns. Furthermore, symptoms of lactose malabsorption may also be the result of bacterial fermentation of undigested carbohydrate in the colon. The type and extent of the colonic bacterial profile and the absorption of hydrogen and the volatile fatty acids will influence individual reports of symptoms associated with lactose intolerance. Clearly, lactose malabsorption is not a homogeneous event. Neither is it an all-or-none phenomenon having its origins in a single etiology. Clinical expressions of lactose malabsorption, lactose intolerance, and milk rejection find their origins in one or more of the causes outlined previously (**Table 1**).

Prevalence

Children

A review of reported data on diverse populations supports the conclusion that in later childhood and adolescence an important transition in lactose digestion occurs. Older children and young adults are increasingly unable to digest even modest amounts of lactose. This results in increased symptom

Table 1 Patterns of lactose digestion by lactase status

Lactase status	*Test results*	*Symptoms*	*Lactose intolerance*	*Milk consumption*
Adequate	Normal (−)	0	0	Average (+)
Marginal lactase Deficiency	+	0/+	0/+	+ −
Moderate lactase Deficiency	+	0/+	0/+	+ −
Severe lactase Deficiency	+++	++	++	−

Adapted from Paige DM, Davis L, Bayless TM *et al.* (1979) The effects of age and lactose tolerance on blood glucose rise with whole cow and lactose hydrolyzed milk. *Journal of Agriculture and Food Chemistry* **27**(4): 667.

production, recognition of discomfort, and avoidance of lactose-containing products that provoke symptoms.

A progressive decrease in lactase is noted from approximately 1 to 5 years of age through adolescence. Reported rates in US African American children ranged from 27% lactose maldigestion following lactose testing using a lactose load equivalent to two 8-oz. glasses of milk at 1 or 2 years to 74% in 11- or 12-year-old children. The progressive decrease in the ability to hydrolyze a lactose challenge was observed in children of both high and low socioeconomic status. Studies of white children 1–12 years of age identified only 17% of children maldigesting a lactose challenge. Signs and symptom production associated with a reduction in lactose digestion in a child population is difficult to measure due to the nature of the symptoms being reported, the signs observed, and the subjective nature of the reports. This is reinforced by a report on 21 African American girls 11–15 years of age indicating 82% had evidence of lactose maldigestion with reports of gastrointestinal symptoms being negligible, and breath hydrogen excretion, while remaining high, varied between two time periods. Consistent with the previous data, milk consumption studies, both observed and reported, suggest a progressive decline in milk intake with increasing age in the African American population of children and parallel reports of children from other populations with a high prevalence of lactose maldigestion.

Adults

The progressive increase in prevalence of lactose maldigestion increases with age, reaching reported adult levels of approximately 70% of the world's adult population. The exceptions are populations of Northern and Central Europeans and some Middle Eastern populations as well as groups of primarily European descent in Australia, New Zealand, and North America. Thus, minority populations in North America and Europe, as well as adult populations in most developing countries, are lactose maldigesters (**Table 2**).

Reported milk drinking patterns of individuals classified as maldigesters vary considerably in adults. Data range from 50% reporting symptoms with one 8-oz. glass of milk to 75% reporting symptoms with two 8-oz. glasses of milk and 30% reporting not drinking any milk. Nevertheless, caution

Table 2 Prevalence of lactose maldigestion in selected populations

Population	*Country*	*% Lactose maldigestion*	*Population*	*Country*	*% Lactose maldigestion*
African American, 18–54 years	US	75	General, 21–65 years	Finland	15
Asian, 23–39 years	US	100	General, 20.3 years	Germany	70
Native American, 18–54 years	US	81	General, 16–54 years	Chile	80
African American, 13–19 years	US	69	Non-Caucasian	Peru	94
Mexican, 18–94 years	US	53	General, 38–49 years	Brazil	80
Vietnamese, 22–63 years	US	100	Arab adult	Israel	81
Sicilian, 25 years average	Italy	71	General, male 14–34 years	Egypt	73
Northern, 28.7 years average	Italy	52	General, 15–78 years	Greece	45
Central, 36 years average	Italy	18	Bantu, 13–43 years	Uganda	100
Romai	Hungary	56	Yoruba, 13–70 years	Nigeria	83
Austrian, 22 years average	Austria	20	General, adult	India	61
General, 20.3 years average	Finland	17	General, 17–83 years	Korea	75
Aboriginal	Australia	84	General, 15–64 years	Japan	100

must be exercised in interpreting reported symptoms and making the diagnosis of lactose intolerance. There can be considerable crossover between individuals who self-identify as intolerant to lactose and are not diagnosed as lactose intolerant versus those in whom the diagnosis was carefully established. More attention to identifying and categorizing symptoms better may help. A Finnish study noted flatulence as the most severe symptom in maldigesters, whereas abdominal bloating was most frequently reported by individuals self-identifying as lactose digesters. Moreover, microbiota may play a role in the presence and intensity of lactose-related symptoms. Data suggest that increased levels of colonic bacteria, as well as their diversity, may play a role as a result of increased fermentative capacity in reducing the symptoms associated with lactose intolerance.

Pregnant Women

The role of lactose digestion in pregnant women is of special interest. Despite the nutritional value of milk during pregnancy, the lactase levels in some individuals in a number of racial and ethnic groups may be insufficient to hydrolyze commonly consumed amounts of lactose, resulting in lactose maldigestion and possibly milk intolerance. The Institute of Medicine report notes that "lactose intolerance among pregnant African American women may result in their subsequent avoidance of milk." Other populations may also experience lactose maldigestion and intolerance to milk during pregnancy.

Studies of lactose maldigestion in pregnant women, as measured by breath hydrogen response to 240 ml of low-fat (1%) milk, reinforce the Institute of Medicine's concern with lactose digestion among pregnant African American women. The prevalence of lactose maldigestion in early (13–16 weeks), late (30–35 weeks), and 8 weeks postpartum was 66, 69, and 75%, respectively, and that of nonpregnant control women was 80% (**Table 3**).

Accordingly, health care providers instructing African American women on the optimal dietary pattern during pregnancy need to be mindful of a high rate of lactose maldigestion. Implications for fetal growth and development remain to be determined by further study. Also, health providers need to be aware that the presence or absence of symptoms may be unevenly reported by pregnant African American women, and symptoms do not represent a reliable guide to lactose digestion. Less than 25% of pregnant lactose maldigesting women reported any symptoms with 240 ml of low-fat (1%) milk. Symptoms may be further reduced when milk is consumed with other foods. Unanswered is the level of digestion and absorption of a range of nutrients in the consumed milk. Health care providers should discuss with the pregnant woman her ability to tolerate milk and, when appropriate, should educate her as to other food options. In this regard, Kingfisher and Millard report that "Euro-American staff tended to give advise that was biologically appropriate for them but not for many of their patients, a process reflecting biocentrism."

Table 3 Lactose maldigestion[a] in pregnant and nonpregnant African American women

African American women	*% Lactose maldigestion*
Early pregnancy (13–16 weeks)	66
Late pregnancy (30–35 weeks)	69
Postpartum (8 weeks)	75
Nonpregnant	80

[a]Breath hydrogen increase >20 ppm following the consumption of 240 ml of low-fat (1%) milk containing 12 g of lactose following an overnight fast.

Age-Specific Prevalence

Age-Specific prevalence data suggest a progressive decrease in lactose absorption with age in African American children studied in the United States. This progressive decrease was seen in a study of 409 African American children 13 months to 12 years of age. The population was stratified by age to have approximately equal representation in each 12-month category. The mean age of the children studied was 6.6 years. The study subjects were drawn from four well child clinic sites and a private pediatrician's office in Baltimore, Maryland. All subjects were in good health as determined by history and a review of recent clinic visits. The children were free of any overt intestinal or allergic disorders and had no recent history of gastroenteritis.

Secondary Lactase Deficiency

Secondary lactase deficiency is distinct from genetically determined loss of lactase with age. Secondary lactase deficiency is frequently associated with diseases of the small intestine. Enteric viruses, such as rotavirus and Norwalk agent, can induce lactase deficiency by penetration of the enterocyte in the small intestine Rotaviruses are a principal cause of diarrhea and lactose intolerance in infancy. Denudation of the brush border of the jejunal mucosa associated with diarrhea can lead to the loss of the other two disaccharides, maltase and sucrose. Continued diarrhea may also lead to severe complications such as monosaccharide intolerance. Giardiasis have also

been implicated as contributing to lactose maldigestion. An additional infection resulting in an interference with lactose digestion is *Ascaris lumbricoides*. Severe protein malnutrition is frequently associated with lactose maldigestion. Other disease conditions that give rise to secondary lactose maldigestion are celiac disease, gluten-induced enteropathy, and tropical and nontropical sprue. The mucosal brush border of the small intestine is severely damaged in each case.

Lactose Digestion and Diet

Calcium

Dietary calcium is an important element in skeletal development. Dairy products can account for up to three-fourths of dietary calcium in some populations. Milk is a rich source of calcium. Nevertheless, many minorities in the United States and population groups throughout the world drink decreasing amounts of milk after early childhood and little milk as adults. Given the high prevalence of lactose intolerance, alternatives to cow's milk should be identified for those who need them. Lactose-intolerant individuals ultimately attribute their discomfort to lactose-containing foods and voluntarily reduce or eliminate their milk intake. Data from national studies in the United States indicate that African American and Hispanic women have lower intakes of calcium compared to non-Hispanic women. An Institute of Medicine report concludes that the disparity in calcium intake "may be explained in part by the much higher prevalence of lactose intolerance among African Americans and Hispanics, sometimes resulting in their subsequent avoidance of milk." In general, populations at risk for lactose intolerance report a lower calcium intake as a result of the decline in the intake of milk and milk products. One solution to this problem is to educate lactose-intolerant groups about alternative calcium-containing foods, reinforce appropriate cultural patterns and dietary practices that include alternatives to milk, and identify other culturally acceptable calcium-containing foods. Meeting the calcium requirement with an alternative diet is a challenge but nevertheless is required for many in the community. Although milk may serve as a primary source of calcium, appreciable quantities of calcium can be found in nondairy foods.

Clearly, it is more difficult to meet the published calcium recommendation with a diet low in whole cow's milk. A review of the tables of food composition reveals a variety of foods that contain acceptable levels of calcium per 100 g portion or other standard portions (**Table 4**). Other lactose-modified dairy products, including hard cheeses, yogurts, and lactose-modified milk, are good calcium sources.

Table 4 Calcium content in milligrams per 100-g portion or as noted

Food	*mg*
Canned sardines (3 oz.)	372
Buckwheat pancakes	249
Kale (raw)	225
Mustard greens	220
Muffins[a]	206
Waffles[a]	192
Figs (dry)	186
Canned salmon (3 oz. with bones)	167
Collard greens	162
Oat breakfast cereal[a]	160
Wheat pancakes	158
Almonds	152
Tofu (8 oz.)	143
Egg yolk	147
Corn bread[a]	139
Kale (frozen)	134
Filberts	120
Beet greens	118
Oysters (1/2 cup)	113
Whole cow's milk (100 g)	113
Swiss chard	105
Rhubarb (cooked 1/2 cup)	105
Canned shrimp (3 oz.)	98
Okra	92
Soy beans (1 cup)	90
Sunflower seeds	88
Broccoli	88
Sauerkraut (1 cup)	85
Potato salad (1 cup)	80
Peanut butter	74
Spinach	73
Dates (dry)	72
Brewer's yeast (2 tbs)	66
Lobster	65
Green beans	63
Flounder	61
Bran flakes	61
Canned apricots (1 cup)	57
Gingerbread (1 piece)	57
Plain rolls[a]	55
Toaster pastry (1 piece)[a]	54
Prunes (dry)	54
Orange	54
Whole egg	54
Peanuts	54
Artichokes	51
Cod	50
Brussels sprouts	50
Clams (3 oz.)	47
Lima beans	47
Puffed wheat[a]	46
Whole wheat bread (2 slices)	46
Sweet potato	46
Fruit cocktail (1 cup)	46

Continued

Table 4 Continued

Food	*mg*
Raisins (1/2 cup)	45
Apricots	44
Farina (1 cup)	44
Fig bars (4 cookies)	44
Pecans	43
White bread (2 slices)	42
Pecans	43
White bread (2 slices)	42
Tangerine	40
Raspberries (raw)	40
Apple sauce	21

[a]Enriched, fortified, or restored to legal standard when one exists.

From Oski FA and Paige DM (1994) Cow's milk is a good food for some and a poor choice for others: Eliminating the hyperbole. *Archives of Pediatric and Adolescent Medicine* **148**: 104–107.

In addition, lactose digestive aids are available and are increasingly used, including lactase tablets, lactase preparations, lactose-free milk, and prehydrolyzed milk. Live culture yogurt is another alternative to milk. Lactose in yogurt is better digested than lactose in milk. Tolerance to yogurt is thought to be due to the microbial β-galactosidase activity that digests the lactose.

Osteoporosis

The role of lactose maldigestion, calcium intake, and osteoporosis has been studied. Osteoporosis and osteoporitic fractures are major public health problems. The role of lactose maldigestion and osteoporosis remains unsettled. For example, minority populations consuming small amounts of milk should be at greater risk for osteoporosis. Nevertheless, African American and Hispanic populations in the United States appear to have a lower risk of developing osteoporosis. Caucasian and Asian women were found to have the highest risk for osteoporosis, with fracture rates of 140.7/100 000 and 85.4/100 000, respectively. Hispanic and African American females had lower age-adjusted rates at 49.7/100 000 and 57.3/100 000, respectively. The paradox reinforces the complexity of the disease and the importance of biologic, genetic, and as yet undetermined factors in the eitology of osteoporosis.

Nutrition Policy

Apart from the nutritional implications outlined previously, there are policy considerations that require attention. Clearly, milk has important economic, nutritional, and emotional significance in Western culture, a culture strongly committed to the concept that milk is an ideal food. However, lactose digestion should be an important consideration in developing a suitable policy regarding the use of milk and dairy products by the lactose malabsorber and by ethnic or racial groups, among whom high rates of malabsorption prevail. Accordingly, a balance must be struck between dietary guidance and the interests of a diverse population with a large number of lactose maldigesters. For many, the continued use of a limited amount of milk may be appropriate and comfortable. For others, dietary modification and lactose reduction or elimination may be warranted. The substitution of low-lactose products or alternative foods may be nutritionally beneficial. The successful introduction of a lactose-reduced milk, Lact-Aid, into the US market in the 1970s by Alan Kligerman is one important example of a well-accepted milk product alternative. Traditional diets among lactose-maldigesting populations, using little or no milk or dairy products, should be respected.

Summary

The principles of genetics and evolution help to explain the emergence of the aberrant phenomenon of lactose tolerance. Darwin referred to food as a major factor in selective pressures. Lactose digestion is most effective in illustrating how a certain food, by indirectly favoring the survival of those able to digest that substance, can influence the evolutionary process of man.

Clinical and nutritional consequences of lactose digestion in adults must be examined in relation to malabsorption, intolerance, milk rejection, and symptoms and their recognition. Estimates of how frequently milk intolerance will be a clinically significant problem in adults vary with the nature of the associated gastrointestinal disorders and the format of the individual studies.

There is a series of interrelated physiologic events affecting the amount of undigested sugar and fluid that the small intestine, and subsequently the colon, must metabolize or reabsorb. A balance of these factors tends to prevent symptoms when the stomach, small intestine, and colon can compensate for the increased solute load, but abdominal discomfort or diarrhea occur when these small intestinal and colonic physiologic mechanisms are loaded beyond their capacity. The role of the colonic flora in metabolizing unabsorbed sugar and the importance of colonic salvage of unabsorbed carbohydrate are important variables in the symptom complex. Secondary lactase deficiency due to infectious gastroenteritis and malnutrition represents a distinct clinical syndrome and must be distinguished from lactose intolerance.

Dietary recommendations must take account of lactose maldigestion. Milk and dairy product

consumption will vary among lactose-maldigesting and milk-intolerant individuals. Lactose-reduced or lactose-free products are available to lactose-intolerant individuals who wish to drink milk and milk-based products. Nevertheless, dietary recommendations must be modified and respectful of those who do not drink milk. Accordingly, appropriate alternatives to milk and other lactose-containing foods must be identified and guidance provided in developing nutritionally equivalent diets.

See also: **Calcium**. **Celiac Disease**. **Dairy Products**. **Osteoporosis**. **Pregnancy**: Nutrient Requirements.

Further Reading

Enattah NS, Sahi T, Savilahti E *et al.* (2002) Identification of a variant associated with adult-type hypolactasia. *Nature Genetics* **30**(2): 233–237.

Hollox EJ, Poulter M, Zvarik M *et al.* (2001) Lactase haplotype diversity in the Old World. *American Journal of Human Genetics* **68**(1): 160–172.

Institute of Medicine (US) Subcommittee on Nutritional Status and Weight Gain During Pregnancy (1990) *Nutrition During Pregnancy*. National Academy of Sciences.

Kingfisher CP and Millard AV (1998 Dec) Milk makes me sick but my body needs it: conflict and contradiction in the establishment of authoritative knowledge. *Med Anthropol Q* **12**(4): 447–66.

Labayen I, Forga L, Gonzalez A *et al.* (2001) Relationship between lactose digestion, gastrointestinal transit time and symptoms in lactose malabsorbers after dairy consumption. *Alimentary Pharmacology and Therapeutics* **15**(4): 543–549.

Oski FA and Paige DM (1994) Cow's milk is a good food for some and a poor choice for others: Eliminating the hyperbole. *Archives of Pediatric and Adolescent Medicine* **148**: 104–107.

Paige DM (1981) Lactose malabsorption in children: Prevalence, symptoms, and nutritional considerations. In: Paige DM and Bayless TM (eds.) *Lactose Digestion: Clinical and Nutritional Implications*, pp. 151–161. Baltimore: Johns Hopkins University Press.

Paige DM, Davis L, Bayless TM *et al.* (1979) The effects of age and lactose tolerance on blood glucose rise with whole cow and lactose hydrolyzed milk. *Journal of Agriculture and Food Chemistry* **27**(4): 667.

Paige DM, Witter F, Bronner YL *et al.* (2003) Lactose intolerance in pregnant African-Americans. *Public Health Nutrition* **6**(8): 801–807.

Perman JA, Barr RB, and Watkins JB (1978) Sucrose malabsorption in children; Non-invasive diagnosis by interval breath hydrogen determination. *Journal of Pediatrics* **93**: 17–22.

Pribila BA, Hertzler SR, Martin BR, Weaver CM, and Savaiano DA (2000) Improved lactose digestion and intolerance among African-American adolescent girls fed a dairy-rich diet. *Journal of the American Dietetic Association* **100**(5): 524–528.

Rao DR, Bello H, Warren AP, and Brown GE (1994) Prevalence of lactose maldigestion. *Digestive Diseases and Sciences* **39**: 1519–1524.

Scrimshaw NS and Murray EB (1988) The acceptability of milk and milk products in populations with a high prevalence of lactose intolerance. *American Journal of Clinical Nutrition* **48**: 1083–1159.

Suarez F, Savaiano D, and Levitt M (1995) A comparoson of symptoms after the consumption of milk or lactose-hydrolyzed milk by people with self-reported severe lactose intolerance. *New England Journal of Medicine* **333**: 1–4.

Swallow DM (2003) Genetics of lactase persistence and lactose intolerance. *Annual Review of Genetics* **37**: 197–219.

Tursi A (2004) Factors influencing lactose intolerance. *European Journal of Clinical Investigation* **34**: 314–315.

Vonk RJ, Priebe MG, Koetse HA *et al.* (2003) Lactose intolerance: Analysis of underlying factors. *European Journal of Clinical Investigation* **33**: 70–75.

LEGUMES

M A Grusak, Baylor College of Medicine, Houston, TX, USA

Published by Elsevier Ltd.

Legumes have been an important component of the human diet for several millennia and are used throughout the world today. They are a diverse group of plants that belong to the Fabaceae family (sometimes also referred to as the Leguminosae) and are estimated to include approximately 20,000 species in 700 genera. However, only a handful of these species have been developed as crops that are in common culture. Some of the more extensively grown legumes are listed in **Table 1**.

Legumes are consumed primarily as seed foods, but pods, leaves, and roots or tubers of various species are also eaten. The pod is an enveloping structure that protects the seeds as they develop and mature, and it is a characteristic feature of this group of plants. In fact, the name legume comes from the Latin word *legumen*, which means seeds that are harvested from pods. Other names used for legume seeds are pulse, which is derived from the Latin word *puls*, meaning pottage, or the phrase grain legume, used in reference to leguminous seeds. The more general phrase, food legume, is used to represent any vegetative or reproductive structures from legume plants that are utilized for human food.

An important nutritional aspect of legume foods is their high concentration of protein, which in most

Table 1 Commonly cultivated legume species

Scientific name	*Common names*
Arachis hypogea L.	Peanut, groundnut
Cajanus cajan (L.) Millsp.	Pigeon pea, red gram, Congo pea
Cicer arietinum L.	Chickpea, garbanzo, Bengal gram
Glycine max (L.) Merr.	Soybean, soya, edamame
Lablab purpureus (L.) Sweet	Hyacinth bean, Indian bean, Egyptian bean
Lathyrus sativus L.	Grass pea, chickling pea
Lens culinaris Medik.	Lentil
Lupinus albus L.	White lupine
Macrotyloma uniflorum (Lam.) Verdc.	Horse gram, Madras gram
Phaseolus lunatus L.	Lima bean, butter bean
Phaseolus vulgaris L.	Common bean, black bean, kidney bean, pinto bean, snap bean, string bean, French bean
Pisum sativum L.	Pea, garden pea, English pea
Psophocarpus tetragonolobus (L.) DC.	Winged bean, Goa bean, four-angled bean
Vicia faba L.	Broad bean, fava bean
Vigna aconitifolia (Jacq.) Marechal	Moth bean, mat bean
Vigna mungo (L.) Hopper	Urd bean, black gram
Vigna radiata (L.) Wilczek	Mung bean, green gram, golden gram
Vigna subterranea (L.) Verdc.	Bambara groundnut
Vigna umbellata (Thumb.) Ohwi and Ohashi	Rice bean, Mambi bean
Vigna unguiculata (L.) Walp. ssp. *unguiculata*	Cowpea, black-eyed pea, southern pea

Source: Rubatzky VE and Yamaguchi M (1997) *World Vegetables: Principles, Production, and Nutritive Values*. New York: Chapman & Hall.

legume seeds is at least twice that of cereal seeds. Legumes can produce more protein because the plants are generally well nourished with nitrogen, even in soils with limited inorganic nitrogen. Legume roots have the ability to form symbiotic associations with particular microbial species, in a structure called the root nodule. This symbiosis allows the plant to readily acquire atmospheric nitrogen and use it for the synthesis of amino acids. These protein precursors are transported to the developing seeds and are deposited there for later use. Legume seeds also contain a broad mix of energy reserves (starch or oil), minerals, and various phytochemicals—all of which are stored in seeds to provide nourishment to the young developing seedling.

As omnivores, humans have been able to take advantage of the nutrient and phytochemical reserves in legume seeds for dietary requirements and health benefits. This is especially important in the developing world, where malnutrition is an ever-present concern, and legumes can provide an inexpensive source of dietary protein (relative to animal food products), among other nutrients. The protein in legume seeds, although somewhat lacking in sulfur amino acids and tryptophan, is still an important complement to energy-rich carbohydrate staples, such as rice, wheat, maize, and various root and tuber crops. However, when eating legumes, we also must deal with the various antinutrients and toxic compounds found in seeds. These seed components include various enzyme inhibitors, tannins, phenolics, alkaloids, and neurotoxins. Some of these can cause debilitating consequences in humans, although cooking and other processing techniques can be used to reduce or alleviate their negative effects.

Legume Types

Legumes are grown throughout the world, with some adapted to warmer tropical and subtropical climates and others preferring temperate to cooler climates. The 20 species listed in **Table 1** are some of the more commonly cultivated legumes and include those whose annual production reaches levels that allow for worldwide marketing. In developing countries, many locally adapted legume species are cultivated on a small scale or are harvested from wild sources. These less cultivated legumes are usually harvested as mature seeds, but immature pods, leaves, roots, or tubers can also be collected.

Most of the common legume species are grown agronomically and harvested as mature seeds. These can be cooked and consumed in their entirety, or they are cracked and used as split seeds with the hulls (seed coats) removed. Seeds of some species are milled to produce a flour product, or they can be processed to yield protein isolate (e.g., soybean and lupine), extracted oils (e.g., soybean and peanut), or starch (e.g., pea).

For those legumes also cultivated as vegetable crops, immature seeds or immature pods can be harvested. These are canned, frozen, or sold as fresh products. Immature pods are nutritionally similar to leafy vegetables in that they contain various carotenoids and other phytochemicals; however, they also contain immature seeds that can provide a modest amount of protein. For some species, young tender leaves or whole shoots are also collected and used as vegetable greens that are eaten fresh or cooked. More detailed information is given on some of the common legume types in the following sections.

Bambara Groundnut

Bambara groundnut (*Vigna subterranea* (L.) Verdc.) is indigenous to west central Africa. Most of its current production is in Africa, but the plant is also cultivated in India, Southeast Asia, Australia, and Central and South America. The plant has an interesting growth habit in that after pollination, the developing pod and seeds are pushed into the ground, where they grow until full maturity. Plants are typically uprooted at harvest to collect the seeds and pods; because of this subterranean growth, they have acquired the common name groundnut. Mature seeds are boiled and consumed as a cooked seed, prepared as porridge, or milled into a flour to form cakes. Immature seeds are also harvested and cooked as a fresh vegetable.

Broad Bean

Broad bean (*Vicia faba* L.), also known as fava bean, is grown from tropical to temperate regions, with production occurring in North and South America, Europe, Africa, and China. This legume is grown for its enlarged, succulent, immature seeds that are removed from its thick, fleshy pod. Mature dry seed is also harvested. Although broad beans are widely consumed, they do contain storage proteins (vicine and convicine) whose metabolites can lead to acute hemolytic anemia in individuals with a deficiency in glucose-6-dehydrogenase (found predominantly in people of Mediterranean or African descent). Additionally, broad beans contain high levels of L-DOPA, a phenolic compound that can be converted to dopamine. Because of their L-DOPA content, broad beans should be avoided by individuals using monoamine oxidase inhibitors (MAOI-type drugs). The use of these drugs, in combination with high intakes of dopamine (or dopamine precursors), can lead to dangerous increases in blood pressure.

Chickpea

Chickpea (*Cicer arietinum* L.) is grown worldwide and is best adapted to cool, dry climates. Thus, it is a winter crop in some regions of the world. Two seed types are recognized: the large-seeded kabuli type, characterized by its beige-colored seed coat and ram's head shape, and the desi type, with its smaller size and dark-colored irregularly shaped seeds. Kabuli varieties are preferred for consumption as whole seeds, whereas desi types are typically processed into flour. Immature green pods and young tender leaves are also cooked and eaten as vegetables, especially in India.

Common Bean

Common bean (*Phaseolus vulgaris* L.) is grown in temperate zones as well as in temperate regions within the subtropics. As a dry seed, it is an important crop in Africa and in Central and South America. Many bean types are cultivated that exhibit vast differences in seed coat coloration and pod characteristics. Mature seeds are harvested as dry beans (e.g., black bean, pinto bean, and kidney bean); immature pods are used as a vegetable (e.g., snap bean and French bean). Pod types have been bred to have few fibers in the pod wall.

Cowpea

Cowpea (*Vigna unguiculata* (L.) Walp. ssp. *unguiculata*) is grown throughout the tropics and subtropics. It is an important crop in Africa, its probable center of origin, but is also grown in Brazil, India, Southeast Asia, and the United States. There are three major subspecies of *V. unguiculata*; in addition to ssp. *unguiculata*, there is appreciable production of ssp. *cylindrica* (common names: catjang cowpea and Bombay cowpea) and ssp. *sesquipedalis* (common names: yardlong bean and asparagus bean), especially in Asia. All types are harvested as vegetables (shoots, leaves, and immature pods) or as dry, mature seeds.

Grass Pea

Grass pea (*Lathyrus sativus* L.) is a hardy, cool-weather adapted legume that is cultivated in India, Africa, the Middle East, and South America. It is harvested primarily as a dry, mature seed, although young leaves and immature pods are edible. Grass pea is quite tolerant of limited moisture and does well in nutrient-poor soils; thus, in times of drought it is one of the few legumes that produces a harvest, and it is widely consumed by low-income populations during times of famine. Unfortunately, excessive or prolonged consumption of grass pea can lead to lathyrism, a debilitating muscle paralysis that is caused by a neurotoxin in the seeds.

Hyacinth Bean

Hyacinth bean (*Lablab purpureus* (L.) Sweet) is grown in India and in many tropical regions of the world. Mature seeds are consumed as a cooked food or a sprouted seed. The immature pods and seeds are also harvested as vegetable foods. Although this plant is cultivated as an annual, it will persist as a perennial, and when cultivation is extended it will form large, starchy roots that can be eaten. Some varieties (mostly dark-seeded types) contain high levels of a cyanogenic glycoside in their seeds.

When cyanogenic glycosides are hydrolyzed by plant enzymes during cooking, or possibly by intestinal enzymes after ingestion, cyanide can be released and lead to cyanide poisoning.

Lentil

Lentil (*Lens culinaris* Medik.) is another of the world's important pulse crops, especially for populations in developing countries. The plant is adapted to cool climates; Canada, India, and Turkey account for nearly 70% of its production. Lentils are harvested primarily as a dry, mature seed, but immature pods are also used as a vegetable in India.

Mung Bean

Mung bean (*Vigna radiata* (L.) Wilczek) is grown in tropical climates and is an important legume in India, China, and other Asian countries. Dry seeds are harvested and consumed as split, whole, boiled, or roasted forms. Immature pods are eaten, and there has been interest in developing the tuberous root as a food because of its high protein content (nearly 15%).

Pea

Pea (*Pisum sativum* L.) is grown primarily in cooler regions of the world. Different varieties have been developed to produce mature, dry seeds; succulent, well-developed immature seeds; or succulent, immature edible pods. Dry seed varieties are sometimes referred to as field peas. The names garden pea and English pea are used for the varieties harvested as immature seeds, whereas the edible pod types are commonly known as snow pea or sugar snap pea. In some Asian cuisines, the shoots of pea plants are also used as vegetable greens.

Peanut

Peanut (*Arachis hypogea* L.) is grown throughout the tropics, much of the subtropics, and even in some temperate zones. As with Bambara groundnut, its pods have a subterranean growth habit, and thus it also has acquired the common name, groundnut. Peanut is one of the few commonly grown legumes whose seeds contain high levels of oil. Most legume seeds have less than 5% oil, but for some peanut cultivars seed oil content is as high as 40–50%. Roasted seed and extracted oil is used and marketed worldwide; in some regions, young shoots and leaves of the plant are used as greens, and immature pods are consumed as a cooked vegetable. Although a nutritious legume, peanut has recently gained much attention and scientific interest due to the low, but nonetheless significant, incidence of individuals who are allergic to peanut proteins. For those extremely hypersensitive to this food, violent and life-threatening reactions can occur in response to exposure to as little as 0.1 mg of peanut. In fact, peanut is believed to be the most common cause of death due to foods.

Pigeon Pea

Pigeon pea (*Cajanus cajan* (L.) Millsp.) is broadly adapted to many climatic regions and soil types, and thus its production occurs over a huge area of crop land. It is an important food legume in India, other Asian countries, Africa, and South America. Mature grains are usually consumed as split, dehulled seeds. Immature seeds and pods are also consumed in large quantities.

Soybean

Soybean (*Glycine max* (L.) Merr.) is undoubtedly the most important food legume today, being a major source of protein and extracted oil. Soybean is believed to have originated in eastern Asia as a subtropical plant, but plant breeders have helped develop varieties adapted to several climatic zones. The crop is grown in many countries, but more than 70% of the world's production comes from the United States, Brazil, and China. Most soybeans are harvested as dry seed; a typical variety contains 20% seed oil and 35% protein (although some varieties can be as high as 45% protein). Both soy oil and soy protein isolate are found as ingredients in many processed foods. In eastern Asia, the immature seed is also harvested extensively and used as a vegetable.

Winged Bean

Winged bean (*Psophocarpus tetragonolobus* (L.) DC.) is adapted to tropical conditions and is grown in Southeast Asia, Papua New Guinea, various Pacific Islands, and Africa. The tender pods are the most widely consumed part of the plant, especially throughout Asia, but the leaves, stems, flowers, seeds, and tuberous roots are all nutritionally valuable and are used as food. Winged bean is another of the legumes with elevated seed oil content; varieties typically average 15% oil, with protein levels of 30–37%. The tuberous roots are a good source of energy in the form of starch, and they contain 8–10% protein.

Grain Legume Nutritional Value

As noted previously, many parts of legume plants are consumed by humans. However, the seeds are

the predominant food type across all species, and their nutritional value is discussed in the following sections.

Protein

Protein content in legume seeds is governed both by genotype and by environment. Seed protein levels can vary across varieties of a given species and even among seeds on an individual plant. In general, however, food legumes contain 20–30% protein by proximate analysis (**Table 2**). The exceptions to this are soybean and winged bean, which contain up to 37 and 45% protein, respectively.

Legume proteins are primarily of two types: storage proteins, which account for approximately 70% of total seed nitrogen, and enzymatic, regulatory, and structural proteins, which are present for normal cellular activities, including the synthesis of storage proteins. Legume storage proteins are soluble in dilute salt solutions but insoluble in water and therefore fall into the classical globulin group of protein fractions. Legume protein types are further characterized by their sedimentation coefficients, which in most species approach 11S and 7S; these are commonly referred to as the legumins and vicilins, respectively. Most legumes contain both types of storage protein, but the proportion of the two types varies from species to species.

In terms of protein quality, as defined by an optimal proportion of amino acids required by humans, legume proteins are deficient in the sulfur-containing amino acids and tryptophan but are rich in lysine. Cereals, on the other hand, are relatively deficient in lysine; thus, the combination of legumes with cereals often can improve the overall protein quality of the mixed foods. The nutritive value (or biological value) of legume proteins has been investigated quite extensively and has been shown to be rather low in some legumes, with the amount of utilizable protein ranging from 32 to 78%. In other words, not all of the protein available in a given legume (see **Table 2**) is converted into new protein when consumed by humans. The reasons for this are the general deficiency of essential amino acids (sulfur-containing and tryptophan) and the presence of many inhibitors of protease activity that are found in legume seeds. These enzyme inhibitors are primarily proteinaceous in character, and many have an effect on the digestive enzymes trypsin or chymotrypsin. The inhibition of these enzymes leads to a reduction in protein digestibility and thus the gut's ability to absorb amino acids. Fortunately, because many of these inhibitors are proteinaceous, cooking, heating, fermenting, and, in some cases, germination can inactivate and significantly lower their inhibitory effect. However, not all of the inhibitors found in legume seeds are proteins (e.g., other inhibitors include tannins and polyphenols).

Table 2 Protein contents of food legume seeds

Legume	*Protein range (% dry weight)*
Broad bean	22.9–38.5
Chickpea	14.9–29.6
Common bean	21.1–39.4
Cowpea	20.9–34.6
Grass pea	22.7–29.6
Horse gram	18.5–28.5
Lentil	20.4–30.5
Moth bean	21.0–31.3
Mung bean	20.8–33.1
Pea	21.2–32.9
Peanut	23.5–33.5
Pigeon pea	18.8–28.5
Rice bean	18.4–27.0
Soybean	33.2–45.2
Urd bean	21.2–31.3
Winged bean	29.8–37.4

Source: Salunkhe DK, Kadam SS and Chavan JK (1985) *Postharvest Biotechnology of Food Legumes*. Boca Raton, FL: CRC Press.

Lipids

Grain legumes generally contain higher concentrations of lipids than cereals. In legumes, lipids are stored in oil bodies in the cotyledons (the bulk of the seed), whereas most oils in cereals are limited to the outer bran layer. Most common legumes contain 1–7% lipid, based on proximate analysis. Exceptions to this range are soybean, peanut, and winged bean, which average 20, 45, and 15%, respectively. Legumes are good lipid sources for humans because they contain high amounts of essential fatty acids. Although composition varies across species, most legumes contain some quantity of oleic, linoleic, and linolenic acids. Phospholipids and glycolipids are also found in legume seeds.

Carbohydrates

Legume seeds contain starch, mono- and oligosaccharides, and other polysaccharides. Total carbohydrates range from 25 to 65% across the commonly grown legume species. Starch is the predominant carbohydrate in most cases, with exceptions in the oilseeds soybean and peanut. Legumes generally contain low amounts of monosaccharides (usually 1% or less) and only slightly higher amounts of

disaccharides, such as sucrose (1–3%). However, some soybean varieties have been reported to contain as much as 7% sucrose.

Various oligosaccharides have been characterized in legume seeds, including raffinose, stachyose, and verbascose, which are galactosides of sucrose. Because humans do not express the enzyme α-galactosidase, these compounds remain undigested in the small bowel and pass through to the large bowel, where they can be fermented by anaerobic microbes. This leads to flatulence, or gas production, which is experienced following the consumption of some legumes. The concentration of raffinose-type oligosaccharides varies among legume species and, not surprisingly, the capacity to induce flatulence also varies.

Fiber

Legume seeds are a source of dietary fiber, containing both crude fiber and neutral detergent fiber. Most legumes contain 3–5 g of fiber per 100 g of dry seed, with most of the fiber found in the seed coat fraction. Exceptions are grass pea and hyacinth bean, which contain 8 and 10 g of fiber, per 100 g of dry seed, respectively. Compositionally, legume seeds contain varying quantities of lignin, cellulose, hemicelluloses, pectins, gums, and mucilage.

Minerals

Legume seeds contain a broad mix of minerals, many of which are essential both for plants and for animals. In fact, almost all essential minerals for humans can be found stored in the seeds. In comparison to cereals, legumes tend to have higher concentrations of calcium and potassium, as well as the micronutrients iron, zinc, and copper. Most of the calcium is sequestered as calcium oxalate crystals, however, and this form of calcium has extremely low bioavailability. Also, the majority of phosphorus in legume seeds is stored as phytic acid, which can complex calcium, iron, and zinc and thereby diminish their bioavailability. Other compounds found in legume seeds, including tannins, phenols, organic acids, protein, and fiber, can also interact with minerals and lower their bioavailability. Fortunately, certain processing procedures, such as fermenting or sprouting seeds, can reduce the levels of some of these mineral chelators. Due to these various problems, there is a significant effort under way in the plant science community to increase the absolute mineral levels in various legume seeds as well as to lower the levels of several major inhibitory compounds.

Vitamins

Most food legumes are good sources of thiamin, riboflavin, and niacin but are poor sources of ascorbic acid. This vitamin is present at only low levels in newly harvested dry seeds, and it disappears after long storage. In some species, varieties exist that produce green- or orange-colored cotyledons, and β-carotene, a pro-vitamin A carotenoid, can be found in some cases. The amounts of this vitamin precursor, however, are generally quite low. Tocopherols (vitamin E) are also found in some legume seeds, and folate, which is present in all legumes, can be quite high in certain species (e.g., lentil). Because folate is important in the prevention of neural tube defects, legume consumption is recommended for women of childbearing age, especially in regions of the world where folate fortification is limited.

Health-Promoting Phytochemicals

There is much interest in the role of various phytochemicals to promote good health and to reduce the risk of various cancers. As with many plant foods, legume seeds contain a number of these types of compounds. Prominent in this group are the isoflavonoids, such as genistein and daidzein, which are found at high levels in soybeans. Epidemiological studies have suggested a positive association between the consumption of soy isoflavones and reduced risk of breast and prostate cancer in humans. These and other related isoflavones are found in seeds of most of the commonly grown legumes. In addition, various saponins, catechins, epicatechins, and anthocyanidins have been measured in various legume seeds, and these compounds have also been suggested to have health-promoting qualities. Plant biochemists and human nutritionists are actively working to manipulate the levels of these and other compounds in legumes.

See also: **Bioavailability**. **Cereal Grains**. **Protein**: Quality and Sources. **Vegetarian Diets**. **Whole Grains**.

Further Reading

Aykroyd WR (1982) *Legumes in Human Nutrition* Rome: Food and Agriculture Organization of the United Nations.

Deshpande SS, Salunkhe DK, Oyewole OB *et al.* (2000) *Fermented Grain Legumes, Seeds and Nuts: A Global Perspective* Rome: Food and Agriculture Organization of the United Nations.

Duke JA (1981) *Handbook of Legumes of World Economic Importance* New York: Plenum Press.

Duranti M and Gius C (1997) Legume seeds: Protein content and nutritional value. *Field Crops Research* 53: 31–45.

Hedley CL, Cunningham J, and Jones A (2000) *Carbohydrates in Grain Legume Seeds: Improving Nutritional Quality and Agronomic Characteristics* New York: CABI.

Rubatzky VE and Yamaguchi M (1997) *World Vegetables: Principles, Production, and Nutritive Values* New York: Chapman & Hall.

Salunkhe DK and Kadam SS (1989) *CRC Handbook of World Food Legumes: Nutritional Chemistry, Processing Technology, and Utilization*, vols. 1–3. Boca Raton, FL: CRC Press.

Salunkhe DK, Kadam SS, and Chavan JK (1985) *Postharvest Biotechnology of Food Legumes* Boca Raton, FL: CRC Press.

Summerfield RJ and Roberts EH (1985) *Grain Legume Crops* London: Collins.

Wang TL, Domoney C, Hedley CL *et al.* (2003) Can we improve the nutritional quality of legume seeds? *Plant Physiology* **131**: 886–891.

LIPIDS

Contents

Chemistry and Classification

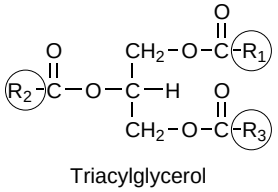
J L Dupont, Florida State University, Tallahassee, FL, USA

Lipids are generally known as fats and oils in food and nutrition. They are unique in nutrition, as they are in all of biology, in that they are not soluble in water. Early work on the chemistry of living organisms led to the discovery that fatty substances were soluble in organic solvents, such as chloroform, ethyl ether, alcohols, and hydrocarbons. Those solubility characteristics are dependent on the neutral or polar attributes of particular lipids and define the structural and functional aspects of lipids in living systems. This article presents the classification of lipids in their chemical groupings, their characteristic chemical and physical properties, and their nomenclature. Major groups of lipids include fatty acids, acylglycerols, phospholipids and sphingolipids, and sterols. Some lipid compounds, such as fat-soluble vitamins and waxes, are not included.

Fatty Acids and Acylglycerols

Nomenclature

Fatty acids are hydrocarbons of chain length two or greater with a carboxyl group at one end. Hydrocarbon chains are termed acyl lipids, and fatty acids occur most abundantly esterified to glycerol as triacylglycerols (**Figure 1**). Nomenclature for fatty acids has evolved from studies of food or organ sources of the lipid, extraction and identification methods, and attempts at classification. **Table 1** lists fatty acids important in food and nutrition. The accepted shorthand description shows the number of carbons: number of double bonds, location of double bonds from the carbon at the methyl (n or omega) position, and *cis* or *trans* configuration (**Figure 2**). Saturated fatty acids (SFAs) have hydrogen atoms at every possible carbon site, and unsaturated fatty acids have double bonds. Fatty acids with double bonds may occur in isomeric forms. Geometric isomers are referred to as *cis* and *trans* rather than the convention Z and E preferred by chemists. For example, linoleic acid is 18:2 n-6, having 18 carbons, two double bonds located at the n (or omega) minus 6 and n minus 9 positions on the chain. Conventional carbon numbering is from the carboxyl end; therefore, linoleic acid can be written as *cis* 18:2 $\Delta^{9,12}$. Delta indicates numbering from the carboxyl carbon and the atom number from the carboxyl is sometimes used (C-9). Desaturase enzymes are named according to the delta number (i.e., Δ-9-desaturase). Commonly, the *cis* configuration is not noted because almost all natural fatty acids are in the *cis* configuration. Also, unless otherwise specified, the



Triacylglycerol

Figure 1 Stereochemical numbering of lipids derived from glycerol. R_1, R_2, and R_3 refer to *sn* nomenclature.

Table 1 Fatty acids important in nutrition

Symbol[a]	*Systematic name*[b]	*Common name*	*Melting point (°C)*	*Sources*
Saturated fatty acids (SFAs)				
2:0	*n*-Ethanoic	Acetic	16.7	Many plants
3:0	*n*-Propanoic	Propanoic	−22.0	Rumen
4:0	*n*-Butanoic	Butyric	− 7.9	Rumen and milk fat
6:0	*n*-Hexanoic	Caproic	− 8.0	Milk fat
8:0	*n*-Octanoic	Caprylic	12.7	Milk fat, coconut
10:0	*n*-Decanoic	Capric	29.6	Milk fat, coconut
12:0	*n*-Dodecanoic	Lauric	42.2	Coconut, palm kernel
14:0	*n*-Tetradecanoic	Myristic	52.1	Milk fat, coconut
16:0	*n*-Hexadecanoic	Palmitic	60.7	Most common SFA in plants and animals
18:0	*n*-Octadecanoic	Stearic	69.6	Animal fat, cocoa butter
20:0	*n*-Eicosanoic	Arachidic	75.4	Widespread minor
22:0	*n*-Docosanoic	Behenic	80.0	Minor in seeds
24:0	*n*-Tetracosanoic	Lignoceric	84.2	Minor in seeds
Monounsaturated (monoenoic) fatty acids				
10:1 n-1	*cis*-9-Decenoic	Caproleic		Milk fat
12:1 n-3	*cis*-9-Dodecenoic	Lauroleic		Milk fat
14:1 n-5	*cis*-9-Tetradecenoic	Myristoleic		Milk fat
16:1 n-7*t*	*trans*-Hexadecenoic	Palmitelaidic		HVO[c]
16:1 n-7	*cis-9*-Hexadecenoic	Palmitoleic	1	Most fats and oils
18:1 n-9	*cis*-9-Octadecenoic	Oleic	16	Most fats and oils
18:1 n-9*t*	*trans*-9-Octadecenoic	Elaidic	44	Ruminant fat, HVO[c]
18:1 n-7*t*	*trans*-11-Octadecenoic	*trans* Vaccenic	44	Ruminant fat
20:1 n-11	*cis*-9-Eicosaenoic	Gadoleic		Fish oils
20:1 n-9	*cis*-11-Eicosaenoic	Gondoic	24	Rapeseed, fish oils
22:1 n-9	*cis*-13-Docosaenoic	Erucic	24	Rapeseed, mustard oil
Polyunsaturated (polyenoic) fatty acids				
Dienoic				
18:2 n-9	*cis*,*cis*-6,9-Octadecadienoic		−11	Minor in animals
18:2 n-6	*cis*,*cis*-9,12-Octadecadienoic	Linoleic	− 5	Most plant oils
Trienoic				
18:3 n-6	All-*cis*-6,9,12-octadecatrienoic	γ-Linolenic		Evening primrose, borage oils
18:3 n-3	All-*cis*-9,12,15-octadecatrienoic	α-Linolenic	−11	Soybean and Canola oils
20:3 n-6	All-*cis*-8,11,14-eicosatrienoic	Dihomogammalinolenic		
Tetra; penta; hexaenoic				
20:4 n-6	All-*cis*-8,11,14-eicosatetraenoic	Arachidonic	−49.5	Meat
20:5 n-3	All-*cis*-5,8,11,14,17-eicosapentaenoic	EPA, Timnodonic		Fish oils
22:4 n-6	All-*cis*-7,10,13,16-docosatetraenoic	Adrenic		Brain
22:5 n-6	All-*cis*-7,10,13,16,19-docosapentaenoic	DPA, Clupanodonic		Brain
22:6 n-3	All-*cis*-4,7,10,13,16,19-docosahexaenoic	DHA		Fish

[a]Number of carbons:number of double bonds, location of first double bond from the methyl carbon.
[b]Geometric isomer-Δ positions of double bonds.
[c]HVO, hydrogenated vegetable oil.
t, *trans*.

double bonds are 3 carbons apart, referred to as 'methylene interrupted,' as contrasted with conjugated double bonds (**Figure 3**).

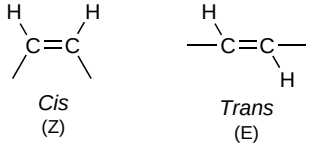


Figure 2 Structure of *cis* and *trans* double bonds.

Physical and Chemical Properties

Nonesterified fatty acids or free fatty acids have a polar (acidic) component and a neutral hydrocarbon component. The ratio of carbon to oxygen depends on the chain length and accounts for the solubility properties as well as the energy density of the lipid molecule. The hydrocarbon chain of the fatty acid is hydrophobic and the carboxyl end is hydrophilic, making the molecule amphipathic. This causes a dispersion of oil in water to form a mono molecular

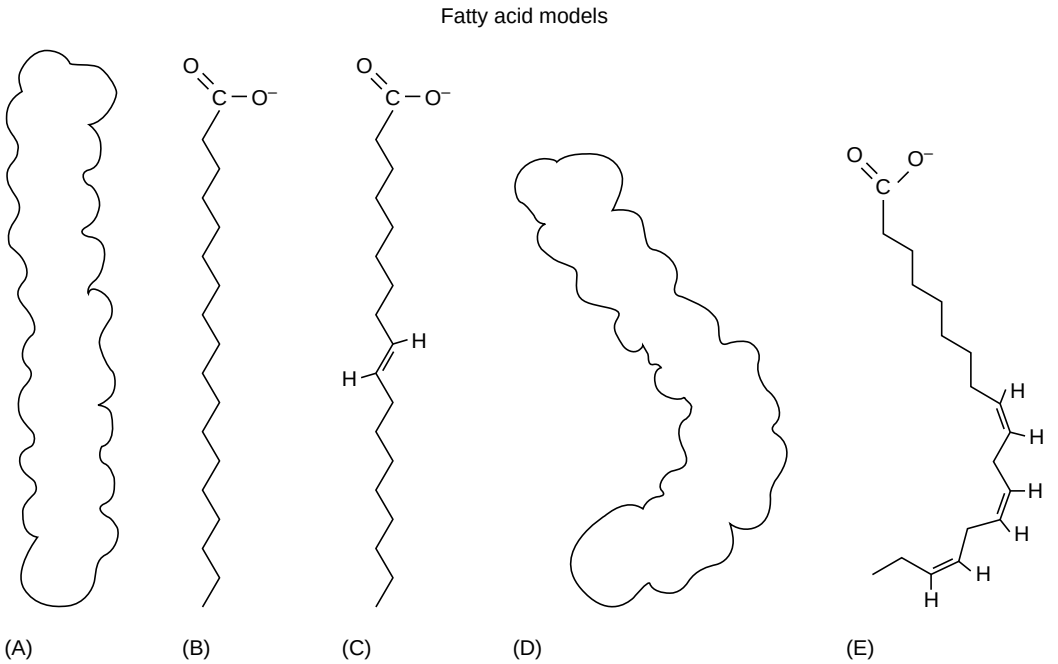


Figure 3 Space-filling and conventional models of fatty acids: (A) stearic acid (18:0), space-filling; (B) stearic acid, conformational; (C) elaidic acid (18:1n-9*t*) *trans*, conformational; (D) α-linolenic acid, all-*cis*, space-filling; (E) α-linolenic acid, conformational.

layer at the surface with the carboxyl end in contact with water and the hydrocarbon extending out of the water. The fatty acids may form a micelle (**Figure 4**) to separate the oil and water phases. These orientations of fatty acids and more complex lipids are a primary aspect of their participation in biological structures and functions. Furthermore, there is free rotation about the carbon–carbon bonds so the fatty acids and acylglycerols are capable of assuming a number of configurations.

Differences in the physical characteristics of fatty acids, particularly saturated compared with unsaturated, are extremely important in food and nutrition. SFAs with a chain length longer than 12 carbons are solid at usual ambient temperatures. As the chain lengthens, the melting point increases. On the contrary, mono- and polyunsaturated fatty acids (PUFAs) are liquid at room temperatures. Uses of fats in food products are based on these properties. Salad oils, margarines, and shortenings are examples of such differences. Hydrogenation of oils containing PUFAs was introduced to provide food fats that are resistant to rancidity and of a desirable plasticity. The chemical hydrogenation process yields mixtures of *cis* and *trans* fatty acids (**Figure 5**). The physical conformation of *trans* fatty acids is important for their functions in foods and nutrition (**Figure 3**). The melting points are similar to those of SFAs of similar length and their shapes are linear rather than bent as forced by the *cis* configuration. These physical characteristics affect their space-filling functions and the mobility of the molecule.

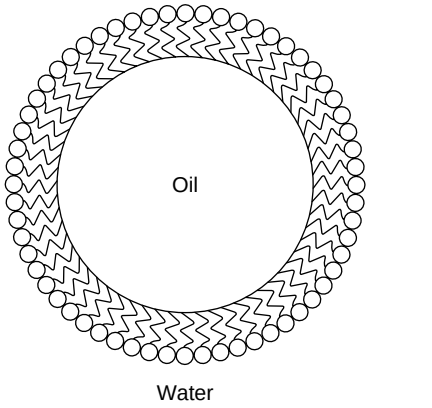


Figure 4 Micelle formed by oil dispersion in water.

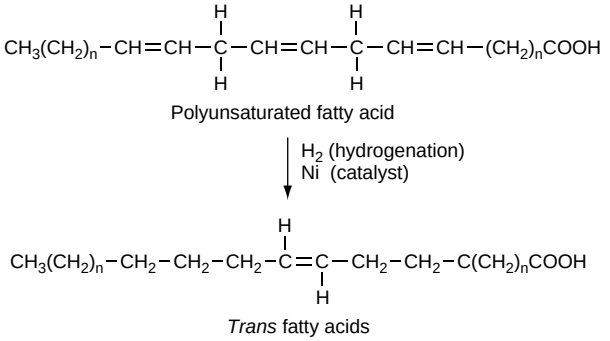


Figure 5 Hydrogenation of polyunsaturated fatty acids. *trans* fatty acids are produced when hydrogenation results in incomplete saturation of double bonds during chemical processing.

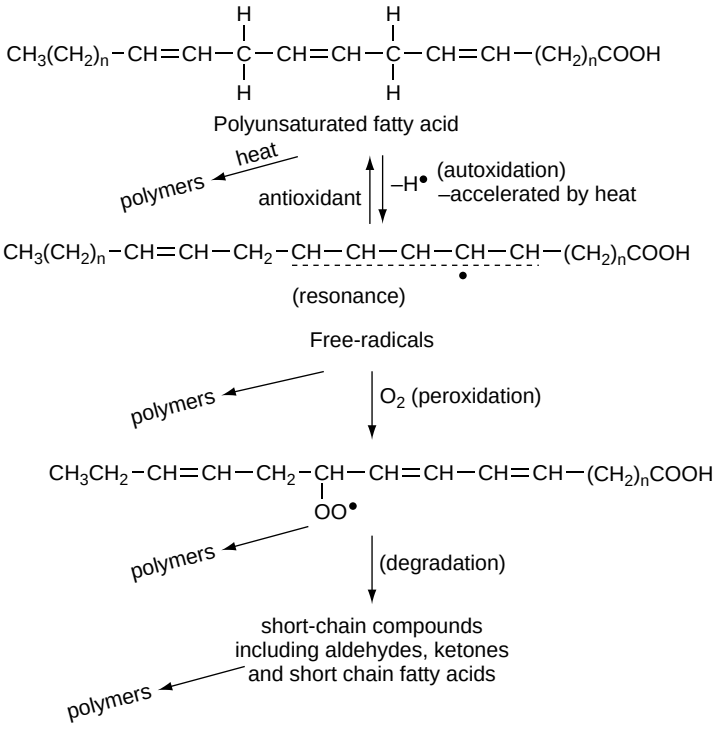


Figure 6 Autoxidation is caused by removal of a hydrogen from the methyl group between double bonds in polyunsaturated fatty acids. Resonating free radicals are produced and propagate peroxidation, degradation, and formation of polymers.

Polyunsaturated fatty acids with the methylene-interrupted double bond are also susceptible to oxidation (**Figure 6**). The hydrogen atoms on the methyl group between double bonds are susceptible to sequestration by oxidizing agents, such as iron or free radicals. This autoxidation results in a resonating free radical that is self-propagating and, with exposure to oxygen, yields peroxides. The peroxides may polymerize or degrade to smaller molecules. In foods, this process results in the condition of rancidity characterized by off flavors. In living systems, the products of peroxidation may cause reactions that damage proteins, membranes, and DNA resulting in pathological processes. Antioxidants are compounds that are capable of interrupting free radical propagation by reducing the peroxide to an alcohol without itself becoming a free radical. Tocopherols are a major antioxidant group in living systems and chemical antioxidants such as BHT (3,5-di-*t*-butyl-4-hydroxytoluene) are used in food products.

Another primary characteristic of naturally occurring PUFAs is that they cannot be synthesized by animals but are necessary for metabolism; therefore, they are an essential component of the diet. Animal organisms can introduce a double bond at the C-9 position but lack the enzymes to insert double bonds between the C-9 position and the methyl terminal carbon. The fatty acids are therefore considered to be in three families in relation to their biological functions: the monounsaturated (n-9 or omega-9) family and the polyunsaturated n-6 (omega-6) and n-3 (omega-3) families.

Phospholipids

The glycerol backbone is the central structure of phospholipids, as it is for acylglycerols. They are characterized by a phosphate group at the *sn*-3 position making phosphatidic acid (**Figure 7**).

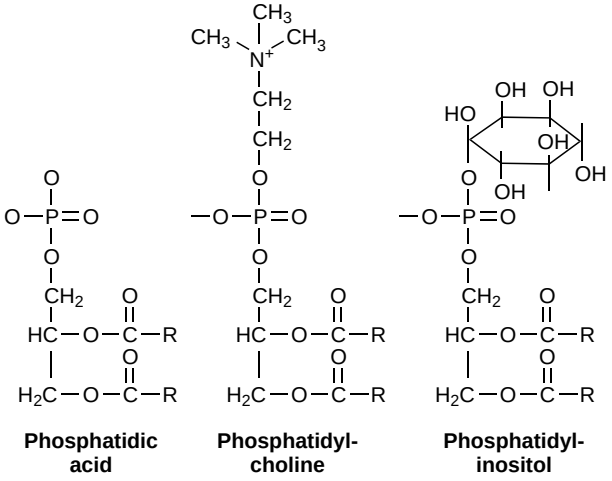


Figure 7 Structures of phospholipids.

Phosphoglycerides have fatty acids esterified at the 1 and 2 positions. A number of compounds may be esterified to the phosphate moiety, including choline, ethanolamine, serine, and *myo*-inositol. The compounds are called phosphatidylcholine, etc. These molecules are obviously amphipathic, having very polar constituents at the *sn*-3 position and acyl chains at the 1 and 2 positions. This attribute is very important to their function in biological membranes.

Sphingolipids

A group of acyl lipids has a sphingosine-based structure (**Figure 8**). The derivatives of sphingosine are cerebrosides and ceramides. They are characterized by having sugar molecules as part of their structure and thus are called glycosphingolipids. The most common sugar moiety is galactose. There is a large family of mono-, di-, and triglycosyl ceramides. Some of the glycosphingolipids have sialic (*N*-acetyl neurominic) acid linked to one or more of the sugar residues of a ceramide oligosaccharide. These are called mono-, di-, and trisialogangliosides and are abundant in membranes, particularly in nervous tissue. Their amphipathic structures are functional in membranes and in the water impermeability of skin.

Steroids

Sterols

Sterols are monohydroxy alcohols with a four-ring core structure or steroid nucleus (**Figure 9**). Cholesterol is the most abundant sterol in animal tissues. The tetracyclic structure is uniquely compact and rigid. The unesterified molecule has only one polar site, the hydroxyl on the number 3 carbon. When it is esterified to an acyl group, usually oleic acid, the cholesteryl ester is extremely hydrophobic. The free hydroxyl enables the cholesterol molecule to orient in membranes, a major function of

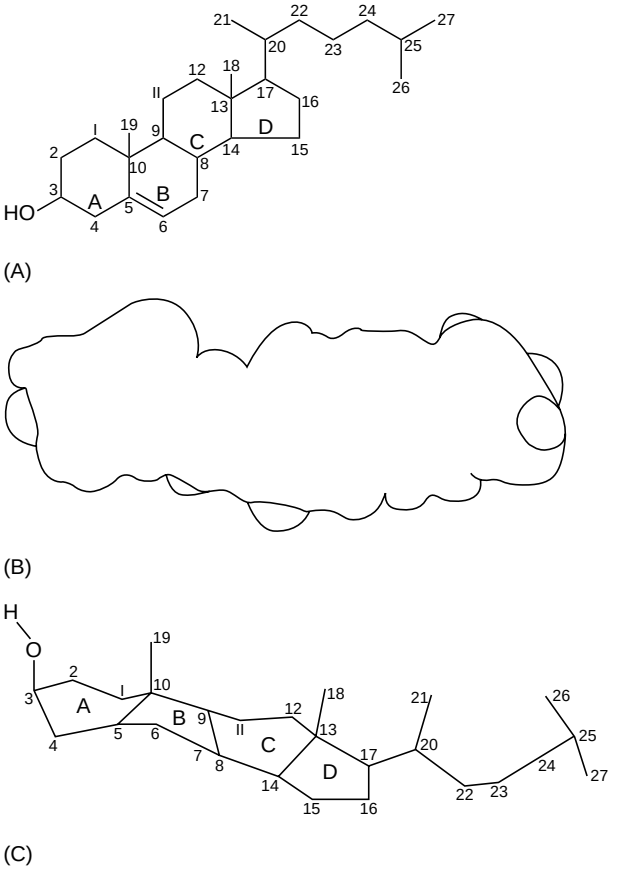


Figure 9 Cholesterol structural models: (A) conventional, (B) space-filling, and (C) conformational.

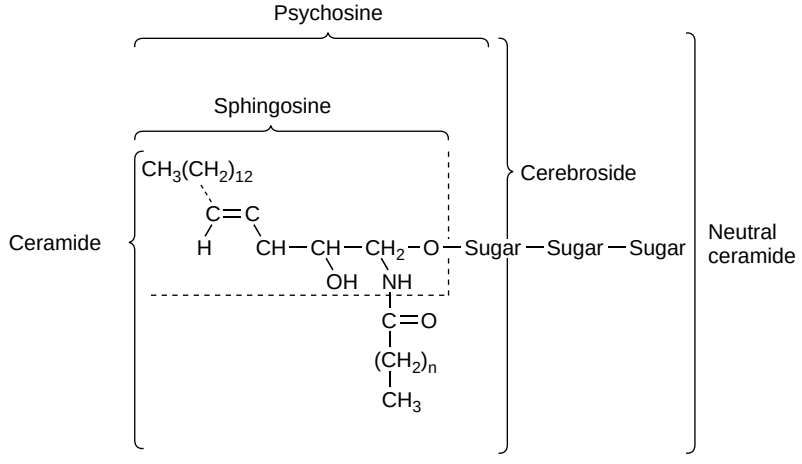


Figure 8 Structure of a sphingolipid.

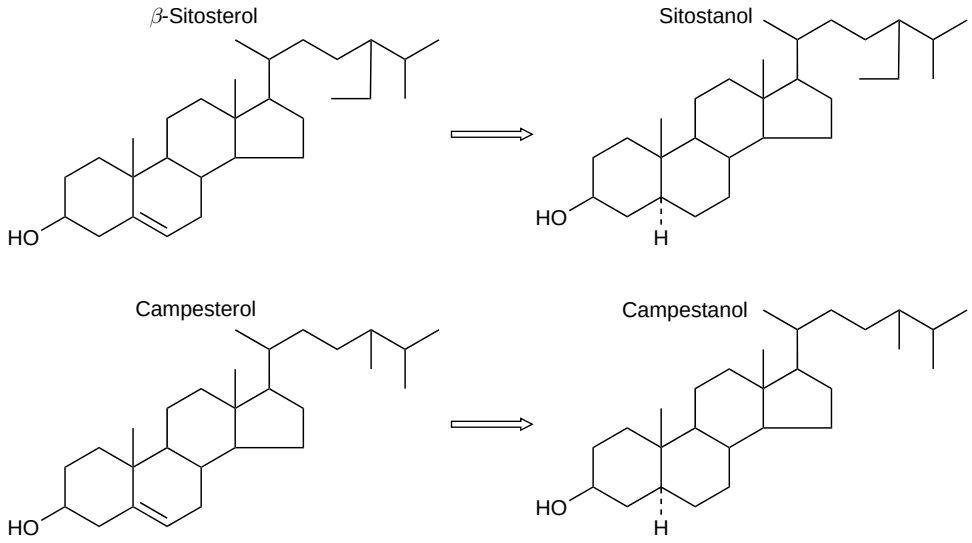
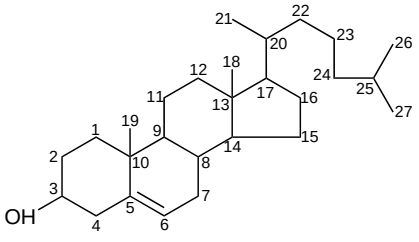
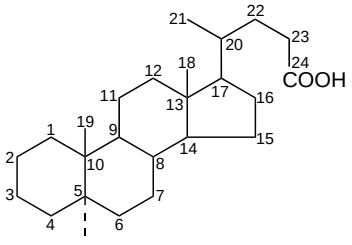


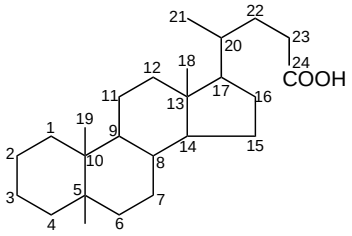
Figure 10 Plant sterols.



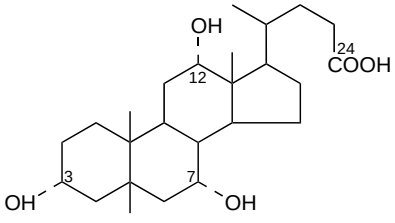
Cholesterol



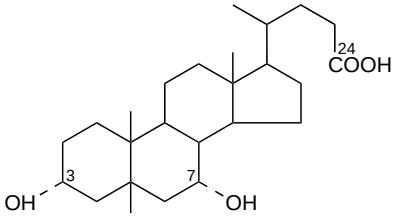
5α-Cholanic acid



5β-Cholanic acid



Cholic acid
3α,7α,12α-Trihydroxy-5β-Cholanic acid



Chenodeoxycholic acid
3α,7α-Dihydroxy-5β-Cholanic acid

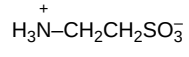
Figure 11 Bile acid synthesis from cholesterol.

cholesterol. In brain and other nervous tissues, free cholesterol is the major component of myelin and renders the myelin sheath impermeable to electron transfer (dielectric).

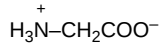
Phytosterol and campesterol are the major sterols in plant tissues (**Figure 10**). The plant sterols and their stanol derivatives (saturated at the 5–6 carbons) along with cholestanol are active in regulating cholesterol absorption. All these sterols are consumed in the diet, and some are being added to foods as positive adjuncts to regulation of cholesterol metabolism.

Bile Acids

The most abundant derivatives of cholesterol are bile acids. Cholesterol is important in metabolism and is biosynthesized as well as consumed in the diet. It is the precursor for vitamin D (cholecalciferol) and for the adrenocortical hormones, such as estrogen, androgens, and progesterone. These compounds require very small amounts of the cholesterol precursor. The sterol nucleus cannot be broken by mammalian enzymes, except for the formation of cholecalciferol. Bile acids constitute approximately 50% of the excretion products of cholesterol metabolism and perform essential functions in digestion and absorption of dietary lipids. They are synthesized in the liver (**Figure 11**) and exist in metabolism as conjugates with taurine and glycine (**Figure 12**). As with other lipids in metabolism, the amphipathic properties of the compounds characterize their functions. The planar sterol moiety with an acid group at the 24 carbon is capable of separating water and lipid interfaces and is important in facilitating interaction between lipids and enzymes in digestion. The even greater contrast of polar and neutral within the same molecule is exemplified by the conjugated bile acids. They are also active in digestion and exist in all tissues, so they may have additional functions as amphipathic facilitators between enzymes and lipids.



Taurine



Glycine

Figure 12 Structures of taurine and glycine.

See also: **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Fats and Oils**. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Fertility**. **Hyperlipidemia**: Overview; Nutritional Management. **Lipids**: Composition and Role of Phospholipids.

Further Reading

Dupont J (1990) Lipids. In: *Present Knowledge of Nutrition*, pp. 56–66. New York: International Life Sciences Institute/Nutrition Foundation.

Dupont J, White PJ, and Feldman EB (1991) Saturated and hydrogenated fats in food in relation to health. *Journal of the American College of Nutrition* **10**: 577–592.

Gropper SS, Smith JL, and Groff JL (2005) *Advanced Nutrition and Human Metabolism*, 4th edn., pp. 128–171. Belmont, CA: Thomson Wadsworth.

Gurr ML, Harwood JL, and Frayn KN (2002) *Lipid Biochemistry*, 5th edn. Oxford: Blackwell Science.

Composition and Role of Phospholipids

A D Postle, University of Southampton, Southampton, UK

Phospholipids are amphipathic (amphipathic describes molecules with regions that are both water seeking (hydrophillic) and water repellent (hydrophobic). This is the fundamental physical property that drives the formation of biological membranes) lipids consisting of hydrophobic and hydrophilic regions. This amphipathic nature, which enables phospholipid molecules to assemble into bilayer and hexagonal membrane structures, is critically important for the functional viability of all eukaryote cells. Cellular membranes, composed primarily of phospholipid, separate the intracellular milieu from the extracellular environment and facilitate the formation of specialised intracellular organelles. For many years, phospholipids were considered to be important but relatively inert structural components of the cell. Recently, the central role of membrane phospholipid composition and turnover in the regulation of a wide range of cellular functions has become widely recognized. For instance, all membrane receptor events take place

within a phospholipid-rich environment, and it is therefore not surprising that cells have adopted hydrolysis of membrane phospholipids as a major signaling mechanism. Phospholipids have multiple roles, including the following:

1. They provide a structural framework to maintain cellular integrity and to compartmentalize diverse events within the cell.
2. They provide the appropriate physicochemical environment to optimize the activities of membrane-associated receptors, enzymes, and proteins.
3. They act as substrate molecules for a variety of phospholipase enzymes involved in signaling mechanisms.
4. They provide sites for binding of proteins involved in cellular signaling processes.
5. They exert a physicochemical detergent-like action to facilitate the physiological function of a variety of tissues, including the lungs, stomach, and synovial surfaces.
6. They regulate the synthesis and secretion of lipoproteins from the liver.

Phospholipid Structures

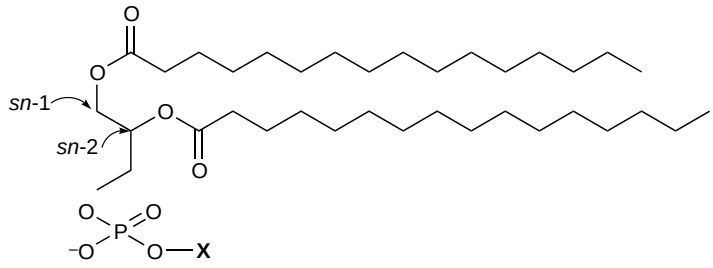
There are two major classes of phospholipid, depending on whether they contain a glycerol or sphingosyl backbone. Glycerophospholipids are molecules based on phosphatidic acid (3-*sn*-phosphatidic acid); the nature of the esterified group X defines the class of phospholipid (**Figure 1**). The most common substituent groups include nitrogenous bases, such as choline and ethanolamine, and polyalcohols, such as myoinositol and glycerol. Sphingophospholipids contain sphingosine (*trans*-D-erythro-1,3-dihyroxy 2-amino-4-octadecene). Sphingomyelin is the most abundant sphingophospholipid class, and it is the phosphorylcholine ester of *N*-acylsphingosine, also called ceramide. Sphingophospholipids are important components of all cell membranes and are structurally and metabolically closely related to glycosphingolipids such as glycosylceramides, gangliosides, and cerebrosides. Sphingomyelin is recognized as a major substrate for sphingomyelinase enzymes involved in generating intracellular ceramide and sphingosine, which are intimately involved in the regulation of programmed cell death (apoptosis). Sphingophospholipids contain principally saturated and monounsaturated fatty acids; little information is available on the nutritional effects on sphingophospholipid composition, and sphingomyelin metabolism has been recently reviewed.

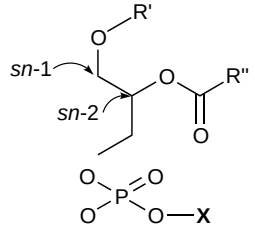
The distribution of phospholipids is heterogeneous within any cell both between different subcellular membranes and within individual membranes. For instance, mammalian cells maintain an enriched distribution of neutral lipids, such as phosphatidylcholine (PC) and sphingomyelin, in the outer leaflet of the plasma membrane and hence present an uncharged surface to the exterior of the cell (**Figure 2**). It is critically important to restrict the distribution of uncharged phospholipids to the interior of the cell because increased concentration of phosphatidylserine (PS) in the outer leaflet of the plasma membrane is the initial signal for both programmed cell death (apoptosis) and the clotting cascade.

Classification and Nomenclature of Glycerophospholipids

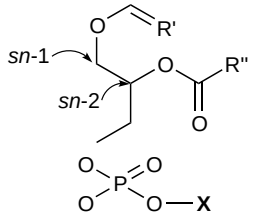
Glycerophospholipid classes are commonly referred to as phosphatidylcholine, phosphatidylethanolamine, etc. They are composed of a spectrum of molecular species (phospholipid molecular species are the individual different molecules within any different class of phospholipid determined by the combination of fatty acids esterified to the glycerol backbone. Any given mammalian cell contains up to 1000 individual phospholipid molecular species) defined by the substituent fatty acid groups attached to the *sn*-1 and *sn*-2 positions of the glycerol backbone. For example, the individual molecular species palmitoyloleoyl phosphatidylcholine can be named formally as either glycerol 1-hexadecanoate 2-9-octadecaenoate 3-phosphocholine or 1-hexadecanoyl-2-octadeca-9-enoyl-3-glycerophosphocholine. One shorthand designation for this molecule, adopted in this article, is PC16:0/18:1, where PC designates the phospholipid class, in this case phosphatidylcholine, and 16:0 and 18:1 (fatty acid nomenclature is based on total number of carbon atoms in the acyl chain, followed by total number of double bonds. For instance, 16:0 is saturated 16-carbon palmitic acid, whereas 20:4 is poplyunsaturated 20-carbon arachidonic acid) designate the fatty acids esterified at the *sn*-1 and *sn*-2 positions.

For phospholipids from cell membranes, saturated fatty acids are generally located at the *sn*-1 and unsaturated fatty acids at the *sn*-2 position, with notable exceptions. For instance, dipalmitoyl PC (PC16:0/16:0) is a major component of lung and surfactant PC, whereas significant amounts of didocosahexaenoyl phosphatidylethanolamine (PE22:6/22:6) are present in retinal PE. In addition, PC species with 18:1n-9 at the *sn*-1 position are minor components of many cells.





sn-1-alkyl-*sn*-2acyl species



sn-1alkenyl-*sn*-2-acyl species

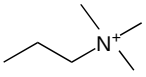
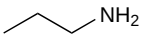
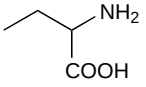
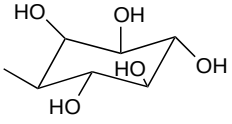
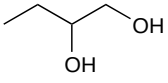
Phospholipid	Headgroup(X)
Phosphatidylcholine (PC)	
Phosphatidylethanolamine (PE)	
Phosphatidylserine (PS)	
Phosphatidylinositol (PI)	
Phosphatidylglycerol (PG)	
Phosphatidic acid (PA)	

Figure 1 Molecular structures of phospholipids. The class of phospholipid is defined by the nature of the nitrogenous base or polyol esterified to the phosphate group (X). The species distribution within any phospholipid class is determined by the fatty acyl substituents at the *sn*-1 and *sn*-2 positions of the glycerol backbone. The dipalmitoyl species would be designated PC16:0/16:0 if X was choline. If arachidonic acid was esterified at *sn*-2, then the molecule would be designated as PC1 6:0/20:4. In the diacyl species, fatty acids are attached by ester linkages. For *sn*-1-alkyl-*sn*-2-acyl species, the *sn*-1 fatty acid is attached by an ether bond. For *sn*-1-alkenyl-*sn*-2-acyl species, the *sn*-1 fatty acid is attached by a vinyl ether linkage.

In addition to diacyl species, with both fatty acids attached by ester bonds, there are a number of species with ether-linked fatty acids, principally in the *sn*-1 position. These ether phospholipids include 1-alkyl-2-acyl species, largely present in PC, and 1-alk-1-enyl species (plasmalogens), largely present in PE. These ether lipids are major components of many cell membranes, particularly neuronal and inflammatory cells, and there have been significant advances in understanding the biochemical pathways for their synthesis and catabolism. Some alkyl acyl PC species are substrates for the generation of the potent bioactive lipid platelet-activating factor (1-alkyl-2-acetyl-glycero-3-phosphocholine (PAF)), but the function of most ether lipids is largely unclear. One possibility is that generation of 1-alkyl-2-acyl-glycerol as a second messenger rather than diacylglycerol may contribute to differential regulation of protein kinase C isoforms, and antioxidant properties have been reported for plasmalogens.

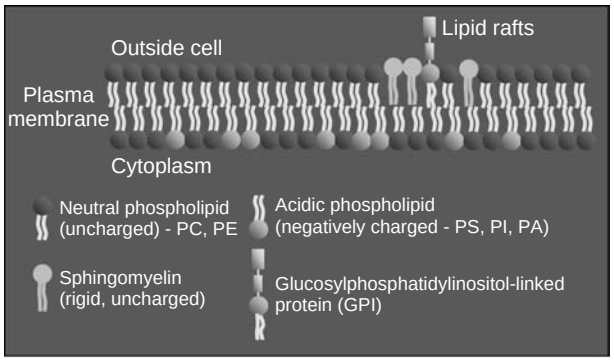


Figure 2 Topology of distribution of phospholipids within the plasma membrane of eukaryotic cells. The outer leaflet of the membrane bilayer is enriched in neutral PC and rigid components, such as sphingolipids and glucosylphosphatdylinositol-linked proteins. The distribution of charged acidic phospholipids, such as PS and PI, to the inner leaflet of the membrane is actively regulated by a combination of enzymes called flipases and scramblases.

Phospholipid Analysis

Historically, phospholipid compositions have been determined by thin-layer chromatography of different classes followed by gas chromatography of fatty acids. Such traditional analysis provides no direct information about the individual molecular species compositions of phospholipids, which are the functional, biologically relevant molecules. For instance, a fatty acid analysis of a phospholipid mixture as 50% 16:0, 50% 18:1 could represent either 16:0/18:1 or an equivalent combination of 16:0/16:0 and 18:1/18:1, which all have very different physical and functional properties. A variety of techniques have been established to provide such information, including high-performance liquid chromatography (HPLC), nuclear magnetic resonance, and mass spectrometry. In this article, compositional data are provided in terms of individual molecular species largely determined by sensitive electrospray ionization mass spectrometry (ESI-MS) methodologies. (Electrospray ionization mass spectrometry is a soft ionization technique that resolves intact molecular ions with minimal fragment formation. Best known for proteomic analysis, when applied to lipid analysis ESI-MS provides direct, very sensitive analysis of molecular species composition with a high degree of resolution.)

Phospholipid Composition

The glycerophospholipid composition of most cell types in the body is regulated within relatively restricted limits and is often specialized for the function of the cell involved. Moreover, most cells maintain distinct and different compositions of the various phospholipid classes. For example, ESI-MS analysis of total lipid extracts from a variety of mouse tissues shows that PC exhibits a wide variation of composition but that of phosphatidylinositol (PI) is relatively constant (**Figure 3**). PC species range from predominantly polyunsaturated in liver (**Figure 3A**) to disaturated and monounsaturated in lungs (**Figure 3B**) and brain (**Figure 3C**). Liver PC contains substantial amounts of species with either n-6 fatty acids (20:4n-6) or n-3 fatty acids (22:6n-3), whereas PC from pancreas, for instance, is essentially composed solely of species containing n-6 fatty acids (**Figure 3D**). This inherent variation in PC composition is emphasised by that of spleen (**Figure 3E**), which, like lung, is dominated by PC16:0/16:0 but also contains increased concentrations of monounsaturated and polyunsaturated species. In contrast, PI from all tissues measured was dominated by the single polyunsaturated stearoyl arachidonoyl species (PI18:0/20:4) (**Figures 3F–3J**). This diversity of composition is mirrored for the distinctive and different compositions of all the other phospholipid classes in most cell types and emphasizes the highly specific mechanisms that regulate phospholipid compositions *in vivo*. PE is typically enriched in arachidonoyl-containing species, whereas PS is generally dominated by the monounsaturated species PS18:0/18:1. It is important to recognize, for nutritional studies *in vitro*, that many of these tissue-specific distributions are lost or reduced for cells maintained in cell culture supplemented with fetal calf serum.

Phospholipid Synthesis

These compositions are mediated by interactions between phenotypic expression and cellular nutrition, which determine the specificities of the enzymes of phospholipid synthesis and hydrolysis and of the transfer proteins that exchange phospholipid species between different membranes. Regulation of synthesis is best characterized for the formation of PC in rat hepatocytes, where PC synthesis is essential for assembly and secretion of very low-density lipoprotein particles, and in the lung epithelial cells responsible for synthesis of pulmonary surfactant phospholipid. Phosphatidylcholine species synthesized *de novo* from diacylglycerol by the enzyme cholinephosphotransferase are subsequently modified by acyl remodeling mechanisms involving sequential actions of phospholipase and acyltransferase activities. The rate of PC synthesis is thought to be dependent on the activity of CTP:choline phosphate cytidylyltransferase (CCT), which is subject to complex regulatory mechanisms involving phosphorylation and reversible enzyme

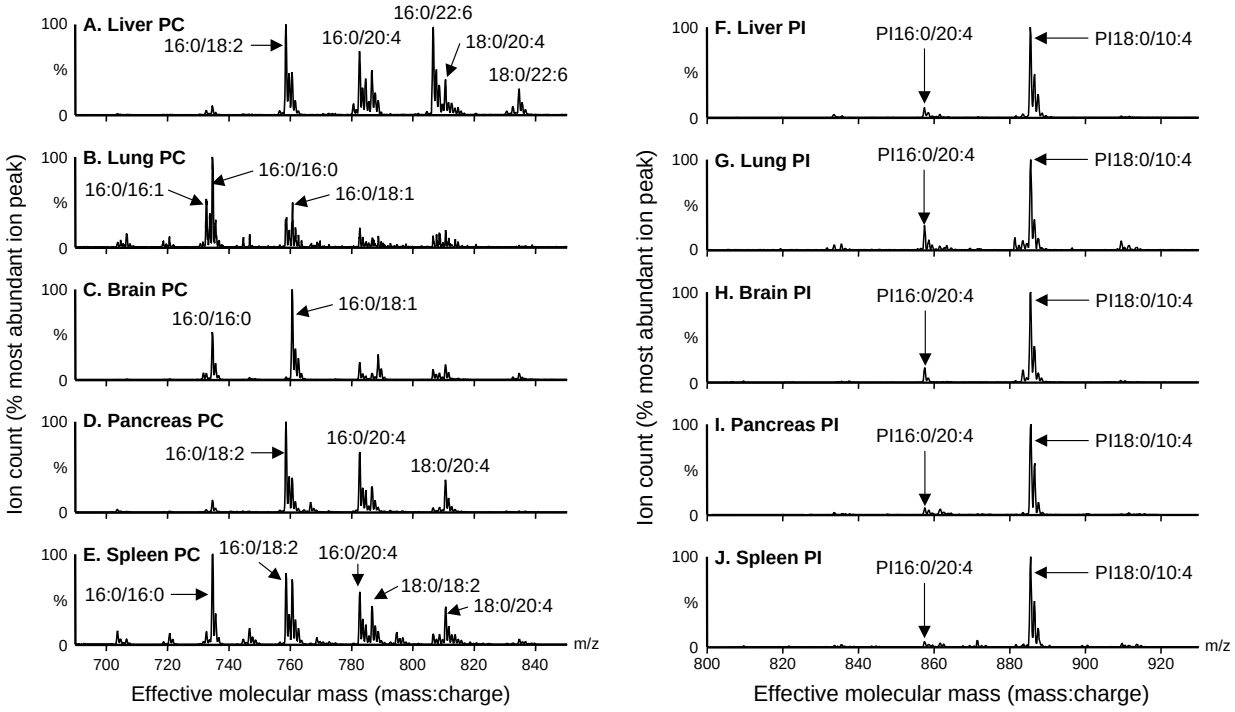


Figure 3 Electrospray ionization mass spectrometry analysis of phospholipid compositions of selected mouse tissue. Total lipids were extracted from liver, lungs, pancreas, and spleen using chloroform:methanol and then analyzed for PC (A–E) and PI (F–J) using diagnostic mass spectrometry scans. The distribution of the phospholipid molecular species in these illustrative spectra is given by the response of the individual ions, presented relative to the predominant ion on display. The identities of the major PC and PI species identified were confirmed by diagnostic fragmentation analysis by tandem MS/MS. (Dombrowsky H, Bernhard W, Rau G, Clark G and Postle A, unpublished results.)

translocation between cytosol and membrane fractions of the cell. In this context, CCT acts as a sensor for the physical structure of the endoplasmic reticulum membrane. Hydrolysis of PC alters the inherent curvature of the membrane and decreases its stored elastic energy, enabling CCT to bind and thus replenish membrane PC.

The spatial pathway of phospholipid synthesis is illustrated schematically in **Figure 4** for the type 2 epithelial cell of the lung alveolus, which synthesizes and secretes lung surfactant. Initial synthesis of phospholipids on the endoplasmic reticulum is followed by a complex series of events that include modification of esterified fatty acid groups by a process of acyl remodeling, selective transport between different intracellular membranes, and uptake of selected phospholipids into lamellar bodies. These lamellar bodies are intracellular stores of surfactant that, when secreted in response to cell stretch, are actively secreted into the alveolar space where they adsorb to the air–liquid interface, oppose surface tension forces within the lungs, and prevent alveolar collapse. In addition, inactive surfactant is recycled by type 2 cells into endosomes that fuse into multivesicular bodies and subsequently into lamellar bodies. Although the phospholipid metabolism of the type 2 cell is complex compared to that of most cell types, it demonstrates very well the various stages in phospholipid synthesis, transport, and metabolism with potential for modification of molecular compositions.

A limited number of conditions are known in which alterations to the processes of phospholipid synthesis and metabolism have profound effects on health and survival. In human subjects, the inability to synthesise the major phospholipids, such as PC and PE, is incompatible with life, so most genetic abnormalities have been identified in abundances of more minor phospholipids. For instance, Barth's syndrome is an X-linked recessive disorder characterised by childhood onset of cardiomyopathy, neutropenia, and abnormal mitochondrial structure and function. The gene affected is the *tafazzin* gene, responsible for acyl remodeling in cardiolipin synthesis. Cardiolipin is a minor phospholipid enriched within the heart and in mitochondria that contains four fatty acid and two phosphate moieties and is synthesised on the endoplasmic reticulum predominately with four oleoyl ($C_{18:1}$) chains. Patients with Barth's syndrome are unable to convert this tetraoleoyl form into the more functional tetralinoleoyl ($C_{18:2}$) cardiolipin species. This is the only condition

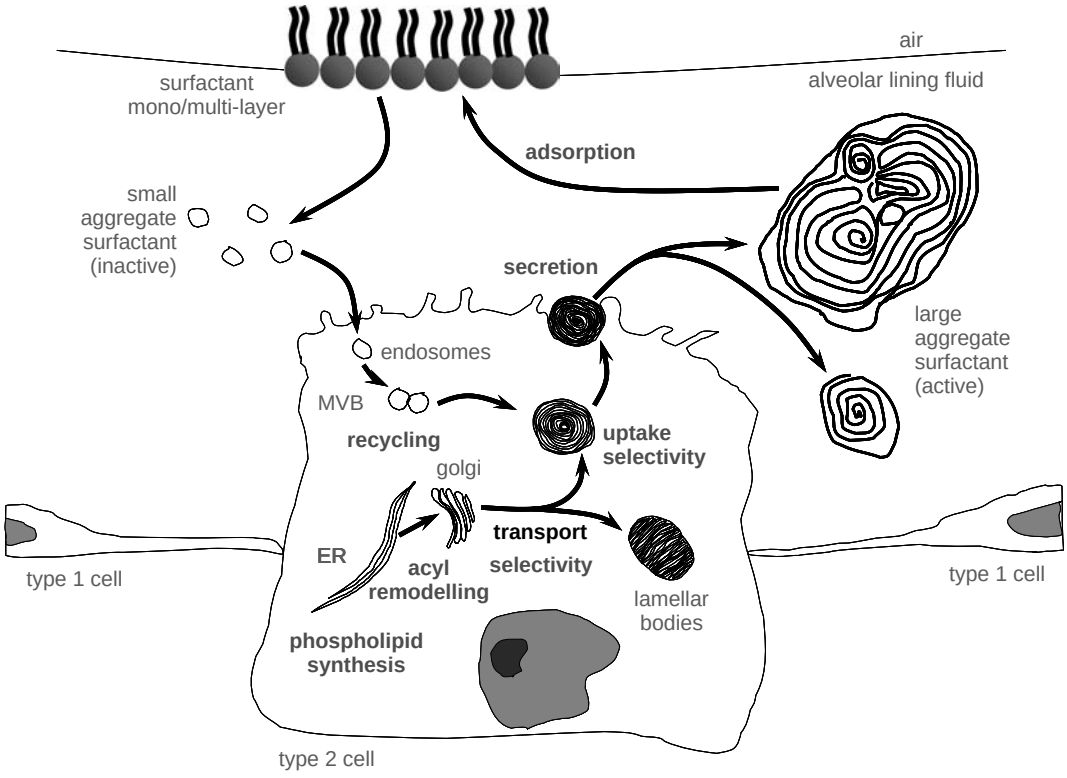


Figure 4 Synthesis and secretion of lung surfactant phospholipid by the type 2 epithelial cell of the lung alveolus. Phospholipid synthesized in the endoplasmic reticulum (ER) is routed through the Golgi apparatus for uptake and packaging into intracellular storage vesicles called lamellar bodies. In response to cell stretch, lamellar bodies fuse with the plasma membrane and secrete their contents into the alveolar space. After processes of adsorption and desorption from the air–liquid interface, inactive surfactant is recycled into lamellar bodies via endosomes and multivesicular bodies (MVB). Metabolically active type 2 cells occupy only approximately 5% of the surface area of the alveolus, with the thin type 1 cells responsible for gas exchange contributing the other 95%.

identified in which the inability to synthesize a precise composition of an individual phospholipid class is apparently responsible for clinical symptoms.

In addition to modification of synthetic mechanisms, alterations to transport and uptake processes can result in severe disease and mortality. ABCA3 is a membrane protein member of the ATP-binding cassette (ABC) family of proteins, which includes the multidrug resistance protein and the ABCA1 protein responsible for reverse cholesterol transport. ABCA3 is though to be involved in the selective uptake and processing of phospholipids destined for lung surfactant assembly within lamellar bodies. Mutations in the *ABCA3* gene cause fatal surfactant deficiency in newborn infants and have been recognized as major contributors to lung disease and respiratory failure in infants delivered full term.

Phospholipid Breakdown

There is also a considerable diversity of specificity of phospholipase enzymes responsible for phospholipid hydrolysis, in terms of both positional and molecular species selectivity. Phospholipase A activity in rat liver will act selectively to remove *sn*-1 16:0 from PC species containing *sn*-2 18:2, whereas cytosolic phospholipase A_2 (PLA_2) is specific for species containing *sn*-2 20:4n-6. In contrast, secretory PLA_2 must be bound to negatively charged phospholipids for activation, but it is not acyl specific. Mammalian phospholipase Cs (PLCs) act preferentially on phosphatidylinositol-4,5-bisphosphate rather than on PI or PC, whereas agonist-stimulated phospholipase Ds (PLDs) are selective for PC species. However, although the distribution of phospholipases is tissue specific, the contribution of their activities to the regulation of phospholipid compositions in most tissues has not been well defined.

Phospholipid Composition in Development

The most extensive changes to phospholipid composition occur during fetal and neonatal development and have been best characterized for PC in liver,

lung, and brain. These changes illustrate clearly the limitations of dietary manipulation on phospholipid composition. During human pregnancy, the polyunsaturated fatty acids (PUFAs) 20:4n-6 and 22:6n-3 are supplied across the placenta from maternal to fetal circulations in increasing quantities toward term. At birth, the onset of milk feeding is characterized by increased intake of the PUFA precursor 18:2n-6. This sequence of nutritional supply is reflected in fetal and neonatal liver PC composition. Immature fetal human liver contains a high proportion of monounsaturated PC species, particularly PC16:0/18:1, and tends to become enriched with species containing 20:4n-6 and 22:6n-3 toward term (**Table 1**).

Postnatally, the content of 18:2n-6 species increases, and fetal and neonatal plasma PC composition directly mirrors this changing pattern. However, these alterations in development are regulated primarily by metabolic and hormonal rather than by nutritional considerations. The increased supply of PUFA from mother to fetus in later gestation is independent of any change in maternal dietary lipid intake and instead is a consequence of hormonal effects on the specificity of PC synthesis and lipoprotein export by the maternal liver. Similarly, although switching from placental to enteral feeding is the major factor causing the dramatic changes to plasma PC at birth, this composition is still dependent on the metabolic regulation of the specificity of hepatic PC synthesis. The programmed nature of this regulation is shown clearly by food restriction in newborn guinea pig pups, which still display equivalent postnatal alterations to plasma and liver PC composition as their fed litter mates, even in the total absence of enteral nutrition.

Table 1 Phosphatidylcholine molecular species composition of human liver during fetal and postnatal development[a]

Molecular species	*Liver phosphatidylcholine concentration (nmol/pg wet weight)*		
	Fetal (15 weeks of gestation, n = 4)	*Stillborn (term, n = 4)*	*Infant (43–64 weeks old, n = 6)*
16:0/16:0	992 ± 156	1004 ± 81	538 ± 121
16:0/18:1	2007 ± 250	2240 ± 173	2353 ± 496
16:0/18:2	466 ± 52	1259 ± 139	2202 ± 273
16:0/20:4	1402 ± 98	1784 ± 38	1062 ± 219
16:0/22:6	431 ± 110	953 ± 82	614 ± 512
18:0/18:2	308 ± 56	443 ± 68	1239 ± 252
18:0/20:4	1298 ± 288	953 ± 89	448 ± 403
18:0/22:6	115 ± 31	210 ± 50	221 ± 267

[a]Molecular species were analyzed by reverse-phase HPLC and quantified by postcolumn fluorescence detection with 1,6-diphenyl-1,3,5-hexatriene. Concentrations expressed as mean ± SD.

In contrast, immature fetal lung PC also contains a high concentration of PC16:0/18:1 but becomes more, rather than less, saturated with progression of gestation due to increased synthesis of the disaturated species PC16:0/16:0 and PC16:0/14:0. PC16:0/16:0 is a major component of pulmonary surfactant that acts to oppose surface tension forces in the lungs and prevent alveolar collapse. Infants who are born preterm with immature surfactant are at high risk of death and disability caused by neonatal respiratory distress syndrome. In contrast to fetal liver, the phospholipid composition of fetal lung is only marginally affected by the changes to lipid nutrition in utero. Nevertheless, some nutritional influence is evident, even though PUFA-containing species are minor components of lung PC. Comparison of PC compositions in prenatal human lung shows a postnatal increase in the content of PC16:0/18:2, which reflects the increased dietary supply of 18:2n-6. The situation in developing lung reflects that of most other tissues in the body, in which dietary lipid modulation causes relatively modest changes to the specificity of phospholipid compositions. Such subtle alterations to membrane composition, however, can exert profound effects on cellular function.

Finally, adult brain PE contains approximately 50% of 22:6n-3 species, enriched in neuronal synapses and possibly involved in synaptic transmission. Failure to acquire sufficient 22:6n-3 in brain PE during neuronal differentiation in early development can lead to permanent suboptimal neurological function. Many of the changes to maternal lipid metabolism in pregnancy represent adaptations to ensure adequate supply of PUFA to the developing fetal brain. Increased synthesis and secretion of PC16:0/22:6 in livers of pregnant rats and guinea pigs correlates with the period in fetal brain growth of maximal accumulation rate of 22:6n-3 into brain PE. Once incorporated into brain or retinal PE, 22:6n-3 is retained throughout life, even in periods of prolonged nutritional deprivation. Infants who are born preterm and with inadequate reserves of 22:6n-3 are recognized to be in danger of nutritional deficiency if fed milk formula lacking preformed long-chain PUFA. For instance, 22:6n-3 content of brain PE was decreased in infants fed such formula and who had died suddenly from sudden infant death syndrome. For this reason, supplementation of preterm infant milk formula with preformed PUFA has been recommended by the European Society for Pediatric Gastroenterology and Nutrition.

Phospholipid Composition in Adult Tissues

Information about the detailed molecular species compositions of phospholipids from adult human tissues is surprisingly haphazard. There have been many isolated reports of extensive characterizations of selected phospholipid classes in individual tissues, but such studies have generally measured compositions of bulk preparations from relatively large tissue samples. Very few clinical or nutritional studies have characterized phospholipid compositions in molecular terms. In reality, each cell type contains in excess of 1000 glycerophospholipid species, with differential compositions between different membranes in the same cell and even between different regions in the same membrane. Such regions of microheterogeneity may occur either because of the physical properties of the lipids themselves (e.g., forming hexagonal rather than bilayer structures) or because of sequestration by membrane proteins. Phase transitions within the membrane can also exert significant effects, and interactions of cytoskeletal components of the cell have been described with relative solid gel-phase phospholipids in the plasma membrane. One additional important factor is the transmembrane phospholipid distribution between the two leaflets of the cell bilayer. For practically all cell types, PC is relatively more concentrated in the outer leaflet, whereas PE is located primarily in the inner (cytoplasmic) leaflet. Importantly, PS is almost totally restricted to the side of the plasma membrane facing the cytoplasm, where it acts as an activator of protein kinase C. Redistribution of PS to the outer leaflet of the plasma membrane is a signal of cell senescence and is a potent activator of the clotting cascade. Finally, there has been considerable interest in the concept of lipid rafts, subfractions of membranes that are resistant to extraction with detergent and have been extensively implicated in transmembrane signaling particularly in immune cells. The compositional aspects of many such studies must be interpreted with caution; recent analyses have indicated that detergent solubility is more an intrinsic property of individual lipids than a property dependent on membrane organization.

Examples of recent ESI-MS analyses of phospholipid molecular species compositions of a variety of human tissues are summarized in **Figure 5**, which compares ESI-MS spectra of PC from human blood lymphocytes, monocytes, and neutrophils. As for most hematopoeic cell types and in contrast to the mouse compositions shown in **Figure 2**, the PC composition of these cells is dominated by monounsaturated PC

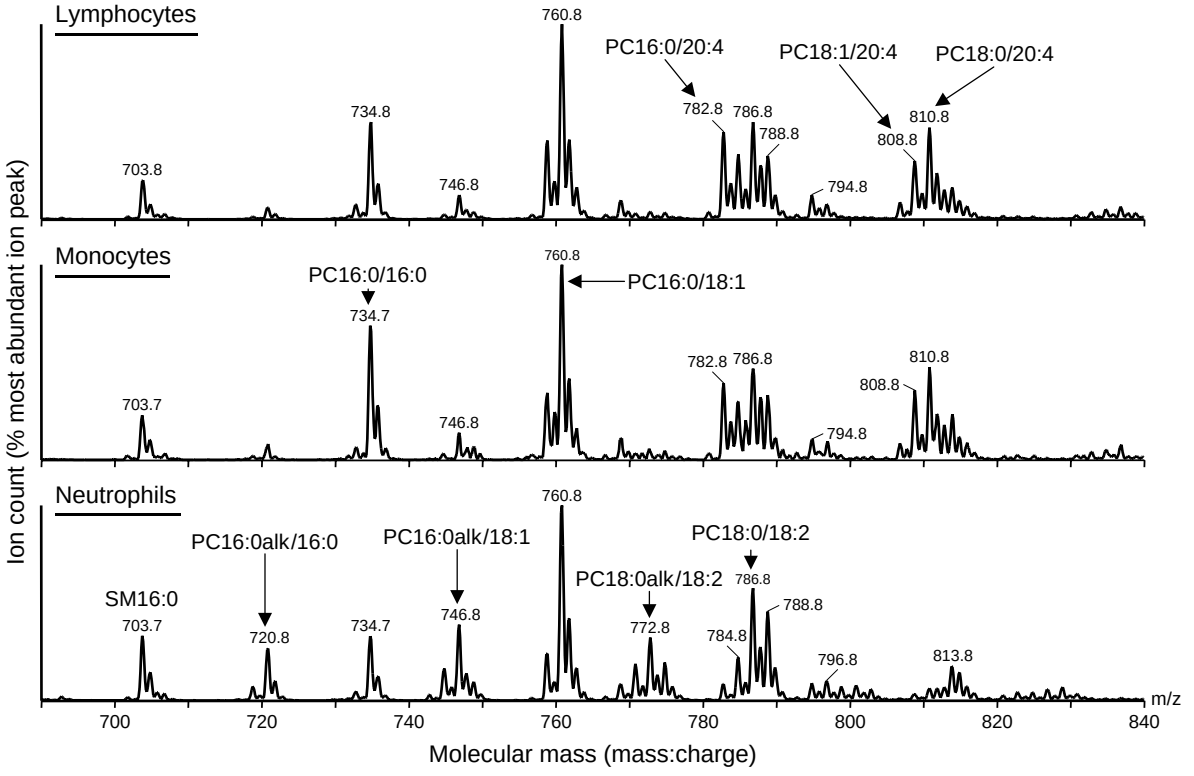


Figure 5 Electrospray ionization mass spectrometry analysis of PC compositions of human blood leukocytes. Total lipids were extracted from lymphocytes, monocytes, and neutrophils and analyzed as described in the legend for **Figure 2**. (Madden J, Wright S, Clark G and Postle A, unpublished observations.)

species, especially PC16:0/18:1 (m/z = 760), but the distribution of polyunsaturated species is considerably variable between cell types. Both lymphocytes and monocytes are relatively enriched in species containing 20:4n-6 (PC16:0/20:4 m/z = 782; PC18:0/20:4 m/z = 810), with an increased content of PC16:0/16:0 (m/z = 734). In contrast, neutrophils are relatively depleted in both PC16:0/16:0 and arachidonoyl species, but they contain considerably higher proportions of *sn*-1-alkyl-*sn*-2-acyl species (PC16:01alk/16:0 m/z = 720; PC16:0alk/18:1 m/z = 746; PC18:0alk/18:2 m/z = 772). This comparison illustrates an important role for phenotypic expression as one contributor toward the specificity of cell PC composition.

Different phospholipid classes from the same tissue generally exhibit considerable variation in composition, shown in **Figure 2** for mouse tissues and also in the analysis of the white matter of human brain. Although brain PC was highly enriched in monounsaturated species, diacyl PE was enriched in species containing PUFA. The distribution of such species, however, was highly asymmetric, with 22:6n3 and 20:4n-6 species containing 16:0 at the *sn*-1 position being present in much lower abundance than the same species containing *sn*-1 18:0. In contrast, both alkenylacyl PE and PS were characterized by a predominance of monounsaturated species. However, whereas PC was enriched in PC16:0/18:1, alkenylacyl PE was enriched in PE18:1alk/18:1 and PS was enriched in PS18:0/18:1. This comparison illustrates the tight regulation of the composition of individual phospholipid classes and emphasizes potentially important differences in molecular compositions that could not be predicted from total fatty acid analysis.

Nutritional Effects on Phospholipid Molecular Species

Practically all nutritional studies of dietary lipid effects on cellular phospholipid compositions have reported fatty acid compositions, with no molecular information. Due to significant differences in the detailed regulation of their phospholipid composition and metabolism, nutritional data obtained from laboratory animals generally have only a restricted application to human nutrition. The data in **Figure 6** are from one study in which human volunteers were fed fish oil supplements for 4 weeks and the change in their erythrocyte PE molecular species composition was measured. Of interest, the extent of increase in species containing

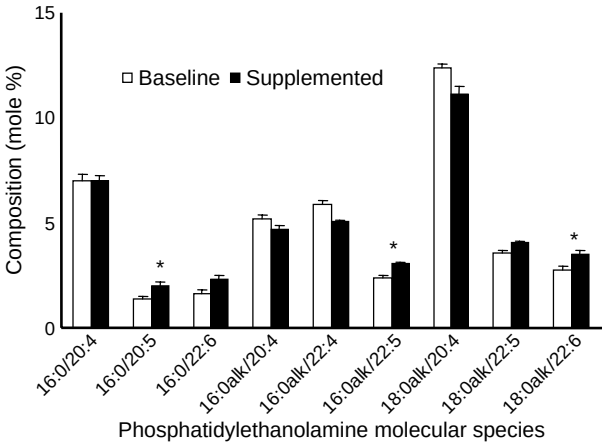


Figure 6 Dietary lipid and the composition of human erythrocyte phosphatidylethanolamine. Erythrocyte PE species were analyzed from six volunteers before and after consumption of fish oil containing 9 g eicosapentaenoic acid (22:5n-3) and 6 g docosahaexenoic acid (22:6n-3) per day for 4 weeks. Results are expressed as mean ± SEM; $*p < 0.05$. (From Knapp HR, Hullin F and Salem N Jr (1994) Asymmetric incorporation of dietary n-3 fatty acids into membrane aminophospholipids of human erythrocytes. *Journal of Lipid Research* **35**: 1283–1291.)

n-3 fatty acids was variable, and the extent of such changes was modest. This comparison illustrates a general observation that although manipulation of cultured cell phospholipid compositions by medium lipid supplementation is relatively easy, phospholipid compositions of similar cell types in vivo are considerably more resistant to dietary manipulation.

Functions of Phospholipids

Phospholipid composition is a significant factor in most cellular processes. This section, however, is restricted to selected examples of the role of molecular species composition with regard to physiological functions.

Membrane Structure

One frequently addressed role of phospholipids is to maintain an appropriate membrane structure for optimal cell function. The term 'membrane fluidity' is often used but is imprecise. It generally describes the combined effects of lateral and rotational movement of lipids within the plane of the membrane. Other concepts are perhaps more useful, such as the stress, termed 'stored elastic energy,' when phospholipids are prevented by their location within the membrane from assuming their lowest energy configuration. Typically, phospholipids such as PC prefer to adopt a convex membrane curvature, whereas molecules such as PE and PA will spontaneously adopt a concave configuration. In these paradigms, alterations to dietary lipid intake

may exert their modulatory effect on cell function by changing phospholipid molecular composition and hence altering these physiochemical properties. Although such effects are evident in model systems, extensive measurement by fluorescence polarization suggests that processes of homeoviscous adaptation restrict the extent of adaptations observed *in vivo*. For instance, increased incorporation of PUFA into membrane phospholipid, which would be expected to have a fluidizing effect, is invariably balanced by compensatory increases in the membrane content of cholesterol and more rigid phospholipid molecules.

Lung Surfactant

Maintenance of the essential composition of lung surfactant phospholipid is critical for the survival of all mammalian species. Lung surfactant is secreted from specialized type 2 epithelial cells in the lung alveolus and forms a continuous lining layer at the air–liquid interface throughout the lungs. To provide adequate gas exchange surface area in the lungs to support respiration, alveolar diameters must be very small, giving a large surface area:volume ratio. One consequence of the small dimensions of the alveolus is that surface tension forces contribute significantly to the dynamics of lung function. Surfactant opposes surface tension in the lungs. It is the absence of adequate surfactant that leads directly to lung collapse and the high incidence of morbidity and mortality associated with neonatal respiratory distress syndrome.

Lung surfactant has a unique phospholipid composition, containing PC16:0/16:0 as 40–60% of total PC and monounsaturated phosphatidylglycerol (PG) species as 10–15% of total phospholipid. Phosphatidylglycerol is not found at such high concentration in any other membrane of the body. PC16:0/16:0 is the principal surface-active component of lung surfactant, has a gel:liquid crystalline transition temperature of 41 °C, and consequently is, in effect, solid at a body temperature of 37 °C. It has been suggested that the compressed PC16:0/16:0 monolayer at the air–liquid interface survives the high surface pressures within the lungs by forming a solid monomolecular sheet, and it thus prevents any surface tension effects. At the same time, PC16:0/16:0 is metabolically inert, and one proposed specialized role for PG is to fluidise PC16:0/16:0 and facilitate its metabolic processing, secretion, and adsorption to the air–liquid interface.

This composition of lung surfactant is restricted to air-breathing animals. Comparative studies with reptiles, amphibia, and lower vertebrates have shown that concentration of PC16:0116:0 in surfactant correlates with the ratio of lung:body surface area as a measure of an animal's reliance on lung-mediated respiration. Lung surfactant from amphibia, by comparison, also contains phospholipid, but this is largely cholesterol and unsaturated PC, which is thought to serve an antiglue function. By analogy with lung surfactant, phospholipid-rich surfactants have been described for other epithelial surfaces, including the stomach, eustachian tube, and synovial surfaces, where they are thought to create a protective hydrophobic lining layer. The comparison with lung surfactant is somewhat misleading, however, because the PC fraction of these other epithelial secretions contains minimal PC16:0/16:0 and high contents of mono- and diunsaturated species.

Signal Transduction

Phospholipids are substrate molecules for a wide range of lipid-derived signaling molecules, including diacylglycerol (DAG), phosphatidic acid (PA), 20:4n6, eicosanoid products, PAF, and lysophosphatidic acid, generated by the action of PLA_2, PLC, and PLD. The activation of these enzymes is complex, partly because of the large number of isoforms present within a cell and also because of the interdependence and coordination of their regulation. For instance, the bacterial peptide formyl-methionyl-leucyl phenylalanine (FMLP) binds to its receptor on neutrophils and activates the G-protein-regulated PLCβ. PLCβ hydrolyses PI-4,5-P_2 to form DAG, an activator of traditional Ca^{2+}-dependent PKC isoforms, and inositol trisphosphate, which stimulates intracellular Ca^{2+} mobilisation. In addition, FMLP activates PC-specific PLD and cytoplasmic PLA_2. PLD generates PA, which also has signaling responses, including stimulation of NADPH oxidase activity, but which is also readily interconverted with DAG. Alkenyl species of PE are probably the major substrates for cytoplasmic PLA_2, which is specific for molecular species containing 20:4n-6. This multitude of responses to a single agonist is highly coordinated and is typical of lipid signaling mechanisms in general. The activation of the various phospholipase enzymes is tightly regulated by a variety of protein kinases, phosphatases, and regulatory proteins, such that their responses are sequential rather than simultaneous.

Evidence suggests that phospholipid structure contributes to the coordinated regulation of phospholipase activation. PI-4,5-P_2, the substrate for PLC, is an obligate activator of ADP ribosylation factor-dependent PLD; consequently PI-4,5-P_2 must be regenerated after the transient activation of PLC,

before maximal activation of PLD can be achieved. In addition, individual phospholipids can act as binding sites for a wide range of signaling proteins and enzymes, enabling their coordinated regulation at the membrane. Perhaps the best characterised of these systems is the generation of trace amounts of 3-phosphorylated PI, typically PI-3,4,5-P_3, when PI-3-kinase is activated by insulin and growth factors. Signaling proteins containing appropriate binding motifs (plecstrin homology or PH domains) then bind to PI-3,4,5-P_3 and initiate signaling cascades. The prototype of such protein is protein kinase B (PKB) also know as Akt. PKB undergoes a conformational change when bound to PI-3,4,5-P_3, becomes phosphorylated, and then is active in the regulation of cell proliferation.

The mechanisms of action of dietary lipid modulation on these signaling pathways are largely unknown. There is good evidence that eating a diet rich in fish oil (containing 22:6n-3 and 22:5n-3) attenuates neutrophil-mediated inflammatory reactions. Part of this antiinflammatory nutritional effect may be to reduce the content of phospholipid species containing 20:4n-6, thus decreasing available substrate for synthesis of eicosanoid and leukotriene products derived from 20:4n-6. Alternatively, it may also result in part from the modulation of the spectrum of molecular species of DAG and PA generated by the various PLC and PLD enzymes. In this paradigm, altering the composition of substrate phospholipid will result in the formation of different DAG or PA species, which then have differential actions on target kinase enzymes. Because inositol phospholipids are generally composed of the 18:0120:4 species, activation of PLC1 will form DAG18:0/20:4, whereas hydrolysis of PC will generate predominately monounsaturated DAG species. It has been suggested, for instance, that individual isoforms of protein kinase C can be differentially regulated in response to different molecular species of DAG, thus providing a molecular basis for many nutritional effects on a wide range of cellular functions.

Despite extensive studies since the 1960s, remarkably little is understood about the fundamental reasons why cells expend considerable energy maintaining lineage-specific molecular species compositions of membrane phospholipids. Even for cell lines in culture, which can be grown successfully over many generations with grossly nonphysiological membrane phospholipid compositions, a degree of lineage specificity is maintained. The detailed metabolic processes that control membrane phospholipid composition are slowly being defined, and studies of the specificities and activities in intact cells of acyltransferase and phospholipid synthetic enzymes using gene transfection and sensitive analytical techniques such as ESI-MS will increase understanding of the fundamental mechanisms involved.

See also: **Brain and Nervous System**. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated. **Lipids**: Chemistry and Classification.

Further Reading

Gunstone FD, Harwood JL, and Padley FB (1994) *The Lipid Handbook*, 2nd edn. London: Chapman & Hall.

Han X and Gross RW (2003) Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: A bridge to lipidomics. *Journal of Lipid Research* **44**: 1071–1079.

Hazel JR and Williams EE (1990) The role of alterations in membrane lipid composition in enabling physiological adaptations of organisms to their physical environment. *Progress in Lipid Research* **29**: 167–227.

Lee AG (1991) Lipids and their effects on membrane proteins: Evidence against a role for fluidity. *Progress in Lipid Research* **30**: 323–348.

Neuringer ME, Anderson GJ, and Connor WE (1988) The essentiality of n-3 fatty acids for development and function of the retina and brain. *Annual Review of Nutrition* **8**: 517–541.

Shulenin S, Nogee LM, Annilo T *et al.* (2004) ABCA3 gene mutations in newborns with fatal surfactant deficiency. *New England Journal of Medicine* **350**: 1296–1303.

Spiegel S and Merrill AH Jr (1996) Sphingolipid metabolism and cell growth regulation. *FASEB Journal* **10**: 1388–1397.

Valianpor RF, Wanders RJA, Barth PG, Overmars H, and van Geenup AH (2002) Quantitative and compositional study of cardiolipin in platelets by electrospray ionisation mass spectrometry: Application for the identification of Barth syndrome patients. *Clinical Chemistry* **48**: 1390–1397.

LIPOPROTEINS

J M Ordovas, Tufts University, Boston, MA, USA

Cholesterol and triacylglycerol are transported in blood as lipoproteins. Lipoproteins are generally spherical particles, with a surface layer composed of phospholipid with the fatty acids oriented toward the core of the particle. Included in this phospholipid layer are specific proteins known as apolipoproteins and free cholesterol. The core of the lipoprotein particles is made up of cholesteryl ester and triacylglycerol molecules.

The classification of serum lipoproteins has evolved historically through several phases corresponding with the development of different laboratory methodologies, including electrophoretic, ultra-centrifugal, and immunological techniques. By using these techniques, lipoproteins can be classified based on their electrophoretic mobility, hydrated density, and protein content.

Classification of Lipoproteins

Classification of Serum Lipoproteins According to Their Electrophoretic Mobilities

With the development of techniques to separate proteins according to their electrophoretic behavior, it could be demonstrated that most of the lipid present in serum was associated with proteins migrating with α_1- and β-globulin mobilities. This resulted in the first classification of lipoproteins as α_1- and β-lipoproteins. The ratio of lipid to protein on the α_1-lipoproteins was approximately 1:1, whereas the β-lipoproteins had a greater relative content of lipids. Application of more advanced electrophoretic techniques resulted in further discrimination among the lipoprotein classes and for many years lipoproteins were classified as β-, pre-β-, and α-lipoproteins. Careful observation of the electrophoretic lipoprotein profiles in normals and subjects with familial lipoprotein disorders gave rise to the first classification of lipoprotein disorders by Fredrickson and colleagues. The equivalence between electrophoretic and ultracentrifugal separation is presented in **Table 1**.

Several electrophoretic supports have been used to separate plasma lipoproteins. These include paper, cellulose acetate, agarose, and polyacrylamide. Agarose gel electrophoresis remains the most commonly used for easy and rapid assessment of lipoprotein patterns in the clinical laboratory. This technique is especially useful for identifying the presence of a broad β band in the diagnosis of type III hyperlipidemia. Gradient agarose-polyacrylamide gel electrophoresis under nondenaturing conditions has been an essential tool to analyze low-density lipoprotein (LDL) and high-density lipoprotein (HDL) subclasses, providing a greater resolution than ultracentrifugation. LDL subfractions have been resolved by nondenaturing polyacrylamide gradient gel electrophoresis (2–16%) in up to seven LDL subclasses with densities ranging from 1.020 to 1.063 g ml^{-1} and diameters ranging from 22.0 to 28.5 nm. Usually a major subpopulation and several (one to four) minor LDL subpopulations are found in most subjects examined. A predominance of smaller, more dense LDL, versus larger, more buoyant LDL particles in plasma has been associated with increased coronary heart disease (CHD) risk. There is evidence supporting the genetic origin of the

Table 1 Classification of plasma lipoproteins

Lipoprotein	*Diameter (nm)*	*Density (g ml^{-1})*	*Electrophoretic mobility*	*Major lipids*	*Major apolipoproteins*
Chylomicrons	80–500	<0.95	Origin	Dietary triacylglycerols, cholesteryl esters	A-I, A-II, A-IV, B-48, C-I, C-II, C-III, E
Remnants	>30	<1.006	Origin	Dietary cholesteryl esters	B-48, E
VLDL	30–80	<1.006	pre-β	Endogenous triacylglycerols	B-100, C-I, C-II, C-III, E
IDL	25–35	1.006–1.019	pre-β and β	Cholesteryl esters, triacylglycerols	B-100, E
LDL	18–28	1.019–1.063	β	Cholesteryl esters	B-100
HDL_2	9–12	1.063–1.125	α	Cholesteryl esters, phospholipids	A-I, A-II
HDL_3	5–9	1.125–1.210	α	Cholesteryl esters, phospholipids	A-I, A-II

distribution of LDL subfractions; however, age, gender, and environmental factors strongly influence the penetrance. HDL subfractions have been resolved using a similar technique, with a polyacrylamide gradient ranging from 4 to 30%, into five subclasses (HDL_{3c}, HDL_{3b}, HDL_{3a}, HDL_{2a}, and HDL_{2b}). More recently 11–14 subclasses have been described, including β-migrating particles, using an improved electrophoresis technique. The clinical importance of these subfractions is still under investigation.

Classification of Serum Lipoproteins According to Their Ultracentrifugal Characteristics

The presence of lipids within the lipoprotein particles confers these macromolecular complexes with a lower density compared with other serum proteins. With the arrival of the analytical ultracentrifugation in the 1940s, this characteristic allowed its initial separation as a discrete peak using this technique. During the following years, it was demonstrated that this fraction was made up of a wide spectrum of particle sizes and densities (d) ranging from 0.92 to 1.21 g ml^{-1}.

Lipoproteins were classically separated into four major classes designated as chylomicrons (exogenous triacylglycerol-rich particles of $d < 0.94$ g ml^{-1}), very low-density lipoproteins (VLDL, endogenous triacylglycerol-rich particles of $d = 0.94$–1.006 g ml^{-1}), LDL (cholesteryl ester-rich particles of $d = 1.006$–1.063 g ml^{-1}), and HDL (particles containing approximately 50% protein of $d = 1.063$–1.21 g ml^{-1}). With subsequent improvements to the ultracentrifugation techniques, further heterogeneity was detected within each of those major lipoprotein classes; this resulted in the need for further subdivision into several density subclasses such as HDL_{2a} ($d = 1.10$–1.125 g ml^{-1}), HDL_{2b} ($d = 1.063$–1.10 g ml^{-1}), and HDL_3 ($d = 1.125$–1.21 g ml^{-1}).

There is no doubt that the separation of lipoproteins by ultracentrifugation has been esential for the advances in this field; however, this technique is very labor intensive and the isolated lipoproteins are usually modified due to the high g force and salt concentrations used in this process. The development of new vertical and near vertical rotors has shortened considerably the runs and thus diminished some of these negative effects.

Classification of Serum Lipoproteins According to Their Apolipoprotein Composition

Recent interest on the study of lipoprotein subfractions has resulted in an increased use of methods of separation based on affinity chromatography, specially those using immunoaffinity. By using columns containing antibodies against specific apolipoproteins (**Table 2**), a large number of HDL subpopulations have been resolved. Similarly, this technique

Table 2 Classification and properties of apolipoproteins

Apolipoprotein	*Amino acids*	*Tissue expression*	*Chromosomal localization*	*Functions*
apo A-I	243	Liver Intestine	11	Major structural component of HDL Ligand for HDL binding Activator of LCAT Reverse cholesterol transport
apo A-II	77	Liver	1	Structural component of HDL Activator of hepatic lipase
apo A-IV	377	Intestine Liver	11	Regulator of LPL activity Activator of LCAT Intestinal lipid absorption
apo B-48	2152	Intestine	2	Structural component of TRL Secretion of chylomicrons
apo B-100	4536	Liver	2	Structural
apo C-I	57	Liver Intestine	19	Activator of LCAT Inhibitor of the LRP
apo C-II	79	Liver Intestine	19	Activator of LPL
apo C-III	79	Liver Intestine	11	Inhibits LPL
apo D	169	Most tissues	3	Radical scavenger Reverse cholesterol transport Binding of haem-related compounds
apo E	299	Liver Macrophage	19	Ligand for the LDL receptor Ligand for the LRP Reverse cholesterol transport
apo(a)	Variable	Liver	6	?

allows the separation of several triacylglycerol-rich lipoproteins subfractions.

Lipoproteins containing apo A-I can be separated into two major species: those containing both apo A-I and apo A-II, known as LpAI:AII, and those containing apo A-I but not apo A-II (LpAI). Small numbers of particles containing apo A-II, but not apo A-I, have been detected in normal subjects; however, these particles could become predominant in the presence of rare genetic disorders associated with HDL deficiency. Another HDL species containing apo A-I and apo E is important in reverse cholesterol transport by transporting cholesterol from the cell membranes to the liver for elimination from the body.

Lipoproteins containing apo B consist of four lipoprotein families. Lipoproteins containing apo B only (Lp(B)) are cholesteryl ester-rich and are found primarily within the LDL density range, but they have also been detected within the VLDL range. Particles containing both apo B and apo C (LpB:C), apo B and apo E (LpB:E), and all three apolipoprotein groups (LpB:E:C), are triacylglycerol-rich and are found within the VLDL and IDL density range. The apo C and apo E content decreases as density increases.

More recently, the affinity for lectins of Lp(a), a lipoprotein containing apo B-100 as well as an antigenically unique apolipoprotein [apo(a)], has been used to develop a new technique to measure the levels of this lipoprotein in plasma.

Synthesis and Catabolism of Lipoproteins

Metabolism of Lipoproteins Carrying Exogenous Lipids

Dietary fats absorbed in the intestine are packaged into large, triacylglycerol-rich chylomicrons for delivery through the bloodstream to sites of lipid metabolism or storage. These lipoproteins interact with lipoprotein lipase (LPL) and undergo lipolysis, forming chylomicron remnants. The major sites of LPL activity are adipose tissue, skeletal muscle, the mammary gland, and the myocardium. In these sites, the fatty acids from the trcacylglycerols are used for storage, oxidation, or secretion back to the circulation. The triacylglycerol-depleted particles resulting from the lipolysis, known as chylomicron remnants, pick up apo E and cholesteryl ester from HDL and are rapidly taken up by the liver via a process mediated by the apo E receptor. This is a fast process and chylomicron particles are not usually present in the blood after a prolonged fasting period. The occurrence of chylomicronemia can be easily detected by the presence of a creamy supernatant floating on top of the plasma or serum kept several hours at 4 °C.

Transport of Endogenous Lipids

The liver cell secretes triacylglycerol-rich VLDL, which can be converted first to intermediate-density lipoprotein (IDL) and then to LDL through lipolysis by a mechanism similar to that described for chylomicrons. The excess surface components are usually transferred to HDL, and the triacylglycerol-depleted VLDL becomes an IDL. Some of these particles may be taken up by the liver via an apo E receptor, whereas others are further depleted of triacylglycerols, becoming cholesteryl ester-enriched particles known as LDL, which contain apo B as their only apolipoprotein. Consumption of fat-rich meals or glucose enhances VLDL production.

Some primary causes of elevated VLDL or IDL levels are familial endogenous hypertriglyceridemia (type IV according to Fredrickson's classification) and familial dysbetalipoproteinemia (type III hyperlipidemia). Genetic mutations at the apo E gene locus are responsible for the type III phenotype. Some secondary causes for elevated VLDL levels are obesity, diabetes mellitus, alcohol consumption, as well as the use of high doses of certain drugs (e.g., thiazide diuretics and estrogens). The presence of elevated levels of IDL has been associated with an increased atherosclerotic risk.

LDL particles are major carriers of cholesteryl ester in the blood. An LDL receptor that recognizes apo B-100 and apo E, but not apo B-48, allows the liver and other tissues to catabolize LDL. High-fat and high-cholesterol diets can decrease the activity of the LDL receptor, leading to increased levels of circulating LDL. These particles supply cholesterol to cells in the periphery for synthesis of cell membranes and steroid hormones. Modified or oxidized LDL can also be taken up by the scavenger receptor on macrophages in various tissues, including the arterial wall. This process is a potential initiator of foam cell formation and atherosclerosis.

Several LDL subclasses have been identified using gradient gel electrophoresis. Large, less dense LDL particles are commonly found in premenopausal women and men at low risk for CHD, whereas the small, more dense particles have been associated with a significant increased risk for myocardial infarction. The distribution of these particles appears to have a significant genetic component modulated by age and environmental factors.

Reverse Cholesterol Transport

HDL is synthesized by both the liver and the intestine. Its precursor form is discoidal in shape and matures in circulation as it picks up unesterified cholesterol from cell membranes and other lipids (phospholipid and triacylglycerol) and proteins (A-I, E, and C apolipoproteins) from triacylglycerol-rich lipoproteins (chylomicron and VLDL) as these particles undergo lipolysis. The cholesterol is esterified by the action of the lecithin–cholesterol acyltransferase (LCAT) and the small HDL_3 particle becomes a larger HDL_2 particle. The esterified cholesterol is either delivered to the liver or transferred by the action of cholesteryl ester transfer protein (CETP) to other lipoproteins (such as chylomicron, VLDL remnants, or LDL) in exchange for triacylglycerols. This cholesterol may then be taken up by the liver via receptors specific for these lipoproteins, or it can be delivered again to the peripheral tissues. The triacylglycerol received by HDL_2 is hydrolyzed by hepatic lipase and the particle is converted back to HDL_3, completing the HDL cycle in plasma. In the liver, cholesterol can be excreted directly into bile, converted to bile acids, or reutilized in lipoprotein production.

Several genetic disorders have been identified associated with low levels or total deficiency of HDL.

Effects of Dietary Fats and Cholesterol on Lipoprotein Metabolism

The cholesterolemic effects of dietary fatty acids have been extensively studied. The saturated fatty acids $C_{12:0}$, $C_{14:0}$, and $C_{16:0}$ have a hypercholesterolemic effect, whereas $C_{18:0}$ has been shown to have a neutral effect. Monounsaturated and polyunsaturated fatty acids in their most common *cis* configuration are hypocholesterolemic in comparison with saturated fatty acids. The effects of *trans* fatty acids on lipid levels are under active investigation. Our current knowledge shows that their effect is intermediate between saturated and unsaturated fats. The effect of dietary cholesterol on lipoprotein levels is highly controversial. This may be due in part to the dramatic interindividual variation in response to this dietary component. Specific effects of dietary fats and cholesterol on each lipoprotein fraction are the focus of other articles and they are only briefly summarized below and in **Table 3**.

Effects of Diet on Chylomicron Metabolism

Diets very high in saturated fat have been associated with increased postprandial chylomicrons and chylomicron remnants compared with diets rich in n-6 polyunsaturated fats; however, human experiments carried out using moderate to high fat intake have not shown significant effects of different types of dietary fat or dietary cholesterol on postprandial lipoproteins.

The effects of dietary carbohydrates on postprandial lipoproteins have also been studied. Most protocols have used diets very high in simple carbohydrates. In general, high carbohydrate intake has been associated with increased levels of fasting triacylglycerols and increased postprandial levels of chylomicrons and chylomicron remnants.

Effects of Diet on VLDL Metabolism

It is well-known that diets high in simple carbohydrate increase hepatic secretion of VLDL. This carbohydrate induction of hypertriglyceridemia is the source of the current controversy regarding the optimal diet for subjects at high risk for cardiovascular disease. Some authors have demonstrated that the increased hepatic triacylglycerol secretion induced by high-carbohydrate diets was not accompanied by parallel increases in apo B-100 secretion. In other words, the consumption of low-fat, high-carbohydrate diets did not affect the number of particles but resulted in larger, more triacylglycerol-enriched VLDL particles.

Intake of saturated fat results in an increased secretion of the number of VLDL particles by the liver, whereas the opposite effect is observed with polyunsaturated fat. Of special note are the dramatic effects on VLDL production found following high intakes of n-3 fatty acids. These diets are associated with marked decreases in triacylglycerol secretion by mechanisms not fully understood. It

Table 3 Effects induced on the major lipoprotein fractions by different dietary components following isoenergetic replacement of saturated fatty acids

	MUFA	*PUFA n-6*	*PUFA n-3*	*trans FA*	*Simple carbohydrate*	*Carbohydrate plus fiber*
VLDL-C	≈	≈/↓	↓	↑	↑	≈
LDL-C	↓	↓	≈/↓	↑	↓	↓
HDL-C	≈/↑	≈/↓	↓	↓	↓	≈/↓

≈ equivalent effect; ↓ concentration reduced; ↑ concentration increased.

has been speculated that n-3 fatty acids may stimulate intracellular degradation of apo B in hepatocytes. Dietary cholesterol, within the physiological range, appears to play a minor role in hepatic VLDL production.

Effects of Diet on LDL Metabolism

The effects of dietary fat and cholesterol on LDL metabolism have been extensively studied. However, the effects of dietary cholesterol are still highly controversial. Whereas some studies have demonstrated increased LDL production and decreased catabolism associated with high cholesterol intakes, others have failed to find such associations.

Replacement of saturated by polyunsaturated fats has been associated with decreased LDL apo B production in some studies, whereas in other studies, increased ratios of polyunsaturated to saturated fats resulted in increased LDL apo B catabolism. Unlike the effects described for VLDL metabolism, intake of n-3 fatty acids appears to play a minor role on LDL metabolism.

Effects of Diet on HDL Metabolism

Diets high in simple carbohydrates reduce HDL cholesterol levels. This effect appears to be mediated by increases in the catabolism of apo A-I; however, one study has also demonstrated an additional decrease in apo A-I production.

Disorders of Lipoprotein Metabolism

For historical reasons the classification of disorders of lipoprotein metabolism will be presented according to the classical Fredrickson's classification (**Table 4**).

Type I or Familial Chylomicronemia

This disorder is characterized by greatly elevated levels of exogenous triacylglycerols and it is the result of impaired lipolysis of chylomicrons due to a deficiency of LPL or its activator, the apo C-II. Several genetic mutations at the structural genes for both LPL and apo C-II have been reported. These are autosomal recessive traits. In the heterozygous state, subjects have normal to slightly elevated plasma triglycerides, whereas homozygotes have triacylglycerol levels that may exceed $1000\,mg\,dl^{-1}$ in the fasting state. The diagnosis of the homozygous state takes place during the first years of life from the presence of recurrent abdominal pain and pancreatitis. Eruptive xanthomas and lipemia retinalis may also occur.

The recommended treatment includes a diet low in simple carbohydrates and with a fat content below 20% of total energy. The use of medium-chain triglycerides (MCT) has also been reported to be efficacious. Body weight should be maintained within the normal limits and alcohol consumption should be avoided.

Other secondary causes leading to the presence of chylomicrons in the fasting state include uncontrolled diabetes mellitus, alcoholism, estrogen use, and hypothyroidism.

Fasting chylomicronemia has not been clearly associated with increased risk for atherosclerosis; however, there is considerable evidence supporting the atherogenic properties of chylomicron remnants.

Type II or Familial Hypercholesterolemia

Familial hypercholesterolemia (FH) is an autosomal dominant disorder characterized by elevation of plasma LDL cholesterol levels. Mutations at the LDL receptor gene locus on chromosome 19 are

Table 4 Classification of hyperlipidemias according to Fredrickson

Type	*Plasma cholesterol*	*Plasma triacylglycerol*	*Lipoprotein fraction(s) affected*	*Atherosclerosis risk*	*Genetic disorder*
I	Normal to elevated	Very elevated	Chylomicrons	No	Familial LPL deficiency Apo C-II deficiency
IIa	Elevated	Normal	LDL	High	Familial hypercholesterolemia Familial combined hyperlipidemia Polygenic hypercholesterolemia
IIb	Elevated	Elevated	LDL and VLDL	High	Familial hypercholesterolemia Familial combined hyperlipidemia
III	Elevated	Very elevated	IDL	High	Familial dysbetalipoproteinemia
IV	Normal or elevated	Elevated	VLDL	Moderate	Familial hypertriglyceridemia Familial combined hyperlipidemia
V	Normal or elevated	Very elevated	VLDL and chylomicrons	Moderate	Familial hypertriglyceridemia

responsible for this disorder. Multiple different mutations have been described at this locus resulting in the FH phenotype. In the heterozygous state, subjects develop tendinous xanthomas, corneal arcus, and CHD. Elevations of LDL can result from well-characterized genetic disorders such as FH or familial defective apo B-100.

The ranges of LDL cholesterol levels in plasma of FH subjects are 200–400 mg dl^{-1} in heterozygotes and above 450 mg dl^{-1} in homozygotes. The frequency of defects at the LDL receptor locus is about 1 in 500 for the heterozygous state and 1 in a million in the homozygous state.

Inhibitors of 3-hydroxy-3-methylglutaryl (HMG) coenzyme A are useful in the treatment of hypercholesterolemia. Most pharmacological therapies are ineffective in the homozygous state. FH homozygotes may be treated with LDL apheresis, liver transplantation, and portacaval shunt. More recently, encouraging results have been obtained using *ex vivo* gene therapy.

The genetic defect(s) associated with a common form of hypercholesterolemia present in most subjects with cholesterol levels between 250 and 300 mg dl^{-1} has not been elucidated. This disorder may be due to a combination of minor gene defects (i.e., presence of apo E-4 allele) that in combination with the environment (i.e., diet, lack of exercise) predispose individuals to moderately elevated LDL cholesterol levels. This disorder has been also named polygenic hypercholesterolemia.

Familial Defective apo B-100

Familial defective apo B-100 is an autosomal dominant genetic disorder that presents with a phenotype similar to FH. The frequency of this disorder may be similar to FH; however, it varies considerably depending on the ethnicity of the population studied. The specific mutation responsible for this disorder is a point mutation at amino acid 3500 of the mature apo B. The diagnosis of this disorder requires molecular biology techniques.

Type III or Familial Dysbetalipoproteinemia

In this disorder both plasma triacylglycerol and cholesterol are increased. Several mutations within the apo E gene locus have been found to be responsible for this disease; however, in most patients the complete expression of the clinical genotype needs additional interactions such as age, obesity, and diabetes. In addition to the accumulation in plasma of VLDL remnants and chylomicrons, other characteristics of this disorder are tuboeruptive xanthomas and in some cases also planar xanthomas. Therapies include diet and hypolipidemic agents such as fibrates, statin, or nicotinic acid. In most cases, diagnosis can be carried out first by agarose gel electrophoresis, followed by molecular biology techniques to detect the presence of the apo E-2 allele.

Familial Type IV and Type V Hypertriglyceridemias

These two disorders may have overlapping phenotypes. In type IV or familial endogenous hypertriglyceridemia, triacylglycerol levels are increased and HDL is usually decreased. This disorder appears to be autosomal dominant and relatively frequent in populations consuming high-fat diets. The precise molecular defect has not been defined; however, the increase in triacylglycerol is associated with overproduction of triacylglycerol by the liver and often with consequent reduced clearance. Diet should be the first step in therapy, followed if necessary by pharmacotherapy using fibrates or nicotinic acid. Premature CHD has been seen in some but not all cases presenting with this phenotype.

Type V hyperlipidemia is a much more rare disorder. Usually the first signs of this abnormality are abdominal pain or pancreatitis. VLDL levels are high and chylomicrons are present in the fasting state. This abnormality has not been linked to any specific molecular defect. Besides the primary genetic defect, other secondary causes of type V hyperlipidemia are poorly controlled diabetes mellitus, nephrotic syndrome, hypothyroidism, glycogen storage disease, and pregnancy. Recent data indicate increased susceptibility to atherosclerosis.

Familial Dyslipidemia

Familial dyslipidemia may be a variant of the familial hypertriglyceridemias described previously. It is characterized by hypertriglyceridemia in combination with low HDL cholesterol. Patients are generally overweight, with male pattern obesity, insulin resistance, diabetes, and hypertension. These subjects have both increased hepatic triacylglycerol secretion and increased HDL apo A-I catabolism.

Familial Combined Hyperlipidemia

Familial combined hyperlipidemia (FCH) was initially described as the combination of hypercholesterolemia and hypertriglyceridemia within the same kindred, and with kindred members having one of these abnormalities or both. Moreover, most subjects with FCH have HDL cholesterol levels below the 10th percentile. Affected subjects have elevation in VLDL, LDL, or both. This disorder has a frequency of approximately 10% in survivors of

premature myocardial infarction (less than 60 years of age) and about 14% in kindred with CHD.

It has been reported that affected subjects have overproduction of apo B-100. The precise molecular defect has not been elucidated, although there are already several candidate gene loci, including the LPL. The expression of this disorder may be triggered by other factors, such as overweight, hypertension, diabetes, and gout. The treatment should include diet and exercise and, if necessary, niacin, HMC CoA reductase inhibitors, or fibrates, depending on the major lipid present in excess.

Familial Hyperapobetalipoproteinemia

Familial hyperapobetalipoproteinemia is characterized by apo B values above the 90th percentile in the absence of other lipid abnormalities; it has been suggested to be a variant of FCH. This disorder is relatively common (~5%) in kindreds with premature CHD. The molecular defect is not known, but metabolic studies suggest overproduction of apo B-100.

Familial Hypoalphalipoproteinemia

Severe HDL deficiency, characterized by HDL cholesterol levels $<10\,mg\,dl^{-1}$ is rare and may be due to Tangier disease, apo A-I deficiencies, LCAT deficiency, or fish-eye disease. The apo A-1 deficiency states are due to rare deletions, rearrangements, or point mutations within the apo A-I/C-III/A-IV gene complex. Familial hypoalphalipoproteinemia is relatively common and is characterized by HDL cholesterol levels below the 10th percentile of normal. These subjects have been reported to have either decreased HDL production or increased HDL apo A-I catabolism. This phenotype is present in about 4% of kindred with premature CHD.

The genetic defect or defects are not known; however, it has been suggested that familial combined hyperlipidemia, familial hyperapobetalipoproteinemia, familial dysbetalipoproteinemia, and familial hypoalphalipoproteinemia may be variants of a single disorder. This disorder is characterized by a genetic predisposition in subjects consuming high-fat, high-cholesterol diets to an increased secretion of apo B-containing lipoproteins and an increased catabolism of apo A-I-containing lipoproteins. The expression of the phenotype is usually enhanced by the presence of male pattern obesity.

Familial Lipoprotein (a) Excess

Lipoprotein (a) (Lp(a)) is an LDL particle with one molecule of apolipoprotein (a) attached to it. Elevated levels of Lp(a) (>35–$40\,mg\,dl^{-1}$ or 90th percentile) have been associated with premature CHD. This increased risk appears to result from two different mechanisms: cholesterol deposition in the arterial wall and inhibition of fibrinolysis.

Lp(a) concentrations are highly variable among individuals; however, they tend to remain constant during a person's lifetime. Between 80 and 90% of the variability appears to be of genetic origin, owing, for the most part, to variations at the structural apo(a) gene locus. Lp(a) concentrations are inversely associated with a size polymorphism of apo(a). This polymorphism is due to differences in the number of a multiple repeat of a protein domain highly homologous to the kringle 4 domain of plasminogen. Diets and medications used to lower LDL cholesterol levels do not appear to have a significant effect on Lp(a) concentrations; however, niacin has been reported to decrease Lp(a) levels. There have been reports suggesting that diets high in *trans* fatty acids have some raising effects on Lp(a) levels, whereas estrogen replacement lowers Lp(a) in postmenopausal women.

General Guidelines for the Treatment of Lipoprotein Abnormalities for CHD Prevention

There is a clear benefit from lowering LDL cholesterol with diet or drug therapy in patients with hyperlipidemia or CHD or both. Dietary therapy includes using diets that are restricted in total fat (<30% of calories), saturated fat (<7% of calories), and cholesterol ($<200\,mg\,day^{-1}$). Pharmacological therapies include anion exchange resins, niacin, and HMG CoA reductase inhibitors. The latter agents have been demonstrated to also lower CHD mortality. It should be noted that dramatic interindividual variations have been demonstrated in response to diet and drug therapies. Consequently the efficacy of hypolipidemic therapies will vary from individual to individual. More information is needed about the benefits of HDL cholesterol raising in patients with low HDL cholesterol levels as well as the benefits of lowering triacylglycerol plasma concentrations, and more specifically the triacylglycerol carried in lipoprotein remnants. This is also true regarding the benefits of Lp(a) lowering using niacin in patients with elevated Lp(a) levels.

See also: **Body Composition**. **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Coronary Heart Disease**: Hemostatic Factors; Lipid Theory; Prevention. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3

Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Fertility**.

Further Reading

Alaupovic P (1996) Significance of apolipoproteins for structure, function, and classification of plasma lipoproteins. In: Bradley WA, Gianturco SH, and Segrest JP (eds.) *Methods in Enzymology, Plasma Lipoproteins*, part C, vol. 263, pp. 32–60. San Diego: Academic Press.

Austin MA, Breslow JL, Henneckens CH *et al.* (1988) Low density lipoprotein subclass patterns and risk of myocardial infarction. *Journal of the American Medical Association*, **260**: 1917–1921.

Li Z, McNamara JR, Ordovas JM, and Schaefer EJ (1994) Analysis of high density lipoproteins by a modified gradient gel electrophoresis method. *Journal of Lipid Research* **35**: 1698–1711.

National Cholesterol Education Program (1994) Second Report of the Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel II). *Circulation* **89**: 1329–1445.

Ordovas JM (1991) Molecular biological approaches to the understanding of lipoprotein metabolism. In: Witiak DT, Newman HAI, and Feller DR (eds.) *Medical, Chemical and Biochemical Aspects of Antilipidemic Drugs*, pp. 97–121. Amsterdam: Elsevier.

Ordovas JM (1993) Metabolism of triglyceride-rich lipoproteins: Genetic mutations associated with its pathology. *Cardiovascular Risk Factors* **3**: 1–8.

Ordovas JM (1994) Genetic and environmental factors: Effects on plasma lipoproteins. In: Serrano Rios (ed.) *Dairy Products in Human Health and Nutrition*, pp. 303–307. Rotterdam: Balkema.

Ordovas JM, Civeira F, Genest J, and Schaefer EJ (1990) Genetic high density lipoprotein deficiency states. In: Lenfant C, Albertini A, Paoletti R, and Catapano A (eds.) *Atherosclerosis Reviews. Biotechnology of Dyslipoproteinemias. Application in Diagnosis and Control,* vol. 20, pp. 261–274. New York: Raven Press.

Ordovas JM, Lopez-Miranda J, Mata P *et al.* (1995) Gene–dict interaction in determining plasma lipid response to dietary intervention. *Atherosclerosis* **118**: S11–S27.

Schaefer EJ and Ordovas JM (1992) Diagnosis and management of HDL deficiency states. In: Miller NE and Tall AR (eds.) *High Density Lipoproteins and Atherosclerosis,* vol. III, pp. 235–251. Amsterdam: Elsevier.

Schaefer EJ, Genest Jr JJ, Ordovas JM, Salem DN, and Wilson PWF (1993) Familial lipoprotein disorders and premature coronary artery disease. *Current Opinion in Lipidology* **4**: 288–298.

Schaefer EJ, Lichtenstein AH, Lamon-Fava S, McNamara JR, and Ordovas JM (1995) Lipoproteins, nutrition, aging, and atherosclerosis. *American Journal of Clinical Nutrition* **61**(supplement): 726S–740S.

Zannis VI, Kardassis D, and Zanni EE (1993) Genetic mutations affecting human lipoproteins, their receptors, and their enzymes. *Advances in Human Genetics* **21**: 145–319.

LIVER DISORDERS

J Hampsey and K B Schwarz, Johns Hopkins School of Medicine, Baltimore, MD, USA

This article covers the role of the liver in normal nutrition, including the important functions of bile salt production, macronutrient metabolism, and fat-soluble vitamin absorption, metabolism, and storage. Next, the pathogenesis of malnutrition in liver disease is discussed, starting with the mechanisms of malnutrition in both acute and chronic liver failure. Specific nutritional issues in liver failure are addressed, including metabolic disturbances of carbohydrates, protein, and fats. Nutritional disturbances in the major types of specific liver diseases are reviewed: hepatocellular, metabolic liver disease, and biliary tract disorders. Nutritional assessment and management of patients with acute, chronic liver disease and end stage liver disease are discussed.

Liver in Normal Nutrition

Bile Salts

A normal functioning liver will secrete 600–1200 ml of bile to the gall bladder on a daily basis. Bile is made up of bile salts, lecithin, conjugated bilirubin, phospholipids, cholesterol, electrolytes, and water. Bile salts, which are the predominant component of bile, are synthesized from cholesterol in the hepatocyte. The primary function of bile salts lies in their interaction with lipid digestion. Bile salts bind with large fat particles, which alone are insoluble in water, and act on them as an emulsifier, breaking down into smaller particles called micelles. Micelles, the product of the fat particle and bile salt structure, aid in the transport of fat to the mucosal membrane for absorption. Fat-soluble vitamins and cholesterol are also incorporated into mixed micelles for proper absorption.

Micellar solubilization is only required for long-chain fatty acids. Short-chain fatty acids (10 carbons

or less) do not require micelle formation for absorption; instead, they enter the portal circulation directly, bound to albumin, and are transferred to the liver for oxidation. Approximately 94% of the micelle forming bile acids are reabsorbed in the ileum and shuttled via the portal hepatic vein bound to albumin back to the liver. Only 6% of bile acids are lost in excretion.

Macronutrient Metabolism

Carbohydrates The liver is responsible for maintaining normal blood glucose concentrations under various metabolic conditions. Among the several metabolic processes that allow this fine regulation are glycogenesis, gluconeogenesis, and glycolysis. The end product of carbohydrate digestion is 80% glucose, with the remaining 20% being fructose and galactose; the latter two are quickly converted into glucose in the liver. Once transported into the hepatocyte, the glucose molecule is phosphorylated (via glucokinase) and cannot leave the cell unless dephosphorylated with glucose phosphatase. Glucose is either used for immediate energy release or stored as glycogen.

Proteins The liver plays a major role in protein metabolism in the deamination of amino acids, urea formation for removal of ammonia, plasma protein synthesis, and in the interconversions among amino acids. Ingested protein is the sole source of the 10 essential amino acids and the primary source of nitrogen necessary for the synthesis of other amino acids. Protein is digested and broken down to amino acids that are absorbed into the circulation and taken to cells throughout the body, primarily the liver, and quickly become combined by peptide linkages. The plasma level of amino acids is tightly controlled and maintained near a constant level. Once the cellular limit of protein storage is met, excess amino acids are degraded and used for energy or stored as fat or glycogen. The liver is the primary site of all amino acid catabolism with the exception of branch-chained amino acid catabolism, which occurs in the muscle cells. The urea cycle, in which the toxic compound ammonia is converted to urea, occurs solely in the liver. The synthesis of the plasma proteins albumin, fibrinogen, and globulin also occurs in the liver.

Lipids The liver plays a role in fat metabolism in four key processes: fatty acid oxidation for energy, lipoprotein syntheses, the synthesis of cholesterol and phospholipids, and the conversion of carbohydrate to fat for storage. Digested fat is a major source of energy in which after splitting into fatty acids and glycerol, the fatty acid components further split via beta oxidation into acetyl-CoA. Two molecules of acetyl-CoA become paired together to form acetoacetic acid and are transported to other cells to provide energy in the citric acid cycle.

Fat-soluble vitamins The liver plays a key role in the absorption of the fat-soluble vitamins—A, D, E, and K—as they are only successfully absorbed in association with fat and sufficient quantities of bile salts. The liver is also the primary storage site for several vitamins, including A, E, K, and B^{12}. Vitamin A is stored in the largest quantity in a sufficient amount to prevent deficiency for 5–10 months. Vitamin D is stored in amounts sufficient for 2–4 months. Vitamin B_{12} is stored in amounts sufficient for at least 1 year. The liver is responsible for the hydroxylation of vitamin D to its storage form, 25-hydroxy vitamin D. It is released into circulation and thence delivered to the kidney where it is converted to its active form, 1, 25-dihydroxy vitamin D.

Specific Nutritional Issues

Carbohydrates As discussed earlier, the liver plays a major role in the maintenance of normal blood sugar levels and overall glucose metabolism. Not surprisingly, in the patient with liver disease, glucose intolerance and insulin resistance are common. Cirrhotic patients are prone to developing diabetes. Energy from carbohydrates plays an important role in protein sparing mechanisms, preventing the use of protein as energy.

Lipids In cholestatic liver disease there is malabsorption of dietary lipid and consequent malnutrition. There are experimental data from primates showing that chronic ethanol consumption results in a decrease of liver phospholipids and of phosphatidylcholine (PC). Consequently, the total phospholipid content of the mitochondrial membranes is decreased; mitochondria are altered both structurally and functionally. There is diminished mitochondrial oxidation because of decreased cytochrome oxidase activity, which can be restored by administration of PC. The extent to which chronic liver disease of etiologies other than chronic ethanol consumption results in similar perturbations is unknown.

Protein Plasma proteins such as albumin and coagulation factors constitute approximately 50% of the proteins synthesized in the liver. In liver disease, decreased synthesis of these proteins has important clinical consequences, including ascites from

hypoalbuminemia and coagulopathy from decreased synthesis of coagulation factors. In end stage liver disease, hypoglycemia can result from decreased hepatic gluconeogenesis from amino acids. Decreased activity of the urea cycle enzymes results in hyperammonemia and hepatic encephalopathy, the ultimate expression of which can be cerebral edema.

Fat-soluble vitamins Deficiencies of fat-soluble vitamins are common in liver disease associated with steatorrhea due to the concomitant malabsorption of fat. Vitamin A deficiency can result in anorexia, growth failure, decreased resistance to infections, and night blindness. Vitamin D deficiency results in osteopenia or osteoporosis as well as rickets. The prevalence of fractures is increased in women being treated for alcohol abuse and also following sobriety; deficiencies of vitamin D as well as calcium, phosphorus, and fluoride may play a role. The deficiency of vitamin E results in neuraxonal dystrophy, clinically manifesting as peripheral neuropathy and cerebellar disturbances. Vitamin K deficiency results in hemorrhage because of reduced synthesis of clotting factors.

Trace elements Zinc deficiency in cirrhotics may contribute to hypoalbuminemia and dermatitis as well as anorexia from hypogeusia. Deficiency of selenium can lead to decreased synthesis of important antioxidant selenoproteins such as glutathione peroxidase. Little is known about the effect of acute or chronic liver disease on other trace elements.

Liver in Specific Hepatobiliary Disorders

Hepatocellular Diseases

Alcoholic liver disease The term 'alcoholic liver disease' refers to a spectrum of types of hepatic injury associated with continuous alcohol ingestion, ranging from alcoholic fatty liver to alcoholic steatohepatitis, fibrosis, and cirrhosis. Nutritional disturbances in alcoholics are an important cause of morbidity and mortality; all classes of nutrients are affected. Anorexia leads to decreased food intake and subsequent protein-calorie malnutrition. Maldigestion and malabsorption can occur secondary to chronic alcohol injury to small intestinal mucosa. Alcohol consumption is often associated with chronic pancreatic insufficiency, which results in steatorrhea and decreased absorption of dietary protein, fat, and fat-soluble vitamins. Chronic alcohol ingestion also results in impaired hepatic amino acid uptake and protein synthesis.

In alcoholics, utilization of lipids and carbohydrates is markedly compromised due to an excess of reductive equivalents and impaired oxidation of triglycerides. Alcoholics are often resistant to insulin and exhibit impaired uptake of glucose into muscle cells. Insulin-dependent diabetes is common. Heavy alcohol consumption is frequently associated with deficiencies of a wide variety of micronutrients, including the fat- and water-soluble vitamins, particularly folate, pyridoxal-5′-phosphate, thiamine, and vitamin A.

Table 1 summarizes the five published controlled trials of the effect of oral or enteral nutritional supplements on patients with alcoholic hepatitis. In most, nitrogen balance and/or protein synthesis improved, although no effect on mortality was shown, perhaps because of the small number of patients studied and/or the duration of follow-up. In the largest study, at 1-year follow-up, the experimental group had a significantly better survival: 2/24 (8%) died compared to 10/27 (37%) of the controls. In general, the effects of parenteral nutrition in alcoholic liver disease are similar to those noted the studies of enteral nutritional supplements.

Many studies have examined the effect of oral or enteral nutritional supplementation in patients with alcoholic cirrhosis. Results are summarized in **Table 2**. Many studies are small and of short duration, so it is not surprising that results are inconclusive. Most studies demonstrated an improvement in nitrogen balance and protein synthesis; only one showed increased survival in the treated group. Taken together, these studies suggest that there are benefits to nutritional supplementation in this population.

A variety of international associations have made nutritional recommendations for patients with various types of alcoholic liver disease. The primary recommendation is of course abstinence, which may be all that is needed in patients with fatty liver. Patients with alcoholic hepatitis should take 40 kcal/kg, 1.5–2.0 g protein/kg, 4–5 g/kg of carbohydrates, and 1–2 g/kg of lipids per day. Those with cirrhosis without malnutrition should take 35 kcal/kg, 1.3–1.5 g protein/kg, and carbohydrates and lipids as recommended for patents with alcoholic hepatitis. Those with cirrhosis and malnutrition should take higher amounts of protein (1.5–2.0 g/kg) and lipids (2.0–2.5 g/kg) and lower amounts of carbohydrates (3–4 g/kg). Fluid should be restricted to 2–2.5 l/day and all eight B vitamins, including folate and thiamine, as well as vitamins C and K should be routinely supplemented. In addition, patients with cholestasis should take 50% of their dietary lipids as medium-chain triglycerides

Table 1 Studies on therapy of alcoholic hepatitis with oral or enteral nutritional supplements

Reference	*Design*	*Patients (No.)*	*Duration (days)*	*Experimental treatment (EXP)*	*Control treatment (CTR)*	*Mortality*	*Secondary end points*
Galambos *et al.* (1979)[a]	Open label	16	16–42	Oral (standard hospital diet) or intravenous supplement (51.6–77.4 g protein)	None	Not assessed	Nitrogen balance + albumin improved in EXP, CTR not assessed Improvement of albumin, transferrin, RBP
Mendenhall *et al.* (1985)[b]	Historical controls	57	30	Standard hospital diet (2500 kcal/day) + 2200 kcal/day BCAA	Standard hospital diet	NS	Positive nitrogen balance in EXP, delayed hypersensitivity improved
Calvey *et al.* (1985)[c]	Randomized, controlled	64	21	Standard diet (~2000 kcal/day) + 65 g standard AA or BCAA	Standard diet, 80 g protein/day	NS	Positive nitrogen balance in EXP, delayed hypersensitivty improved
Soberon *et al.* (1987)[d]	Crossover	14	6	Nasoduodenal tube, 35 kcal/kg/day, fat/carbohydrate/protein 45/40/15%	3 days standard hospital diet (35 kcal/kg/day)	0/6 controls 3/8 treatment	Nitrogen balance improved five-fold at 2 weeks
Cabre *et al.* (2000)[e]	Randomized, controlled	71	28	Nasogastric tube, 2000 kcal/day, 72 g protein/day, 31% BCAA	Standard diet (1 g protein/kg) + 40 mg/day prednisolone	11/35 TEN 9/36 PRED NS FU: 2/24 TEN 10/27 (P = 0.04)	No dropouts in PRED, 8 dropouts in TEN; equal improvements of albumin, Child score, Maddrey score; equal rate of infections

[a]Galambos JT, Hersh T, Fulenwider JT *et al.* (1979) Hyperalimentation in alcoholic hepatitis. *American Journal of Gastroenterology* **72**: 535–541.

[b]Mendenhall CL, Bongiovanni G, Goldberg S *et al.* (1985) VA cooperative study on alcoholic hepatitis. III: Changes in protein-calorie malnutrition associated with 30 days of hospitalization with and without enteral nutritional therapy. *Journal of Parenteral and Enteral Nutrition* **9**: 590–596.

[c]Calvey H, Davis M, and Williams R (1985) Controlled trial of nutritional supplementation, with and without branched chain amino acids enrichment, in treatment of acute alcoholic hepatitis. *Journal of Hepatology* **1**: 141–151.

[d]Soberon S, Pauley MP, Duplantier R *et al.* (1987) Metabolic effects of enteral formula feeding in alcoholic hepatitis. *Hepatology* **7**: 1204–1209.

[e]Cabre E, Rodriguez-Iglesias P, Caballeria J *et al.* (2000) Short-term and long-term outcome of severe alcohol-induced hepatitis treated with steroids or enteral nutrition: A multicenter randomized trial. *Hepatology* **32**: 36–42.

AA, amino acids; BCAA, branched-chain amino acid; FU, follow-up; NS, not significant; PRED, prednisolone group; TEN, total enteral nutrition group.

From Stickel F, Hoehn B, Schuppan D, and Seitz HK (2003) Review article: Nutritional therapy in alcoholic liver disease. *Alimentary Pharmacology & Therapeutics* **18**: 357–373.

Table 2 Studies on treating alcoholic cirrhosis with oral and enteral nutritional therapy[a]

Reference	*Design*	*Patients (No.)*	*Duration (days)*	*Experimental treatment (EXP)*	*Control treatment (CTR)*	*Mortality*	*Secondary end points*
Smith *et al.* (1982)[a]	Open label	10	10–60	Three different formulae: Oral 76–143 g protein, 2000–3716 kcal/day	None	None	Positive nitrogen balance, improved albumin, transferrin, creatinine/height, midarm muscle, fat areas
Keohane *et al.* (1983)[b]	Open label	10	3–23	Oral BCAA formula: 80 g protein/day through nasogastric tube	None	1 death (HRF)	Positive nitrogen balance, improved albumin
McGhee *et al.* (1983)[c]	Randomized, double-blind, crossover	4	11	20 g casein + 30 g BCAA formula	50 g casein/day	None	EXP equal to CTR, positive nitrogen balance
Christie *et al.* (1985)[d]	Randomized, double-blind, crossover	8	12	BCAA (50%) formula	Standard diet (18% BCAA)	1 death (infection)	EXP equal to CTR, positive nitrogen balance
Okita *et al.* (1985)[e]	Open label	10	4	40 g protein + 40 g BCAA formula/day	2100 kcal/day 80 g protein/day	None	EXP equal to CTR, positive nitrogen balance
Bunout *et al.* (1989)[f]	Randomized, controlled	36	28	50 kcal/kg, 1.5 g protein/day	Standard diet	EXP 2/17 CTR 5/19 (NS)	No differences
Cabre *et al.* (1990)[g]	Randomized, controlled	35 (23 alc.)	23–35	2115 kcal/day including 71 g BCAA formula	Standard diet	Improved ($p = 0.02$)	Child score improved, albumin improved
Marchesini *et al.* (1990)[h]	Randomized	64	90	Standard diet + BCAA supplement (0.24 g/kg)	Standard diet + casein supplement	None	Nitrogen balance improved in both, BCAA better than standard diet
Kerans *et al.* (1992)[i]	Randomized	31	28	Casein supplement (1.5 g protein/day, 40 kcal/day/kg/day)	Standard diet	NS	Both groups improved nitrogen balance and albumin
Hirsch *et al.* (1993)[j]	Randomized, controlled	51	12 (months)	Standard diet + casein supplement (1000 kcal/day, 34 g protein/day)	Standard diet	EXP 3/26 CTR 6/25 (NS)	Fewer hospitalizations, improved albumin and visceral protein
Nielsen *et al.* (1995)[k]	Open label	15	38	Increasing amounts of protein via standard diet (1.0–1.8 g/kg/day)	None	None	Increased protein retention through gradual or protein intake

Continued

Table 2 Continued

Reference	*Design*	*Patients (No.)*	*Duration (days)*	*Experimental treatment (EXP)*	*Control treatment (CTR)*	*Mortality*	*Secondary end points*
Campillo *et al.* (1995)[l]	Open label	26	30	Standard diet	None	None	Anthropometric ratios improved
Hirsch *et al.* (1992)[m]	Open label	31	6 (months)	Standard diet + casein supplement (1000 kcal/day, 34 g protein/day)	None	6 deaths/31	Increased albumin, improved cellular immunity

[a]Smith J, Horowitz J, Henderson JM *et al.* (1982) Enteral hyperalimentation in undernoursished patients with cirrhosis and ascites. *American Journal of Clinical Nutrition* **2**: 1209–1218.
[b]Keohane PP, Attrill H, Brimble G *et al.* (1983) Enteral nutrition in malnourished patients with hepatic cirrhosis and acute encephalopathy. *Journal of Parenteral and Enteral Nutrition* **7**: 34–50.
[c]McGhee A, Henderson JM, Millikan WJ *et al.* (1983) Comparison of the effects of hepatic-aid and casein modular diet on encephalopathy, plasma amino acids, and nitrogen balance in cirrhotic patients. *Annals of Surgery* **197**: 288–293.
[d]Christie ML, Sack DM, Pomposelli J *et al.* (1985) Enriched branched-chain amino acid formula versus a casein-based supplement in the treatment of cirrhosis. *Journal of Parenteral and Enteral Nutrition* **9**: 671–678.
[e]Okita M, Watanabe A, and Nagashima H (1985) Nutritional treatment of liver cirrhosis by branched-chain amino acid-enriched nutrient mixture. *Journal of Nutritional Science and Vitaminology* **31**: 291–303.
[f]Bunout D, Alcardi V, Hirsch S *et al.* (1989) Nutritional support in hospitalized patients with alcoholic liver disease. *European Journal of Clinical Nutrition* **43**: 615–621.
[g]Cabre E, Gonzalez-Huix F, Abad-Lacruz A *et al.* (1990) Effect of total enteral nutrition on the short-term outcome of severely malnourished cirrhotics. A randomized controlled trial. *Gastroenterology* **98**: 715–720.
[h]Marchesini G, Dioguardi FS, Bianchi GP *et al.* (1990) Long term oral branched-chain amino acid treatment in chronic hepatic encephalopathy. A randomized double-blind casein-controlled trial. The Italian Multicenter Study Group. *Journal of Hepatology* **11**: 92–101.
[i]Kearns PJ, Young H, Garcia G *et al.* (1992) Accelerated improvement of alcoholic liver disease with enteral nutrition. *Gastroenterology* **102**: 200–205.
[j]Hirsch S, Bunout D, De la Maza MP *et al.* (1993) Controlled trial on nutritional supplementation in outpatients with symptomatic alcoholic cirrhosis. *Journal of Parenteral and Enteral Nutrition* **17**: 119–124.
[k]Nielsen K, Kondrup J, Martinsen I *et al.* (1995) Long-term oral refeeding of patients with cirrhosis of the liver. *British Journal of Nutrition* **74**: 557–567.
[l]Campillo B, Bories PN, Leluan M *et al.* (1995) Short-term changes in energy metabolism after 1 month of a regular oral diet in severely malnourished cirrhotic patients. *Metabolism* **44**: 765–770.
[m]Hirsch S, de la Maza MP, Gattas V *et al.* (1999) Nutritional support in alcoholic cirrhotic patients improves host defenses. *Journal of the American College of Nutrition* **18**: 434–414.
BCAA, branched-chain amino acid; HRF, hepatorenal failure; NS, not significant.
From Stickel F, Hoehn B, Schuppan D, Seitz HK (2003). Review article: Nutritional therapy in alcoholic liver disease. *Alimentary Pharmacology & Therapeutics* **18**: 357–373.

and should be supplemented with the fat-soluble vitamins—A, D, E, and K. The major strategy in the management of alcoholic cirrhotics with ascites and edema is to restrict fluids to 1–1.5 l/day and to restrict sodium as well.

Autoimmune liver disease The two major categories of autoimmune liver disease are primary biliary cirrhosis (PBC), a disease generally presenting in young female adults, and autoimmune hepatitis, which also most frequently presents in adults but can affect both sexes and present at any time from young childhood to mid-adulthood. PBC results in steatorrhea and malabsorption of the fat-soluble vitamins. Osteoporosis and osteopenia are common.

The nutritional consequences of autoimmune hepatitis, which can evolve into cirrhosis, are similar to those of alcoholic hepatitis and cirrhosis secondary to alcoholic liver disease, and thus the management is similar as well. Occasionally, autoimmune hepatitis can be accompanied by intestinal diseases such as inflammatory bowel disease or celiac disease, and the nutritional management should take both organ systems into account. Although mild liver function abnormalities are common in celiac disease, there are reports of celiac disease in patients with severe liver disease, all of whom demonstrated an improvement in their liver disease with introduction of a gluten-free diet.

Neonatal cholestasis The major differential diagnosis of conjugated hyperbilirubinemia in the first 30 days of life is extrahepatic biliary atresia and the neonatal hepatitis syndrome, for which a large number of specific genetic disorders have been identified. These include α-1 antitrypsin deficiency, inborn errors of bile salt synthesis or transport, cystic fibrosis/liver disease, Alagille syndrome, hypothyroidism, and panhypopituitarism. The nutritional consequences are similar for all: steatorrhea and malabsorption of the fat-soluble vitamins and failure to thrive. Nutritional management is also similar for all: use of an elemental formula rich in medium-chain triglycerides (MCTs) and supplementation with vitamins A, D, E, and K. Water-miscible vitamin E is poorly absorbed; administration of vitamin E solubilized in polyethylene glycol succinate is a more effective way to administer vitamin E to cholestatic infants.

Nonalcoholic fatty liver disease and nonalcoholic steatohepatitis Nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH) have become very important causes of liver disease in both children and adults, particularly because obesity is being diagnosed in epidemic proportions in both age groups and both liver disorders are most commonly associated with obesity. Children with NAFLD may present before their fifth birthday. The disorder is more common in males. Hepatic fibrosis is common and may even evolve into cirrhosis during childhood. Treatment consists of weight reduction and aerobic exercise. Vitamin E may be beneficial.

In adults, NASH and NAFLD have been recognized for at least 25 years as chronic liver diseases associated with obesity (with or without non-insulin-dependent diabetes mellitus and with or without hyperlipidemia). NAFLD may account for as much as 80% of cases of elevated liver enzymes in the United States. Most adults with the disorders are 110–130% above ideal body weight. The prognosis of NAFLD is good if weight reduction is achieved. NASH is usually slowly progressive but can lead to cirrhosis and the need for liver transplantation in the minority of individuals affected.

In many patients, NAFLD is a component of the insulin-resistance syndrome known as the 'metabolic syndrome,' which is characterized by central obesity, hypertension, hypertriglyceridemia, low levels of high-density lipoprotein-cholesterol, and hyperglycemia. In patients with this syndrome, it is hypothesized that there is greater insulin resistance in muscles and adipose tissue than in liver. As shown in **Figure 1**, in adult patients with NAFLD, the body mass index class $>30\,kg/m^2$ is associated with an increased prevalence of each of the five components of the metabolic syndrome.

As shown in **Table 3**, compared to controls, patients with NASH exhibited a higher intake of

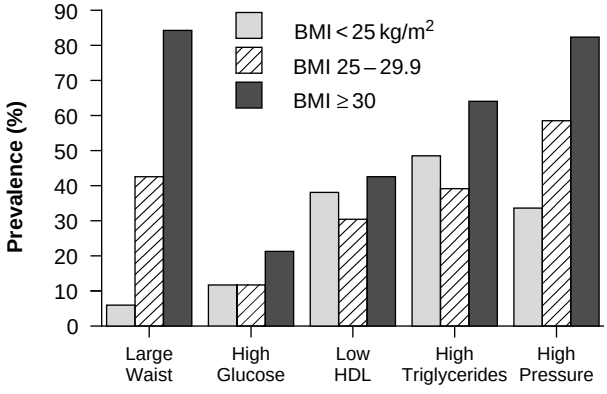


Figure 1 Prevalence of metabolic alterations fitting the criteria of the metabolic syndrome in patients with NAFLD according to classes of body mass index. (Reproduced with permission from Marchesini G, Buigianesi E, Forlani G *et al.* (2003) Nonalcoholic fatty liver, steatohepatitis, and the metabolic syndrome. *Hepatology* **37:** 917–923.)

Table 3 Daily intake of main dietary constituents in NASH patients and controls[a]

	NASH patients (n = 25)	*Controls (n = 25)*	*p value*
Total energy intake (kcal)	2638 ± 444	2570 ± 739	0.695
kcal/kg body weight	33 ± 5	32 ± 6	0.580
Dietary fat (g)	102.8 ± 31.6	92.1 ± 35.2	0.264
Dietary carbohydrate (g)	295.1 ± 53.7	315.2 ± 101.9	0.387
Dietary protein (g)	121.2 ± 25.2	107.2 ± 32.7	0.096
Alcohol (g)	13.3 ± 7.3	13.5 ± 8.9	0.705
Dietary fat (% kcal)	35.1 ± 7.1	32.3 ± 6.7	0.158
Dietary carbohydrate (% kcal)	44.7 ± 8.7	48.6 ± 9.1	0.128
Simple carbohydrate (% total carbohydrate)	30.3 ± 6.4	32.5 ± 5.1	0.185
Fiber (g)	12.9 ± 4.1	23.2 ± 7.8	0.000
Dietary protein (% kcal)	20.2 ± 3.7	16.7 ± 4.3	0.003
SFA (g)	40.2 ± 12.7	28.7 ± 11.1	0.001
MUFA (g)	52.1 ± 17.4	47.8 ± 16.7	0.377
PUFA (g)	103.9 ± 4.9	13.4 ± 4.1	0.019
Cholesterol (mg)	506 ± 108	405 ± 111	0.002
SFA (% total kcal)	13.7 ± 3.1	10.0 ± 2.1	0.000
MUFA (% total kcal)	17.7 ± 4.4	16.7 ± 5.1	0.462
PUFA (% total kcal)	3.5 ± 1.3	4.7 ± 2.0	0.015
SFA (% total fat)	39.1 ± 4.8	31.1 ± 5.2	0.000
MUFA (% total fat)	50.9 ± 6.5	51.9 ± 5.9	0.572
PUFA (% total fat)	10.0 ± 3.5	14.5 ± 4.0	0.000
(P:S ratio)	0.24 ± 0.10	0.46 ± 0.12	0.000
Vitamin A (μg)	582.6 ± 383.7	647.1 ± 507.3	0.614
Vitamin C (mg)	84.3 ± 43.1	144.2 ± 63.1	0.000
Vitamin E (mg)	5.4 ± 1.9	8.7 ± 2.9	0.000
Iron (mg)	12.1 ± 2.3	14.5 ± 3.9	0.011

[a]Data are presented as mean ± SD.
SFA, saturated fat intake; PUFA, polyunsaturated fat intake; MUFA, monounsaturated fat intake; P:S ratio, polyunsaturated to saturated fat.
From Musso R, Gambino R, DeMichiele F *et al.* (2003) Dietary habits and their relations to insulin resistance and postprandial lipemia in nonalcoholic steatohepatitis. *Hepatology* **37**(4): 909–915.

saturated fatty acids, total fat, and cholesterol and a lower intake of polyunsaturated fat, fiber, and the antioxidant vitamins C and E. These findings provide a strong rationale for specific dietary modifications in NASH patients.

Pregnancy and liver disease Liver diseases that predominantly affect females, such as PBC and autoimmune hepatitis, decrease the chances of conception and demand that pregnant women with these disorders should be managed in high-risk obstetric facilities. Liver diseases that can evolve as a consequence of pregnancy include intrahepatic cholestasis of pregnancy, acute fatty liver of pregnancy, and HELLP (hemolysis, elevated liver enzymes, and low platelets syndrome). The latter has been associated with disorders of fatty acid oxidation in offspring. Successful pregnancies are the rule for women who have undergone liver transplantation, but preterm delivery and low-birth-weight infants are common. Careful attention to the nutritional management of the pregnant female with liver disease is necessary to ensure the best outcome for the fetus.

Total parenteral nutrition-associated liver disease Premature infants and children with short gut syndrome are particularly prone to develop this disorder, and in the pediatric age group total parenteral nutrition (TPN) liver disease is usually cholestatic. The cholestasis can be solely intrahepatic or can be associated with cholelithiasis. TPN liver disease can be seen at any age and with any disease etiology resulting in long-term dependence on parenteral nutrition; in older children and adults, steatosis is more common as an initial presentation rather than cholestasis. Potential pathogenetic mechanisms include the gastrointestinal dysfunction associated with the lack of enteral nutrients as well as components of the parenteral nutrition solutions as potential hepatotoxins, including amino acids, glucose, lipids (particularly peroxidizable lipids), and photoexposed multivitamins. The most effective management is aggressive administration of enteral

nutrients and a decrease and/or discontinuation of parenteral nutrition as early as possible.

Viral hepatitis Hepatitis A virus infection never results in chronic liver disease, so there are no specific nutritional recommendations for patients with this disorder. Hepatitis B virus infection evolves to chronic hepatitis in ~95% of neonates who acquire the infection perinatally but only ~5% of adults. Hepatitis C virus (HCV) infection has a much higher rate of chronicity in adults—up to 80% of those infected will develop chronic infection. Approximately 20–30% of those will progress to cirrhosis over 10–20 years and a smaller proportion of those will develop hepatocellular carcinoma. There is much less information about nutritional disturbances and nutritional management of patients with chronic viral hepatitis than there is for patients with alcoholic liver disease.

In general, the nutritional recommendations for management of alcoholic hepatitis, cirrhosis, or cholestasis detailed previously can be applied to patients with these various manifestations of chronic viral hepatitis. For example, it has been shown that thiamine deficiency is common in patients with cirrhosis secondary to either chronic alcohol consumption or chronic HCV and thiamine supplementation is indicated for patients with either type of liver disease.

Metabolic Disorders

Galactosemia Galactosemia is secondary to the deficiency of galactose-1-phosphate uridyl transferase. Galactose-1-phosphate is toxic and accumulates in liver and other organs, causing liver failure in early infancy. The usual presentation is hypoglycemia and encephalopathy in the first few days of life. Vomiting, diarrhea, jaundice, and failure to thrive are common. Treatment is by elimination of galactose (and, consequently, lactose) from the diet for life. Liver function improves by this maneuver, but long-term complications such as mental disability, speech defects, ovarian failure, and neurologic syndromes are common despite dietary restriction.

Glycogen storage disease Glycogen storage disease (GSD) I, II, and III are the most common glycogen storage diseases to present with hepatomegaly. Type IV (amylopectinosis) is the only one of these disorders to present with cirrhosis; however, most of the disorders originate from deficiency of a key hepatic enzyme of glycogenolysis. The main nutritional management strategy, most important for types I and III, is the prevention of hypoglycemia; night time administration of cornstarch is often effective. Since restricted diets are key to the management of most inborn errors of metabolism, patients with these disorders are at high risk for nutrient deficiencies.

Hemochromatosis This disorder is among the most common autosomal recessive diseases in the world, occurring as frequently as 1/300. Two mutant alleles of the HFE gene are responsible for essentially all cases. Hepatomegaly and hepatic dysfunction as manifested by elevation of serum aminotransferases are common. Pancreatic dysfunction and darkening of the skin may occur. Transferrin saturation, serum iron concentration, and ferritin levels are the usual tests for iron overload, but molecular testing is rapidly becoming the diagnostic modality of the future. Liver biopsy may still be necessary to determine the degree of hepatic iron overload; management is by dietary iron restriction and phlebotomy.

Hepatorenal tyrosinemia I This disorder, which is secondary to deficiency of fumarylacetoacetate hydrolase, is the most common and severe of the genetic defects of tyrosine metabolism. Initial management is with a phylalanine- and tyrosine-restricted diet. The current intervention with 2-(2-nitro-4-trifluoro-methylbenzoyl)-1,3 cyclohexenedione (NTBC) has improved the quality of life of patients suffering from this disorder because it decreases the frequency of episodes of acute liver failure and coagulopathy.

Hereditary fructose intolerance This disorder results from a deficiency of the enzyme fructose-1-phosphate aldolase, which results in the accumulation of fructose-1-phosphate in the liver. This substance is a competitive inhibitor of phosphorylase, which regulates the conversion of glycogen to glucose. With inhibition of this enzyme, hypoglycemia and lactic acidosis result. The clinical presentation is with vomiting, diarrhea, choelestasis, and hepatomegaly, usually in early infancy. Renal injury and growth retardation are common. Standard treatment is restriction of fructose (and sucrose) from the diet and early treatment results in an excellent clinical outcome.

Urea cycle disorders There are six of these disorders, all of which present with varying degrees of hyperammonemia. In the neonatal period, these disorders present dramatically with somnolence, poor feeding, vomiting, lethargy, seizures, and even hyperammonemic coma. In older children and adults, the presentation may be more subtle and

begin with chronic vomiting, developmental delay, seizures, psychiatric illness, postpartum decompensation, and hyperammonemia associated with valproate therapy. Nutritional management includes restriction of upstream essential nutrients to prevent intoxication and supplementation of downstream nutrients to prevent secondary deficiency. In addition, alternative routes of disposal of precursor metabolites can be stimulated. For the severe deficiencies, including severe neonatal ornithine transcarbamylase deficiency and carbamyl phosphate synthetase deficiency, more aggressive strategies, such as liver transplantation and gene therapy, are being investigated.

Wilson's disease Wilson's disease is an autosomal recessive disorder of copper accumulation secondary to mutations in ATP7B, a copper-binding ATPase primarily expressed in the liver. The clinical expression in children and adolescents is often dramatic subacute hepatic necrosis or fulminant liver failure accompanied by hemolysis. In adults, the hepatic presentation is more subtle—manifestations of portal hypertension and cirrhosis, such as fatigue and ascites, and neuropsychiatric manifestations are common. The management of severe liver disease is liver transplantation. However, in the absence of severe liver disease, treatment is with a copper-restricted diet and copper chelating agents.

Hepatobiliary Disorders

Biliary atresia This disorder is the prototypic biliary tract disorder in infancy, accounting for ~50% of all liver transplants in the pediatric age group and ~10% of all liver transplants. It presents with cholestasis in early infancy; there is a palliative surgical procedure called the Kasai hepatic portoenterostomy that, if performed before 60 days of age, may at least delay disease progression. In ~20–25% of infants in whom the procedure is done in a timely fashion, liver transplantation may never be necessary. Severe steatorrhea and malnutrition are common and malabsorption of the fat-soluble vitamins is profound, sometimes requiring parenteral administration (particularly of vitamin K) to achieve sufficiency. Nutritional deficiency disorders such as osteoporosis are common.

Primary sclerosing cholangitis This disorder most commonly presents in association with ulcerative colitis and less commonly with Crohn's disease or as an isolated entity. The nutritional management of the disorder is essentially like that of other cholestatic disorders; in patients with Crohn's disease of the small bowel, aggressive administration of an elemental diet rich in medium-chain triglycerides may be beneficial. It is accepted, however, that endoscopic interventions should be used as needed in the case of significant biliary obstruction. For prevention of severe osteoporosis, supplementation with vitamin D and calcium is needed. Vitamin K and alendronate may be beneficial in increasing bone mineral density. Serum levels of the fat-soluble vitamins should be monitored in high-risk patients and vitamins replaced as appropriate.

Nutritional Management

Acute Liver Failure

The nutritional status of someone with acute liver failure versus chronic liver failure can differ greatly. The primary goal of the nutritional management in acute liver failure is supportive. An increase in nausea, vomiting, and anorexia may be associated with acute liver disease, which may result in decreased oral intake. If normal nutritional status prior to the insult is assumed, the patient will have a much higher nutritional reserve than that of a patient in chronic liver failure. Energy needs can be met by providing the Dietary Reference Intakes for infants and children and approximately 30 kcal/kg for adults. The provision of adequate protein is crucial in fulminant hepatic failure and encephalopathy. Adequate protein must be provided to minimize catabolism, which may exacerbate any hyperammonemia present. Excessive protein intake should be avoided because it may increase ammonia levels.

Protein recommendations for adults and teenagers are 0.5–1.0 g/kg/day and for infants and children 1.2–1.5 g/kg/day. Additional protein restrictions or an increase in the intake of branched-chain amino acids intake may be beneficial. In health, the ratio of branched-chain amino acids/aromatic amino acids (leucine + isoleucine + valine/phenylalanine + tyrosine) = ~3:1, and in liver failure the ratio may decline to ~1, often in association with some degree of hepatic encephalopathy. There are data indicating that normalization of this ratio by administration of branched-chain amino acid formulae can improve hepatic encephalopathy.

Chronic Liver Disease

Chronic liver disease is often accompanied by nutritional deficiencies. The goals of nutritional management are to provide adequate energy and protein to prevent energy deficits and protein catabolism and to

Table 4 Management of chronic liver failure in children

Nutritional support
Energy intake, 120–150% (recommended daily amount)
Carbohydrate, 15–20 g/kg/day
Protein, 3–4 g/kg/day
Fat, 8 g/kg/day (50% medium-chain triglyceride)
Fat-soluble vitamins
Fluid balance
Avoid excess sodium (<2 mmol/kg)
Ascites: spironolactone (3 mg/kg), furoseimide (0.5–2 mg/kg), albumin infusion, paracentesis
Encephalopathy
Low protein (2 g/kg)
Lactulose (5–20 ml/day)
Coagulopathy
Vitamin K (2–10 mg/day)
Fresh frozen plasma, cryoprecipitate, platelets

From Kelly DA (2002) Managing liver failure. *Postgraduate Medical Journal* **78**: 660–667.

promote hepatic cell growth. Recommendations for nutritional management of children with chronic liver disease are presented in **Table 4**. The energy need for adults with chronic liver disease is 30–35 kcal/kg/day. Energy requirements are increased to compensate for the weight loss that often occurs in cirrhosis. Protein should be provided as 0.8–1 g/kg for adults; unnecessary protein restriction should be avoided because it may only worsen total body protein losses. Energy from fat is best delivered as MCTs due to malabsorption of long-chain fatty acids. Several infant, pediatric, and adult formulas are available with a large percentage of fat in the form of MCTs.

Supplementation with fat-soluble vitamins (A, D, E, and K) in water-miscible solutions is necessary due to the potential for deficiencies associated with fat malabsorption. Serum levels should be monitored regularly to ensure appropriate levels and prevent toxicity. Supplementation with zinc, selenium, iron, and calcium should be given as needed. Copper and manganese should not be supplemented because they are excreted via the bile and may build to toxic levels. Sodium and/or fluid restrictions may be necessary in cirrhosis characterized by ascites and edema. This can impose difficulty because this restriction decreases the palatability of the diet, further decreasing oral intake.

End Stage Liver Disease Pre- and Post-Liver Transplantation

Maintaining optimal nutritional status is important in the patient with end stage liver disease both pre- and post-transplant. However, nutritional assessment in end stage liver disease is particularly problematic. In the pretransplant setting, fluid retention, ascites, and hepatosplenomegaly make body weight an unreliable nutritional index. True decreases in body weight, due to loss of fat stores and lean body mass, may not be fully appreciated solely following weight trends. In the pediatric population, linear growth is often a better indicator of nutritional status. Chronic malnutrition is often present, as reflected in a decrease in linear growth velocity.

Although anthropometric measurements, 24-h creatinine, bioelectric impedance analysis, and indirect calorimetry have all been used, they are affected by ascites and peripheral edema. *In vivo* neutron activation analysis and isotope dilution techniques are more accurate ways of assessing body composition but are time-consuming and costly. For practical purposes, the indirect assessments of 24-h urinary creatinine excretion to determine body muscle mass and mid-arm muscle area can be used for patients without high volumes of extracellular fluid; in those with ascites, the creatinine–height index is a better way of assessing body muscle mass.

Visceral proteins, including albumin, transferrin, prealbumin, and retinol binding protein, are typically used in monitoring nutritional status due to the decrease seen in inadequate dietary protein intake. However, they should be used with caution in liver disease because the synthesis of these proteins is also decreased in end stage liver disease. Serum levels of fat-soluble vitamins should be monitored closely as well.

Improving nutritional status prior to transplant is imperative because malnutrition affects morbidity and mortality post-transplant. Although it may not be possible to reverse the degree of malnutrition, aggressive nutrition support should be implemented to prevent further worsening of the nutrition state and possibly reduce pre- and post-transplant infection and complications.

Post-transplant nutrition support should not be overlooked because the nutrition deficit is not cured merely by the transplant. Additionally, the surgery poses increased nutritional demand for post-surgery healing and support. Nutrition repletion may occur at a more rapid rate than pretransplant because the patient now has a functional liver in which metabolism and digestion of macro- and micronutrients will be improved.

See also: **Celiac Disease**. **Cystic Fibrosis**. **Obesity**: Definition, Etiology and Assessment; Fat Distribution; Complications; Prevention; Treatment. **Osteoporosis**. **Pregnancy**: Role of Placenta in Nutrient Transfer; Nutrient Requirements; Energy Requirements and Metabolic Adaptations; Weight Gain; Safe Diet for

Pregnancy; Prevention of Neural Tube Defects; Pre-eclampsia and Diet. **Vitamin A**: Biochemistry and Physiological Role; Deficiency and Interventions. **Vitamin D**: Physiology, Dietary Sources and Requirements; Rickets and Osteomalacia. **Vitamin E**: Metabolism and Requirements.

Further Reading

Guyton AC and Hall JE (1996) Digestion and absorption in the gastrointestinal tract. In *Textbook of Medical Physiology*. Philadelphia: WB Saunders.

Guyton AC and Hall JE (1996) The liver as an organ. In *Textbook of Medical Physiology*. Philadelphia: WB Saunders.

Hoffmann GF, Nyhan WL, Zschocke J, Kahler SG, and Mayatepek E (2002) *Inherited Metabolic Diseases*. Baltimore, MD: Lippincott Williams & Wilkins.

Lowell JA and Shaw BW (2001) Critical care of liver transplant recipients. In: Maddrey WC (ed.) *Transplantation of the Liver*. Philadelphia: Lippincott Williams & Wilkins.

Marchesini G, Buigianesi E, Forlani G *et al.* (2003) Nonalcoholic fatty liver, steatohepatitis, and the metabolic syndrome. *Hepatology* **37**: 917–923.

Mckiernan PJ (2002) Neonatal cholestasis. *Seminars in Neonatology* **7**: 153–165.

Musso G, Bambino R, De Michieli F *et al.* (2003) Dietary habits and their relations to insulin resistance and postprandial lipemia in nonalcoholic steatohepatitis. *Hepatology* **37**: 909–916.

Rigby SH and Schwarz KB (2001) Nutrition and liver disease. In *Nutrition in the Prevention and Treatment of Disease*. San Diego: Academic Press.

Roberts EA (2002) Steatohepatitis in children. *Best Practice and Research Clinical Gastroentrology* **16**: 749–765.

Sandhu BS and Sanyal AJ (2003) Pregnancy and liver disease. *Gastroenterology Clinics of North America* **32**(1): 407–436.

Stickel F, Hoehn B, Schuppan D, and Seitz HK (2003) Review article: Nutritional therapy in alcoholic liver disease. *Alimentary Pharmacology & Therapeutics* **18**: 357–373.

LOW BIRTHWEIGHT AND PRETERM INFANTS

Contents
Causes, Prevalence and Prevention
Nutritional Management

Causes, Prevalence and Prevention

M Merialdi and M de Onis, World Health Organization, Geneva, Switzerland

It is widely accepted that weight at birth is a key indicator of fetal and neonatal health, both for individuals and for populations. The strong association between low birthweight and perinatal mortality and morbidity is now well recognized by health care providers, as are the different determinants and health consequences of low birthweight. These epidemiological associations became progressively evident during the past century. In the United States, the practice of weighing infants at birth was introduced at the end of the nineteenth century when low birthweight infants were categorized as 'premature' and usually left unattended with minimal or no intervention attempted to prevent their deaths. When information on birth weight and gestational age was introduced in the birth certificate in mid-twentieth century, it became apparent that prematurity was the most important cause of infant deaths at the national level.

With progressive awareness of the importance of low birthweight as a predictor of infant mortality, it appeared that being born small could be due either to a restriction of the normal process of fetal growth or to delivery before the term of gestation. Thus, the World Health Organization (WHO) made a distinction between the condition of low birthweight (birth weight less than 2500 g) and prematurity (delivery at less than 37 completed weeks, i.e., 259 days). A further development was the introduction of the concept of small for gestational age (SGA) that better describes infants affected by intrauterine growth restriction (IUGR). According to this classification, infants with birth weight below the 10th percentile of a reference population are considered SGA. Although these distinctions and definitions are commonly applied in developed countries, their use is

more difficult in developing countries where information on gestational age is often nonexistent or unreliable. This is unfortunate because low birthweight conditions due to growth restriction or preterm birth have different determinants and prognosis, as well as different epidemiological distributions that vary by country and socioeconomic status. Thus, before discussing the causes, prevalence, and prevention of low birthweight, it is important to understand how its two components (gestational age and fetal growth) can be correctly identified and quantified for epidemiological purposes and what are the major limitations in doing so.

Assessment of Gestational Age and Fetal Growth: Methods and Limitations

Preterm birth is defined as delivery before 37 completed weeks (259 days). To accurately differentiate between preterm and term delivery it is crucial to have a reliable estimate of gestational age. Sonographic determination is the most accurate method to estimate gestational age. When ultrasonography is not available, gestational age can be determined by patient's recall of the time of last menstrual period, physical examination of the size of the uterus, and examination of the neonate. These methods can be used alone or in combination but are inaccurate.

Early pregnancy sonographic estimation of gestational age is crucial also for estimation of fetal growth *in utero*, which is assessed by evaluating the size of several fetal anatomical parameters and comparing those measurements with the normal ranges at specific gestational ages obtained from reference populations with growth that can be considered unaffected by pathological conditions. Alternatively, fetal growth can be assessed by the anthropometrical evaluation of the neonate. Several classification systems have been proposed for newborn birth weight. The simplest is categorizing newborns <2500 g as having a low birthweight, but this classification does not differentiate between infants born small for their gestational age and infants who are small because they are born preterm. Reference charts of birthweight at different gestational ages classify infants as SGA, a proxy for IUGR; adequate for gestational age; and large for gestational age. WHO defines SGA as a birth weight below the 10th percentile for a given gestational age based on the sex-specific reference by Williams *et al*. Because it is based on percentile distributions, this classification categorizes some normal, constitutionally small, newborns at the lower end of the normal fetal growth distribution as growth restricted. In addition, the interpretation of the reference data is complicated by inaccuracies in the estimation of gestational age at delivery and by the pathological processes that may affect the size of infants born early in gestation.

Causes

Recognizing that low birthweight may be due to either IUGR or preterm delivery, and, in some cases, a combination of the two, the scientific community has progressively started to consider that IUGR and preterm delivery are two conditions likely caused by various and possibly independent etiopathological factors.

Several complications of pregnancy, such as preeclampsia, fetal distress, fetal growth restriction, abruptio placenta, fetal death, placenta previa, and multiple gestations, are associated with preterm delivery, either spontaneous or induced. Importantly, developments in obstetric and neonatal care, and the consequent increase in obstetric interventions, are likely to be associated with the increase in rates of preterm delivery observed in recent years. Although several lifestyle factors and conditions have been implicated as possible causes, a definitive etiology has not been determined, making it difficult to identify women at risk and to implement preventive strategies. Poor nutrition, cigarette smoking, and alcohol and drug abuse have been indicated as possible risk factors, as well as young maternal age, poverty, short stature, occupational factors, and psychological stress. In addition, genetic factors are likely to be involved in the etiopathogenesis of preterm delivery, as suggested by the fact that the condition tends to recur in families and that prevalence varies across races. The possible role of infection in triggering preterm delivery has been suggested by several studies that have shown associations between delivery before term and amniotic fluid and chorioamniotic infection, bacterial vaginosis, genitourinary clamydial infection, and periodontal disease. Despite the biological plausibility of these associations, their causal relationship has not been definitely proved by unequivocal scientific and epidemiological evidence.

Several conditions have been associated with intrauterine growth restriction. However, present knowledge of the process of fetal growth is limited by the difficulty of differentiating between constitutional and environmental determinants of fetal growth. This limitation complicates the investigation of an important determinant of fetal growth such as maternal size. Small women tend to have smaller

babies. There is evidence that intergenerational effects on birth weight are transmitted through the maternal line, thus suggesting a genetic effect. However, poor maternal nutrition and social deprivation have been related to impaired fetal growth and may also be related to small maternal size. Similarly, the relationship between fetal size and race may be mediated by genetic and environmental factors. Specifically designed studies are necessary to determine the contribution of genetic and environmental determinants to the process of fetal growth.

Other factors that have been associated with fetal growth restriction are fetal infections, congenital malformations, chromosomal abnormalities, chemical teratogens, vascular disease such as preeclampsia, chronic renal disease, chronic hypoxia, placental and cord abnormalities, and multiple fetuses.

Health Consequences

Low birthweight, due to either preterm delivery or IUGR, is associated with increased neonatal mortality. Mortality tends to increase with decreasing gestational age at delivery and birth weight (**Figures 1 and 2**). Preterm delivery is the most important obstetric complication in developed

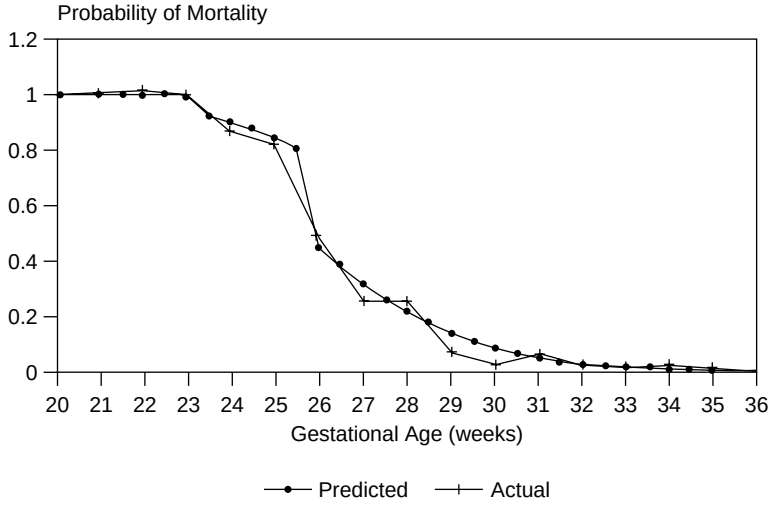


Figure 1 Probability of neonatal mortality as a function of gestational age at delivery in 3386 births between 20 and 37 weeks. The predicted mortality curve is smoothed using statistical methods. Reproduced with permission from Copper RL *et al.* (1993) A multicentre study of preterm birth weight and gestational age-specific neonatal mortality. *American Journal of Obstetrics and Gynecology* **168**: 78–84.

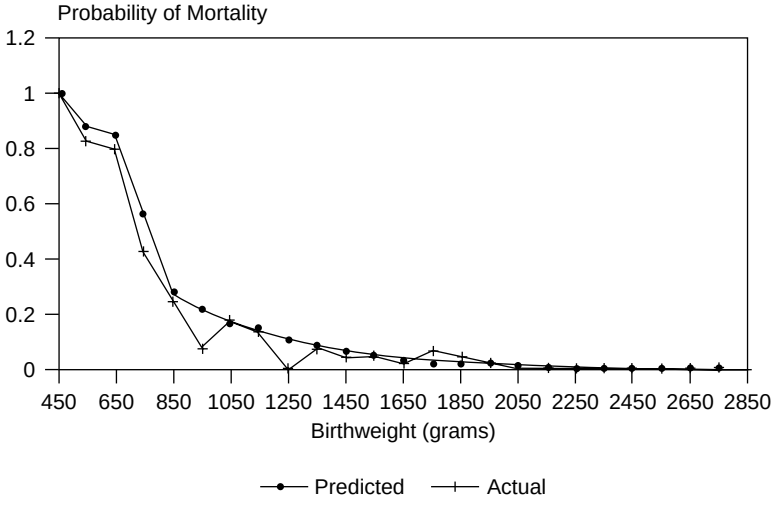


Figure 2 Probability of neonatal mortality as a function of birthweight in 3386 births between 20 and 37 weeks. The predicted mortality curve is smoothed using statistical methods. Reproduced with permission from Copper RL *et al.* (1993) A multicentre study of preterm birth weight and gestational age-specific neonatal mortality. *American Journal of Obstetrics and Gynecology* **168**: 78–84.

countries and, together with IUGR, a major cause of neonatal deaths both in developed and in developing countries. However, the burden of disease of preterm delivery and IUGR in terms of neonatal death is disproportionally heavier for developing countries. Ninety-eight percent of neonatal deaths occur in developing countries, and they account for 33% of all deaths of children younger than 5 years of age. Importantly, 50% of all neonatal deaths are due to being born underweight.

Morbidity is also increased in low birthweight newborns and the negative effects of preterm delivery and/or IUGR tend to persist in infancy as deficits in growth and neurological development. In addition, there is evidence that the negative effects of IUGR may persist long after infancy because low birthweight has been associated with the development of cardiovascular disease, high blood pressure, obstructive lung disease, diabetes, high cholesterol concentrations, and renal damage in adulthood.

Epidemiology

Table 1 shows estimates of prevalence of low birthweight, IUGR, and preterm delivery from three major multicountry studies: the WHO Antenatal Care Trial, the WHO Collaborative Study on Pregnancy Outcomes, and the WHO Misoprostol Trial. The data presented demonstrate that the distribution of low birthweight varies across populations, and the observed differences are due to the varying contribution of preterm delivery and IUGR to the total rates of low birthweight. Rates of low birthweight are higher in developing than in developed countries, as shown by the data presented in **Table 2**. Data from 11 developed areas and 25 developing ones indicate that rates of low birthweight steadily increase as the level of development decreases. The observed differences in low birthweight rates between geographical areas are enormous, with rates ranging from as low as 3.6% in Sweden to as high as 43% in Mumbai, India. Importantly, a global review of the magnitude of the problem demonstrates that many countries currently exceed the internationally recognized cutoff levels for triggering public health action (IUGR >20% and low birthweight (LBW) >15%). Most of low birthweight infants born in developed countries are the result of preterm delivery, whereas in developing countries they are more likely to be IUGR infants. In addition, differences in the distribution of low birthweight are observed at country level, as indicated by the higher prevalence of low birthweight (due mainly to preterm delivery) among African Americans compared to other ethnic groups in the United States.

Prevention

Results from clinical trials provide the most powerful scientific evidence to guide policy and programmatic public health strategies. Interventions aimed at preventing low birthweight either acting toward preterm delivery or IUGR have usually not proven to be effective by randomized clinical trials. The multicausal nature of these conditions is likely responsible for the fact that single interventions do not show an effect of enough magnitude to be detected by medium-sized clinical trials. Thus, appropriate combinations of interventions should be a priority for evaluation in the context of large, methodologically sound trials. Evidence shows that some interventions may be effective and their combined implementation may have a significant public health impact. Interventions likely to be beneficial in preventing IUGR are smoking cessation, antimalarial chemoprophylaxis in primigravide, and balanced protein energy supplementation. Treatment of urinary tract infection, placement of circumferential stitches on a structurally weak uterine cervix (cerclage), and treatment of bacterial vaginosis in high-risk women have been shown to be effective in preventing preterm birth. Unfortunately, these interventions are applicable only to a small number of high-risk women, and their overall effect in the general population is likely to be limited.

In the following sections, nutritional interventions to prevent preterm delivery and IUGR are reviewed with the aim of identifying potentially effective interventions and suggesting possible mechanisms that may link maternal nutritional status to low birthweight. The focus is on the review of randomized clinical trials that provide the most unbiased epidemiological evidence on the effectiveness of interventions. Clinical trials testing the same or similar interventions can be pooled together to estimate an overall effect by means of a systematic review of published and unpublished studies and the meta-analysis of the trials' results.

Nutritional Interventions to Prevent Preterm Delivery

Of the nutritional interventions conducted during pregnancy that have been tested by clinical trials to prevent preterm delivery, only calcium and fish oil supplementation appear promising. In addition, nutritional advice and magnesium supplementation are likely to be effective; however, methodological problems in the analysis of the trials' results prevent from drawing definitive conclusions. Most of the other

Table 1 Prevalence of LBW, IUGR, and preterm delivery in different countries

Country	*Location*	*Years*	*Sample size*	*LBW (% of live births)*	*IUGR (% of live births)*	*Preterm (% of live births)*	*Data source*
Argentina	City of Rosario	1996–1998	6,789	7.0	9.0	9.7	ANC
Argentina	City of Rosario	1984–1986	5,634	6.3	9.7	7.2	CSPO
Argentina	City of Rosario	1998–1999	2,709	4.0	6.0	6.3	MiT
China	Six subdistricts of Nanshi in Shanghai	1981–1982	4,753	4.2	9.4	7.5	CSPO
China	Shanghai	1998–1999	2,195	3.5	5.3	5.4	MiT
Colombia	City of Cali	1989	4,598	16.1	17.8	15.7	CSPO
Cuba	Havana	1996–1998	5,573	6.9	14.4	5.2	ANC
Cuba	Mixed urban and rural centres	1981	4,779	8.1	14.7	7.2	CSPO
Gambia	Keneba village	1976–1984	379	12.1	13.5	13.5	CSPO
Guatemala	Four highland rural villages	1969–1977	286	12.5	25.3	15.8	CSPO
India	Pune	1990	4,307	28.2	54.2	9.7	CSPO
Indonesia	City of Bogor and surrounding villages	1983	1,647	10.5	19.8	18.5	CSPO
Ireland	Dublin	1979–1980	6,424	5.6	6.9	6.2	CSPO
Ireland	Dublin	1998–1999	447	3.1	7.8	5.1	MiT
Malawi	Three rural communities	1986–1989	938	11.6	26.1	8.2	CSPO
Myanmar	Communities in rural and urban areas	1981–1982	3,542	17.8	30.4	24.6	CSPO
Nepal	Rural areas	1990	NA	14.3	36.3	15.8	CSPO
Saudi Arabia	Jeddah	1996–1998	3,923	7.5	17.1	7.9	ANC
Sri Lanka	Rural areas	1990	1,851	18.4	34.0	14.0	CSPO
Switzerland	Zurich	1998–1999	353	1.1	5.6	2.5	MiT
Thailand	Khon Kaen province	1996–1998	6,289	8.3	20.9	8.0	ANC
Thailand	Rural and urban centres	1979–1980	4,124	9.6	17.0	21.3	CSPO
Thailand	Khon Kaen province	1998–1999	1,816	5.5	13.3	7.5	MiT
United Kingdom	Aberdeen	1971–1976	4,803	6.2	12.3	4.6	CSPO
USA/CDC (Black)	17 states and District of Columbia	1989	4,614	10.6	11.2	16.6	CSPO
USA/CDC (Hispanic)	17 states and District of Columbia	1989	2,205	4.8	5.8	10.2	CSPO
USA/CDC (White)	17 states and District of Columbia	1989	16,481	6.0	6.9	9.3	CSPO
Vietnam	City of Hanoi and one rural district	1982–1984	4,428	5.2	18.2	13.6	CSPO
Vietnam	Ho Chi Min City	1998–1999	3,001	11.1	27.8	6.2	MiT

ANC, WHO Antenatal Care Trial; CSPO, WHO Collaborative Sudy on Pregnancy Outcomes; IUGR, intrauterine growth restriction; LBW, low birthweight; MiT, WHO Misoprostol Trial; NA, not available.

Table 2 Low birthweight rates in developed and developing areas

Location	*Year*	*No. of newborn infants*	*Incidence of LBW (≤2500 g)*	*Proportion of IUGR–LBW (≤2500 g, ≥37 weeks) (%)*	*Source of gestational age other than LMP*
Developed areas					
Sweden	1973	107,717	3.6	46	
Japan	1973	206,629	4.7	62	
Finland	1966	11,931	4.2	24	
Italy	1965	963,653	4.3	49	
New Zealand	1976	59,568	4.9	50	
Austria	1973	48,758	5.7	35	
Canada	1977	348,000	6.4	64	
United Kingdom	1970	16,815	6.9	57	
United States	1977	3,148,910	7.4	45	
Hungary	1973	152,996	10.8	28	
United States	1968–1974		6.0	30	
Developing areas					
Bogota, Colombia	1979	407	9.8	65	Early pregnancy detection[a]
Rosario, Argentina	1975	689	10.0	50	Neonatal examination[b]
Cuba	1973	208,503	10.1	38	
Tanga, Tanzania	1966	1,000	11.2	56	Neonatal examination
Nairobi, Kenya	1971	3,160	13.6	34	Neonatal examination
Eastern Guatemala	1977	1,276	16.0	73	Prospective amenorrhea detection[c]
Dominican Republic	1975–1976	304	17.8	68	
Nairobi, Kenya	1974	3,700	18.9	34	Neonatal examination
Johannesburg, South Africa	1971–1972	1,800	19.5	73	Neonatal examination
Colombo, Sri Lanka	1971	1,988	21.0	76	Neonatal examination
Vellore, India	1971–1972	2,626	22.1	76	Neonatal examination
Central Guatemala	1977	1,000	23.0	77	Neonatal examination
Jakarta, Indonesia	1967–1968	2,210	23.1	72	
New Delhi, India	1964–1966	2,273	24.1	79	
Tirupati, India	1968–1969	1,000	24.7	73	
Ibadan, Nigeria	1973–1974	1,290	24.9	43	Neonatal examination
Berhampur, India	1972	986	26.3	65	
New Delhi, India	1968–1969	4,100	30.0	78	
10 states, India	1969	10,739	30.5	77	
Rajasthan, India	1970	1,651	30.7	71	Neonatal examination
Tamil Nadu, India	1969–1975	4,420	31.8	77	Prospective amenorrhea detection
Hyderabad, India	1969	846	33.1	88	Neonatal examination
Pondicherry, India	1970–1971	1,279	34.0	66	Neonatal and amniotic fluid examination
Western Guatemala	1964–1972	415	41.6	83	Prospective amenorrhea detection
Mumbai, India	1966–1967	10,000	43.0	96	

[a]During the first trimester and follow-up during pregnancy to validate gestational age.
[b]The use of one or more recognized methods to measure neonatal maturation.
[c]Periodic home visits to detect the first lack of menstruation.
IUGR, intrauterine growth restriction; LBW, low birthweight.
From Villar J and Belizán JM (1982) The relative contribution of prematurity and fetal growth retardation to low birthweight in developing and developed societies. *American Journal of Obstetrics and Gynecology* **143**: 793–798.

interventions that have been hypothesized to have potential to prevent preterm delivery, such as protein and energy supplementation, protein and energy restriction, salt restriction, iron and/or folate supplementation, zinc supplementation, and vitamin A supplementation, have not been proved to be effective.

Nutritional Interventions to Prevent IUGR

Among the interventions that have been tested by randomized clinical trials to prevent IUGR, balanced energy protein supplementation has been shown to reduce the risk of SGA by approximately 30%. On

the basis of these results, it has been proposed that universal balanced energy supplementation should be provided to women in areas with a high prevalence of maternal undernutrition to prevent impaired fetal growth. There is evidence that magnesium supplementation and calcium supplementation may be effective, even though for the latter it is not clear if the observed effect on low birthweight is due to a direct effect on fetal growth or mediated by a prolongation of gestational age at delivery. Other interventions, such as nutritional advice, energy protein restriction, salt restriction, iron and/or folate supplementation, fish oil supplementation, zinc supplementation, and vitamins E, C, and D supplementation, did not show any effect in preventing IUGR. Interestingly, high protein supplementation in women of low socioeconomic status in the United States has been associated with an increase in the rate of SGA infants, suggesting that nutritional supplementation may, in some cases, have potentially harmful effects.

Conclusion

Low birthweight, due to either preterm delivery or IUGR, represents a major public health problem for developing and developed countries. In developed countries, access to adequate obstetrics and neonatal care prevents most of the negative short- and long-term outcomes associated with low birthweight that are observed in developing countries in terms of both mortality and morbidity. Thus, public health efforts should be aimed at improving the level and accessibility of health care in developing countries. This is particularly important because most preventive strategies have been shown by clinical trials to be ineffective. Among nutritional interventions to prevent low birthweight, only balanced energy protein supplementation has been shown to be effective in reducing the risk of SGA and has been proposed to be provided to women in areas with a high prevalence of maternal undernutrition.

Research efforts should focus on the determination of the etiological factors responsible for preterm delivery and IUGR. Despite the considerable burden of disease related to these conditions, very little progress has been made in identifying their causes, thus limiting the possibility to implement effective preventive and primary care therapeutic interventions that would particularly benefit the populations of developing countries with limited access to secondary and tertiary health care.

See also: **Infants**: Nutritional Requirements; Feeding Problems. **Low Birthweight and Preterm Infants**: Nutritional Management. **Pregnancy**: Role of Placenta in Nutrient Transfer; Nutrient Requirements; Safe Diet for Pregnancy; Dietary Guidelines and Safe Supplement Use; Pre-eclampsia and Diet. **Supplementation**: Dietary Supplements.

Further Reading

Barker DJ (1992) *Fetal and Infant Origins of Adult Disease*. London: BMJ.

Battaglia FC and Lubchenco LO (1967) A practical classification of newborn infants by weight and gestational age. *Journal of Pediatrics* **71**: 159–163.

Cunningham FG, Gant NF, Leveno KJ *et al.* (2001a) Preterm birth. In: *Williams Obstetrics*, 21st edn., pp. 689–727. New York: McGraw-Hill.

Cunningham FG, Gant NF, Leveno KJ *et al.* (2001b) Fetal growth disorders. In: *Williams Obstetrics*, 21st edn., pp. 743–764. New York: McGraw-Hill.

de Onis M, Blossner M, and Villar J (1998) Levels and patterns of intrauterine growth retardation in developing countries. *European Journal of Clinical Nutrition* **52**(supplement 1): S5–S15.

Goldenberg RL and Rouse DJ (1998) Prevention of premature birth. *New England Journal of Medicine* **339**: 313–320.

Gulmezoglu M, de Onis M, and Villar J (1997) Effectiveness of interventions to prevent or treat impaired fetal growth. *Obstetrical and Gynecological Survey* **52**: 139–149.

Kramer MS and Victora CG (2001) low birthweight and perinatal mortality. In: Semba RD and Bloem MW (eds.) *Nutrition and Health in Developing Countries,* pp. 57–70. Totowa, NJ: Humana Press.

Merialdi M, Carroli G, Villar J *et al.* (2003) Nutritional interventions during pregnancy for the prevention or treatment of impaired fetal growth: An overview of randomized controlled trials. *Journal of Nutrition* **133**: 1626S–1631S.

Villar J and Belizan JM (1982) The relative contribution of prematurity and fetal growth retardation to low birthweight in developing and developed societies. *American Journal of Obstetrics and Gynecology* **143**: 793–798.

Villar J, Merialdi M, Gulmezoglu AM *et al.* (2003) Nutritional interventions during pregnancy for the prevention or treatment of maternal morbidity and preterm delivery: An overview of randomized controlled trials. *Journal of Nutrition* **133**: 1606S–1625S.

World Health Organization (1961) *Public Health Aspects of Low Birth Weight. WHO Expert Committee on Maternal and Child Health*. Geneva: WHO.

World Health Organization (1995) The newborn infant. In *Physical Status: The Use and Interpretation of Anthropometry. WHO Expert Committee on Physical Status.* Geneva: WHO.

Nutritional Management

J M Cox, Johns Hopkins Hospital, Baltimore, MD, USA

Introduction

Thanks to advances in modern medicine and technology the outcome of preterm infants has improved dramatically. Many infants are now surviving who are born as young as 23 weeks' gestation and as small as 450 g. These infants enter life with their maternal nutrient supply abruptly disconnected and with only minimal nutrient stores. There is no other time in the life cycle when nutrition is more crucial. Additionally, nutrition in this early neonatal period may have an impact upon health throughout life.

These infants are vulnerable to poor growth and abnormal developmental outcome if not nourished appropriately. Since the preterm infant lacks the ability to voluntarily consume and process nutrients, all of the infant's needs must be provided through enteral and frequently parenteral nutrition. Preterm infants have numerous nutritional risk factors. Nutrient stores are accumulated during the third trimester; therefore, preterm infants have low energy reserves as well as minimal reserves of other nutrients. In fact, infants with birth weights less than 1000 g have energy reserves of less than 200 kcal kg^{-1} (836 kJ kg^{-1}). The metabolic rate of the preterm infant is elevated due to the predominance of metabolically active tissue and minimal fat stores. Protein, fat, and glucose needs are very high to provide adequate energy for metabolism, fat deposition, and growth. The preterm infant has excessive evaporative losses and increased urinary losses, which greatly increase fluid needs. The gastrointestinal tract of the preterm infant is very immature with minimal production of enzymes and growth factors, poor gastric emptying, and discoordinated peristalsis. To further complicate the provision of nutrients, preterm infants have episodes of metabolic instability including hypo- and hyperglycemia, poor lipid clearance, and electrolyte disturbances. The preterm infant also has high rates of stressful events including respiratory distress, hypoxemia, hypercarbia, and sepsis.

Usually, the goal is to provide sufficient nutrients to achieve the fetal growth rate. However, since the fetus and newborn differ in both physiology and metabolism this may not be an appropriate goal and in actuality this goal is rarely achieved both in regards to growth as well as body composition. Most preterm and low-birth-weight infants show significant delays in growth due to the inability to provide adequate nutrients especially in the first few weeks following birth. Over the past several years, improvements in neonatal management and a more aggressive approach to nutrition have accelerated growth but it still lags behind the fetal growth rate. However, the growth potential of preterm infants may actually be greater than even that of the normal infant. Growth velocity in the infant is greatest between 25 and 30 weeks' gestation, greater than at 40 weeks. If the infant is undernourished during this key growth period, adequate catch-up growth may never be achieved. Protein and energy are the key nutrients for growth, but they must be provided in appropriate proportions for the optimal utilization of both. Vitamins, minerals, and electrolytes must also be supplied in adequate amounts and proportions to contribute to growth. During the first few weeks after birth most preterm infants are usually undernourished due to instability so that once stability is achieved an increased supply of nutrients may be necessary to achieve catch-up growth. Nutrients are usually supplied parenterally in the initial period, gradually transitioned to a combination of parenteral and enteral nutrition, and finally when stability has been achieved full enteral nutrition.

Energy Needs

Most estimates of energy expenditure in preterm infants have been done using indirect calorimetry in relatively healthy infants weighing greater than 1200 g. The energy needs of the smaller and more preterm infant may differ somewhat. Basically, energy needs of the preterm infant involve several energy-requiring functions. Resting metabolic rate accounts for the greatest percentage of energy needs. Resting metabolic rate is equivalent to basal metabolic rate plus some of the energy used for growth; estimates have ranged from 45 to 60 kcal kg^{-1} day^{-1}(188–251 kJ kg^{-1} day^{-1}). The energy cost of activity ranges between 2 and 12% of the total energy expenditure. The smaller more premature infants are probably at the lower end of the range while the larger less preterm infant has increased activity and therefore a higher expenditure. Although preterm infants are cared for in a thermoneutral environment, there is, nevertheless, energy lost to thermoregulation during nursing care and medical procedures. There may also be energy lost to thermoregulation in a stable growing infant during bathing, feeding, and when weaned to a bassinette. The energy cost of growth includes that needed for

Table 1 Energy needs of the growing preterm infant

Energy factor	*kcal kg^{-1}day^{-1}*	*kJ kg^{-1}day^{-1}*
Resting metabolic rate	45–60	188–250
Activity	10–15	42–63
Thermoregulation	10	42
Thermic effect of food	8	33
Fecal losses	12	50
Growth	25	105
Total	**110–130**	**460–545**

tissue synthesis as well as the energy stored in tissues. The estimates for growth needs vary widely probably depending on the composition of weight gain in the infant. For the enterally fed infant the thermic effect of food and fecal losses also contribute to total energy need. The total energy needs of the growing preterm infant are summarized in **Table 1**.

Parenteral Nutrition

Since the gastrointestinal tract of the preterm infant is immature, substantive enteral nutrition is not possible in the first 2–3 weeks after birth, especially in those infants whose birth weights are less than 1500 g; therefore, the preterm infant is dependent on intravenous fluid for the bulk of fluid needs. Parenteral nutrition (PN) is basically the infusion of a nutrient solution into the circulation. Its development has allowed for the provision of nutrients during the time that enteral nutrition cannot meet nutrient needs. The use of PN has reduced the catabolism that occurs until full enteral nutrition can be achieved. PN should definitely be considered in infants whose birth weights are $<$1500 g and/or gestational age $<$30 weeks. It may also be needed for the infant whose birth weight is between 1500 and 2000 g and/or gestational age 30–32 weeks especially if the initiation or progression of enteral feeding is likely to be prolonged.

Historically, parenteral nutrition was delayed for several days after birth, probably due to metabolic instability of the infant and concern for tolerance of the components in the solution. More recently, the early use of PN has been recommended within 24 h after birth. This practice minimizes the interruption of nutrient delivery and the catabolism that occurs when only dextrose solutions are infused.

Parenteral nutrition can be administered by two different routes. There are both risks and benefits associated with each route. In the early days of parenteral nutrition it was always infused via an indwelling, surgically placed catheter into a central vein. Since some of the complications with this method were related to the catheter, the use of peripheral veins for infusion became popular and is still employed today. The dextrose concentration of peripheral PN is limited to ~10%; thus, the nutrient intake by this route is somewhat limited without excessive fluid intake. Peripheral parenteral nutrition is usually recommended when its use will be of short duration. While peripheral lines are considered less risky complications can occur. If the intravenous line infiltrates some infants have experienced serious deep sloughing, sometimes requiring skin grafts. These lines require vigilance on the part of nursing to prevent infiltrates and some infants will have multiple intravenous attempts daily because the line needs to be replaced. The advent of the percutaneously inserted central catheter and its liberal use in the last few years has improved and stabilized the delivery of parenteral nutrition to the preterm infant. Central parenteral nutrition is recommended when it is anticipated that it will be used for $>$5–7 days, usually in infants weighing $<$1000–1250 g. If the infant tolerates glucose and clears lipids well it is possible to meet estimated nutrient needs using this route. Complications such as pneumothorax, pleural effusions, and increased risk of sepsis are associated with central lines (**Table 2**).

Components of Parenteral Nutrition

Parenteral nutrition solutions contain dextrose, amino acids, lipids, electrolytes, vitamins, and minerals.

Glucose Glucose, provided as a dextrose solution, is the predominant energy source in PN. It is the main energy substrate for the fetus as well as the

Table 2 Risks and benefits of parenteral nutrion routes

Peripheral	*Central*
Adequate for short-term use	Recommended when PN needed $>$7 days
Dextrose limited to 10–12.5%	Requires placement of central line/PICC line
Can provide 80–85 kcal kg^{-1} day^{-1} if adequate fluid available	Able to meet estimated needs if adequate fluid available
Possible complications	**Possible complications**
Intravenous line can infiltrate and cause deep skin sloughing	Sepsis
Requires nursing vigilance to care for intravenous line	Line complications (pleural effusions, pneumothorax)
Can require multiple intravenous attempts	

neonate after birth. Preterm infants often require more glucose than the term infant secondary to the higher brain to body weight ratio and the need for additional energy for central nervous system energy requirements. Measurements of glucose utilization in the preterm infant range from 6 to 10 $mg\,kg^{-1}\,min^{-1}$(0.033–0.055 $mmol\,kg^{-1}\,min^{-1}$). Glycogen stores are very limited in the preterm infant; therefore, it requires a large and continuous source of glucose. This should be initiated at a rate of ~6 $mg\,kg^{-1}\,min^{-1}$ (0.033 $mmol\,kg^{-1}\,min^{-1}$) and can be advanced 1–2 $mg\,kg^{-1}\,min^{-1}$ (0.0055–0.011 $mmol\,kg^{-1}\,min^{-1}$) each day to an optimum of 12–14 $mg\,kg^{-1}\,min^{-1}$ (0.066–0.78 $mmol\,kg^{-1}\,min^{-1}$) as long as the infant does not become hyperglycemic. Above this rate, glucose is not used for energy but rather fat deposition. This is an inefficient process that can result in increased energy expenditure and carbon dioxide production.

Difficulties with glucose metabolism are a common problem in preterm infants. This may be due to decreased energy stores, increased gluconeogenesis secondary to stress, decreased insulin secretion, or insulin resistance. When hyperglycemia occurs the glucose infusion rate should be decreased, however the rate should not be decreased below 4–6 $mg\,kg^{-1}\,min^{-1}$ (0.022–0.33 $mmol\,kg^{-1}\,min^{-1}$) as this is the minimum supply rate necessary to provide adequate energy to the brain. Usually, the infusion of amino acids improves glucose tolerance by decreasing glucose production, stimulating insulin secretion, and enhancing insulin action. The use of continuous insulin infusions to treat hyperglycemia is controversial. If used, the insulin infusion should be initiated at a rate of 0.05 $U\,kg^{-1}\,h^{-1}$ and titrated to achieve and maintain a plasma glucose concentration between 80 and 120 $mg\,dl^{-1}$ (4.44–6.66 $mmol\,l^{-1}$).

Protein The early administration of protein in the form of crystalline amino acids to the preterm infant is one of the changes to have occurred over the last decade. Early studies of amino acid administration in preterm infants in the 1960s and 1970s raised the concern for protein toxicity because these infusions were associated with acidosis, azotemia, and hyperammonemia; this caused a delay in the routine administraton of protein. However, the above conditions were probably the result of the preparations being casein or fibrin hydrolysates and of suboptimal quality. Since the 1980s crystalline amino acid solutions have been used. In the late 1980s amino acid solutions specifically for use in infants were designed to produce a plasma amino acid level comparable to that of a postprandial breast-fed infant.

The early administration of amino acids is crucial because studies have shown that the preterm infant suffers protein losses of between 0.8 and 1.2 $g\,kg^{-1}\,day^{-1}$. A number of studies have demonstrated that the infusion of amino acids along with glucose decreased protein catabolism. As little as 1–1.5 $g\,kg^{-1}\,day^{-1}$ of amino acids have been shown to prevent negative nitrogen balance. Studies have also shown that the infusion of 3 $g\,kg^{-1}\,day^{-1}$ within the first 2 days of life resulted in increased protein synthesis, suppressed protein breakdown, and produced plasma aminograms similar to the breast-fed infant.

The provision of adequate energy is needed for protein metabolism and deposition. Most infants can achieve positive nitrogen balance at 2 $g\,kg^{-1}\,day^{-1}$ of protein intake when given 50–60 $kcal\,kg^{-1}\,day^{-1}$ (209–251 $kJ\,kg^{-1}\,day^{-1}$) of energy. Additionally, approximately 22 kcal (92 kJ) per g protein (15–20% of kcal) results in reasonable amino acid utilization.

Therefore, protein should be started if possible on the first day of life at 1.5–2 $g\,kg^{-1}\,day^{-1}$ and advanced to 3.5–4 $g\,kg^{-1}\,day^{-1}$ to achieve *in utero* accretion rates.

Cysteine The amino acid cysteine is a conditionally essential nutrient in the preterm infant because they have low cystathionase activity. Cystathionase, an enzyme, is necessary to convert methionine to cysteine. However, this amino acid is unstable in liquid solutions so commercially available crystalline amino acid solutions do not contain cysteine. Plasma levels of cysteine are low in infants receiving cysteine-free PN. Cysteine hydrochloride however is soluble and is stable in aqueous solutions for a short period of time so 40 $mg\,g\,protein^{-1}$ is often added to PN solutions when prepared. The addition of cysteine may result in acidosis necessitating an increase in acetate. However, an additional advantage is that the addition of cysteine decreases the pH of the PN solution, which allows the addition of more calcium and phosphorous.

Lipids Lipids are the most concentrated source of calories in the PN solution. They are available as lipid emulsions of soy bean and safflower oil; 20% emulsions are recommended for use because they contain less phospholipid than the 10% emulsion. Lipids are critical for central nervous system development. Additionally, when infused with the PN solution they may also prevent phlebitis. Lipids are usually infused to prevent essential fatty acid deficiency and as an

energy source. Maximum lipid clearance occurs when lipids are infused over 24 h. Starting recommendations vary; but it is generally accepted to start with 0.5–1 g kg^{-1} day^{-1} between 1–3 days of life. Lipids should be advanced to an optimum of 3 g kg^{-1} day^{-1}. Studies have shown that preterm infants have optimal protein retention when approximately 30–40% of calories are provided as lipids. Plasma triglycerides can be used to monitor lipid clearance. It is generally accepted that levels below 150–200 mg dl^{-1} indicate adequate clearance. Lipoprotein lipase and hepatic lipase are the major enzymes for clearance of intravenous lipid. These activities are inducible by low-dose heparin, which is usually present in central PN solutions. Administration of heparin should be considered in those infants receiving peripheral PN showing poor lipid clearance. In infants with hypertriglyceridemia the provision of 0.5–1 g kg^{-1} day^{-1} of lipid is adequate to prevent essential fatty acid deficiency and is a dose likely to be tolerated by most infants.

Carnitine Carnitine is necessary for the transport of free fatty acids into the inner mitochondrial membrane, the site of oxidation. Since the preterm infant has decreased carnitine synthesis capability and low plasma and tissue concentrations, carnitine may be an essential nutrient. Studies are conflicting as to whether there is benefit to adding it to parenteral nutrition. Its use should be considered in infants with birth weights <1000 g, those receiving long-term parenteral nutrition without enteral feedings, and those with hypertriglyceridemia.

Electrolytes The electrolyte content of parenteral nutrition solutions is usually similar to that found in normal intravenous solutions: usually 3–4 mmol kg^{-1} day^{-1} of sodium and 2–3 mmol kg^{-1} day^{-1} of potassium. Very immature infants and those on diuretics may require additional amounts to maintain normal plasma concentrations. Chloride and acetate need to be dosed based on electrolyte levels. The very young preterm infant may need a higher proportion of acetate secondary to urinary bicarbonate losses. Later, when chronic diuretics are used a greater proportion of chloride may be needed.

Calcium, phosphorous, and magnesium Calcium and phosphorous are relatively insoluble in solution together. This makes it difficult to provide adequate levels of these minerals to meet the needs of the preterm infant. When parenteral nutrition solutions are advanced to 10% dextrose and 2 g protein per 100 ml usually 60–80 mg (1.5–2 mmol) calcium and 40–60 mg (1.3–1.9 mmol) phosphorous can be added to the solution. Since the accretion rate of calcium in the fetus is normally 100 mg kg^{-1} day^{-1} (2.5 mmol kg^{-1} day^{-1}), infants on prolonged parenteral nutrition may develop osteopenia and fractures. The usual dose of magnesium is 0.3–0.5 mEq kg^{-1} day^{-1} (0.3–0.5 mmol kg^{-1} day^{-1}).

Trace minerals Zinc and copper deficiencies occurred in some preterm infants before these trace elements were routinely added to parenteral nutrition solutions. There is very little research that defines the parenteral requirements of trace minerals in preterm infants. The current recommendations for trace minerals are summarized in **Table 3**.

Vitamins Like trace minerals, the recommendations for intake of vitamins are not based on randomized trials but are based on the best information available. Infants receiving these parenteral intakes do not develop deficiencies or evidence of excessive intake (**Table 4**).

Table 3 Suggested parenteral intakes of trace minerals

Trace mineral	*μg kg^{-1} day^{-1}*
Zinc	400
Iron	200
Copper	15–20
Selenium	1.5–2
Manganese	1
Iodide	1
Molybdenum	0.25
Chromium	0.2

Table 4 Suggested parenteral intake of vitamins

Vitamin	*Amount/per kg per day*
Vitamin A (μg)	280–500
Vitamin E (mg)	2.8
Vitamin K (μg)	100
Vitamin D (μg)	4
Vitamin (IU)	160
Ascorbic acid (mg)	25
Thiamin (μg)	350
Riboflavin (μg)	150
Pyridoxine (μg)	180
Niacin (mg)	6.8
Pantothenate (mg)	2
Biotin (μg)	6
Folate (μg)	56
Vitamin B_{12}	0.3

Total dose should not exceed the amounts provided by 5 ml of reconstituted MVI Pediatric (Armor Pharmaceutical Co., Chicago, IL, USA): 700 μg vitamin A, 7 μg vitamin E, 200 μg vitamin K, 10 μg vitamin D, 80 mg ascorbic acid, 1.2 mg thiamin, 1.4 mg riboflavin, 1.0 mg pyridoxine, 17 mg niacin, 5 mg pantothenic acid, 20 μg biotin, 140 μg folic acid, 1 μg vitamin B_{12}.

Table 5 Suggested initiation and advancement of parenteral nutrition for the preterm infant

Component	*Initial*	*Advancement/day*	*Goal*
Dextrose	6–8 m kg^{-1} min^{-1} (0.033–0.044 mmol kg^{-1} min^{-1})	1–2 mg kg^{-1} min^{-1} (0.0055–0.011 mmol kg^{-1} min^{-1})	12–14 mg kg^{-1} min^{-1} (0.066–0.077 mmol kg^{-1} min^{-1})
Protein	1.5–2 g kg^{-1}	1 g kg^{-1}	3.5
Lipids	0.5–1 g kg^{-1}	0.5–1 g kg^{-1}	3

Suggested initiation and advancement of parenteral nutrition in the preterm infant is shown in **Table 5**.

Enteral Nutrition

The provision of adequate enteral nutrition is the goal of those caring for the preterm infant. However, a fear of the development of necrotizing enterocolitis, a serious intestinal disease of preterm infants associated with enteral feedings, has influenced feeding practices. Necrotizing enterocolitis (NEC) is a major cause of morbidity and mortality in preterm infants. The incidence of this disease is estimated to be between 8 and 10% of preterm infants. The cause of NEC is considered multifactorial including enteral feeds, hypoxia, ischemia, patent ductus arteriosus, and infection. Approximately 90% of infants who develop NEC have been enterally fed and several studies have shown that the rapid advancement of enteral feedings is associated with NEC. With the advent of parenteral nutrition, the tendency was to delay enteral feeding for prolonged periods of time in order to prevent this disease and to use parenteral nutrition as the sole source of nutrition. However, it is known that delayed enteral feeding has a negative effect on gastrointestinal structure and function. Lack of enteral nutrition induces gastrointestinal atrophy, depresses gut hormone secretion, and delays the maturation of gastrointestinal motility. There are now numerous studies that demonstrate the benefits of early enteral feeding including the promotion of endocrine adaptation, the accelerated maturation of gut motility patterns, the provision of luminal nutrients, and possible benefits to the immune system. In fact, early enteral nutrition may enhance feeding tolerance and may actually decrease the incidence of NEC.

Trophic Feedings

Even though it is recognized that early enteral feeding is beneficial, there is still hesitation to begin feedings in the early days following birth. One of the strategies that has been extensively studied since the late 1980s is trophic feeding, also referred to as minimal enteral nutrition or gut priming. This method involves giving the infant small volumes of feedings, approximately 10–20 ml kg^{-1} day^{-1}, for a period of 10–14 days before advancing to full enteral feedings. The benefits found are greater energy intake, earlier attainment of full enteral feedings, improved growth, less PN-related complications, reduced risk of infection, and earlier hospital discharge. Furthermore, infants who received trophic feedings had no increased incidence of NEC. Many clinicians have adapted variations of this practice, some with a shortened period of trophic feeds, others reserving this practice for the smallest, most preterm infants while employing advancement of feeds in larger, more stable infants. Once minimal enteral nutrition has been established and the infant is stable enough to advance feedings, it is generally considered a safe practice to increase feedings by 20 ml kg^{-1} day^{-1} while using PN for the balance of intake until an adequate enteral intake has been established and tolerated. Although fast feeding advancement has been associated with NEC, one study has shown no increase in the incidence of NEC amongst preterm infants whose feeds were advanced by 35 ml kg^{-1} day^{-1}.

Feeding Route

Because preterm infants lack the ability to coordinate sucking, swallowing, and breathing, tube feedings must be used. Jejunal feeding was a popular method for feeding infants during the 1970s to early 1980s. It was felt that this method would minimize the risk of reflux and aspiration. This method is now generally reserved for infants in whom reflux and aspiration is complicating chronic lung disease or those who have poor gastric emptying. Now, most infants are fed using an orogastric or nasogastric tube; the former usually selected for the tiniest babies as the feeding tube may occlude one naris and impair nasal breathing.

Feeding Selection

Breast milk expressed by the infant's mother is the preferred type of feeding for most preterm infants. It is nutritionally superior to artificial formula in many respects. There is improved gastric emptying, more

stool frequency, and improved fat absorption when breast milk is used. There are many trophic factors found in human milk that enhance the development of the gastrointestinal tract. Human milk contributes to host defense and reduces the risk of NEC. Preterm infants who have been fed expressed human milk also show a neurodevelopmental advantage but it is difficult to isolate this from the social variables also associated with mothers willing to express their milk. The use of expressed breast milk also enhances mother-infant bonding as this is one task that only the baby's mother is able to perform. However, there are also nutritional concerns related to the use of breast milk in infants born less than 33 weeks' gestation. Protein supplementation is necessary for optimal growth and maintenance of optimal protein status. Supplementation of calcium and phosphorous is also needed for adequate bone mineralization. There are multinutrient fortifiers available that can be added to breast milk to improve nutrient intake. The use of these fortifiers has been associated with improved intake of protein and minerals and growth and bone mineralization, and balance studies show improved nutrient retention.

If breast milk is not available, the feeding of choice becomes preterm infant formulas. These formulas have greater protein content and are cow's milk whey predominant. The carbohydrate is a mixture of lactose and glucose polymers, and the fat a mixture of both long-chain and medium-chain triglycerides for improved nutrient absorption. The concentration of minerals, electrolytes, and vitamins is increased to meet the estimated nutrient needs of the preterm infant when fed in an amount to provide 120 kcal kg^{-1} day^{-1}. Studies have shown that infants fed preterm infant formulas have improved growth over those fed term formula or even fortified breast milk.

Feeding Rate

The decision regarding how to feed must also be made: continuous versus bolus feeding. The preferred method is controversial. Some clinicians feel that continuous feedings are better tolerated while others feel that bolus feedings are more physiologic. In studies, bolus feedings have been associated with improved gastric emptying, and more mature intestinal motility patterns. It is difficult to compare feeding tolerance between continuous and bolus feeds due to differences in the criteria used. Comparison feeding studies have found fewer gastric residuals in those infants given bolus feedings than those fed continuously. A more recent study has found that feeds given as a slow bolus, over 2 h, resulted in a normal duodenal motility pattern, suggesting that some infants may benefit from slow intermittent feedings. Regardless of the method chosen, if an infant does not tolerate one method it may be beneficial to try a different one.

Monitoring Feeding Tolerance

Feeding tolerance among preterm infants must be closely monitored since NEC is associated with enteral feedings. The presence of gastric residuals is one factor that is frequently used, but because preterm infants have poor gastric emptying amounts less than 50% of a previous feed should not be considered significant. Other indicators that should be used in conjunction with gastric residuals include the increase in abdominal girth, the absence of active bowel sounds, the presence of blood in the stool, a change in the number or quality of stools, and the presence of emesis. A careful exam is warranted if these symptoms are present.

Estimated calorie and protein intakes to achieve fetal weight gain are shown in **Table 6**.

Monitoring Nutritional Status

The nutritional status and growth of the preterm infant should be monitored throughout the hospitalization. The daily fluid and caloric intake should be monitored daily, body weight should be recorded daily, length and head circumference should be measured weekly, and all three measurements plotted on standardized growth charts. If growth is inadequate the volume or caloric density of feeds and or the protein content should be increased. Biochemical measurements should also be assessed periodically.

Preparation for Discharge

Approximately 1 week prior to discharge, preterm infants should be converted to the feeding regimen that will be used at home. Infants who have been fed expressed breast milk should demonstrate the ability to directly breast-feed and/or to feed supplemented breast milk or formula from the bottle as needed to gain adequate weight. The infant who weighs less than 2500 g at discharge, especially those infants born at less than 30 weeks' gestation, may require the supplementation of some breast-milk feedings with post discharge formula powder or the feeding of a concentrated post discharge formula for some of the daily feedings.

Table 6 Estimated calorie and protein intakes to achieve fetal weight gain

	Body weight (g)				
	500–700	*700–900*	*900–1200*	*1200–1500*	*1500–1800*
Protein (g)					
Parenteral	3.5	3.5	3.5	3.4	3.2
Enteral	4.0	4.0	4.0	3.9	3.6
Energy (kcal/(kJ) kg^{-1} day^{-1})					
Parenteral	89 (372)	92 (385)	101 (422)	108 (451)	109 (456)
Enteral	105 (440)	108 (451)	119 (500)	127 (530)	128 (535)
Protein/energy (g/100 kcal or 418 kJ)					
Parenteral	3.9	4.1	3.5	3.1	2.9
Enteral	3.8	3.7	3.4	3.1	2.8

Adapted from Ziegler EE, Thureen PJ, and Carlson SJ (2002) Aggressive nutrition of the very low birthweight infant. *Clinical Perinatology* **29**(2): 225–244.

For those infants who were fed preterm formula, conversion to a nutrient-enriched post discharge formula is recommended. These formulas contain additional protein, vitamins, and minerals compared to term formulas. Studies have shown that infants fed these formulas for the first 9–12 months of life have improved gains in weight, length, and head circumference.

If growth is inadequate with either feeding regimen then alteration in caloric density may be needed. Arrangements should be made for the nutritional status of these infants to be monitored after discharge.

Conclusions

Preterm infants have specialized nutritional needs and each infant must be carefully and continuously assessed to ensure that the best possible nutritional support is provided to promote optimal growth without causing additional morbidity and mortality.

Table 7 Periodic monitoring of nutritional status

Indicator	*Frequency*
Weight	Daily
Length	Weekly
Head circumference	Weekly
Electrolytes (PN)	Daily until stable then 2 times weekly
Electrolytes (enteral)	Weekly
Albumin	Weekly
Bili/transaminases (PN)	Weekly
Calcium, phosphorous, magnesium, alkaline phosphatase	Weekly
Hemoglobin/hematocrit	Weekly

See also: **Breast Feeding**. **Energy**: Requirements. **Growth and Development, Physiological Aspects**. **Growth Monitoring**. **Infants**: Nutritional Requirements; Feeding Problems. **Low Birthweight and Preterm Infants**: Causes, Prevalence and Prevention. **Nutritional Support**: Infants and Children, Parenteral.

Further Reading

Berseth CL (2001) Feeding methods for the preterm infant. *Seminars in Neonatology* **6**: 417–424.

Denne SC (2001) Protein and energy requirements in preterm infants. *Seminars in Neonatology* **6**: 377–382.

Embleton NE, Pang N, and Cooke RJ (2001) Postnatal malnutrition and growth retardation: an inevitable consequence of current recommendations in preterm infants? *Pediatrics* **107**(2): 270–273.

Griffinc IJ (2002) Postdischarge nutrition for high risk neonates. *Clinical Perinatology* **29**(2): 327–344.

Hay WW, Lucas A, Heird WC *et al.* (1999) Workshop summary: nutrition of the extremely low birth weight infant. *Pediatrics* **104**: 1360–1368.

Hay WW (1996) Assessing the effect of disease on nutrition of the preterm infant. *Clinical Biochemistry* **29**(5): 399–417.

Heird WC and Gomez MR (1996) Parenteral nutrition in low birthweight infants. *Annual Review of Nutrition* **16**: 47–99.

Heird WC (2002) Determination of nutritional requirements in preterm infants, with special reference to "catch-up" growth. *Seminars in Neonatology* **6**: 365–375.

Neu J and Koldovsky O (1996) Nutrient absorption in the preterm neonate. *Clinical Perinatology* **23**(2): 229–243.

Newell SJ (2000) Enteral feeding of the micropremie. *Clinical Perinatology* **27**(1): 221–234.

Thureen PJ and Hay WW (2000) Intravenous nutrition and postnatal growth of the micropremie. *Clinical Perinatology* **27**(1): 197–219.

Tyson JE and Kennedy KA (1997) Minimal enteral nutrition in parenterally fed neonates. www.nichd.nih.gov/cochraneneonatal.

Ziegler EE, Thureen PJ, and Carlson SJ (2002) Aggressive nutrition of the very low birthweight infant. *Clinical Perinatology* **29**(2): 225–244.

LUNG DISEASES

A MacDonald, The Children's Hospital, Birmingham, UK

Respiratory disease covers a wide range of disorders and interest has grown in the provision of its nutritional support. Epidemiological studies suggest that dietary habits influence lung function and the tendency to common lung diseases, such as asthma, chronic obstructive pulmonary disease (COPD), and lung cancer. Malnutrition and weight loss are commonly reported in patients with COPD, and intensive, specialist nutritional support is required. For cystic fibrosis (CF), nutritional therapy has been shown to improve the prognosis. Novel diet therapies, including exclusion of food allergens and reduction of salt intake, have been tried as possible treatments for asthma. This article outlines the rationale for and describes methods of nutritional support and therapy in the treatment of lung disease.

Chronic Obstructive Pulmonary Disease

COPD is a term used to describe a spectrum of disorders characterized by reduced maximal expiratory flow and slow forced emptying of the lungs (**Table 1**). It is associated with symptoms of obstructive lung diseases, such as cough, mucus production, breathlessness, airflow limitation, and wheezing. COPD can be present with or without substantial physical impairment or symptoms. The World Health Organisation states that it is the fourth leading cause of global mortality. In the United Kingdom, it is estimated that 18% of males and 14% of females aged 40–68 years may have developed features of COPD. Severe exacerbations remain the largest single cause of emergency admissions for respiratory disease, with a mean hospital stay of 10 days.

Table 1 Consequences of disease-related malnutrition in COPD

Reduced lung function and dyspnea
Reduced maximal O_2 consumption
Decreased peripheral muscle function
Decreased exercise performance
Decreased quality of life
Increased need for hospitalization
Increased postoperative complications during lung volume reduction surgery
Increased mortality

Adapted from Stratton *et al.* (2002).

The prevalence of COPD is greatest in socioeconomically deprived people. It is probably underdiagnosed partly because many people do not consult their general practitioners or do not reveal all their symptoms. In general, the major cause of COPD is smoking. Other causes include α_1-antitrypsin deficiency, cystic fibrosis, air pollution, occupational exposure, and bronchietasis.

Chronic bronchitis and emphysema are two of the major diseases grouped under COPD. Patients with COPD have features of both conditions, although one may be more prominent than the other.

Chronic Bronchitis

Definition and etiology Chronic bronchitis is defined by the presence of chronic bronchial secretions sufficient to cause expectoration occurring on most days for a minimum of 3 months for 2 consecutive years. It became recognized as a distinct disease in the late 1950s associated with the great British Smog. It develops in response to long-term irritants on the bronchial mucosa. Important irritants include cigarette smoke, dust, smoke, and fumes; other causes include respiratory infection, particularly in infancy, and exposure to dampness, sudden changes in temperature, and fog. In the United Kingdom, it affects 10% of older people, and it is more common in industrial countries. Chronic bronchitis is a slowly progressive disorder unless the precipitating factors are avoided and it is treated.

Clinical features Symptoms include productive cough and frequent and recurrent chest infections. The disease progresses over many years from a troublesome cough producing a little clear sputum to marked wheezing, severe breathlessness leading to poor exercise tolerance, and copious and purulent sputum. It may cause right heart failure (i.e., cor pulmonale), such as oedema and cyanosis.

Pathology There is hypertrophy of the mucus-secreting glands. The structural changes described in the airways include atrophy, focal squamous metaplasia, ciliary abnormalities, variable amounts of airway smooth muscle hyperplasia, inflammation, and bronchial wall thickening. The respiratory bronchioles display a mononuclear inflammatory process, lumen occlusion by mucus plugging, goblet cell metaplasia, smooth muscle hyperplasia, and distortion due to fibrosis. These changes combined

with loss of supporting alveolar attachments cause airflow limitation by allowing airway walls to deform and narrow the airway lumen.

Emphysema

Diagnosis and etiology Emphysema means 'inflation' in the sense of abnormal distension with air. It is a condition in which there is permanent destructive enlargement of the airspace distal to the terminal bronchioles without obvious fibrosis. In the general population, emphysema usually develops in older individuals with a long smoking history. However, other causes include exposure to heavy metals such as cadmium, and 5% of early presenting cases are caused by the autosomal recessive disorder α_1-antitrypsin deficiency. It affects almost 5% of older people, and it is more common in industrialised countries. The prognosis is variable. Progression is slow, provided it is treated.

Clinical features Patients may be very thin with a barrel chest and have little or no cough expectoration. Symptoms include intense dyspnea with purse-lip breathing and overinflation of the chest. Breathing may be assisted by pursed lips and use of accessory respiratory muscles. The chest may be hyperresonant, and wheezing may be heard. Heart sounds are very distant and overall appearance is more like classic COPD exacerbation.

Chemical Pathology There are three types of emphysema:

Panacinar emphysema: A generalized destruction of the alveolar walls. As a consequence, the elastic network of the normal lung is badly disorganized and the lung becomes floppy, leading to a severe degree of airways obstruction, particularly during expansion. It generally develops in patients with α_1-antitrypsin deficiency.

Centriacinar emphysema: Distension and damage affect the respiratory bronchioles; the more distal alveolar ducts and alveoli tend to be well preserved. This is very common and not necessarily associated with disability.

Paraseptal emphysema: The least common form, involving distal airway structures and alveolar ducts and sacs.

Nutritional Management in COPD

Malnutrition is an important clinical problem in a subpopulation of patients with COPD. Emaciation with emphysema was reported as early as 1898. Studies suggest malnutrition occurs in between 27 and 71% of all patients, increasing with the severity of airways obstruction. In one of the larger studies examining patients with stable COPD, one-fourth of 779 men were less than 90% of their ideal body weight. However, in patients who need hospital admission, malnutrition approaches or exceeds 50%. When acute respiratory failure complicates the clinical course, severe malnutrition is observed in 60% of cases.

Malnutrition has a considerable impact on both morbidity and mortality. Reduced respiratory muscle mass and function as well as increased susceptibility to infection are recognised as deleterious consequences of malnutrition in patients with or without lung disease. In a necropsy study, diaphragm muscle mass was reduced by 43% in malnourished patients whose weights were 71% of ideal body weight. Decreases in body weight, creatinine–height index, total lymphocyte count, serum transferrin, and retinol binding protein have been documented.

Nutritional depletion is an independent risk factor for mortality and hospitalization in patients with COPD. Studies have indicated a hospital stay of approximately 30 days for patients with a body mass index (BMI) of less than 20 compared with 18 days for those with a BMI of less than 30. If a patient with COPD begins to lose weight progressively, the average reported life expectancy is only 2.9 years, and it is considerably less in malnourished patients who have survived an episode of acute respiratory failure during an acute exacerbation of their disease. However, it is not certain whether this implies a casual relationship or whether low weight is a marker for more severely impaired lung function.

Reasons for Malnutrition

The cause of progressive weight loss in patients with COPD is not well understood but two factors have been implicated.

Increased resting energy expenditure The relationship between resting energy expenditure (REE), lung function, oxygen cost of breathing, and malnutrition in COPD has been the focus of much attention in recent years. Results of studies on REE are conflicting. In two early studies on malnourished patients with COPD, Goldstein and coworkers described 10 patients whose REE was 113% of predicted, and Wilson and colleagues described 7 patients with a REE 115% of predicted. In contrast, two later studies found REE to be only 94 and 104% of predicted in stable fasted COPD patients, respectively. It has therefore been hypothesized that if patients

with COPD are not hypermetabolic, malnutrition is related more to impaired gas exchange (as evidenced by a low diffusing capacity of carbon monoxide) than to airflow obstruction. The impaired gas exchange results from loss of the pulmonary capillary bed and may result in an inability to augment cardiac output in response to the stress of even minimal effort, leading to lack of oxygen delivery to the tissues and nutritional depletion. An alternative hypothesis is that malnutrition is precipitated by acute illnesses, leading to a combination of anorexia and hypercatabolism causing significant weight loss.

Reduced energy intake Many studies have examined the energy intake of malnourished COPD patients and have shown it to be either similar to that of well-nourished patients with COPD or higher than the respective dietary recommendations. Most of these studies were conducted on stable patients, but several factors may adversely affect energy and nutrient intake:

Hypoxia-related appetite suppression or anorexia due to acute exacerbation
Chronic sputum production and frequent coughing, which may alter desire for and taste of food
Fatigue, which can interfere with the ability to shop for food and prepare and eat meals
Depression
Side effects of medications, including nausea, vomiting, diarrhea, dry mouth, and gastric irritation, which may limit dietary intake
Raised plasma tumor necrosis factor-α levels
Arterial oxygen desaturation due to altered breathing pattern during chewing and swallowing

Nutritional Support

Several controlled studies have evaluated the effect of nutritional support in COPD in either outpatients or inpatients, and their outcome was related to the overall energy intake achieved. Weight gain was only achieved by substantially increasing energy intake by more than 30% above the usual intake, amounting to more than 45 kcal/kg per day. Moreover, improvement in muscle function or exercise tolerance occurred only with concomitant weight gain. In one of these controlled studies, oral supplementation was given for 3 months to ambulatory malnourished patients with chronic obstructive lung disease. Daily energy intake increased by 48% above the usual intake and corresponded to 47 kcal/kg on average. The authors reported a mean weight gain of 4.2 kg, an increase in maximal respiratory pressures and in handgrip and stemomastoid strength, and a decrease in stemomastoid muscle fatigability; similar improvements were not observed in a control group. However, these improvements were not maintained when oral supplementation was discontinued. In another study, six malnourished patients received an additional 1000 kcal via a nasogastric tube for 16 days, whereas a control group of four received only an additional 100 kcal. A weight gain of 2.4 kg, and improvements in respiratory muscle strength and endurance were seen in the fed patients but not in the control group. Other studies have not demonstrated improvements in weight gain or muscle performance and have been less successful in increasing energy intake.

Type of Nutritional Support

If the patient with COPD is less than 90% ideal body weight for height, nutritional support should be considered. This can be difficult in COPD, but it can be provided on three levels.

Normal high-energy, high-protein diet For many patients with COPD, advising small, frequent nutrient-dense meals, regular snacks, and food fortification using high-energy and -protein foods, such as milk, yoghurt, butter, and cream, may provide adequate nutritional support. Foods of low nutritional value, such as tea, squash, and clear soup, should be discouraged. For some patients who lack vitality, use of readily prepared microwave dinners with a rest prior to mealtime is helpful. A daily multivitamin and minerals supplement may be indicated.

Use of high-energy, high-protein supplements These products can be used to augment a patient's dietary intake. They are available in the form of milk, sweet and savoury drinks, fortified fruit juices, milkshake powder, glucose polymer powders and liquids, and puddings. Patients and caregivers need to be given complete instructions regarding their use to optimise this form of nutritional supplementation. Unfortunately, many studies found that in COPD the use of these supplements led to a reduction in usual energy intake and caused symptoms such as bloating, nausea, and early satiety. Oral supplements are probably less effective in older patients with a systematic inflammatory response.

Tube feeding Overnight tube feeding should be considered in patients with COPD when oral methods of maintaining nutritional status have failed, although few studies have investigated this method. The composition of tube feeds for patients with COPD has received attention. Carbon dioxide

production ($V\dot{C}O_2$) is higher when carbohydrates are the main energy sources and lower when fat is mainly oxidised. However, patients with COPD who are in a stable clinical state usually appear to tolerate carbohydrates without difficulty. Respiratory failure has not been reported in studies of patients with COPD receiving nutritional support with enteral feeds and nutritional supplements containing up to 54% carbohydrate, but further work is needed to determine the optimal and safest feeding regimens for these patients.

Feeding Patients on Artificial Ventilation

The artificial ventilator may be used to control the breathing patterns of patients who have acute breathing problems (e.g., respiratory failure with worsening blood gases). If the patient can be enterally fed, the composition of feed has a profound effect on gas exchanges, especially CO_2 production, and therefore respiratory quotient. This is expressed as the ratio of CO_2 produced to oxygen consumed. Because CO_2 production is greater during carbohydrate metabolism, a diet high in carbohydrate requires increased ventilation to eliminate the excess CO_2, whereas high-fat feeds reduce CO_2 production and are therefore potentially beneficial. Overfeeding negates any beneficial response to high-fat feeds because the conversion of energy to fat involves a disproportionately large production of CO_2.

Cystic Fibrosis

Definition and Etiology

In CF, there is widespread dysfunction of exocrine glands that causes chronic pulmonary disease; pancreatic enzyme deficiency; intestinal obstruction in the neonate (distal intestinal obstruction syndrome); liver disease; infertility, especially in males; and abnormally high concentrations of electrolytes in sweat, resulting from the failure of salt reabsorption in the sweat gland ducts. This is the most common inherited disease in Caucasian populations. A gene located on chromosome 7, coding for the protein called cystic fibrosis transmembrane regulator (CFTR), is defective. CFTR acts as a cyclic-AMP-activated chloride channel blocker. More than 800 mutations of the gene have been identified, and they are categorized into five classes on the basis of CTFR alterations. The most predominant mutation, which accounts for approximately 70% of all the CTFR genes worldwide, is $\Delta 508$, but there is geographical variation and it is less common in non-white races.

Although previously this disease was considered lethal in childhood, the median survival for newborns in the 1990s is predicted to be 40 years. Survival is largely dependent on the severity and progression of lung disease, and more than 90% of mortality is due to chronic bronchial infections and their complications. Patients with pancreatic insufficiency have a worse prognosis in terms of growth, pulmonary function, and long-term survival. The mortality of females is generally greater than that of males.

Incidence

CF affects 1 in 2500 births in Caucasian populations, 1 in 20 000 in black populations, and 1 in 1 million in Oriental populations. It is extremely rare in Japan, China, and black Africa.

Clinical Features

Most children with CF present with malabsorption and failure to thrive accompanied by recurrent or persistent chest infections. In the lungs, viscid mucus in the smaller airways predisposes to chronic infection, particularly with *Staphylococcus aureus* and *Haemophilis influenzae*, and subsequently with *Pseudomonas* species. This leads to damage of the bronchial wall, bronchietasis, and abscess formation.

Approximately 90% of CF patients have pancreatic insufficiency, requiring pancreatic enzyme supplements. Untreated patients pass frequent, large, pale, offensive, greasy stools. Ten to 15% of infants present with a meconium ileus resulting from the blockage of the terminal ileum by highly proteinaceous meconium at birth. Distal intestinal obstruction syndrome may occur later in childhood or adult life. In addition to these symptoms, a number of other complications may occur in CF that are identified in **Table 2**.

Chemical Pathology

CTRF regulates the chloride channel in the cell at its luminal surface, and its absence or dysfunction results in an abnormally high concentration of sodium in sweat and in a low water content in the mucus produced by airways, pancreas, and intestine. In the lung, this leads to ciliary dysfunction and repeated infection and colonisation with bacteria, resulting in a vicious cycle of bacterial colonization, 'lung inflammation,' and scarring. These in turn result in severe bronchiectasis, which progressively destroys lung function.

Table 2 Complications of cystic fibrosis

Respiratory
Pneumothorax
Asthma/wheezing
Hemoptysis
Nasal polyps
Respiratory failure
Cor pulmonale
Allergic bronchopulmonary aspergillosis
Gastrointestinal
Meconium ileus
Rectal prolapse
Distal intestinal obstruction syndrome
Abdominal distension
Colonic strictures
Intussusception
Gastro-oesophageal reflux
Biliary cirrhosis
Hepatomegaly
Portal hypertension
Cholelithiasis
Cholecystitis
Obstructive jaundice
Pancreatitis
Other
Diabetes
Male infertility
Amyloidosis
Arthropathy
Salt depletion
Growth failure/weight loss/failure to thrive
Delayed puberty
Osteopenia

Nutritional Management

Nutritional intervention is associated with better growth and improvement or stabilization of pulmonary function and possibly may improve survival in CF. Malnutrition has several adverse effects, including poor growth, impaired muscle function, decreased exercise tolerance, increased susceptibility to infection, and decreased ventilatory drive. Studies indicate that BMI strongly correlates with lung function, but the exact mechanism of this relationship has not been fully determined. Achieving optimum nutrition and growth may minimize the progressive decline in pulmonary function commonly seen in CF. As early as the 1970s, the Toronto CF clinic was able to show that a high-fat diet promoted a normal growth pattern and improved survival.

Reasons for Malnutrition

A variety of complex organic and psychosocial factors contribute to malnutrition in CF.

Malabsorption Pancreatic exocrine secretions contain less enzymes and bicarbonate, have a lower pH, and are of a smaller volume, and the physical properties of proteins and mucus within the lumen are affected. This results in obstruction to the small ducts and secondary damage to pancreatic digestive enzyme secretions causing malabsorption. Other problems, such as gastric hypersecretion, reduced duodenal bicarbonate concentration and pH, disorders of bile salt metabolism, disordered intestinal motility and permeability, liver disease, and short bowel syndrome after intestinal resection in the neonatal period, may contribute to malabsorption. The severity of malabsorption is variable, and there can be significant malabsorption of protein and fat-soluble vitamins despite adequate use of enzyme supplements.

Increased energy expenditure Resting energy expenditure, an estimate of basal metabolic rate, is 10–20% greater than in healthy controls and may contribute to energy imbalance. Increased REE appears to be closely associated with declining pulmonary function and subclinical infection. Bronchial sepsis leads to local release of leukotrienes, free oxygen radicals, and cytokines, including tumour necrosis factor-α. Antibiotics have been shown to reduce energy requirements of moderately ill patients with chronic *Pseudomonas aeruginosa*.

Anorexia and low energy intake Inadequate energy intake is often the main reason for growth failure in CF. Factors associated with a reduced appetite include

- Chronic respiratory infection and other complications of CF, such as distal ileal obstruction syndrome, abdominal pain, GOR resulting in oesophagitis, pain, and vomiting
- Behavior feeding problems in preschool and school-age children
- Media pressure to eat a healthy low-fat, low-sugar diet
- Inappropriate concepts regarding body image
- Depression
- Eating disorders in teenagers
- Poor use of dietary supplements
- Dislike of high-energy foods

Nutritional Support

Nutritional requirements will vary according to the clinical state as well as the age, sex, and activity of the individual. Because of the heterogeneity of CF, it is impossible to give universal recommendations. Crude estimates suggest an energy intake of 120–150% of estimated average requirements; it is better to assess and monitor energy intake and equate this

to the nutritional status of the patient. If weight gain or growth is poor, the usual energy intake is increased by an additional 20–30% of total intake. Likewise, exact protein requirements are unknown; but it is generally accepted that the protein intake should be increased to compensate for excessive loss of nitrogen in the feces and sputum and increased protein turnover in malnourished patients. Protein should provide 15% of the total energy intake.

High-energy/high-protein diet The encouragement of a high-calorie, high-protein diet will produce growth in the majority of children and adults with CF. A good variety of energy-rich foods should be encouraged, such as full cream milk, cheese, meat, full cream yoghurt, milk puddings, cakes, and biscuits. Extra butter or margarine can be added to bread, potatoes, and vegetables. Frying foods or basting in oil will increase energy density. Extra milk or cream can be added to soups, cereal, desserts, or mashed potatoes and used to top canned or fresh fruit. Regular snacks are important. Malnourished children achieve higher energy intake when more frequent meals are offered. Attention should be given to psychological, social, behavioral, and developmental aspects of feeding. A meta-analysis of differing treatment interventions to promote weight gain in CF demonstrated that a behavioural approach was as effective in promoting weight gain as evasive medical procedures.

Dietary supplements Although almost half of adult patients take dietary supplements, there are few published data to demonstrate their efficacy in CF. One study was unable to show any improvement in height and weight z scores when up to 30% of energy requirements were supplied by a supplement for 3 months. As a consequence, dietary supplements should be reserved for weight loss, any decline in height z score, if intake of a range of nutrients does not meet dietary reference values, or during acute chest infections. They should complement normal food intake and not replace food. In order to avoid reducing the intake of normal food, the recommended quantities are age dependent and are given in **Table 3**.

Enteral nutrition Enteral feeding is more commonly used in teenagers and adults, reflecting their deterioration in nutritional status. It is considered if the patient is less than 85% expected weight for height, the patient's weight has declined by two centile positions, the patient has failed to gain weight over a 6-month period, or the patient has a BMI less than 19. Enteral feeding is associated with improvements in body fat, height, lean body mass, muscle mass, increased total body nitrogen, improved strength, and development of secondary sexual characteristics. To produce lasting benefit, numerous studies have demonstrated that enteral feeding should be continued long term. The choice of route used is influenced by the duration of feeding and the preference of the patient and family, but gastrostomies, sited by endoscopic placement, are usually chosen for long-term feeding (**Table 4**).

Table 3 Recommended dosage of dietary supplements in cystic fibrosis

Age (years)	*Daily dosage (Kcal)*
1–2	200
3–5	400
6–11	600
>12	800

Adapted from MacDonald A (2001). Cystic fibrosis. In: Shaw V and Lawson M (eds.) *Clinical Paediatric Dietetics*, pp. 137–157. Oxford: Blackwell Scientific.

It is common practice to give enteral feeding for 8–10 h overnight, with at least 40–50% of the estimated energy requirement given via the feed. Most patients tolerate an energy-dense polymeric feed providing at least 1.5 kcal/ml with additional pancreatic enzymes. However, there is some support for the use of chemically defined elemental or short-chain peptide feeds. These are generally low in fat and are administered without the use of pancreatic enzymes, although there is little evidence to support this practice and it is disputed by some. Monitoring for glucose intolerance is important. Patients receiving supplemental feeds who demonstrate repeated blood sugar levels higher than 11.1 mmol/l during the feed may benefit from insulin given before the feed.

Vitamin and mineral supplements Malabsorption of fat-soluble vitamins is likely in most pancreatic-insufficient patients with CF, and the United Kingdom recommends vitamin A, D, and E supplements. Low fat-soluble vitamin concentrations are associated with poorer clinical status and reduced lung function. Clinical features of vitamin A deficiency include night blindness, conjunctival and corneal xerosis, dry thickened skin, and abnormalities of bronchial mucosal epithelialisation. Vitamin A status is difficult to assess due to lack of a reliable marker and serum levels of retinol do not adequately mirror the concentration of vitamin A in the liver. Some researchers have found liver stores of vitamin A in CF to be 2.5 times higher than those in control subjects, despite lower serum levels of retinol and retinol binding protein.

Table 4 Advantages and disadvantages of enteral feeding routes

Method	*Advantages*	*Disadvantages*
Nasogastric	Short-term feeding	Tube reinsertion may be Distressing to patient/caregiver/nurse Easily removed Risk of aspiration Discomfort to nasopharynx Psychosocial implications
Nasojejunal	Less risk of aspiration Short-term feeding	Difficulty of insertion Radiographic check of position Easily removed Risk of perforation Abdominal pain and diarrhoea unless continuous infusion of feed Discomfort in nasopharynx Reflux of bile is facilitated
Gastrostomy	Cosmetically more acceptable Long-term feeding	Increase reflux if present Local skin irritation Infection Granulation tissue Leakage Gastric distension Stoma closes within a few hours if accidentally removed
Jejunostomy	Reduced risk of aspiration Long-term feeding	Surgical/radiology procedure Risk of perforation Must be constant infusion of feed Bacterial overgrowth Dumping syndrome can occur

Adapted from MacDonald A, Holden C, and Johnston T (2001) Paediatric enteral nutrition. In Payne-James J, Grimble G, and Silk D (eds.) *Artificial Nutrition Support in Clinical Practice*, pp. 347–366. London: Greenwich Medical Media.

Decreased bone mineral density and osteopenia associated with low 25-hydroxyvitamin D levels have been described in patients with CF but may be related to poor nutritional status and delayed puberty. Rickets is rarely seen. Possible contributory factors include low body mass index, disease severity, inadequate calcium intake, delayed puberty, or widespread use of systemic or inhaled steroids.

Blood levels of vitamin E are nearly always low unless supplements are given. In older patients, undetectable serum concentrations of vitamin E have been noted in association with neurological syndromes. Symptoms and signs include absent deep tendon reflexes, loss of position sense and vibration sense in lower limbs, dysarthria, tremor, ataxia, and decreased visual activity.

Some CF centres recommend routine salt supplements to all CF infants on normal infant formula, which is low in sodium, and CF patients during hot weather. Anorexia and poor growth may result from chronic salt depletion. Significant hyponatremia may be accompanied by vomiting.

Pancreatic Enzymes

Approximately 90% of patients with CF require pancreatic enzymes to reduce steatorrhea. They are based on animal pancreatic extracts; presented in powder; tablet, or capsule form; and contain a combination of lipase, protease, and amylase. In addition to enzyme content, many factors affect bioavailability of pancreatic enzymes, including enzyme source, manufacturing process, stability, enteric coating of acid-resistant tablets, formulation as either microspheres or microtablets, and particle size. The smallest dose of pancreatin to control steatorrhoea and achieve a normal pattern of growth and weight gain should be used. The Committee on Safety for Medicines recommends that patients with CF not exceed a daily dose of enzymes equivalent to 10 000 IU lipase/kg/day.

Asthma

Definition and Etiology

The word ‘asthma’ originates from an ancient Greek word meaning panting. It is a chronic obstructive disease characterized by tracheobronchial hyperreactivity leading to paroxysmal airway narrowing, which may reverse spontaneously or as a result of treatment. The smooth muscle surrounding the bronchi has an abnormally increased reaction to stimuli. Specific bronchial stimuli include inhaled

allergens (e.g., house-dust mite, pollen, and moulds). Nonspecific bronchial stimuli include upper respiratory tract infections, cold air, exercise, cigarette smoke, excitement, emotional stress, and chemical irritants. Aspirin and other nonsteroidal antiinflammatory medications provoke asthma in some patients.

Prevalence

Since the 1980s there has been a worldwide increase in the prevalence of asthma in both children and adults. This escalating prevalence has led to significant increases in morbidity and mortality due to the disease. It is the most common chronic respiratory disorder, affecting 3–5% of adults and 10–15% of schoolchildren. Half of the people with asthma develop it before age 10, and most develop it before age 30. In childhood, it is twice as common in boys as in girls, but by adolescence equal numbers are affected. Asthma symptoms can decrease over time, especially in children. Many people with asthma have an individual and/or family history of allergies, such as hay fever (allergic rhinitis) or eczema. Others have no history of allergies or evidence of allergic problems.

It is responsible for 10–20% of all acute medical admissions in pediatric wards in children aged 1–16 years. There are 15–20 deaths from asthma in children each year in the United Kingdom. However, from 1979 to 1999, mortality rates of asthma have decreased in England and Wales in all age groups up to 65 years.

Clinical Features

Asthma is characterized clinically by wheezing, dyspnea, and cough. Coughing commonly produces small amounts of yellowish sputum or bronchial plugs. Some patients present with breathlessness. Most people with asthma have periodic wheezing attacks separated by symptom-free periods. Other asthmatics may have cough as their predominant symptom. Asthma attacks can last minutes to days.

Chemical Pathology

The development and phenotypic expression of allergic airway disease depend on a complex interaction between genetic and environmental factors. Exposure of the sensitized airway to a number of trigger factors results in bronchoconstriction, mucosal oedema, and excessive mucus production that in turn leads to airway narrowing and the clinical features of asthma. Airway inflammation is due to an immune-mediated process in which inflammatory cells and inflammatory mediators enter airway tissues to cause disease. Many cell-mediated immunologic factors participate in the inflammatory process of asthma. The most important inflammatory cells involved are eosinophils, mast cells, and T lymphocytes. The first months of life seem to be a particularly vulnerable period and there is evidence that sensitization is related to the level of allergen exposure during early life.

Dietary Management

There is an increasing interest in the relationship between nutrition and asthma. Associations have been reported between the intake of fruit and the antioxidant vitamins A, C, and E and selenium. Suboptimal nutrient intake may enhance asthmatic inflammation, consequently contributing to bronchial hyperreactivity. There is some suggestion that people who have a diet rich in fruit and vegetables have a lower risk of poor respiratory health, and this may be due to the antioxidant nutrients that food contains. Several issues need to be addressed before causality of these associations can be established. Nevertheless, it appears reasonable to issue dietary guidelines for the primary and secondary prevention of asthma that are in line with a healthy diet for the prevention of coronary heart disease and cancer.

Epidemiological studies also suggest that a diet high in marine fatty acids (fish oil) may have beneficial effects on asthma. However, a Cochrane review of nine randomized controlled trials conducted between 1986 and 2001 indicated there is little evidence to recommend that people with asthma supplement or modify their dietary intake of marine n-3 fatty acids (fish oil) in order to improve their asthma control. Equally, there is no evidence that they are at risk if they do so.

Food intolerance There is much controversy surrounding the role of food in the development and onset of asthma. Evidence suggests that atopic or asthmatic parents, whose children have a high risk of developing asthma, should be advised to avoid smoking during pregnancy; avoid cigarette smoke exposure after the child is born; undertake house dust mite control strategies; exclusively breast-feed their infants for 6 months; and subsequently provide their child with a nutritious, balanced diet. In contrast, there is little to suggest that a low allergen diet for high-risk women during pregnancy is likely to reduce the risk of having an atopic child.

Generally, the incidence of food intolerance in asthma is thought to be small, although there is evidence that intolerance to foods may act as a trigger for some cases of asthma. Common food allergens identified include milk, eggs, nuts, orange

squash, wheat, and red wine. The additives sulfur dioxide, tartrazine, sodium benzoate, and salicylates have been implicated, although in the case of sulfur dioxide its ability to cause asthma depends on the nature of the food to which it is added, the level of residual sulfur dioxide in the food, and the sensitivity of the patient. Foods such as nuts, cola drinks, ice, and those cooked in oil have been found to cause symptoms more frequently in Asian children. Many of the studies that have identified certain foodstuffs as triggering asthma have had limitations or flaws in their design, leading to difficulties in interpreting and extrapolating their results.

Both immediate and delayed-onset symptoms have been reported in asthma. The use of diagnostic diets is difficult, partly because of the variability of asthma, delayed reactions, effect of other precipitating triggers, and dangers of inducing an asthma attack during food challenges. A simple exclusion diet excluding food(s) implicated on history is perhaps the most useful diagnostic diet. Because of the inherent problems with asthma, strict food diets are rarely used.

High-sodium diets Epidemiological studies have suggested that dietary salt may play a role in airway responsiveness and a high salt intake may act as a trigger for asthma. A correlation between regional mortality from asthma and purchase of table salt per person has been reported in England and Wales. Epidemiological studies have also suggested an association between a higher dietary sodium intake and a higher prevalence of self-reported wheeze in adults and children. However, not all of the evidence supports this hypothesis. At least three epidemiological surveys and two experimental studies found no evidence of an association between sodium intake and asthma.

See also: **Cancer**: Epidemiology of Lung Cancer. **Cystic Fibrosis**. **Food Intolerance**. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **Nutritional Support**: Adults, Enteral. **Sodium**: Physiology. **Supplementation**: Role of Micronutrient Supplementation.

Further Reading

Ardern KD and Ram FS (2001) Dietary salt reduction or exclusion for allergic asthma. *Cochrane Database of Systematic Reviews* **2001**(4): CD000436.

Berry JK and Baum CL (2001) Malnutrition in chronic obstructive pulmonary disease: Adding insult to injury. *AACN Clinical Issues* **12**: 210–219.

Couriel J (1997) Respiratory disorders. In: Lissauer T and Clayden G (eds.) *Illustrated Textbook of Paediatrics*, pp. 157–171. London: Mosby.

Felbinger TW, Suchner U, Peter K, and Askanazi J (2001) Nutrition support in respiratory disease. In: Payne-James J, Grimble G, and Silk D (eds.) *Artificial Nutrition Support in Clinical Practice*, pp. 537–552. London: Greenwich Medical Media.

Ferreira IM, Brooks D, Lacasse Y, Goldstein RS, and White J (2002) Nutritional supplementation for stable chronic obstructive pulmonary disease. *Cochrane Database of Systematic Reviews* **2002**(1): CD000998.

Hind CRK and Walshaw MJ (1996) Chest disease. In: Axford J (ed.) *Medicine*. Oxford: Blackwell Science.

Hodge L, Yan KY, and Loblay RL (1996) Assessment of food chemical intolerance in adult asthmatic subjects. *Thorax* **Sl**: 805–809.

Kramer MS (2002) Maternal antigen avoidance during pregnancy for preventing atopic disease in infants of women at high risk. *Cochrane Database of Systematic Reviews* **2000**(2): CD000133.

MacDonald A (2001) Cystic fibrosis. In: Shaw V and Lawson M (eds.) *Clinical Paediatric Dietetics*, pp. 137–157. Oxford: Blackwell Scientific.

MacDonald A, Holden C, and Johnston T (2001) Paediatric enteral nutrition. In: Payne-James J, Grimble G, and Silk D (eds.) *Artificial Nutrition Support in Clinical Practice*, pp. 347–366. London: Greenwich Medical Media.

Manaher S and Burke F (1996) Pulmonary disease. In: Morrison G and Hark L (eds.) *Medical Nutrition and Disease*, pp. 279–287. Oxford: Blackwell Science.

Mickleborough TD, Gotshall RW, Cordain L, and Lindley M (2001) Dietary salt alters pulmonary function during exercise in exercise-induced asthmatics. *Journal of Sports Science* **19**(11): 865–873.

Price D and Duerden M (2003) Chronic obstructive pulmonary disease. *British Medical Journal* **326**(7398): 1046–1047.

Stratton RJ, Green CJ, and Elia M (2003) *Disease-Related Malnutrition: An Evidence Based Approach to Treatment*. Cambridge, UK: CABI.

UK Cystic Fibrosis Trust Nutrition Working Group (2002) *Nutritional Management of Cystic Fibrosis*. London: Cystic Fibrosis Trust.

Woods RK, Thien FC, and Abramson MJ (2002) Dietary marine fatty acids (fish oil) for asthma in adults and children. *Cochrane Database of Systematic Reviews* **2002**(3): CD001283.

LYCOPENES AND RELATED COMPOUNDS

C J Bates, MRC Human Nutrition Research, Cambridge, UK

Introduction

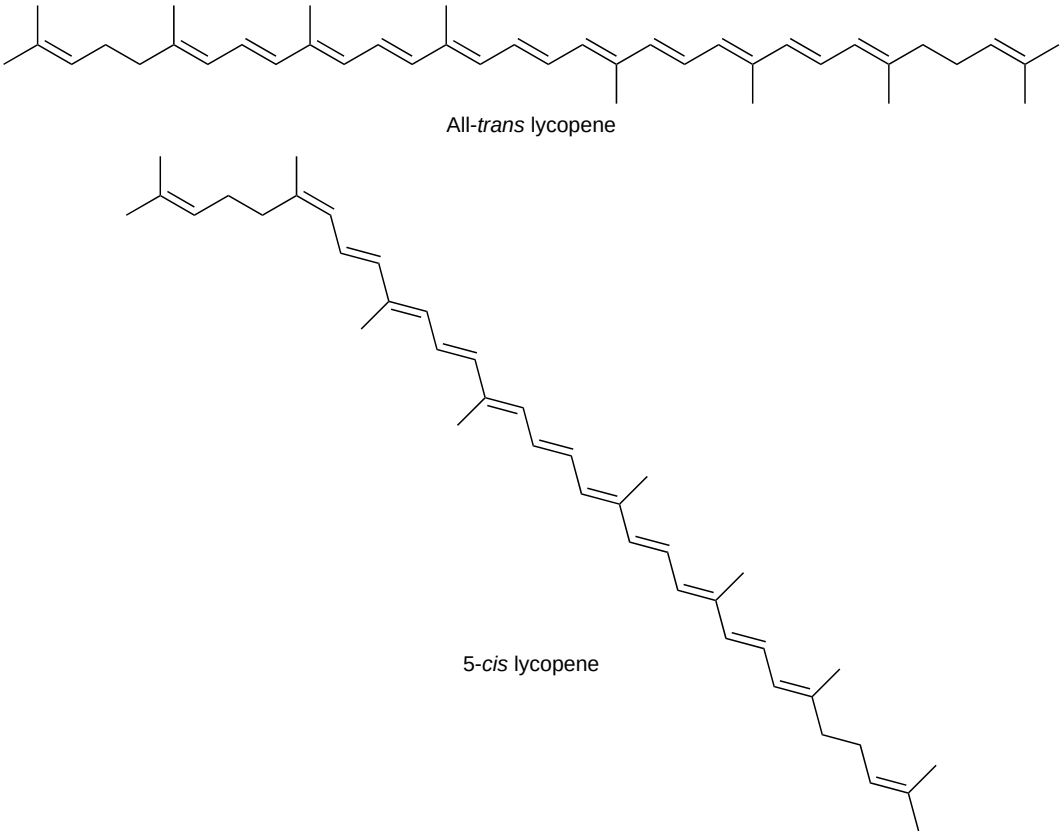
Lycopene, the most abundant pigment in ripe red tomatoes and in a few other fruits, is one of the major carotenoid pigments that is widely present in the diet of the human population in the world today. **Figure 1** illustrates the chemical formula of selected carotenoids that occur widely both in human diets and in the noncellular fraction of human blood in most regions of the world. Carotenoids are yellow-to-red in color, with lycopene being nearer the red end of the carotenoid series. However, unlike the carotenes and cryptoxanthins, it does not possess a beta-ionone ring structure at either end of the molecule, and this precludes it from becoming a precursor of vitamin A in humans and animals. Nevertheless, it is readily transformed from the all-*trans* form that is characteristic of most plants and plant foods for animals and humans, to a range of mono- and di-*cis* forms within the animal's body. In addition, oxidation to epoxides and hydroxylated derivatives occurs, although the control of these oxidation pathways and the nature of their products are not yet well understood or characterized.

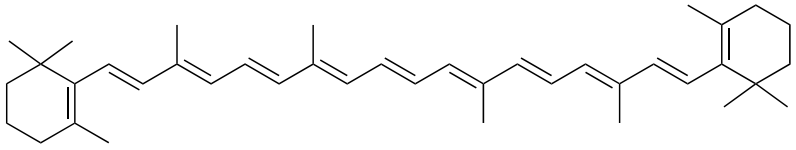
In plant tissues, where it is synthesized, lycopene is thought to help protect vulnerable photosynthetic tissues from light- and oxygen-catalyzed damage. Its role in humans and other animals, which can only obtain the pigment from their diet, is less well understood. Indeed it remains unproven that there is an essential role for lycopene in animal tissues. Nevertheless, considerable research effort is currently being undertaken to test hypotheses that are attempting to link human dietary and tissue lycopene levels to the risk of degenerative diseases, such as vascular diseases, cancers, etc., especially in older people. As discussed in more detail below, this research is being performed in a wide range of tissue culture and animal model systems and human epidemiological studies.

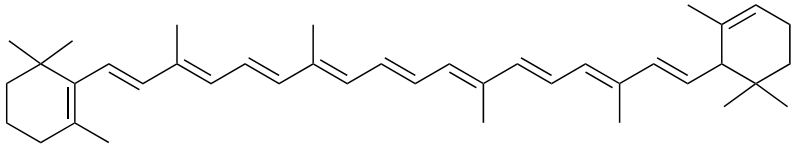
In this article, some key aspects of the chemical and physical properties, the dietary sources, biochemical status indices, and biological significance of lycopene will be described.

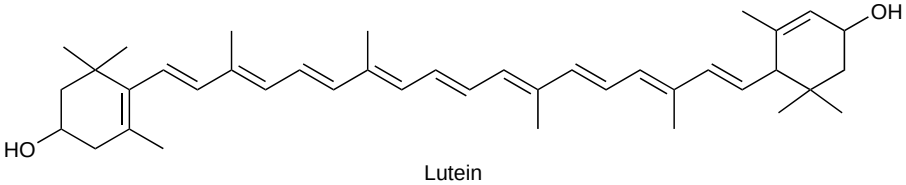
Chemical and Physical Properties of Lycopene; its Food Sources and Enteral Absorption

Lycopene is the most commonly encountered of that subgroup of the naturally occurring carotenoids that have a straight-chain poly-isoprenoid molecule without any terminal β-ionone ring structures (**Figure 1**). The chain length and number of conjugated double bonds determine the absorption spectrum, which peaks at 472 nm with a molar extinction coefficient, $\varepsilon^{1\%}$ of 3450. It is one of the most nonpolar members of the carotenoids, and in organic solution it is also one of the most easily oxidized and thus is easily destroyed, which necessitates the use of rigorous precautions against its oxidative destruction during its extraction and analysis from plants, foods, animal tissues, and body fluids. Currently, such analytical determination is usually based on high-performance liquid chromatography (HPLC), using either its characteristic light absorption property, or its natural fluorescence, or its redox character, for detection and quantitation by absorbance or fluorometric or electrochemical detection. Another characteristic that greatly affects its stability and the problems of its storage and analysis is the phenomenon of *cis-trans* isomerization. Naturally occurring lycopene in tomatoes, the major human food source of this carotenoid, is nearly 100% all-*trans* (**Figure 1**), but during the processing of food, and then during the processes of absorption and accumulation in animal tissues, there is a progressive increase in the proportion of a variety of *cis*-forms. Most of these *cis*-forms contain a single *cis*-bond (mono-*cis*-lycopene), and the 5-, 9-, 13- and 15- mono-*cis*-lycopenes account for more than 50% of the total lycopene in human serum. Smaller quantities of di-*cis*-lycopenes are normally also present. Curiously, another food source of lycopene, red palm oil, has a much higher natural proportion of the *cis*-forms of the pigment. Isomerization is catalyzed by low pH; therefore, stomach acid is believed to be a major factor in the conversion of the all-*trans*-lycopene ingested from tomatoes and their products to a mixture of *cis*-forms in the digestive tract. There is also evidence that further isomerization occurs between the digestive tract and the portal lymphatic lipid micelles. The *cis*-isomers differ from the all-*trans* form in their absorption and intertissue transportation properties, and also in their functional characteristics; for instance, they are more soluble in lipophilic solvents and structures and




All-*trans* β-carotene


All-*trans* α-carotene



Lutein

Figure 1 Structures of lycopene and certain other carotenoids found in human blood and tissues.

are less likely to aggregate into crystalline forms. However, these physicochemical differences and their biological consequences have yet to be adequately explored and described.

Of all the most common naturally occurring carotenoids, lycopene is by far the most efficient in reacting with and quenching singlet oxygen, $^1O_2$, which is a non-free-radical excited and reactive

Table 1 Lycopene content of selected foods

Food category	*Content as summarized by Clinton (1998) (mg per 100 g wet weight)*
Fresh tomatoes	0.9–4.2
Canned tomatoes	
Tomato sauce	6.2
Tomato paste	5–150
Tomato juice	5–12
Tomato ketchup	10–13
Tomato soup	
Grapefruit	3.4
Guava	5.4
Papaya	2–5.3
Watermelon	2.3–7.2

Source: Clinton SK (1998) Lycopene: Chemistry, biology and implication for human health and disease. *Nutrition Reviews* **56**: 35–51.

form of oxygen. This form of oxygen reacts rapidly with lycopene to yield nonexcited triplet oxygen and excited triplet lycopene. The latter then dissipates its extra energy by solvent interactions, thus regenerating nonexcited lycopene and preserving its original structure by recycling. However, another of its chemical interactions with molecular oxygen appears to result in irreversible oxidation to yield one or more cyclic epoxides, which then probably undergo ring-opening. Nevertheless, there are many unresolved questions about the nature and importance of the many degradation and catabolic pathways that are believed to result in the irreversible destruction of lycopene both *in vitro* and *in vivo*.

Lycopene is an essential intermediate in the pathway for synthesis of the β-ionone ring-containing carotenoids such as β-carotene in plant tissues, and in most plant tissues it is present in only minor amounts. However, in a few, including tomato fruit, watermelon, and red grapefruit, this conversion to the β-ionone ring products by the enzyme lycopene cyclase is hindered, so that the intermediate carotenoid forms, lycopene, phytoene and phytofluene, accumulate instead.

In the US, tomato products provide more than 85% of the total quantity of lycopene consumed by the human population. Mean lycopene intakes in the US are considerably greater than they are in the UK, where the mean daily intake is thought to be less than one-third that in the US, while lycopene intakes in Far Eastern countries such as China and Thailand appear to be much lower still. Wild tomatoes originated in Central America and were introduced into Europe following the opening up of the New World, and were later introduced back into North America from Europe. Because tomatoes are the major source of dietary lycopene in many human populations, some epidemiological studies have been designed on the simplistic assumption that tomato consumption can be used as a general proxy for lycopene consumption, and that any disease associations with tomato consumption can be attributed to the biological effects of lycopene. However, tomatoes also contain significant amounts of other carotenoids, vitamin C, bioflavonoids such as naringenin, and phenolic acids such as chlorogenic acid. Much of the existing epidemiological evidence for possible beneficial effects of lycopene (see below) cannot distinguish unequivocally between the biological effects of lycopene and those of the many other bioactive constituents present in tomatoes.

The bioavailability of lycopene from raw tomatoes is low, but it is greatly increased by cooking or by commercial processing such as conversion to soup, sauce, ketchup, etc., and its availability is also increased by increasing the fat content of the food. Interactions with other carotenoids are complex and have only partly been studied, for instance β-carotene in the same dish seems to increase the absorption of lycopene, but large doses of β-carotene given separately seem to decrease the lycopene content of serum lipoproteins. The contribution of several categories of tomato product to intakes in a recent survey of older people in Britain is shown in **Table 2**. The strength of the correlation between dietary lycopene intake and blood (serum or plasma) lycopene concentration varies greatly among studies

Table 2 Tomato products consumed by people aged 65 years and over in Britain

Categories of tomatoes and tomato products	*Percentage of each category consumed*
Raw tomatoes	36.2
Processed tomatoes	
Soups	8.8
Canned tomatoes	7.0
Grilled	5.4
Fried	3.2
Ketchup	0.4
Tomato-based products	
Canned food	29.5
Pizza	2.3
Other	7.1
Total	99.9

Source: Re R, Mishra GD, Thane CW, and Bates CJ (2003) Tomato consumption and plasma lycopene concentration in people aged 65 years and over in a British National Survey. *European Journal of Clinical Nutrition* **57**: 1545–1554. Reproduced with permission from Nature Publishing Group.

and clearly depends on many factors, one of which is the degree of sophistication of the food table values, since subtle differences in food sources and meal composition affect its bioavailability very considerably.

Tissue Contents and Kinetics of Lycopene Turnover

Once absorbed, passively from lipid micelles by the enterocyte, lycopene enters the portal lymphatics and thence the liver, from which it enters the peripheral bloodstream, mainly in association with the β-lipoproteins, in which it is transported to the peripheral tissues. Its half-life in plasma is of the order of 12–33 days; longer than that of β-carotene, which is less than 12 days. Clearly, many of these factors are interdependent, and there is a need for further clarification of the key independent determinants of lycopene status, and whether plasma levels can provide an adequate picture of tissue and whole body status.

Patients with alcoholic cirrhosis of the liver have greatly reduced hepatic lycopene concentrations; indeed, hepatic lycopene seems to offer a sensitive index of hepatic health. Studies of organ concentrations (**Table 3**), suggest a gradient from circulating levels in plasma to different ones in specific tissues. The different carotenoid ratios between organs (not shown) also indicate selective transport and accumulation. However, the mechanisms involved are poorly understood. No lycopene is detectable in the retina or lens of the eye, where lutein and zeaxanthin are found; however, lycopene is present in the ciliary body.

Table 3 Concentrations of lycopene reported in human tissues

Tissue	*Range of mean or median lycopene concentrations (nmol per g wet weight)*
Adrenal	1.9–21.6
Testis	4.3–21.4
Liver	0.6–5.7
Brain	2.5
Lung	0.2–0.6
Kidney	0.1–0.6
Stomach, colon	0.2–0.3
Breast, cervix	0.2–0.8
Skin	0.4
Adipose tissue	0.2–1.3
Prostate	0.1–0.6
Plasma	0.2–1.1

Values were gathered from 11 publications, all based on HPLC analyses.

Functional Properties and Tissue Health

The capacity for quenching of singlet oxygen has been mentioned above; the exceptionally high rate constant, $K = 3.1 \times 10^{10}\,mol^{-1}\,s^{-1}$, renders it one of the most efficient of known quenchers of this powerful oxidant. In the plant, it probably protects chlorophyll, which produces singlet oxygen as a by-product of photosynthesis. In experiments with lymphoid cells, lycopene provided better protection against singlet oxygen damage than several other carotenoids tested. In skin exposed to UV light, lycopene disappears much more rapidly than β-carotene. Lycopene is also able, in model systems, to inhibit the peroxidation of polyunsaturated lipids and the oxidation of DNA bases to products such as 8-hydroxydeoxyguanosine (8-OHdG). It can react directly with hydrogen peroxide and nitrogen dioxide.

Several studies in cell culture have shown a reduction in the formation of oxidation damage products such as malondialdehyde, and have found less injury to cells exposed to oxidants such as carbon tetrachloride, if lycopene (or other carotenoids) are present.

Another characteristic of lycopene and other carotenoids that may be relevant to inhibition of cancer cell growth is the modulation of gap junction cell–cell communication processes. In particular, carotenoids including lycopene have been shown to enhance the efficacy of the protein, connexin43, which helps to ensure the maintenance of the differentiated state of cells and to reduce the probability of unregulated cell division, and which is deficient in many tumors. They may also interact with and enhance the synthesis of binding proteins that downregulate the receptor for the growth-promoting hormone insulin-like growth factor-1 (IGF-1).

In certain circumstances, lycopene can reduce LDL-cholesterol levels, possibly by inhibiting hydroxymethylglutaryl CoA reductase (HMGCoA reductase), the rate-limiting enzyme for cholesterol synthesis (see below). Lycopene was shown to have modest hypocholesterolemic properties in one small clinical trial.

Health, Research Models and Epidemiological Evidence

Table 4 summarizes the various types of evidence that have been used to test the hypothesis that lycopene may have health-promoting or protective properties in man. The ultimate proof of efficacy, which would be long-term controlled intervention studies with clinical diseases and or mortality as the end points, are extremely difficult,

Table 4 Types of evidence being sought, that a nutrient such as lycopene may protect against oxidation-induced or other disease processes

1. Model *in vitro* systems, e.g., oxygen-derived free-radical trapping in pure chemical mixtures.
2. Tissue (cell and organ) cultures, e.g., reduction of optical opacity development in cultured eye lenses; reduced growth rates or apoptosis in tumor cell cultures.
3. Animal studies demonstrating a reduction of oxidation-induced damage or disease with lycopene supplements or with lycopene-rich foods such as tomatoes or tomato products.
4. Human observation studies using intermediate biochemical markers: e.g., inverse relationships between lycopene intakes or its blood levels and biochemical markers, such as lipid or DNA oxidation products.
5. Studies using pathology-related intermediate markers, e.g., arterial thickening or reduced arterial elasticity; precancerous polyposis, etc.
6. Relationships (without intervention) between tomato intakes or estimated lycopene intakes or lycopene contents of serum, plasma, or tissues (e.g., fat biopsies) and actual disease prevalence or incidence in human cross-sectional, case-control, or prospective epidemiological studies.
7. Intervention studies: lycopene supplements producing a reduction in biochemical markers of oxidation damage or in functional markers, or, eventually, in actual human disease incidence or progression.

expensive, and time-consuming to obtain, and cannot address all possible benefits in a single intervention trial.

The two disease categories that have so far received most attention for possible long-term benefits of lycopene have been the amelioration of cancers and of heart disease. Both benefits are plausible in view of the physicochemical and biological properties of lycopene outlined above, because both categories of disease are characterized by tissue damage, which is thought to be induced or exacerbated by reactive oxygen species in the environment or those generated within the body.

Evidence for Possible Anticancer Protection by Lycopene

Most of the indications with respect to cancer comes from human studies linking tomato intake, total estimated lycopene intake, and serum or plasma lycopene concentrations to the subsequent development of cancers (**Table 5**). There is a small amount of evidence from experimental animal studies, for instance, rat and mouse dimethylbenzanthracene-induced mammary tumor studies have supported the hypothesis, as has a model of spontaneous mammary tumor formation in one strain of mice, but many of the animal models of tumor promotion have been criticized as being too dissimilar from the likely processes of spontaneous tumorigenesis in humans.

Partly for historical reasons, there has been a particular interest in prostate cancer (**Table 5**). A large and early trial in the US (US Health Professionals Follow-up Study) reported an impressive difference between groups with high and low intakes of tomatoes and hence of lycopene for subsequent prostate cancer development, which was not shared with other carotenoids. Plausibility was enhanced by the fact that although human prostate lycopene concentrations are not especially high on an absolute basis (**Table 3**), they are higher than those of other carotenoids in this tissue. Subsequent studies have had variable outcomes. A small pilot study reported that tomato oleoresin supplements given for a short period to prostate cancer sufferers who were due for radical prostatectomy resulted in smaller tumor size and other apparent benefits, but this trial now needs to be repeated on a larger scale.

Table 5 Summary evidence for possible lycopene protection against prostate cancer

No. of studies	*Locations*	*Total no. of participants*	*Types of trial*	*Outcome conclusion*
2	Greece, Canada	937	Case–control (intake of tomato or lycopene, or blood level)	Significant association
7	USA, UK, Canada, New Zealand	3824	As above	No significant association
3	USA	954	Prospective studies based on dietary estimates	Significant association
1	Netherlands		As above	No association
3	USA	723	Prospective studies based on serum or plasma lycopene	Inconclusive; one study found a marginal ($P = 0.05$) benefit vs. aggressive cancer

Table 6 Summary evidence of association of relatively high serum or plasma lycopene with lowered risk of cardiovascular disease (CVD)

Study	*Location*	*Sex (total participants)*	*Types of trial and outcome measures*	*Outcome conclusion*
Euramic	Europe, multicenter	M (1379)	C-C, MI	Significant association with protection[a]
ARIC	USA	M + F (462)	C-C, IMT	NS
Street	USA	M + F (369)	NC-C, MI in smokers	NS
Rotterdam	Netherlands	M + F (216)	C-C, PC	Significant association with protection
Bruneck	Italy	M + F (392)	CS + PFU, PC	NS
Linkoping –Vinus	Sweden and Lithuania	M (210)	CS, mortality from heart disease	NS
Kuopio (KHID)	Finland	M (725)	PFU, acute coronary event or stroke	Significant association with protection
Kuopio (ASP)	Finland	M + F (520)	IMT	Males significant; females not significant.

[a]No association with plasma β-carotene in this study.

C-C, case–control study; NC-C, nested case–control study; CS, cross-sectional study; PFU, prospective follow-up study; MI, myocardial infarct; IMT, intima-media thickness estimate; PC, plaque count. NS, no significant evidence for protection.Significance generally after appropriate adjustment for other known CVD risk factors.

Several studies have provided evidence for protection of certain regions of the digestive tract against tumor occurrence or growth. Two studies, one in Iran and another in Italy, found an inverse relationship between esophageal cancer and tomato consumption. Two Italian and one Japanese study reported evidence for protection against gastric cancer, and two studies claimed a reduction in pancreatic cancer. Results with others cancer have been mixed and inconclusive.

Lycopene and Cardiovascular Disease

Table 6 summarizes the evidence. The European Multicentre Euramic Study, which reported that risk of developing myocardial infarct was inversely related to lycopene intake, after appropriate adjustment for other cardiovascular risk factors. Some Scandinavian studies have subsequently supported this claim; moreover, lycopene is capable of reducing LDL-cholesterol levels, possibly by inhibiting hydroxymethylglutaryl CoA reductase (HMGCoA reductase), the rate-limiting enzyme for cholesterol synthesis.

Other Disease-Related Investigations

In an organ culture model, some evidence for protection of rat lenses against induction of cataractogenesis has been reported. There is good reason to believe that carotenoids in general may play a role in the protection of ocular tissues against the damaging effects of UV light and of reactive oxygen substances, whose exposure to light carries some analogy with the known functions of carotenoids in plant tissues. A possible protective role in the ciliary body and iris has been proposed, but not yet tested.

Conclusions

Clearly lycopene possesses chemical and biological properties, which make it a very attractive candidate for tissue protection and reduction of disease, especially degenerative diseases. Lycopene probably interacts more efficiently with one particular reactive oxygen species, singlet oxygen, than any other commonly occurring nutrient. It appears to share with several other carotenoids the capacity to reduce lipid peroxidation and DNA oxidative damage, and to enhance cell–cell gap junction communication and to protect normal IGF-1 function. It may reduce cholesterol formation and its tissue accumulation in some circumstances. Studies related to cancers and cardiovascular disease are ongoing and are attracting increased research interest.

See also: **Alcohol**: Disease Risk and Beneficial Effects. **Antioxidants**: Diet and Antioxidant Defense; Observational Studies; Intervention Studies. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Carotenoids**: Chemistry, Sources and Physiology; Epidemiology of Health Effects. **Fruits and Vegetables**. **Vitamin A**: Biochemistry and Physiological Role; Deficiency and Interventions.

Further Reading

Arab L and Steck S (2000) Lycopene and cardiovascular disease. *American Journal of Clinical Nutrition* **71**(supplement): 1691S–1695S.

Arab L, Steck-Scott S, and Bowen P (2001) Participation of lycopene and beta-carotene in carcinogenesis: Defenders, aggressors or passive bystanders? *Epidemiologic Reviews* **23**: 211–230.

Britton G (1995) Structure and properties of carotenoids in relation to function. *FASEB J* **9**: 1551–1558.

Clinton SK (1998) Lycopene: chemistry, biology, and implications for human health and disease. *Nutrition Reviews* **56**: 35–51.

Gerster H (1997) The potential of lycopene for human health. *Journal of the American College of Nutrition* **16**: 109–126.

Giovanucci E (1999) Tomatoes, tomato-based products, lycopene, and cancer: review of the epidemiologic literature. *Journal of the National Cancer Institute* **91**: 317–331.

International Symposium on the Role of Tomato Products and Carotenoids in Disease Prevention (2002) 14 review articles by different authors, plus 17 symposium abstracts. *Experimental Biology and Medicine* **227**: 843–937.

Nguyen ML and Schwartz SJ (1999) Lycopene: chemical and biological properties. *Food Technology* **53**: 38–45.

Rao AV and Agarwal S (1999) Role of lycopene as antioxidant carotenoid in the prevention of chronic diseases: A review. *Nutrition Research* **19**: 305–323.

Stahl W and Sies H (1996) Lycopene: a biologically important carotenoid for humans? *Archives of Biochemistry and Biophysics* **336**: 1–9.

Weisburger J (1998) International symposium on lycopene and tomato products in disease prevention. *Proceedings of the Society for Experimental Biology and Medicine* **218**: 93–143.

M

MAGNESIUM

C Feillet-Coudray and Y Rayssiguier, National Institute for Agricultural Research, Clermont-Ferrand, France

Magnesium (Mg), the second intracellular cation after sodium, is a essential mineral. It is a critical cofactor in more than 300 enzymatic reactions. It may be required for substrate formation (Mg-ATP) and enzyme activation. It is critical for a great number of cellular functions, including oxidative phosphorylation, glycolysis, DNA transcription, and protein synthesis. It is involved in ion currents and membrane stabilization. Mg deficiency may be implicated in various metabolic disorders, including cardiovascular diseases, immune dysfunction and free radical damage.

Magnesium Metabolism

Distribution of Mg within the Body

The normal adult body contains approximately 25 g of Mg, with more than 60% in bone tissue (**Table 1**). Only a fraction of bone Mg (at the surface of the bone crystal) is exchangeable with extracellular Mg. The muscle contains 25% of total body Mg, and extracellular Mg accounts for only 1%. Plasma Mg is approximately 0.8 mmol/l, half of which is ionised and active in physiological reactions half bound to proteins or complexed to anions. In cells, Mg is associated with various structures, such as the nucleus and intracellular organelles, and free Mg accounts for 1–5% of total cellular Mg. Intracellular free Mg is maintained at a relatively constant level, even if extracellular Mg level varies. This phenomenon is due to the limited permeability of the plasma membrane to Mg and the existence of specific Mg transport systems that regulate the rates at which Mg is taken up by cells or extruded from cells. Mechanisms by which Mg is taken up by cells have not been completely elucidated, and Mg efflux particularly requires the antiport Na^{+}/Mg^{2+}. Various hormonal and pharmacological factors influence Mg transport, and it can be assumed that recent developments in molecular genetics will lead to the identification of proteins implicated in Mg transport.

Intestinal Absorption

Net Mg absorption results from dietary Mg absorption and Mg secretion into the intestinal tract via bile and gastric and pancreatic juice. In healthy adults, 30–50% of dietary Mg is absorbed. The secreted Mg is efficiently reabsorbed and endogenous fecal losses are only 20–50 mg/day. Mg absorption occurs along the entire intestinal tract, but the distal small intestine (jejunum and ileum) is the primary site. It is essentially a passive intercellular process by electrochemical gradient and solvent drag. The active transport occurs only for extremely low dietary Mg intake and its regulation is unknown. Mg uptake in the brush border may be mediated by a Mg/anion complex, and Mg efflux across the basolateral membrane may involve Na^{+}/Mg^{2+} antiport systems. A gene implicated in Mg deficit in humans has been identified. It is expressed in intestine and kidney and appears to encode for a protein that combines Ca- and Mg-permeable channel properties with protein kinase activity. This gene may be implicated in Mg absorption. Because of the importance of the passive process, the quantity of Mg in the digestive tract is the major factor controlling the amount of Mg absorbed.

The possibility of an adaptative increase in the fraction of Mg absorbed as Mg intake is lowered is controversial. In fact, experimental studies indicate that fractional intestinal absorption of Mg is directly proportional to dietary Mg intake. Because only soluble Mg is absorbed, all the factors increasing Mg solubility increase its absorption while formation of insoluble complexes in the intestine may decrease Mg absorption. Most well-controlled studies indicate that high calcium intake does not affect intestinal Mg absorption in humans. In contrast, dietary phytate in excess impairs Mg

Table 1 Magnesium in human tissues

	% distribution	*Concentration*
Bone	60–65	0.5% of bone ash
Muscle	27	6–10 mmol/kg wet weight
Other cells	6–7	6–10 mmol/kg wet weight
Extracellular	<1	
Erythrocytes		2.5 mmol/l
Serum		0.7–1.1 mol/l
Free	55	
Complexed	13	
Bound	32	
Mononuclear blood cells		2.3–3.5 fmol/cell
Cerebrospinal fluid		1.25 mmol/l
Free	55	
Complexed	45	
Sweat		0.3 mmol/l (in hot environment)
Secretions		0.3–0.7 mmol/l

From *Molecular Aspects of Medicine*, vol. 24, Vormann J: Magnesium: nutrition and metabolism, pp. 27–37, Copyright 2003, with permission from Elsevier.

absorption by formation of insoluble complexes in the intestinal tract. Negative effects of a high intake of dietary fiber have often been reported, but these actions have certainly been overestimated. In fact, only the impact of purified fiber was considered, but fiber-rich diets are a major source of Mg and roles of the intestinal fermentation and the large bowel in mineral absorption were neglected. It was demonstrated in animal models that fermentable carbohydrates (oligosaccharides and resistant starch) enhance Mg absorption in the large bowel and that a similar effect exists in humans. Other nutrients may influence Mg absorption but these effects are important only at low dietary Mg intake.

Urinary Excretion

Magnesium homeostasis is essentially regulated by a process of filtration–reabsorption in the kidney. Urinary Mg excretion increases when Mg intake is in excess, whereas the kidney conserves Mg in the case of Mg deprivation. Usually, 1000 mmol/24 h of Mg is filtered and only 3 mmol/24 h is excreted in urine.

A total of 10–15% of the filtered Mg is reabsorbed in the proximal tubule by a passive process. The majority of filtered Mg (65%) is reabsorbed in the thick ascending loop of Henle. The reabsorption in this segment is mediated by a paracellular mechanism involving paracellin-1. It is also related to sodium transport by a dependence on the transepithelial potential generated by NaCl absorption. Thus, factors that impair NaCl reabsorption in the thick ascending loop of Henle, such as osmotic diuretics, loop diuretics, and extracellular fluid volume expansion, increase Mg excretion. At least 10–15% of the filtered Mg is reabsorbed in the distal tubule. The reabsorption occurs via an active transcellular mechanism and is under the control of special divalent cation-sensing receptors. Thus, elevated plasma Mg concentrations inhibit reabsorption of Mg from the distal tubule, leading to an increased magnesuria. Other active transport may also exist since some hormones (parathyroid hormone, glucagon, calcitonin, and insulin) may increase Mg reabsorption. Other factors may also influence Mg reabsorption, such as hypercalciuria or hypophosphatemia, which inhibit the tubular reabsorption of Mg. Metabolic alkalosis leads to renal Mg conservation, whereas metabolic acidosis is associated with urinary Mg wasting. Thus, the chronic low-grade metabolic acidosis in humans eating Western diets may contribute to decreased Mg status.

Dietary Sources of Magnesium

Mg is present in all foods, but the Mg content varies substantially (**Table 2**). Cereals and nuts have high Mg content. Vegetables are moderately rich in Mg, and meat, eggs, and milk are poor in Mg. A substantial amount of Mg may be lost during food processing, and refined foods generally have a low Mg content. In addition to Mg content, it is important to consider the Mg density of food (i.e., the quantity of Mg per unit of energy). Vegetables, legumes, and cereals thus contribute efficiently to daily Mg intake, whereas fat- and/or sugar-rich products have a minor contribution. Some water can also be a substantial source of Mg, but it depends on the area from which the water derives.

Table 2 Mg density of foods

Food	*Magnesium density (mg/MJ)*
Vegetables (lettuce, broccoli)	211
Legumes (bean)	113
Whole cereal (wheat)	104
Nuts (almond)	105
Fruits (apple)	30
Fish (cod)	75
Meat (roast beef)	40
Whole milk	38
Cheese (camembert)	15
Eggs	18
Dessert	
Biscuit	10
Chocolate	52

From Répertoire Général des Aliments (1996).

Requirements

Assessment of Mg Status

Several potential markers for estimating daily Mg requirement have been suggested. Plasma Mg concentration is the most commonly used marker to assess Mg status. In healthy populations, the plasma Mg value is 0.86 mmol/l and the reference value is 0.75–0.96 mmol/l. A low plasma Mg value reflects Mg depletion, but a normal plasma Mg level may coexist with low intracellular Mg. Thus, despite its interest, plasma Mg is not a good marker of Mg status.

Ion-specific electrodes have become available for determining ionized Mg in plasma, and this measurement may be a better marker of Mg status than total plasma Mg. However, further investigation is necessary to achieve a standardized procedure and to validate its use as an appropriate marker of Mg status.

Erythrocyte Mg level is also commonly used to assess Mg status, and the normal value is 2.06–2.54 mmol/l. However, erythrocyte Mg level is under genetic control, and numerous studies have shown no correlation between erythrocyte Mg and other tissue Mg.

The total Mg content of white blood cells has been proposed to be an index of Mg status. However, lymphocytes, polymorphonuclear blood cells, and platelets may have protective mechanisms against intracellular Mg deficiency, and the determination of total Mg content in leukocytes and platelets to assess Mg status is of questionable usefulness.

Mg excretion determination is helpful for the diagnosis of Mg deficit when there is an hypomagnesemia. In healthy populations, the urinary Mg value is 4.32 mmol/day and the reference value is 1.3–8.2 mmol/day. In the presence of hypomagnesemia, normal or high urinary Mg excretion is suggestive of renal wasting. On the contrary, Mg urinary excretion lower than normal values is convincing evidence of Mg deficiency.

The parenteral loading test is probably the best available marker for the diagnosis of Mg deficiency. The Mg retention after parenteral administration of Mg seems to reflect the general intracellular Mg content, and a Mg retention more than 20% of the administered Mg suggests Mg deficiency. However, this test is not valid in the case of abnormal urinary Mg excretion and is contraindicated in renal failure.

Determination of exchangeable Mg pools using Mg stable isotopes is an interesting approach to evaluate Mg status. In fact, Mg exchangeable pool sizes vary with dietary Mg in animals. However, more studies are necessary to better appreciate the relationship between Mg status and exchangeable Mg pool size in humans.

Magnesium Deficit

Two types of Mg deficit must be differentiated. Dietary Mg deficiency results from an insufficient intake of Mg. Secondary Mg deficiency is related to dysregulation of the control mechanisms of Mg metabolism.

Dietary Mg Deficiency

Severe Mg deficiency is very rare, whereas marginal Mg deficiency is common in industrialized countries. Low dietary Mg intake may result from a low energy intake (reduction of energy output necessary for physical activity and thermoregulation, and thus of energy input) and/or from low Mg density of the diet (i.e., refined and/or processed foods). Moreover, in industrialized countries, diets are rich in animal source foods and low in vegetable foods. This leads to a dietary net acid load and thus a negative effect on Mg balance. In fact, animal source foods provide predominantly acid precursors (sulphur-containing amino acids), whereas fruits and vegetables have substantial amounts of base precursor (organic acids plus potassium salts). Acidosis increases Mg urinary excretion by decreasing Mg reabsorption in the loop of Henle and the distal tubule, and potassium depletion impairs Mg reabsorption. Mg deficiency treatment simply requires oral nutritional physiological Mg supplementation.

Secondary Mg Deficiency

Failure of the mechanisms that ensure Mg homeostasis, or endogenous or iatrogenic perturbing factors of Mg status, leads to secondary Mg deficit. Secondary Mg deficiency requires a more or less specific correction of its causal dysregulation.

Intestinal Mg absorption decreases in the case of malabsorption syndromes, such as chronic diarrhoea, inflammatory enteropathy, intestinal resection, and biliary and intestinal fistulas.

Hypermagnesuria is encountered in the case of metabolic and iatrogenic disorders, such as primary and secondary hyperaldosteronism (extracellular volume expansion), hypercalcemia (competition Ca/Mg at the thick ascending loop of Henle), hyperparathyroidism, and phosphate or potassium depletion. Hypermagnesuria may also result from tubulopathy, as the selective defect of the Mg tubular reabsorption (chromosome 11q23), Bartter's

syndrome (thick ascending loop of Henle), or Gitelman's syndrome (distal convoluted tubule).

Administration of medications can be a causal factor in the development of secondary Mg deficiency. Administration of diuretics is the main cause of iatrogenic deficit because it decreases NaCl reabsorption in the thick ascending loop of Henle and thus increases the fractional excretion of Mg.

Causes of Mg Deficit

Complex relations exist between Mg and carbohydrate metabolism. Diabetes is frequently associated with Mg deficit and insulin may play an important role in the regulation of intracellular Mg content by stimulating cellular Mg uptake. Hypomagnesemia is the most common ionic abnormality in alcoholism because of poor nutritional status and Mg malabsorption, alcoholic ketoacidosis, hypophosphatemia, and hyperaldosteronism secondary to liver disease.

Stress can contribute to Mg deficit by stimulating the production of hormones and thus increasing urinary Mg excretion and by impairing neurohormonal mechanisms that spare Mg.

Consequences of Mg Deficit and Implications in Various Metabolic Diseases

Mg deficit causes neuromuscular manifestations, including positive Chvostek and Trousseau signs, muscular fasciculations, tremor, tetany, nausea, and vomiting. The pathogenesis of the neuromuscular irritability is complex, and it implicates the central and peripheral nervous system, the neuromuscular junction, and muscle cells.

Mg deficit perturbs Ca homeostasis and hypocalcemia is a common manifestation of severe Mg deficit. Impaired release of parathyroid hormone (PTH) and skeletal end organ resistance to PTH appear to be the major factors implicated, probably by a decrease in adenylcyclase activity.

Perturbations in the action and/or metabolism of vitamin D may also occur in Mg deficit. Because Mg plays a key role in skeletal metabolism, Mg deficit may be a possible risk factor for osteoporosis. However, epidemiologic studies relating Mg intake to bone mass or rate of bone loss have been conflicting, and further investigation is necessary to clarify the role of Mg in bone metabolism and osteoporosis.

Hypokalemia is frequently encountered in Mg deficit. This is due to an inhibition of Na,K-ATPase activity that impairs K and Na transport in and out of the cell and to stimulation of renin and aldosterone secretion that increases K urinary excretion.

There is increasing evidence that Mg deficiency may be involved in the development of various pathologies. Mg deficit is frequent in diabetes and can be a factor in insulin resistance. It can modify insulin sensitivity, probably by influencing intracellular signaling and processing. Mg deficit has also been implicated in the development or progression of micro- and macroangiopathy and neuropathy.

Mg deficit appears to act as a cardiovascular risk factor. Experimental, clinical, and epidemiological evidence points to an important role of Mg in blood pressure regulation. Mg deficit can lead to cardiac arrhythmias and to increased sensitivity to cardiac glucosides. Mg deficit may also play a role in the development of atherosclerosis. In experimental animal models, dietary Mg deficiency results in dyslipidemia, increased sensitivity to oxidative stress, and a marked proinflammatory effect, thus accelerating atherogenosis. Macrophages and polynuclear neutrophils are activated and synthesize a variety of biological substances, some of which are powerful inducers of inflammatory events (cytokines, free radicals, and eicosanoids). The effect of Mg depletion or Mg supplementation may result in the ability of Mg to modulate intracellular calcium. Pharmacological doses of Mg may reduce morbidity and mortality in the period following infarction. The beneficial effect of Mg may result from calcium-antagonist action, decreased platelet aggregation, and decreased free radical damage.

Magnesium Excess

Magnesium overload can occur in individuals with impaired renal function or during massive intravenous administration of Mg. It is most often iatrogenic. Clinical symptoms such as drowsiness and hyporeflexia develop when plasma Mg is 2- or 3-fold higher than the normal value.

Recommended Dietary Allowances

The Estimated Average Requirement (EAR) is the nutrient intake value that is estimated to meet the requirement of 50% of individuals in a life stage and a gender group. Balance studies and data on stable isotopes suggest an EAR of 5 mg/kg/day for males and females. This value is greater during growth in adolescents and is estimated to be 5.3 mg/kg/day. The Mg requirement is also higher during pregnancy because of Mg transfer to the fetus in the last 3 months; therefore, an additional 35 mg/day is recommended.

In infants, the determination of the Adequate Intake (AI) is based on the Mg content of mother's milk and the progressive consumption of solid food.

Table 3 Recommended dietary allowances of Mg

Age	RDA (mg/day)		AI (mg/day)	
	Male	Female	Male	Female
0–6 months			30	30
6–12 months			75	75
1–3 years	80	80		
4–8 years	130	130		
9–13 years	240	240		
14–18 years	410	360		
19–30 years	400	310		
31–50 years	420	320		
51–70 years	420	320		
<70 years	420	320		
Pregnancy		+40		
Lactation		+0		

From the Institute of Medicine (1997).

Thus, the AI is 30 mg/day during the first 6 months of life and 75 mg/day the second 6 months of life.

The Recommended Dietary Allowance (RDA) is the average daily dietary intake that is sufficient to meet the nutrient requirement of 97.5% of individuals and is set at 20% above the EAR +2 CVs where the CV is 10%. During recent years, dietary reference intakes have been revised by the US Institute of Medicine. The recommended intakes of Mg are given in **Table 3**. It is not known whether decreased urinary Mg and increased maternal bone resorption provide sufficient amounts of Mg to meet increased needs during lactation. Thus, the French Society for Nutrition suggests adding 30 mg/day to intake for lactation.

The intake of Mg has been determined in various populations. Evidence suggests that the occidental diet is relatively deficient in Mg, whereas the vegetarian diet is rich in Mg. For instance, the mean Mg intake of the subjects in the French Supplementation with Antioxidant Vitamins and Minerals Study was estimated to be 369 mg/day in men and 280 mg/day in women. Thus, 77% of women and 72% of men had dietary Mg intakes lower than the RDA, and 23% of women and 18% of men consumed less than two-thirds of the RDA.

Conclusion

Based on evidence of low Mg intake in industrialized countries, intervention studies to improve Mg status and to assess its impact on specific health outcomes are required.

See also: **Calcium**. **Cereal Grains**. **Electrolytes**: Water–Electrolyte Balance. **Fruits and Vegetables**. **Malabsorption Syndromes**. **Vitamin D**: Rickets and Osteomalacia.

Further Reading

Coudray C, Demigné C, and Rayssiguier Y (2003) Effects of dietary fibers on magnesium absorption in animals and humans. *Journal of Nutrition* **133**: 1–4.

Durlach J (1988) *Magnesium in Clinical Practice* London: John Libbey.

Elin RJ (1989) Assessment of magnesium status. In: Itokawa Y and Durlach I (eds.) *Magnesium in Health and Disease*, pp. 137–146. London: John Libbey.

Feillet-Coudray C, Coudray C, Gueux E, Mazur A, and Rayssiguier Y (2002) A new approach to evaluate magnesium status: Determination of exchangeable Mg pool masses using Mg stable isotope. *Magnesium Research* **15**: 191–198.

Galan P, Preziosi P, Durlach V *et al.* (1997) Dietary magnesium intake in a French adult population. *Magnesium Research* **10**: 321–328.

Institute of Medicine (1997) *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D and Fluoride*. Washington, DC: National Academy Press.

Rayssiguier Y, Mazur A, and Durlach J (2001) *Advances in Magnesium Research, Nutrition and Health* London: John Libbey.

Répertoire général des aliments (1996) *Table de Composition Minérale*. Paris: Tec & Doc, Lavoisier.

Rude RK (1998) Magnesium deficiency: A cause of heterogeneous disease in humans. *Journal of Bone Mineral Research* **13**(4): 749–758.

Shils ME (1994) Magnesium. In: Shils ME, Olson JA, and Shike M (eds.) *Modern Nutrition in Health and Disease*, 8th ed, pp. 164–184. Philadelphia, PA: Lea & Febiger.

Vormann J (2003) Magnesium: Nutrition and metabolism. *Molecular Aspects of Medicine* **24**: 27–37.

Wilkinson SR, Welch RM, Mayland HF, and Grunes DL (1990) Magnesium in plants: Uptake, distribution, function and utilization by man and animals. In: Sigel H and Sigel A (eds.) *Compendium of Magnesium and Its Role in Biology, Nutrition and Physiology*, pp. 33–56. New York: Marcel Dekker.

MALABSORPTION SYNDROMES

P M Tsai and C Duggan, Harvard Medical School, Boston, MA, USA

The human gastrointestinal tract has an impressive capacity for water, electrolyte, and nutrient absorption. In some disease states, however, this excess capacity is outpaced by either intestinal secretion or inadequate absorption. Malabsorption is defined as the inability of the gastrointestinal tract to adequately absorb nutrients. Although strictly speaking, malabsorption is distinct and contrasted with maldigestion (inadequate breakdown of nutrients in the intestinal lumen), the therapeutic implications of these two conditions are often similar. Multiple causes of malabsorption exist (e.g., inflammatory bowel disease, cystic fibrosis, and short bowel syndrome). We review the pathophysiology, symptoms, and nutritional therapies for common malabsorption syndromes.

Pathophysiology and Symptoms

Malabsorption can occur when any of the several steps in nutrient digestion, absorption, and/or assimilation are interrupted; see **Table 1** for a list of congenital defects in nutrient assimilation. Carbohydrate malabsorption can occur, for instance, when intestinal disaccharidases are reduced in concentration at the enterocyte. The brush border membrane produces four disaccharidases that are important in carbohydrate digestion. These enzymes are sucrase–isomaltase, maltase–glucoamylase, trehalase, and lactase–phlorizin hydrolase. Worldwide, lactase deficiency is the most common type of acquired disaccharidase deficiency since much of the world's population exhibits a noticeable reduction in intestinal lactase concentration after the age of 2 years. In addition, infants and children with diarrheal disease may suffer from acquired lactase deficiency due to intestinal villous damage that is often temporary. With either congenital or acquired lactase deficiency, malabsorbed carbohydrate remains in the intestinal lumen and exerts an osmotic pull on fluids and electrolytes, leading to abdominal cramping and loose stools. Malabsorbed carbohydrate can be metabolized by gastrointestinal tract bacteria, and the fermented gas produced is associated with flatulence and bloating. Bacterial overgrowth of the small intestine, as seen with short bowel syndrome, can also be associated with carbohydrate malabsorption.

Steatorrhea, excessive fat in the stools, results from fat malabsorption and can have several causes, most notably pancreatic insufficiency due to cystic fibrosis, chronic pancreatitis, Shwachman–Diamond syndrome, and Johanson–Blizzard syndrome. Failure of pancreatic secretion of lipase, amylase, and other digestive enzymes leads to persistence of dietary fat in the intestinal lumen, causing bloating, abdominal pain, and bulky, foul-smelling, oily stools. The stools often float due to a high gas content and test positive for fat. Patients also complain of blunted appetite and nausea. Other causes of fat malabsorption include hepatobiliary disease with inadequate bile salt circulation, severe mucosal disease, and short bowel syndrome.

The most common cause of protein malabsorption is so-called protein-losing enteropathy. Etiologies include diffuse mucosal disease such as celiac disease or Crohn's disease, elevated right heart pressure with resultant dilatation of lymphatics and leakage of lymph into the lumen, and colitides such as *Shigella* or *Salmonella* infections. Since protein is a relatively minor component of dietary energy compared with carbohydrate and fat, symptoms of protein malabsorption can sometimes be minimal. However, infectious colitis or exacerbations of inflammatory bowel disease often present with frequent loose stools, which may be bloody. Rare, congenital etiologies of protein malabsorption include enterokinase and trypsinogen deficiencies (**Table 1**).

Finally, the malabsorption of various micronutrients can occur in conjunction with or separate from the macronutrient malabsorption syndromes noted previously. For instance, steatorrhea can be accompanied by excessive fecal losses of the fat-soluble vitamins A, D, E, and K as well as calcium and other minerals. Alternatively, atrophic gastritis or surgical resection of the terminal ileum can lead to vitamin B_{12} malabsorption in the absence of any symptoms of diarrhea. Proximal bowel resection can result in iron, zinc, and calcium malabsorption. A rare cause of micronutrient inadequacy is abetalipoproteinemia, in which fat-soluble nutrients are normally digested and absorbed by the intestine but are not delivered to the circulation due to defective transepithelial transport. Other rare causes of micronutrient malabsorption are noted in **Table 1**.

Table 1 Congenital defects in nutrient assimilation[a]

Disorder	*Enzyme/protein affected*	*Symptoms*
Carbohydrate digestion		
Congenital lactase deficiency	Lactase	Lactose-induced diarrhea
Hypolactasia	Lactase	Lactose-induced diarrhea
Congenital sucrase–isomaltase deficiency	Sucrase–isomaltase	Sucrose-induced diarrhea
Glucoamylase deficiency	Glucoamylase	Starch-induced diarrhea
Trehalase deficiency	Trehalase	Trehalose-induced diarrhea
Carbohydrate absorption		
Glucose–galactose malabsorption	Sodium–glucose cotransport (SGLT1)	Glucose-induced diarrhea
Fructose malabsorption	Facilitative fructose transport (GLUT5)	Fructose-induced diarrhea
Fanconi–Bickel syndrome	Facilitative glucose transport (GLUT2)	Diarrhea and nephropathy
Protein digestion		
Enterokinase deficiency	Enterokinase	Diarrhea and edema
Trypsinogen deficiency	Trypsinogen	Diarrhea and edema
Fat digestion		
Pancreatic lipase deficiency	Pancreatic lipase	Steatorrhea
Fat assimilation		
Abetalipoproteinemia	Microsomal triglyceride transfer protein	Steatorrhea
Hypobetalipoproteinemia	Apolipoprotein B	Steatorrhea
Chylomicron retention disease	Sar1-ADP-ribosylation factor family GTPases	Steatorrhea
Primary bile acid malabsorption	Sodium–bile acid transporter	Steatorrhea, bile acid diarrhea
Tangier disease	ATP binding cassette transporter 1	Hepatosplenomegaly
Sitosterolemia	ATP binding cassette subfamily G, member 8	Atherosclerosis
Ion and metal absorption		
Congenital sodium diarrhea	Defective Na^+/H^+ exchange	Secretory diarrhea
Congenital chloride diarrhea	Defective $Cl^-/HCO3^-$ exchange	Secretory diarrhea
Cystic fibrosis	CFTR	Pancreatic insufficiency, meconium ileus
Acrodermatitis enteropathica	Zinc and iron-regulated transport proteins (ZIP4)	Diarrhea and dermatitis
Menkes disease	Copper transporter	Developmental delay
Wilson's disease	Copper transporter	Cirrhosis
Primary hypomagnesemia	Paracellin	Seizures, deafness and polyuria
Hemachromatosis	Hepcidin, others	Cirrhosis, cardiomyopathy, diabetes
Vitamin absorption		
Folate malabsorption	?	Macrocytic anemia, diarrhea, developmental delay
Congenital pernicious anemia	Intrinsic factor	Macrocytic anemia, developmental delay
Imerslund–Graesbeck syndrome	Cubilin, amnionless	Macrocytic anemia, proteinuria
Congenital deficit of transcobalamin II	transcobalamin II	Macrocytic anemia, diarrhea, developmental delay
Thiamine-responsive megaloblastic anemia	Thiamine transport protein	Anemia, diabetes, cranial nerve defects
Familial retinol binding protein (RBP) deficiency	RBP-4	Vitamin A deficiency
Selective vitamin E deficiency	α-Tocopheral transport protein	Vitamin E deficiency

[a]Included are congenital defects that are associated with gastrointestinal symptoms and/or nutritional deficiencies. Congenital defects not included here include multiple defects in amino acid absorption.
Adapted from Martin M and Wright EM (2004) Congenital intestinal transport defects. In: Walker WA, Goulet O, Kleinman RE *et al.* (eds.) *Pediatric Gastrointestinal Disease: Pathophysiology, Diagnosis, Management*, 4th edn. Hamilton, Ontario: BC Decker.

General Nutritional Management of Malabsorption

As with all nutritional disorders, a thorough nutritional assessment is needed to plan rational therapy of malabsorption. Important historical points to review include duration of symptoms, underlying etiology of malabsorption, ability to meet nutritional needs by mouth, the presence of food allergies, and concurrent medical and surgical problems. The patient's nutritional status (weight, height, body mass index, and their respective percentiles) should

be determined. Tests of body composition such as arm anthropometrics, bioelectrical impedance, or DXA studies should be considered. If the underlying cause of malabsorption is not known, diagnostic gastrointestinal endoscopy, laboratory studies, and/or imaging studies are indicated.

Specific Nutritional Management of Malabsorption

Fluids and Electrolytes

Diarrhea is usually the most distressing problem for patients with malabsorption and may cause dehydration. Care should be taken to correct fluid losses with appropriately designed oral rehydration solutions. Even in the setting of massive secretory diarrhea, such as seen with cholera infections, oral rehydration solutions are effective at treating dehydration. Data support the safety and efficacy of oral rehydration solutions of reduced osmolarity in children with dehydration from acute diarrhea. An oral rehydration solution composed of glucose 75 mmol/L, sodium 75 mmol/L, potassium 20 mmol/L, base 30 mEq/L, and osmolality 245 mOsm/L is well suited for the rehydration and maintenance therapy during dehydration due to diarrhea.

In some cases of severe diarrhea, parenteral hydration is the mainstay of therapy. Examples include glucose–galactose malabsorption, congenital chloride diarrhea, microvillous inclusion disease, and tufting enteropathy. These cases, as well as other severe causes of more common malabsorptive syndromes, also frequently require the use of parenteral nutrition therapy.

Carbohydrate Malabsorption

Lactose intolerance Lactose intolerance is defined by the occurrence of symptoms due to the inability to digest lactose, the main carbohydrate in milk. These symptoms may include abdominal pain, bloating, diarrhea, or flatulence. Lactose malabsorption is attributed to a relative deficiency of the disaccharidase lactase. Primary lactase deficiency is a condition in which lactase activity declines after weaning. Secondary lactose intolerance is usually due to mucosal injury associated with a condition or disease such as infectious diarrhea, Crohn's disease, or short bowel syndrome.

Although people of Northern European ancestry commonly maintain the ability to digest lactose into adulthood, the majority of the world's population produces less lactase after weaning. In addition to the presence or absence of the lactase enzyme, other factors determine whether a person will have symptoms of lactose malabsorption, including the amount of lactose in the diet, the mixture of lactose with other foods, gastric emptying rate, colonic scavenge of malabsorbed carbohydrate, ethnic origin, and age. Primary lactose intolerance is prevalent in African American, Hispanic, Native American, and Asian populations.

Nutritional management of lactose intolerance consists largely of the removal of lactose from the diet. Lactose is a common ingredient in many foods, including breads, crackers, soups, cereals, cookies, and baked goods. Eliminating or reducing lactose-containing ingredients from one's diet is usually adequate to relieve symptoms. Individuals with primary lactose intolerance may require a permanent dietary change. Individuals with secondary lactose intolerance should eliminate all lactose from their diets for a short period of time ranging from 2 to 6 weeks. If symptoms resolve, lactose may be reintroduced slowly as tolerated by the individual. The amount of lactose that an individual can tolerate is highly variable. Many children can tolerate small amounts of lactose, particularly yogurt, hard cheese, or ice cream, without discomfort. Many adults who consider themselves lactose-intolerant can actually tolerate moderate amounts of milk.

For individuals who choose to restrict lactose in their diets, a variety of lactose-free and low-lactose food choices are available. Lactose-reduced products, containing 70–100% less lactose than standard foods, are available commercially. Individuals may also choose to consume dairy products with concomitant administration of lactase enzyme tablets or drops.

Frequent consumption of milk and other dairy foods has been associated with better bone health in some studies, and a strict lactose-free diet may not contain adequate amounts of calcium and vitamin D. **Table 2** provides a list of some commercially available lactose-free calcium supplements.

Table 2 Commercial calcium supplements

Product	*Manufacturer*	*mg Calcium/ tablet*	*IU vitamin D*
Citracal + D	Mission Pharmacal	315	200
OsCal 500 + D	Marion Lab	500	200
Tums	Smith-Kline Beecham	200	0
Calcium Milk Free (2)	Nature's Plus	500	100
Cal-citrate + D	Freeda	250	100

From DiSanto C and Duggan C (2004) Gastrointestinal diseases. In: Hendricks KM and Duggan C (eds.) *Manual of Pediatric Nutrition*, 4th edn. Hamilton, Ontario: BC Decker.

Sucrose Congenital sucrase–isomaltase deficiency (SID) is the most common congenital disaccharidase deficiency. Patients with this disorder lack functional sucrase, although isomaltase deficiency may be normal or absent. Symptoms of SID can include diarrhea, abdominal pain, and poor weight gain. Dietary avoidance of sucrose or table sugar helps relieve symptoms and can sometimes help with the diagnosis. Sucraid, a sacrosidase produced from *Saccharomyces cerevisiae*, is an enzyme that can be given with meals and allows increased tolerance to sucrose.

Fat Malabsorption: Fat and Fat-Soluble Nutrients

Patients with pancreatic insufficiency are unable to produce and secrete enough enzymes to aid with the breakdown of fats in the intestinal lumen. Studies of normal adults and those with pancreatic insufficiency have demonstrated that pancreatic enzyme secretion needs to be less than 15% of normal levels before significant steatorrhea is seen (**Figure 1**). Once clinically significant steatorrhea is determined, recovery of pancreatic function is unlikely.

Historically, patients with pancreatic insufficiency due to cystic fibrosis (CF) were told to minimize symptoms of steatorrhea by limiting dietary fat. However, epidemiologic studies confirmed that this advice led to negative energy balance, undernutrition, and higher mortality rates compared to communities in which CF patients were treated with high-energy, high-fat diets. The introduction of effective pancreatic replacement therapy has been heralded as one of the most significant breakthroughs in the nutritional management of CF, responsible in part for the substantial increase in life span enjoyed by recent generations of CF patients. In fact, the finding of a lower incidence of growth failure in CF patients diagnosed and treated with aggressive nutritional therapy early in infancy has been used as justification for neonatal screening of this condition.

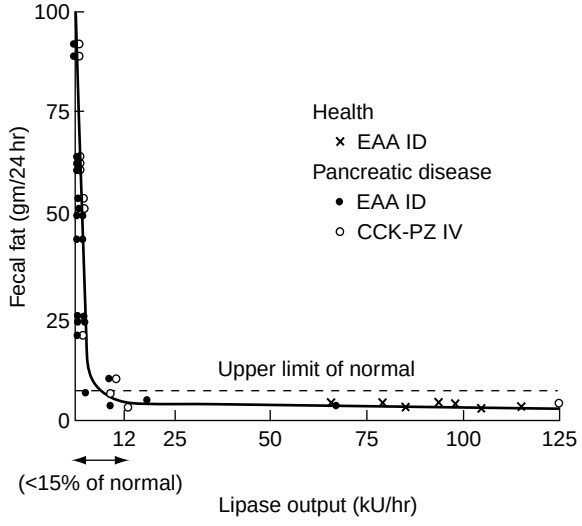


Figure 1 Pancreatic enzyme secretion and steatorrhea. Significant steatorrhea ensues when pancreatic function is less than 15% of normal. (Reproduced with permission from DiMagno EP, Go VLW, and Sumerskill WHJ (1973) Relations between pancreatic enzyme output and malabsorption in severe pancreatic insufficiency. *New England Journal of Medicine* **288**: 814).

Judicious use of pancreatic replacement enzymes is the hallmark of nutritional therapy of CF and other disorders of pancreatic insufficiency. Multiple commercial preparations of porcine pancreatic enzymes are available, most of which contain lipase, amylase, and protease enzymes. The dose is usually titrated to the amount of steatorrhea. If meals take more than 30 min, the dose may be divided, with half given before the meal and half given during the meal. Patients who cannot swallow pills may open the capsules and sprinkle the enzymes into acidic foods.

Another critical aspect of the nutritional management of fat malabsorption is routine supplementation with the fat-soluble vitamins A, D, E, and K. Multiple studies have confirmed that patients with CF, Crohn's disease, and other malabsorptive disorders are prone to micronutrient deficiencies, and some literature suggests that dietary needs for these and other antioxidant nutrients may be increased in settings of infectious and catabolic stress often suffered by these patients. The contribution of fat malabsorption to other important mineral malabsorption, as in the case of calcium or zinc, should also be recognized.

Routine supplementation of fat-soluble vitamins is indicated in patients with fat malabsorption. In addition, serial measurement of fat-soluble vitamin biochemical status is recommended. Since blood nutrient concentrations of these and other nutrients can vary with the concentration of transport proteins, correction for these can aid the interpretation of these lab findings. For instance, vitamin A toxicity should be suspected if the molar ratio of vitamin A:retinol binding protein exceeds 1. Vitamin E concentrations should be corrected for circulating lipids. **Table 3** lists recommendations for therapy of fat-soluble vitamin deficiencies.

Some patients with pancreatic malabsorption may benefit from a diet enriched with medium-chain triglycerides (MCTs). MCTs are absorbed directly into the portal circulation and therefore bypass the steps of intraluminal digestion, reesterification, and enterocyte uptake. Thus, MCTs may be a dietary source of fats more easily absorbed in settings of fat malabsorption due to either pancreatic insufficiency or mucosal disease. However, MCT oils are less energy dense than long-chain fats, more expensive, and do

Table 3 Assessment and treatment of fat-soluble vitamin deficiencies

Vitamin assessment	*Therapy if deficiency*	*Considerations*
A		
Normal: >20 μg/dl	Severe deficiency with xerophthalmia	Serum level is not a good indicator of liver stores.
Marginal stores: 10–19 μg/dl	<6 months old: 50 000 IU PO QD ×2 days, then again at 2 weeks	Low in chronic infection, liver disease, or during an acute phase response.
Deficient: <10 μg/dl	6–12 months: 100 000 IU PO QD ×2 days, then again at 2 weeks	Check retinol binding protein (RBP) circulation in plasma. Assess toxicity by using molar ratio of retinol to RBP:
	>12 months: 200 000 IU PO QD ×2 days, then again at 2 weeks	Retinol(μg/dl) × 0.0349 = μmol/l
	1–8 years: 5000 IU/kg/day ×5 days PO, then 5000–10 000 units/day ×2 months	RBP(mg/dl) × 0.476 = μmol/l
	>8 years and adults: 500 000 IU QD ×3 days, then 50 000 units/day ×14 days, then 10 000–20 000 units/day ×2 months	Molar ratio should be between 0.8 and 1.0. Ratios >1.0 suggest increased levels of free retinol and possible toxicity.
D		
25-OHD:	Vitamin D_2 (Ergocalciferol)	Low in dietary deficiency, decreased absorption, UV light deficiency, prematurity, liver disease, and with certain drugs (anticonvulsants). Higher in summer.
Normal: 9–75 ng/ml	Oral (Drisdol) liquid or capsule	
	Children with malabsorption	
	10 000–25 000 USP units PO/day until normal	Watch for hypercalcemia and hypercalciuria and other signs of toxicity.
	Children with normal absorption	
	1000–5000 units PO ×6–12 weeks	
	IM (Calciferol)—100 000 units/ml	
	10 000–100 000 units IM once. Larger single IM doses may be given.	
	Follow calcium, PTH, 25-OHD concentrations.	
E		
Deficiency if	1 unit/kg/day of water-miscible form plus usual vitamin E supplementation until normal blood levels	Carried exclusively on plasma lipoproteins; thus, vitamin E:total lipid ratio or vitamin E:chol + tryglycerides (TG) is a better indicator of stores than serum levels.
Plasma level <5 mg/L	1 unit = 1 mg dl-α-tocopherol acetate	Conversions
Vitamin E:total lipid ratio <0.6–0.8 mg/g in adults		Total lipids = cholesterol + TG + phospholipids
Vitamin E:chol + TG <1.59 μmol/mmol		Chol (mg/dl) × 0.0259 = chol (mmol/l)
Erythrocyte hemolysis >10%		TG (mg/dl) × 0.0113 = TG (mmol/l)
		Vitamin E (mg/l) × 2.32 = vitamin E (μmol/l)
		Do not give with medications that interfere with vitamin E absorption (vitamin A, cholestyramine, and antacids).
K		
Prothrombin time (PT) deficiency if >13.5 s	Infants and children	Deficiency in malabsorption, long-term antibiotic therapy.
	1–2 mg single IM, SC, or IV dose	
PIVKA-II deficiency if <3.0 ng/ml	Adults	
	10 mg single IM, SC, or IV dose	

From Corrales K (2005) Cystic fibrosis. In: Hendricks KM and Duggan C (eds.) *Manual of Pediatric Nutrition*, 4th edn. Hamilton, ON: BC Decker.

not contain the fatty acids linoleic and linolenic acid, which are essential to humans.

Protein Malabsorption

Protein-losing enteropathy (PLE) can also be treated with a variety of nutritional interventions. PLE due to dilated lymphatics, as with right heart failure, results in leakage of lymphocytes, proteins, and fats into the intestinal lumen. As with fat malabsorption, MCT-supplemented foods and formulas are indicated to allow improved fat absorption in PLE. Fat-soluble vitamin supplementation is indicated. In congenital protein malabsorption syndromes, peptide- or amino acid-based formulas are often helpful.

Mucosal disorders, including inflammatory bowel disease, allergic diseases, and celiac disease, are additional examples of disorders causing protein malabsorption. Once intestinal inflammation is reduced with appropriate medical or nutritional therapy, absorption of protein is usually improved. In Shigella infections, some studies have demonstrated improved nutritional outcomes with a high-protein diet during recovery from the acute symptoms of diarrhea.

Route of Nutrition in Malabsorption

Several factors need to be considered when recommending whether oral, enteral, or parenteral nutrition should be used to provide nutrition to the patient with malabsorption, including etiology of malabsorption, severity of gastrointestinal disease, and underlying nutritional and medical condition. Oral nutrition using modified diets as noted previously is the most customary and desirable by physician and patient alike. In cases of mild lactose malabsorption, modification of a regular, healthy diet to avoid foods high in lactose should be sufficient. In cases in which widespread gastrointestinal disease is leading to severe malabsorption, enteral or 'tube' feeding is helpful for two main reasons: (i) Use of proprietary formulas specially designed for malabsorption is often indicated, and these formulas may be unpalatable, and (ii) enteral feedings, especially slow continuous 'drip' feedings, make efficient use of nutrient transport kinetics, thereby maximizing residual gastrointestinal absorptive function. In severe cases of malabsorption in which tube feedings are unable to achieve adequate nutritional intake, parenteral nutrition may be indicated.

Selection of Enteral Formulas for Malabsorption

A number of commercially available formulas are designed for patients with malabsorption, and these differ with regard to energy density, macronutrient composition, and indicated age. Since infant formulas are often handled in a separate regulatory manner by governments, infant formulas are usually considered separately from formulas designed for older children and adults. In addition, formulas are also conventionally categorized by the extent of the hydrolysis of their protein source. Categories include intact protein formulas, protein hydrolysate formulas, and amino acid-based formulas. Protein hydrolysate formulas are also sometimes referred to as 'semielemental' formulas, and amino acid formulas are sometimes called 'elemental' formulas. However, these terms suffer from vagueness and inaccuracies since not all of their macronutrients are semi- or completely elemental. Marketing strategies often compound the confusion with misleading formula names. These terms should be discouraged, and the terms that refer to the composition and/or biochemical processing should be used instead.

Patients who have carbohydrate malabsorption from lactose intolerance should use lactose-free formula. Fat malabsorption calls for MCT-enriched formula. In cases of protein malabsorption or severe enteropathy, a formula that is a protein hydrolysate or amino acid-based would be most appropriate. Since many malabsorption syndromes overlap in terms of the macronutrient affected, as in cases of severe mucosal disease, some formulas are designed for fat, protein, and carbohydrate malabsorption. For example, all formulas designed for use in adults are lactose-free, and several formulas contain both hydrolyzed proteins and MCT oils.

Clinical Management of Malabsorption

Two of the most clinically challenging scenarios for the management of malabsorption are inflammatory bowel disease (especially Crohn's disease) and short bowel syndrome.

Inflammatory Bowel Disease

Patients with Crohn's disease have widespread and intermittent gastrointestinal inflammation. Some patients with inflammatory bowel disease may require complete bowel rest for several days or even a few weeks to allow time for mucosal healing. In order to provide nutrition during this period, parenteral nutrition may be needed.

Numerous studies have shown that patients with Crohn's disease may safely and effectively achieve clinical remission with primary nutritional therapy. Early literature in the field highlighted the use of protein hydrolysate formulas that, due to unpalatability, often required supplementation with a nasogastric or gastrostomy tube. Recent data have confirmed that intact protein formulas, termed 'polymeric' formulas when describing formulas designed for adults, may work as well as protein hydrolysates, and these formulas can feasibly be given by mouth.

As patients are recovering from an exacerbation and begin advancing their diet, they should temporarily minimize the amount of fiber ingested to decrease trauma to healing mucosa. Patients whose disease affects the small intestine often benefit from temporary avoidance of lactose products as the mucosa heals and brush border membrane enzyme production is restored.

Micronutrients are also needed in the nutritional management of inflammatory bowel disease. Iron supplementation is recommended for anemia that may be secondary to acute or chronic blood loss. Treatment of inflammatory bowel disease frequently requires the use of steroids, which affects bone density. Calcium and vitamin D supplementation is commonly needed to minimize the osteopenic effects of steroid therapy and/or the effects of malabsorption and chronic inflammation.

Short Bowel Syndrome

Patients who have suffered acquired or congenital loss of small intestinal surface area that makes them dependent on specialized enteral or parenteral

support are said to have short bowel syndrome (SBS). Patients with SBS often malabsorb carbohydrates, proteins, fat, as well as numerous micronutrients, depending on the extent and location of bowel resection as well as the presence of mucosal disease in the nonresected bowel.

Special attention should be given to the part of the intestine that remains as well as the length of the intestine. Some patients may have the terminal ileum removed and are unable to absorb vitamin B_{12} and bile acids. Removal of the ileocecal valve increases the risk of bacterial overgrowth. Reduced length also means reduced surface area for the absorption of nutrients and decreased intestinal transit time.

In the immediate postoperative period, parenteral nutrition and gut rest should be used because significant stool output is the norm. Output should be quantified, and electrolytes must be carefully monitored in order to determine appropriate replacement fluids to make up for excess urine, stool, and ostomy losses. Replacement fluids should generally be given separately from standard parenteral nutrition so that they can be adjusted as needed to rapid shifts in fluid and electrolyte status.

As patients recover from surgery, every attempt should be made to feed them enterally as soon as is feasible. Enteral feeds facilitate growth and adaptation of the remaining bowel to allow partial compensation for the missing portion, and several studies have correlated early feeding with better long-term outcome. Attaining independence from parenteral nutrition may take weeks, months, or years. **Table 4** outlines an approach to determining feeding advancement. Whereas some patients are able to grow well or maintain their body weight with enteral feeds, many are dependent on parenteral nutrition. Some patients with SBS also have oral feeding aversion due to prematurity, prolonged mechanical ventilation, and/or prolonged orogastric or nasogastric feeding. Gastrostomy tubes are particularly helpful in this regard.

In infants, breast milk should be used if available. The breast milk may need to be fortified to increase calories, protein, or fat. For older patients or infants who are not receiving breast milk, protein hydrolysates or amino acid-based formulas may be better tolerated since the residual bowel more easily absorbs these nutrients. Lactose-free and MCT-containing formulas are often used as well. Formulas may need to be supplemented with oral rehydration solutions if electrolyte abnormalities persist, particularly with sodium losses through persistent high stool or ostomy output.

Since many patients with SBS are dependent on parenteral nutrition for prolonged periods of time, selenium, carnitine, copper, and zinc blood concentrations should be checked periodically and supplemented if needed. Parenteral nutrition should be cycled off for a few hours each day to help simulate more natural cyclic fluctuations of gastrointestinal hormones. These patients also often have poor absorption of calcium and need calcium supplements to prevent osteopenia, which increases the risk of fractures. Iron may also be needed in patients with anemia from decreased absorption secondary to resection of the duodenum or jejunum. Ultimately, weaning from parenteral and enteral nutrition remains the goal of treatment, although lifelong dietary therapy is often needed.

Table 4 Feeding advancement in short bowel syndrome

Stool output		
If <10 g/kg/day or <10 stools/day	→	Advance rate by 10–20 ml/kg/day
If 10–20 g/kg/day or 10–12 stools/day	→	No change
If >20 g/kg/day or >12 stools/day	→	Reduce rate or hold feeds[a]
Ileostomy output		
If <2 g/kg/h	→	Advance rate by 10–20 ml/kg/day
If 2–3 g/kg/h	→	No change
If >3 g/kg/h	→	Reduce rate or hold feeds[a]
Stool reducing substances		
If <1%	→	Advance feeds per stool or ostomy output
If = 1%	→	No change
If >1%	→	Reduce rate or hold feeds[a]
Signs of dehydration		
If absent	→	Advance feeds per stool or ostomy output
If present	→	Reduce rate or hold feeds[a]
Gastric aspirates		
<Four times previous hour's infusion	→	Advance feeds
>Four times previous hour's infusion	→	Reduce rate or hold feeds[a]

[a]Feeds should generally be held for 8 h and then restarted at three-fourths of the previous rate.
Adapted from Utter SL and Duggan C (2004) Short bowel syndrome. In: Hendricks KM and Duggan C (eds.) *Manual of Pediatric Nutrition*, 4th edn. Hamilton, Ontario: BC Decker.

Summary

Malabsorption can involve any of the macronutrients or micronutrients, and these disorders may be congenital or acquired. Determining the type of malabsorption and root cause is essential to providing appropriate nutritional therapy. Multiple formulas, supplements, and dietary regimens exist to target specific defects in the digestion, absorption, and assimilation of nutrients. In addition, many new nutrients are undergoing investigation that may become a standard part of care in the future, including probiotics, prebiotics, and various amino acids.

See also: **Celiac Disease**. **Colon**: Disorders; Nutritional Management of Disorders. **Cystic Fibrosis**. **Diarrheal Diseases**. **Lactose Intolerance**. **Microbiota of the Intestine**: Prebiotics; Probiotics. **Nutritional Support**: Adults, Enteral.

Further Reading

Basu TK and Donaldson D (2003) Intestinal absorption in health and disease: Micronutrients. *Best Practice and Research: Clinical Gastroenterology* **17**(6): 957–979.

Borowitz D, Baker RD, and Stallings V (2002) Consensus report on nutrition for pediatric patients with cystic fibrosis. *Journal of Pediatric Gastroenterology and Nutrition* **35**(3): 246–259.

Duggan C, Gannon J, and Walker WA (2002) Protective nutrients and functional foods for the gastrointestinal tract. *American Journal of Clinical Nutrition* **75**: 789–808.

Heuschkel RB, Menache CC, Megerian JT *et al.* (2000) Enteral nutrition and corticosteroids in the treatment of acute Crohn's disease in children. *Journal of Pediatric Gastroenterology and Nutrition* **31**: 8–15.

Holt PR (2001) Diarrhea and malabsorption in the elderly. *Gastroenterology Clinics of North America* **30**(2): 427–444.

Schmitz J (2004) Maldigestion and malabsorption. In: Walker WA, Goulet O, Kleinman RE *et al.* (eds.) *Pediatric Gastrointestinal Disease: Pathophysiology, Diagnosis, Management*, 4th edn. Hamilton, Ontario: BC Decker.

Treem WR, McAdams L, Stanford L *et al.* (1999) Sacrosidase therapy for congenital sucrase-isomaltase deficiency. *Journal of Pediatric Gastroenterology and Nutrition* **28**(2): 137–142.

Vanderhoof JA and Young RJ (2003) Enteral and parenteral nutrition in the care of patients with short-bowel syndrome. *Best Practice and Research: Clinical Gastroenterology* **17**(6): 997–1015.

MALNUTRITION

Contents

Primary, Causes Epidemiology and Prevention

A Briend, Institut de Recherche pour le Développement, Paris, France
P Nestel, International Food Policy Research Institute, Washington, DC, USA

Undernutrition is a condition of poor health resulting from an inadequate intake of energy and/or essential nutrients. It can also be caused by an imbalance between energy and nutrient intakes and requirements due to infection that results in malabsorption, anorexia, or excessive losses.

Causes

The determinants of undernutrition can be categorized as being immediate, underlying, or basic (**Figure 1**). The immediate causes include dependence on a diet that is inadequate in quantity and/or quality. This can be due to low food availability or anorexia from recurrent infections and also poor health status that can result in a vicious cycle of ill health and undernutrition. The underlying determinants are food insecurity, inadequate care for mothers and children, and poor sanitation. The basic determinants influence the underlying determinants and include the environmental, technological, and human resources available to a country or community. Access to and use of resources are influenced by both the political and the economic structures as well as cultural and social factors that affect how resources are used to maintain and improve food security, the provision of care, and sanitation. The following discussion is limited to the immediate causes of undernutrition.

Undernutrition is intergenerational, and a cycle of ill health and growth failure frequently occurs in which undernutrition in childhood leads to small body size in adulthood (**Figure 2**). Genetic and environmental influences also affect both maternal height and prepregnancy weight, both of which are important determinants of birth size and, to a lesser extent, later growth and size.

Food Availability and Diet Quality

Poor access to and an inadequate intake of a good quality diet is the major cause of undernutrition in

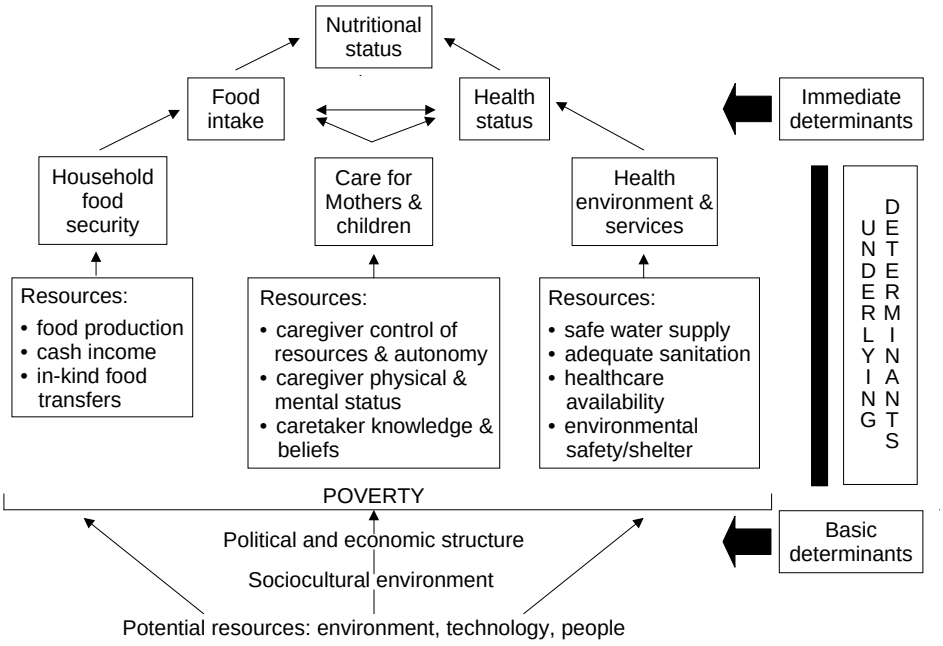


Figure 1 Determinants of undernutrition (Reproduced with permission from Smith LC and Haddad L (2000) *Overcoming Child Malnutrition in Developing Countries: Past Achievements and Future Choices*, Food, Agriculture, and the Environment Discussion Paper No. 30. Washington, DC: International Food Policy Research Institute.)

developing countries. Children are particularly at risk of becoming undernourished due to their high energy and nutrient requirements. Diet diversity, including the intrahousehold distribution of animal source foods, feeding patterns, and child growth are all related to household socioeconomic status.

Low energy intake is associated with growth retardation and often related to a poor quality diet.

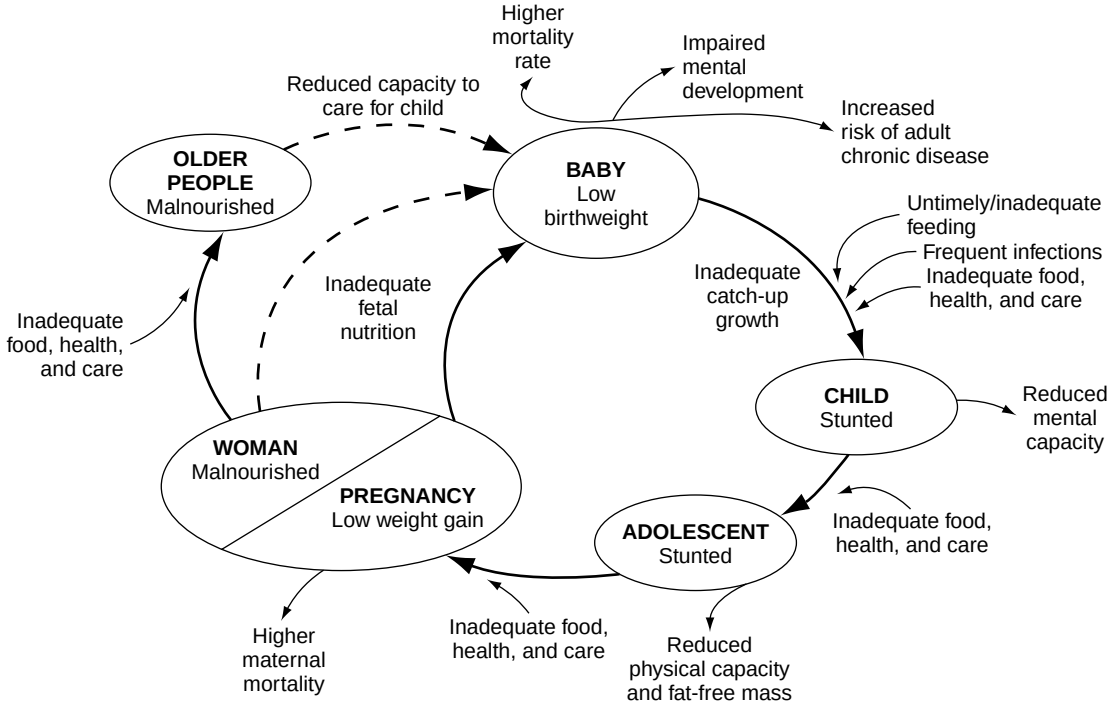


Figure 2 Nutrition throughout the life cycle. (Reproduced with permission from ACC/SCN (2000) *Fourth Report on the World Nutrition Situation*. Geneva: ACC/SCN (in collaboration with the International Food Policy Research Institute).)

Moreover, children can become anorectic when fed monotonous diets, and this can be compounded when superimposed on frequent and repeated bouts of infections. A high-carbohydrate, low-fat diet with a low energy density also precipitates undernutrition in children. Protein deficiency per se is not the major cause of growth retardation, but little is known about the specific effects of essential fatty acid deficiencies on growth. Micronutrient deficiencies, particularly zinc deficiency, cause growth retardation. Zinc supplementation can increase body weight and length/height, but the effect is modest compared with the growth deficit usually observed in growth-faltered children. Children in poor communities are likely to have multiple micronutrient deficiencies, and the combined effect of these coupled with inadequate fat and energy intake will affect growth patterns.

Low income is associated with a low intake of the more expensive animal source foods (meat, fish, dairy, and eggs). The mainly vegetarian diets eaten by children in poor families are frequently associated with growth retardation. Children in developed countries who for cultural reasons eat a diet that lacks any animal products have growth patterns similar to those observed in developing countries.

Women of reproductive age and adolescent girls are also susceptible to micronutrient deficiencies, especially iron, due to their increased physiological requirements for nutrients that can rarely be met from a diet low in animal source foods. In addition, food availability for women is often compounded by sociocultural practices that discriminate against women, and this can start in early childhood and continue throughout the life span.

Infections

Longitudinal studies have shown that infections, especially diarrheal diseases, are associated with growth faltering due to anorexia and/or malabsorption. However, the effect of acute infections on growth is transient, at least in children older than 6 months of age, and no longer apparent after a few weeks. The lack of a long-term effect on growth can be explained by the lower proportion of energy and nutrients needed for growth after 6 months of age, which facilitates catch-up growth after an acute disease. However, chronic infections, even if asymptotic, may lead to anorexia and malabsorption. The frequently observed inverse correlation between markers of chronic infection—such as elevated white blood cell, lymphocyte and platelet counts, C-reactive protein, and gut permeability—and growth supports this hypothesis. A similar mechanism may also explain the delayed growth observed in children infested with worms and the undernutrition observed in adults with chronic diseases, such as HIV infection and tuberculosis.

The edematous form of severe malnutrition, kwashiorkor, has been hypothesized to be due to oxidative stress resulting from insufficient intake of antioxidant nutrients, including selenium, vitamins E, C, and B_2, niacin, and sulfur amino acids. This suggests that infections, leading to an increased production of free radicals, can be one cause of this form of severe undernutrition.

Epidemiology: Assessing the Prevalence of Undernutrition

Food Balance Sheets

The Food and Agriculture Organization (FAO) uses national-level data on the production, export, and import of food in food balance sheets to calculate the average daily per capita energy supply. This value, combined with a measure of the inequality in food distribution, is used to calculate the proportion of undernourished individuals in a population. Globally, 840 million people are undernourished: 11 million in developed countries, 30 million in countries in transition, and 799 million in the developing world. The total number of undernourished people has increased during the past decade, despite the increase in per caput food availability, because of population growth. The proportion of undernourished individuals is highest in Africa, particularly in sub-Saharan Africa, but the greatest number of undernourished people is in Asia (**Figure 3**).

The precision of the FAO estimate for the number of undernourished individuals is dependent on the accuracy and reliability of agricultural statistics in developing countries, many of which include a large sector that produces food for subsistence and not the market economy. FAO is refining its method to adjust for the latter by including data collected at the household level. The food balance method, however, only deals with estimating the deficit in energy but not nutrient intakes. The latter can be obtained from detailed household or individual-level food consumption surveys, although nationally representative dietary intake surveys are rare in developing countries.

Anthropometric Surveys

When energy intake is insufficient to meet requirements, energy is derived by metabolizing fat and lean tissues, mainly muscle. Children first stop gaining height/length and then lose weight, whereas adults lose weight. Weight loss is more rapid in children

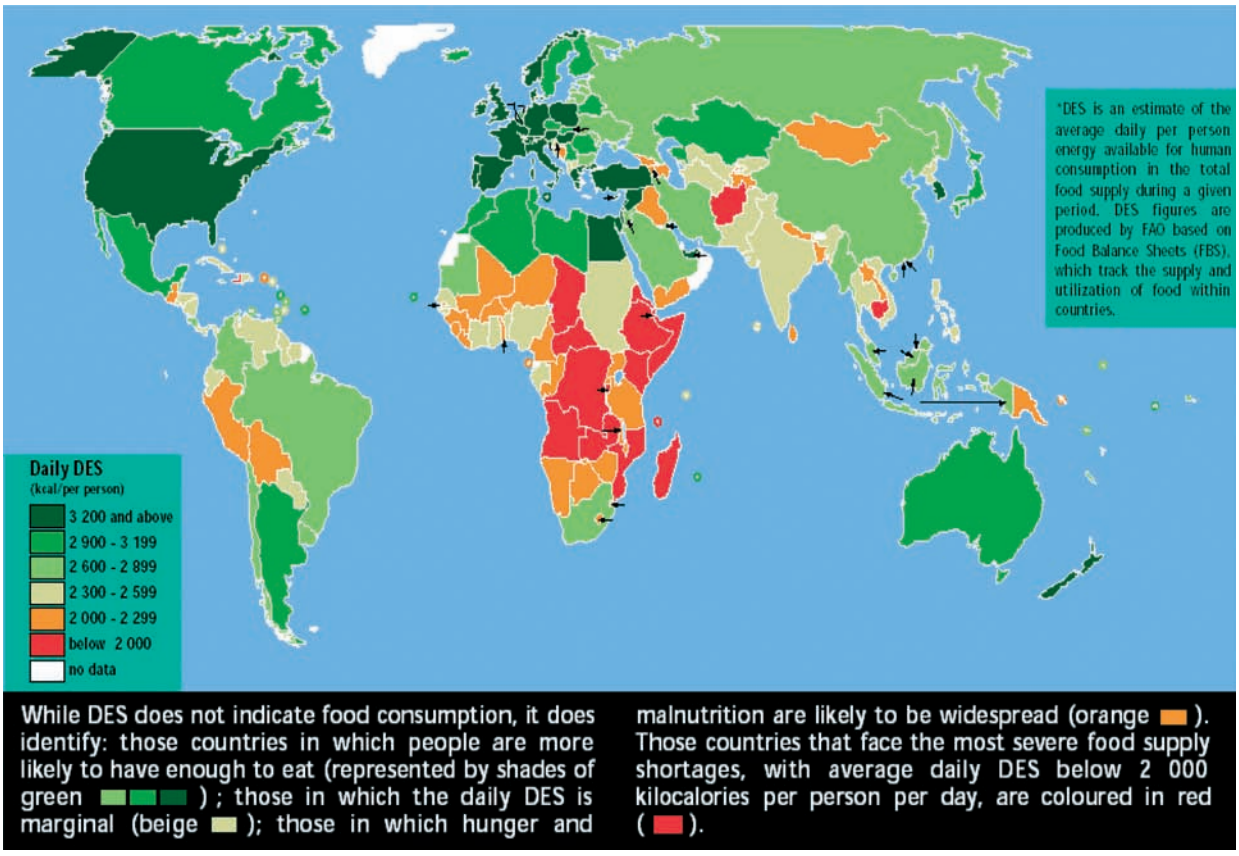


Figure 3 Dietary energy supply (DES), 1994–1996. (Reproduced with permission from the Food and Agriculture Organization (2000) Undernourishment around the world. Counting the hungry: Latest estimates. In: *The State of Food Insecurity in the World*. Rome: FAO.)

than in adults because of their higher energy requirements per kilogram body weight, mainly due to a different body composition. Population-level measurements of body weight and height/length in children indirectly assess the adequacy of food intake on the assumption that a low average body weight and height/length compared with a growth reference reflects an inadequate diet. WHO recommends that the following be used to compare undernutrition in children in different areas of the world:

- Underweight is defined as the proportion of children whose weight in relation to their age is below -2 standard deviations (-2 z scores) of the median of the National Center for Health Statistics (NCHS) reference.
- Wasting is the proportion of children whose weight in relation to their height is below -2 z scores of the median of the NCHS reference.
- Stunting is the proportion of children whose height in relation to their age is below -2 z scores of the median of the NCHS reference.

The previous calculations underestimate the true prevalence of undernutrition because a child can be below his or her optimal weight or height/length yet remain above the -2 z score cutoff point. Wasting is often described as acute malnutrition because it reflects relatively recent weight loss, and stunting is described as chronic malnutrition. Growth retardation is often described as protein–energy malnutrition, which is a misnomer because there is increasing evidence that other nutritional deficiencies besides protein–energy, such as zinc, can lead to growth faltering. Severe undernutrition is defined by a weight in relation to height below -3 z scores of the NCHS reference (marasmus) and/or by the presence of nutritional edema (kwashiorkor or marasmus kwashiorkor), and it is associated with a high risk of dying. WHO and UNICEF maintain a global database on the prevalence of undernutrition among children. Stunting is more prevalent than underweight (**Table 1**) and is often used to monitor long-term trends in undernutrition.

In emergency situations, rapid assessment surveys are often carried out using the mid-upper arm circumference as a proximate indicator of nutritional status in children and, increasingly, adults. This approach is less reliable than methods based on

Table 1 Prevalence of undernutrition by region

UNICEF region	*Under-5 population, 2000*[a]	*Wasting prevalence (%)*		*Underweight prevalence (%)*		*Stunting prevalence (%)*	
		Moderate and severe	*Severe*	*Moderate and severe*	*Severe*	*Moderate and severe*	*Severe*
Sub-Saharan Africa	106 394	10	3	30	9	41	20
Middle East and North Africa	44 478	7	2	15	4	23	9
South Asia	166 566	15	2	46	16	45	22
East Asia and Pacific	159 454	4	—	17	—	21	—
Latin America and Caribbean	54 809	2	0	8	1	16	5
CEE/CIS and Baltic states	30 020	4	1	7	2	16	7
Industrialized countries	50 655	—	—	—	—	—	—
Developing countries	546 471	9	2	28	10	32	17
Least developed countries	110 458	10	2	37	11	43	20

[a]In thousands.
Sources: http://childinfo.org/eddb/malnutrition/database1.htm (underweight), http://childinfo.org/eddb/malnutrition/database2.htm (stunting), and http://childinfo.org/eddb/malnutrition/database3.htm (wasting).

weight and height for epidemiological assessment of undernutrition. Measures of mid-upper arm circumference, however, are useful for screening to quickly identify the severely undernourished, especially children, who are at high risk of dying and need urgent case management.

Anthropometry is also used to assess undernutrition in adults, usually as the body mass index (weight/height^2). A body mass index of less than 18.5 defines chronic energy deficiency, and that less than 16.0 defines severe chronic energy deficiency. A global database on maternal nutrition is not available.

Anthropometric surveys do not give information on the causes (dietary, infectious, or other) of the weight and height deficits they measure. Genetic factors are unlikely to determine child growth at a population level because growth is very similar among well-off children from different countries. Breast-feeding patterns, however, may affect growth patterns, and WHO is developing new growth references based on a longitudinal study of infants from diverse geographic sites who are exclusively or predominantly breast fed for at least 4 months with continued breast feeding throughout the first year and on a cross-sectional study of infants and young children age 18–71 months.

Approximately 55% of all child deaths in developing countries are associated with undernutrition (**Figure 4**), of which at least three-fourths are related to moderate or mild undernutrition rather than severe undernutrition. Some nutritional deficiencies, such as vitamin A, can result in higher mortality without a clear effect on growth. Hence, studies examining the association between undernutrition and mortality, using anthropometry as proxy for undernutrition, are likely to underestimate the strength of this relationship.

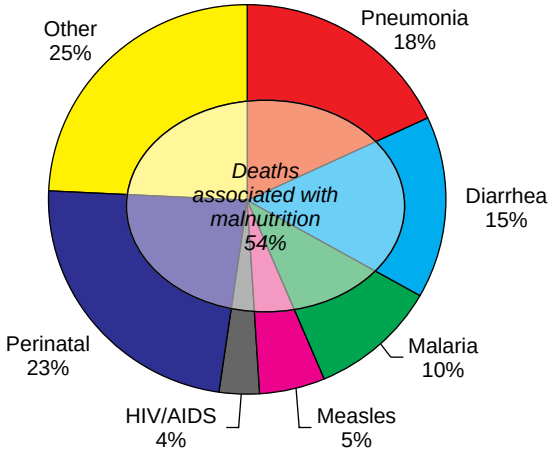


Figure 4 Association between malnutrition assessed by anthropometry and cause-specific mortality in children younger than 5 years of age. (Reproduced with permission from the WHO Department of Child and Adolescent Health and Development (2002) Available at www.who.int/child-adolescent-health/OVERVIEW/CHILD_HEALTH/map_02_world.jpg. Sources: for cause-specific mortality, EIP/WHO; for malnutrition, Pelletier DL, Frongillo EA Jr, and Habicht JP (1993) Epidemiological evidence for a potentiating effect of malnutrition on child mortality. *American Journal of Public Health* **83**: 1130–1133.)

Biochemical Surveys

Because of the interest and focus on controlling iodine, vitamin A, and iron deficiencies, national-level surveys for these micronutrients are conducted in developing countries. Urinary iodine level is a good marker of iodine deficiency, which can cause retarded physical development. However, the biochemical markers of vitamin A and iron are acute phase proteins, which are altered during infection, even subclinical infection. Serum retinol declines during infection, overestimating the prevalence of vitamin A deficiency, whereas serum ferritin rises and overestimates adequate iron stores. Hemoglobin, which is often used as a proxy for iron deficiency, is also affected by the prevalence of infection (as is serum zinc). For these reasons, biochemical surveys need to include markers of infection, which adds to their complexity and cost; thus, they are rarely done.

Prevention

Increasing both the purchasing power of poor families and women's access to and control of household resources can prevent undernutrition. Improving the social and economic status of women and households, however, is insufficient to eliminate undernutrition and interventions that affect each stage of the life cycle and are necessary to break the intergenerational cycle of undernutrition (**Table 2**).

Table 2 Intervention points in the life cycle continuum

Nutritional risks	*Interventions to improve nutritional status and growth*
Infancy and early childhood (<6 months)	
Suboptional breast feeding	Feeding colostrum
	Exclusive breast feeding for 6 months
Complementary feeding period (6–24 months)	
Suboptional breast-feeding	Breast feeding
Inadequate and low-quality diet	Nutrition education
Micronutrient deficiencies	Infection control
Frequent infections and parasites	Food supplements
Gender discrimination in food and health care	Micronutrient supplements
Childhood (2–9 years)	
Inadequate and low-quality diet	Nutrition education
Micronutrient deficiencies	Infection control
Frequent infections and parasites	Micronutrient supplements
Gender discrimination in food and health care	School feeding
Adolescence (10–19 years)	
Inadequate and low-quality diet	Iron supplementation (schools)
Rapid growth	Deworming through schools (endemic areas)
Anemia with onset of menstruation	Nutrition and health education (schools)
Infections (including STDs) and parasites	Prevention and management of STDs and other infections
Early pregnancy and lactation	School meals
Gender discrimination in food and health care	Income-earning skill training
Adult women of reproductive age (20–49 years)	
Food insecurity	Income generation
Micronutrient deficiencies	Family planning/birth spacing
Infections and parasites	Prevention and management of STDs
Gender inequities	Agricultural/gardening
STDs and AIDS	Improved technologies
	Adult literacy
Pregnant women	
Inadequate food intake to meet increased demands for fetal growth	Increase intake of bioavailable nutrients and energy
Micronutrient deficiencies	Malaria chemoprophylaxis (endemic areas)
Maternal mortality	Deworming (endemic areas)
Low birth weight	Iron/folic acid supplementation
Birth defects	Postpartum vitamin A
Lactating women	
Decreased quantity of vitamins in breast milk	Increase intake of bioavailable nutrients and energy
Micronutrient deficiencies	Iron/folic acid supplementation
Weight loss	
Post-reproductive age (49+ years)	
Undernutrition	Food security

Early Infancy (<6 Months)

All newborns should be fed colostrum immediately after delivery. Because breast milk is the best food for infants, exclusive breast feeding is the best way to prevent undernutrition in this age group. Exclusive breast feeding also reduces the risk of diarrhea and other infections that can reduce appetite and absorption or increase nutrient losses.

The advantages of breast feeding should be balanced with the risk of HIV transmission if the mother is known to be HIV positive. In early infancy, breast milk may not easily be replaced, and risks should be carefully assessed by program managers before recommending breast milk substitutes.

Beyond exclusive breast feeding, there is a role for micronutrient supplementation for some high-risk children. For example, low birthweight infants can benefit from zinc, iron, and vitamin A supplementation. Where there is a risk of rickets, such as in areas where young infants do not get any exposure to sunlight or calcium intake is very low, vitamin D supplementation is recommended.

Complementary Feeding Period (6–24 Months)

Undernutrition, especially wasting and micronutrient deficiencies, is most prevalent during the complementary feeding period. Linear growth retardation is usually well established during this stage of the life cycle. This age group is particularly vulnerable because of its high nutrient requirements that cannot be supplied through breast milk alone. A variety of nutrition interventions have been proposed to prevent undernutrition, and the use of multiple approaches is likely to be most successful.

Breast feeding promotion Beyond 6 months of age, breast milk alone is not sufficient to sustain optimal growth and its contribution to energy and nutrient intake progressively declines. After the age of 12 months, breast fed children are not better nourished than non-breast fed children. However, prolonged breast feeding is very important and needs to be promoted because it improves child survival. The poor nutritional status of breast fed children older than 6 months of age is due to late introduction of an appropriately balanced diet. Family planning is an important intervention for promoting prolonged breast feeding because a new pregnancy can be a frequent cause of breast feeding cessation.

Nutrition education Undernutrition among children is often ascribed to ignorance, and nutrition education programs are often proposed to resolve this. However, choosing and recommending foods that are appropriate for promoting child growth at low cost is not easy. Moreover, micronutrient requirements in this age group are high, and it is usually not possible to provide a balanced diet without large quantities of animal source foods or fortified foods that are generally not readily available to the poor.

In theory, advising mothers to use nutrient-dense foods when they are available might improve child nutrition. Some nutrition education programs have be shown to be efficacious in pilot studies, but their effectiveness when scaled up has been disappointing.

Infection control Acute infections have a transient effect on the growth of young children. However, controlling chronic infections, including parasitic and other subclinical infections, can have a positive effect on growth. A general improvement in environmental hygiene and sanitation may be more effective in removing this cause of undernutrition than medical interventions per se, with the possible exception of regular deworming.

Food supplements Food supplements are designed to provide nutrients missing from the diet. They are usually made from a cereal flour mixed with a lysine-rich flour, generally soy flour, or milk powder to improve the amino acid balance (e.g., corn–soy blend). The fat content is usually low because fat mixed with flour is rapidly oxidized and cannot be added to food without costly packaging. Nowadays, such food supplements are usually fortified with micronutrients that are likely to be missing in childrens' diet, especially iron, zinc, retinol, and riboflavin.

The efficacy of food supplements has been tested in pilot programs, and most randomized trials that compared the growth of children receiving food supplements with that of children receiving the usual family diet failed to show a major effect on growth, especially height. The reasons for this are not clear. Biological factors may be involved. For example, most trials tested a low-cost supplement that had a high phytate content, which would limit the bioavailability of minerals such as zinc (known to be important for growth) and iron. Moreover, the food supplements were low in fat, which may be a limiting nutrient for children who usually have a very low fat intake. Finally, even in carefully controlled trials the food supplement may be shared within the family, especially if it requires special preparation, which would dilute its effect.

Besides the pilot trials, the use of donated food supplements has been usually limited to food crises,

such as wars and environmental catastrophes. Food supplements are also available in the commercial market, but their price often precludes their use by those most in need of them.

Micronutrient supplements Vitamin A is routinely distributed through immunization programs and is a standard health package delivered in under-5 clinics to improve child survival. In very few countries, prophylactic liquid iron is given to older infants. Routine iron supplementation to children is not without controversy in malarial areas. Most micronutrient supplementation trials have been carried out using syrups, which are expensive to use on a large scale, or tablets/capsules, which are difficult to administer to young children. New formulations, such as multiple micronutrient sprinkles or water-dispersible tablets, are being tested. Highly fortified spreads that can provide fat along with micronutrients are also being tested.

Childhood (2–9 Years)

The nutritional problems of 2- to 5-year-old children are similar to those of younger children. The critical factors being that these children do not eat enough food and they are still susceptible to repeated bouts of infection. Because there is no obvious contact point with these children (e.g., measles immunization), it is difficult to deliver programs that reach them.

Children 6–9 years old can be reached through school feeding programs. School health programs increasingly include deworming, malaria chemoprophylaxis, and iron supplementation to improve student nutrition and health. Where institutional feeding is provided, the potential exists to provide fortified food in a meal or as a snack food. In situations in which institutional feeding is not provided, community mobilization activities in primary education can incorporate messages to support student health and nutrition.

Adolescents (10–19 Years)

Adolescent girls are at special nutritional risk, especially for iron deficiency anemia, with menstruation taking place while growth is still not complete. Few intervention studies have been done on adolescent girls, and those that have focused on anemia prevention. Well-supervised intermittent iron supplementation to adolescent girls in schools or the workplace can reduce their prevalence of anemia. Although iron supplements have been shown to be important for correcting iron deficiency anemia in adolescent girls, there is only a modest improvement in storage iron status in early pregnancy following supplementation in adolescence if the time between the end of supplementation and conception is more than a few months. Iron supplementation in adolescence cannot build sufficient stores to substitute for the need for iron supplements in pregnancy.

Secondary schools are the easiest way to reach adolescents, although in many countries relatively few girls go to school compared to boys. Supplementary food can also be provided through school meals to induce growth and maximize the pubertal growth spurt, increase school attendance, and serve as an excellent opportunity for nutrition and health education relevant to the age group. Such interventions are likely to be effective because of the potential for good supervision. However, any credible intervention program will have to reach the girls who do not attend school. Currently, the few activities targeted to this age group center around HIV/AIDS awareness and education, including the prevention and management of sexually transmitted diseases. These children, especially girls, are particularly vulnerable and often have no voice in the community.

Women of Reproductive Age (20–49 Years)

Many programs designed to improve food security include adult women of reproductive age, who may not be pregnant or lactating. Managers of these programs recognize that the nutritional benefits are not necessarily direct, and it is not easy to measure or attribute any change in nutritional status to the intervention. Such projects, which invariably strive to empower women, include income-generation or credit schemes, home gardening and agriculture, improved technologies, and adult literacy, alone or in combination.

Health-based interventions, such as family planning and longer birth spacing, are assumed to have a more direct effect on women's nutritional status, but the inputs and outcomes rarely, if ever, include issues related to improving the nutritional status of women. However, the potential is there to change this.

The role of iron/folic acid supplements remains equivocal, except where severe deficiency exists. Insufficient data are available to justify the provision of free multiple micronutrient supplements through the public health system, although there is some rationale to improve micronutrient status before these women become pregnant so that they are in

the best nutritional state possible. The demand side issues for supplementation need to be addressed concurrently; otherwise, it is unlikely that these programs can be successful.

In urban areas, adult women of reproductive age may be reached through the workplace, social/community groups, religious centers, etc., where the possibility exists of getting institutional support for health-related activities that include nutrition. Similar groups could be used in rural areas where they exist, and activities that support or strengthen group 'cohesion' should be seen as a component of an add-on nutrition/health activity. As with the other groups of the life cycle framework, it is important to put nutrition in the context of perceived priorities.

Pregnancy

Prenatal programs focus on identifying and counseling pregnant women on appropriate care and nutrition, including breast-feeding, tetanus toxoid immunization, iron/folic acid supplementation, and referral of high-risk pregnancies. Malaria chemoprophylaxis, especially among primigravidae, and deworming need to be encouraged to prevent anemia in areas where these parasites are public health problems. The provision of postpartum vitamin A supplementation is increasing and needs to be further expanded.

Some programs provide and target supplementary food to at-risk and undernourished women. These programs are effective in increasing weight gain during pregnancy, but they have a significant beneficial effect on birth weight only in women who are genuinely at risk of an inadequate diet, such as rural African women who continue to perform difficult manual work during the hungry season. In other settings, the effect is less clear.

Despite the research evidence that iron supplementation is efficacious, this relatively simple program has not been effective in reducing the prevalence of anemia among women. Most iron/folic acid supplement programs suffer from serious operational constraints related to supply and distribution systems, access to health care services, motivation of health care providers, and compliance. Lack of good quality, low-cost, generic iron supplements, suitable compounds and dispensing mechanisms, and potential side effects are unsolved problems that affect compliance.

Although there is little evidence that an iron supplement program works, it remains one of the few options available to improve iron status of the population, and it is the only program that has the potential to meet the high iron requirements of pregnancy. Operation research has shown that intermittent supplementation is less appropriate for pregnant women and daily iron supplementation should be continued as the intervention of choice.

Lactation

Except under extreme famine-like conditions, undernourished women produce sufficient breast milk, but its micronutrient quality can vary depending on the nutritional status of the mother.

Postnatal programs cover the lactation period and include counseling on nutrition (including breast-feeding) and family planning, although these two programs are not usually integrated. Where the prevalence of anemia is higher than 40%, iron/folic acid supplementation should be extended through the first 3 months of lactation. Clinical signs of vitamin A deficiency are known to exist in lactating women, and postpartum supplementation within 6–8 weeks of giving birth has been shown to be beneficial and safe for both the mother and the newborn.

Some programs include provision of a protein and energy supplement during lactation. Controlled trials have failed to show an effect on milk production, but increased weight gain and a sensation of well-being among mothers are potential positive outcomes.

Post-reproductive Age (49+ Years)

There are no reports on nutrition interventions for older people in developing countries, successful or otherwise. At best, they are covered in food security projects targeted at the household level, although the focus is generally on maternal and child nutrition. This age group deserves more attention, especially in areas with a high prevalence of HIV, where older people often play a key role in sustaining the household.

See also: **Adolescents**: Nutritional Problems. **Anemia**: Iron-Deficiency Anemia. **Breast Feeding**. **Children**: Nutritional Problems. **Dietary Surveys**. **Infants**: Feeding Problems. **Infection**: Nutritional Interactions; Nutritional Management in Adults. **Lactation**: Dietary Requirements. **Malnutrition**: Secondary, Diagnosis and Management. **Nutritional Assessment**: Anthropometry; Biochemical Indices; Clinical Examination. **Older People**: Nutrition-Related Problems. **Pregnancy**: Nutrient Requirements. **Supplementation**: Role of Micronutrient Supplementation; Developing

Countries; Developed Countries. **United Nations Children's Fund**. **Vegetarian Diets**. **World Health Organization**. **Zinc**: Deficiency in Developing Countries, Intervention Studies.

Further Reading

Allen LH and Gillespie SR (2001) *ACC/SCN. What Works? A Review of the Efficacy and Effectiveness of Nutrition Interventions*. Geneva: ACC/SCN (in collaboration with the Asian Development Bank, Manila). Available at www.adb.org/documents/books/nutrition/what_works/default.asp.

Awasthi S, Bundy DA, and Savioli L (2003) Helminthic infections. *British Medical Journal* **327**: 431–433.

Brown KH, Peerson JM, Rivera J, and Allen LH (2002) Effect of supplemental zinc on the growth and serum zinc concentrations of prepubertal children: A meta-analysis of randomized controlled trials. *American Journal of Clinical Nutrition* **75**: 1062–1071.

Checkley W, Epstein LD, Gilman RH, Cabrera L, and Black RE (2003) Effects of acute diarrhea on linear growth in Peruvian children. *American Journal of Epidemiology* **157**: 166–175.

Davidsson L and Nestel P (2003) *Efficacy and Effectiveness of Interventions to Control Iron Deficiency and Iron Deficiency Anemia*. Washington, DC: International Nutritional Anemia Consultative Group. Available at http://inacg.ilsi.org/file/efficacyscreen.pdf.

Food and Agriculture Organization (2002) Undernourishment around the world. Counting the hungry: Latest estimates In: *The State of Food Insecurity in the World*. Rome: FAO. Available at http://ftp.fao.org/docrep/fao/005/y7352e/y7352e01.pdf.

Gera T and Sachdev HP (2002) Effect of iron supplementation on incidence of infectious illness in children: Systematic review. *British Medical Journal* **325**: 1142–1152.

International Vitamin A Consultative Group (2002) *IVACG Statement on Maternal Night Blindness: A New Indicator of Vitamin A Deficiency*. Washington, DC: IVAGC. Available at http://ivacg.ilsi.org/file/Nightblindness.pdf.

Merialdi M, Carroli G, Villar J *et al.* (2003) Nutritional interventions during pregnancy for the prevention or treatment of impaired fetal growth: An overview of randomized controlled trials. *Journal of Nutrition* **133**(5 supplement 2): 1626S–1631S.

Sazawal S, Black RE, Menon VP *et al.* (2001) Zinc supplementation in infants born small for gestational age reduces mortality: A prospective, randomized, controlled trial. *Pediatrics* **108**: 1280–1286.

Steketee RW (2003) Pregnancy, nutrition and parasitic diseases. *Journal of Nutrition* **133**(5 supplement 2): 1661S–1667S.

Tomkins A (2003) Assessing micronutrient status in the presence of inflammation. *Journal of Nutrition* **133**(5 supplement 2): 1649S–1655S.

UNICEF (2003) *UNICEF statistics: Malnutrition*. Available at http://childinfo.org/eddb/malnutrition/index.htm.

World Health Organization (2003) *HIV and Infant Feeding: Framework for Priority Action*. Available at www.who.int/child-adolescent-health/publications/NUTRITION/HIV_IF_Framework.htm.

World Health Organization (2003) *Department of Nutrition for Health and Development. Global Data Base on Child Growth and Malnutrition*. Available at www.who.int/nut-growthdb/.

Secondary, Diagnosis and Management

N Solomons, Center for Studies of Sensory Impairment, Aging and Metabolism (CeSSIAM), Guatemala City, Guatemala

Definitional Considerations

In its broadest context, malnutrition is a state of having an inappropriate nutritional status with respect to one or more macronutrient (water, electrolyte, protein, or fat) or micronutrient (vitamin or mineral) constituent of the body. This imbalance can be a deficit, leading to an insufficient supply or content of the nutrient (undernutrition), or an excess, leading to an excessive content or overloading of the organism with a nutrient (overnutrition).

Victor Herbert enumerated six possible causes for all nutrient deficiencies as: decreased intake; impaired absorption; increased wastage; impaired utilization; increased destruction; and elevated requirements. Correspondingly, with the exception of any utilization defect, overnutrition and excesses can result from the reciprocal defects, that is: hyperphagia; hyperabsorption; increased retention; decreased destruction; and decreased requirements.

As discussed in the previous chapter, the term 'primary' malnutrition relates almost exclusively to the first of these mechanisms, that of the ingestion of nutrients from the diet. It is about food consumption and intake. Secondary malnutrition, by contrast, concerns the disturbed and disordered handling of nutrients. When diseases or abnormal physiological conditions interfere with the normal disposition of nutrients ingested from the diet, this is the basis of a situation of 'secondary' malnutrition. A representative, but not exhaustive, list of diseases and conditions producing secondary undernutrition is provided in **Table 1**. The roster of causes of secondary overnutrition is provided in **Table 2**.

The basis for suspecting the presence of secondary undernutrition emerges when there is evidence of malnutrition (deficiency or excess) but food and nutrients are presumably being consumed with in abundance. Once the suspicion emerges, three distinct diagnostic principles need to be addressed: (1) the confirmation of dietary intake, and estimation of its adequacy; (2) the diagnosis and classification of abnormal nutritional status; and (3) the diagnosis of the functional, physiological, or pathological origins of disordered nutrient disposition. To emphasize the

Table 1 Diseases and conditions associated with secondary macronutrient or micronutrient undernutrition

Decreased nutrient absorption
Gastric atrophy
Pernicious anemia
Celiac disease
Inflammatory bowel disease
Intestinal cryptosporidiasis
Pancreatic insufficiency
Biliary obstruction
Cystic fibrosis
Radiation enteritis
Chronic intestinal pseudoobstruction
Increased nutrient excretion
Hepatic cirrhosis
Laxative abuse
Peptic ulcer
Gastrointestinal fistula
Gastric adenoma
Colonic adenoma
Amebiasis
Hookworm
Schistosomiasis
Diabetes mellitus
Fanconi syndrome
Hypoaldosteronism
Hemodialysis; peritoneal dialysis
Increased destruction or internal consumption nutrients
Hyperthyroidism
Cardiac cachexia
HIV/AIDS
Cancer cachexia
Cystic fibrosis
Bone marrow transplants
Pulmonary tuberculosis
Decreased utilization of nutrients
Lead poisoning
Menkes' copper storage disease

point, one must remain attuned to the nutritional status of patients, clients, or populations, and sensitive to the possibility of a nonprimary origin of any under- or overnutrition.

Table 2 Diseases and conditions associated with secondary macronutrient or micronutrient excess (overnutrition)

Increased nutrition absorption
Wilson's disease
Hemochromatosis
Increased nutrition retention
Prader–Willi syndrome
Hypercorticosterism
Hyperpituitarism
Acute tubular necrosis
Chronic renal failure
Decreased destruction of nutrients
Hypothyroidism

Coexistence of Primary and Secondary Malnutrition

It is important to recognize the potential for the simultaneous coexistence of primary and secondary malnutrition in the same individual. Primary malnutrition in the free-living populations can be associated with famine conditions (crop failure, conflict, natural disaster, refugee crisis), in which sufficient food is simply not available. Alternatively, it can arise from the poverty of landlessness or urban margination, where food is not accessible within the household income. A large number of communicable diseases with consequences for nutrient absorption, retention or utilization, such as parasitoses, tuberculosis or HIV/AIDS, are common in these situations of deprivation and misery.

To the extent that a disease process produces anorexia or dysphagia, or even psychic depression, the net effect is to reduce total intake of dietary energy and nutrients. Whatever, malabsorptive or nutrient-wasting components of the underlying disorders will further compromise the nutritional state.

The Reverse Paradigm: Underlying Pathology Revealed by Detection of Abnormal Nutrition

In clinical medicine, a type of 'reversal of roles' often occurs. Rather than primarily recognizing the presentation of the underlying pathology, recognition of an abnormal nutritional status without a suitable dietary cause leads to the diagnosis of the underlying disorder before any specific (pathognomonic) sign or symptom has yet occurred. For instance, the Prader-Willi syndrome of pathological obesity would initially present as common obesity. Similarly, in hypercorticosteroidism (Cushing's syndrome), abnormal fat deposition and weight gain can be the changes that lead to the recognition of the underlying pituitary or adrenal dysfunction.

Classically, in type 2 diabetes, unexplained weight loss is a presenting complaint when polyuria is mild or absent. Moreover, with common forms of childhood gastrointestinal disorders, such as celiac sprue or Crohn's disease, arrested linear growth is often the first clue that something is clinically awry. It provokes the diagnostic inquiry that leads to the recognition of the bowel lesions. In milder presentations of cystic fibrosis, a similar growth failure occurring in infancy, can indicate an underlying pathological disorder.

In fact, the entire roster of conditions listed in **Table 1** and **Table 2**, as well as others of a similar nature, are subject to being diagnosed as the result

Table 3 Three-level diagnostic principles related to secondary malnutrition

Assessment of dietary and nutrient intake: A quantitative and qualitative evaluation of usual dietary intake by a nutritionally trained practitioner or clinical dietician serves to exclude the possibility that the situation is not primary (low intake) in nature and suggests the secondary basis of the nutritional problem. Caveat: In certain situations, a **combination** of reduced intakes **and** nutritional stress at absorption, retention, or utilization may coexist.
Assessment of nutritional status: This includes the measures of anthropometry and body composition, hematological status, biological indices, and functional indicators, as well as clinical (physical) evaluation.
Diagnosis of underlying cause(s) of secondary nutritional imbalance: It is important, where possible, to identify the underlying entitie(s) that are causing the nutritional problem, to enable (where possible) a direct remedial approach to the cause of malnutrition and to orient management based on any pathophysiological knowledge about the underlying disease.

of a secondary change in nutritional status. The practical message is that the nutritional specialist, physician, or nonphysician may be the first person to whom the secondarily malnourished patient is referred and the acumen of recognising a secondary causation will guide the case to an appropriate clinical diagnostic program to uncover (and hopefully address and remedy) the underlying medical or surgical problems. Overarching guideline principles for uncovering secondary malnutrition states are provided in **Table 3**.

Diagnosis of Secondary Malnutrition

In general terms, a common set of principles applies for assessment of nutrient status whether the bases are primary, secondary, or a combination of both. These principles include: body composition measures, hematological and biochemical findings, functional variables, and clinical signs and symptoms. It is more productive to focus here on the nuances, caveats, and distinctions for the detection of altered nutrition due to background conditions beyond spontaneous food intake.

Caveats for the Diagnosis of Secondary Excess Nutriture

The conditions that cause increased retention of energy and hypometabolism are listed in **Table 2**. When it comes to overweight and obesity, the absence of clear-cut overeating combined with other characteristic signs of the different entities should raise suspicion. Excesses of vitamins and minerals may not easily be detected because the homeostatic control of circulating concentrations confounds biochemical diagnosis. Excessive urinary excretion rates of the nutrients or their metabolites often provide better indications than blood levels when micronutrient overload is the issue.

Caveats for the Diagnosis of Secondary Undernutrition

Undernutrition due to disease and dysfunction obviously requires establishment of the following: (1) the existence of deficiencies; and (2) that factors other than underconsumption are influencing the deficiency states. The body composition standard is a body mass index (BMI) of $<18.5\,\mathrm{kg\,m^{-2}}$. With the worldwide pandemic of overweight, recent weight loss of 10% or more of usual body weight may be a more sensitive and reliable indicator of an incipient undernutrition problem. Weight problems diagnosed in this manner would certainly be detectable well before the BMI will have fallen to the aforementioned criterion.

Ill patients with adequate or excessive body mass indices can manifest metabolic substrate metabolism reminiscent of the severe malnutrition syndromes of adult kwashiorkor or marasmus (inanition). Moreover, fluctuations in weight under acute or semiacute situations often reflect changes in fluid balance. This is also the situation in patients with end-stage renal failure undergoing chronic dialysis. Methods such as bioelectrical impedance, dual X-ray absorbance, or isotope dilution in association with indirect calorimetry can assess true lean- and fat-mass status and macronutrient metabolism in patients of apparently normal body mass.

Hematological evaluation is important in nutritional assessment. A low hemoglobin, hematocrit, or red cell count signifies anemia, but in individuals with associated diseases, anemia can have a series of origins (hemolytic, hypoproliferative) that are nonnutritional and will not respond to nutritional therapy.

Biochemical evaluation for nutrient deficiency status in patients with associated disease is fraught with caveats and limitations. Michael Golden has defined two classes of nutritional deficiency: in type 1 deficiencies, nutritional desaturation of tissue stores occurs, and circulating levels of nutrients reflect the total body nutrient status; in type 2 deficiency, there is homeostatic conservation of tissue and circulating concentrations of nutrients, such that blood concentrations remain virtually unaltered in the face of depletion. Deficiencies of zinc and magnesium, among others, fall into this second category. Inflammation and infection are stimuli that directly alter the circulating concentrations of

nutrient indicators. Ferritin and circulating copper are elevated whereas zinc, iron, and vitamin A concentrations are depressed with activation of the acute-phase response to injury. In liver disease, depressed production of binding proteins can alter the usual indicators of nutritional status as a consequence of hepatic pathophysiology itself, rather than preexisting secondary malnutrition. Finally, it almost goes without saying that attempting biochemical nutrient evaluations from blood samples taken during concurrent infusion of micronutrient solutions in parenteral nutrition regimens – and without a period of distribution and equilibration – will not reflect the tissue stores and total body reserves of the respective nutrients of interest.

Functional indicators of nutritional status have been applied to the assessment of secondary malnutrition and have been plagued by pitfalls. This applies to tests of nitrogen status, immune function, and hepatic protein secretion. Tests such as creatinine excretion, white blood cell counts, and cutaneous delayed hypersensitivity anergy, as well as decreased serum albumin, transferrin, transthyronein (prealbumin), and retinal-binding protein concentrations are sensitive to alteration by stress and injury. Failure to recognize distortion from stress underlies an early fallacy in surgical nutrition, in which low values for albumin, lymphocyte counts, and pre-albumin, together with anergy, predicted poor postoperative outcomes. This misconception justified aggressive preoperative parenteral nutrition and albumin infusions, with little impact on predicted outcomes. In these situations, it was the stress and injury of the underlying disease, rather than nutritional status, that was producing the abnormal values for the biomarkers. Recently, insulin-like growth factor has been advanced as a sensitive indicator of protein status in older patients, but whether it is confounded by nonnutritional features of disease remains to be clarified.

Management of Secondary Malnutrition

Secondary malnutrition has many faces and facets. It may have to be addressed both in a public health sense, for communicable diseases, such as parasitoses or HIV/AIDS, and in a medical care context, for disorders that are particular and clinical in nature, such as hereditary or degenerative diseases.

Principles of Management

The first principle is to identify the underlying functional, physiological, or pathological cause of the malnourished state. If the condition is curable, then the management issues are simplified. For instance, if a person is dehydrated because of hyperglycemic diuresis in uncontrolled diabetes mellitus, the short-term management involves administration of exogenous intravenous fluids to restore normal hydration; however, restoring adequate diabetic control to the patient would be the long-term and definitive solution. The undernutrition and growth failure due to undetected celiac disease is easily eliminated by institution of a gluten-free diet. With deficient nutrition in cystic fibrosis, adequate management of pulmonary problems and digestive-enzyme should allow patients to recover and maintain normal nutrition on a balanced oral diet. Thus, medical or surgical address of the underlying disorder, where possible, is the primary tool for management of secondary undernutrition.

Public Health Approaches

The management of the secondary nutritional deficiency attributable to hookworm or schistosomiasis, i.e., iron deficiency, can be achieved both by anthelmintic medications or by supplemental iron to compensate for parasite-induced losses. In countries where HIV/AIDS is rampant efforts for its prevention are fundamental. A food-security crisis grips the whole society in AIDS endemic areas, and this must be relieved with food and economic assistance. The wasting syndromes produced by tuberculosis are best addressed proactively by prevention of transmission and early detection. However, when primary prevention fails, as in the aforementioned infections, efforts to enhance the enteral intake of infected members of the community are particularly essential for their comfort and well being.

Dietary Management of Secondary Overnutrition

The dietary management of secondary overnutrition would logically be to restrict the intake of the nutrients accruing in excess. This is not always facile or feasible, however, due to the intrinsic complexity of foods and beverages, where most are sources of multiple essential micronutrients. Marked reduction in total energy intake can jeopardize the intake of proteins and essential fats. For the metal-storage afflictions such as Wilson's disease and hemochromatosis, removing copper and iron from the diet, respectively, are the fundamental elements of management. Some additional benefits can be gained by blocking the metals' absorption, as with high doses of zinc in Wilson's disease or with strong black tea (tannins) in hemochromatosis. Fundamentally, however, the management of metal-storage diseases requires some interventions to selectively remove the overload by

chelating agents in Wilson's disease and recurrent phlebotomy in hemochromatosis. In a related variant condition, African hemosiderosis, common among Bantu in southern Africa, removing concentrated iron sources from the diet, specifically the iron-loaded native beers, provides effective long-term control.

Dietary and Nutritional Management of Secondary Undernutrition

The syllogism for dietary and nutritional management is to get enough nutrients into the body to restore nutritional adequacy and balance, taking any chronic barriers to uptake and retention into consideration. The blend of nutrients must be tailored to the specific absorptive or utilization problems, e.g., compensatory fat-soluble vitamins in water-miscible forms with severe fat malabsorption, and extra doses of highly available iron with chronic blood loss. These can be delivered within a dietary context with supplements and fortified vehicles in nonacute conditions. Even nondietary routes have been devised as in the treatment of vitamin D deficiency due to Crohn's disease with tanning bed ultraviolet B radiation.

When accumulated undernutrition is dangerously advanced, absorptive barriers are especially severe, or nutrient losses are excessive more concerted nutritional intervention is required. Intensive therapy can be delivered by three routes: orally, with special diets supplemented by liquid formulas; enterally, with liquid formulas perfused by intragastric or intraintestinal feeding tubes; and parenterally, with intravenous formulas infused into peripheral or central veins. Up to 50% of patients on dialysis have protein-energy malnutrition, which may continue undetected. For end-stage renal patients, intradialytic alimentation (adding nutrients to the dialysis fluids) has been used to reduce nutrient loss. Each approach has its distinct costs, special potential, and limitations and risks, and has been explored and refined in the context of age, physiological status, and specific disease states or surgical indications.

Tailoring of nutrient delivery is required with both enteral and parenteral nutrition, depending upon the pathophysiology of the underlying conditions. Both hypo- and hypermetabolic states can occur; indirect calorimetry with metabolic carts is in vogue for prescribing energy delivery in intensive care. When pulmonary compromise is present, the balance among macronutrients is important to minimize carbon dioxide formation in metabolism.

Maintaining abundant amino acid supply promotes protein-sparing and prevents loss of lean tissue in catabolic states. Enrichment of enteral or parenteral regimens with branched-chain amino acids or keto-analog amino acids has been devised to compensate for the metabolic defects of nitrogen handling in hepatic or renal failure states. The objective of nutritional support in patients with liver failure is to provide adequate macronutrients to ensure the specific substrates for energy and protein synthesis and integrity of normal hepatic tissue function, without inducing or accentuating encephalopathy or otherwise aggravating hepatic insufficiency.

In juvenile cholestasis, large amounts of fat-soluble vitamin supplements and medium-chain triglycerides are usually required for optimum growth. With protracted secretory diarrheal diatheses, fluid and electrolyte balance may be the primary concern, followed by macro- and micronutrient nutriture, invoking the institution of parenteral feeding. Cancer cachexia is a major secondary consequence of disseminated neoplasms. It is tempting to prescribe aggressive nutritional support, but a caveat is that certain nutrients acting with certain neoplasms favor the tumor's growth and dissemination. To the extent that various forms of cachexia are partly driven by catabolic responses mediated by proinflammatory cytokines, antagonists directed at counteracting their action hold promise for retarding the nutrient-wasting in various forms of cachexia.

With intensive nutrition, there are risks and adverse consequences intertwined with the benefits. A variation of the refeeding syndrome, that is hyperalimentation complications from excessive energy substrate perfusion or infusion, can produce hypophosphatemic and hypokalemic episodes. Improper formulation of fluids or liquids with micronutrients can cause deficiency or toxicity states in chronic nutritional support. The hazards of indwelling catheters are multiple, from phlebitis of the veins to sudden dislocation or migration. Fluid overload and sepsis are the most troubling complications of intravenous parenteral nutrition.

For tube-feeding enteral alimentation, tube placement is the crucial element. With nasal placement of the tube, there is a finite risk of respiratory tract inflammation and infection from aspiration of formula and secretions. In hospital, enteral nutrition is a risk factor for nosocomial pneumonia. An alternative site for long-term administration of tube-feeds is percutaneous placement of an intragastric feeding tube under endoscopic control.

Aggressive nutritional support, with its attendant expense and potential morbidity, in critically ill patients remains controversial. In terms of cost–benefit analysis, the use of the intensive formats of enteral artificial nutrition seems to be cost effective to reduce post-hip-fracture hospital stay in underweight women and for preoperative

nutritional support, if carried out at home. Preoperative parenteral nutrition has been judged as prohibitively expensive for the small reduction in postoperative morbidity that it produces.

Conclusions

Dietary intake is the most important determinant of over- or undernutrition, but it is not the only influence on an individual's nutritional status. A series of extrinsic environmental factors or intrinsic clinical or physiological disorders can alter the absorption, retention, utilization, and integrity of nutrients. These give origin to secondary malnutrition states. Primary (dietary origin) and secondary (environmental, pathological) factors often combine within the same individuals. From a public health perspective, the goal is to implement broad policies and programs that increase the availability of specific nutrients imperiled by the local environmental problems, e.g., iron in hookworm infested areas, while addressing the primary diseases. In the clinical setting, management requires diagnosing and managing the underlying pathological states interfering with nutritional health while providing compensatory measures to correct secondary nutritional imbalances.

See also: **Cystic Fibrosis**. **Diabetes Mellitus**: Etiology and Epidemiology. **Handicap**: Prader–Willi Syndrome. **Liver Disorders**. **Malnutrition**: Primary, Causes Epidemiology and Prevention. **Nutritional Support**: Adults, Enteral; Adults, Parenteral; Infants and Children, Parenteral. **Zinc**: Deficiency in Developing Countries, Intervention Studies.

Further Reading

Brooks MJ and Melnik G (1995) The refeeding syndrome: an approach to understanding its complications and preventing its occurrence. *Pharmacotherapy* **15**: 713–726.

Dudrick SJ, Maharaj AR, and McKelvey AA (1999) Artificial nutritional support in patients with gastrointestinal fistulas. *World Journal of Surgery* **23**: 570–576.

Herbert V (1973) The five possible causes of all nutrient deficiencies: Illustrated by deficiencies of vitamin B12 and folic acid. *American Journal of Clinical Nutrition* **26**: 77–88.

McKenzie C, Vicca N, Ward JE, and Coles SJ (2000) Prevalence of malnutrition on admission to four hospitals in England. The Malnutrition Prevalence Group. *Clinical Nutrition* **19**: 191–195.

Murray MJ, Marsh HM, Wochos DN, Moxness KE, Offord KP, and Callaway CW (1988) Nutritional assessment of intensive-care unit patients. *Mayo Clinic Proceedings* **63**: 1106–1115.

Ofman J and Koretz RL (1997) Clinical economics review: nutritional support. *Alimention and Pharmacological Therapy* **11**: 453–471.

Paccagnella A, Calo MA, Caenaro G, Salandin V, Jus P, Simini G, and Heymsfield SB (1994) Cardiac cachexia: preoperative and postoperative nutrition management. *Journal of Parenteral and Enteral Nutrition* **18**: 409–416.

Phang PT and Aeberhardt LE (1996) Effect of nutritional support on routine nutrition assessment parameters and body composition in intensive care unit patients. *Canadian Journal of Surgery* **39**: 212–219.

Solomons NW (1993) Pathways to impairment of nutritional status by gastrointestinal pathogens, with emphasis on protozoal and helminthic parasites. *Parasitology* **107**(supplement): S19–S35.

Solomons NW and Keusch GT (1999) Clinical issues: Childhood illnesses, vaccinations and nutritional status. In: Gershwin ME, German JB, and Keen CL (eds.) *Nutrition and Immunology: Principles and Practice*, pp. 469–474. Totowa, NJ: Humana Press.

Thapa BR (1994) Intractable diarrhoea of infancy and its management: modified cost effective treatment. *Journal of Tropical Pediatrics* **40**: 157–161.

Von Roenn JH, Roth EL, and Craig R (1992) HIV-related cachexia: potential mechanisms and treatment. *Oncology* **49**(supplement 2): 50–54.

Zipf WB (2004) Prader–Willi syndrome: the care and treatment of infants, children, and adults. *Advances in Pediatrics* **51**: 409–434.

MANGANESE

C L Keen, J L Ensunsa, B Lönnerdal and S Zidenberg-Cherr, University of California at Davis, Davis, CA, USA

The essentiality of manganese was established in 1931, when it was demonstrated that a deficit of it resulted in poor growth and impaired reproduction in rodents. Manganese deficiency can be a practical problem in the swine and poultry industries, and it may be a problem in some human populations. Conversely, manganese toxicity can be a significant human health concern. Here, literature related to manganese nutrition, metabolism, and metabolic function is reviewed.

Chemical and Physical Properties

Manganese is the 12th most abundant element in the Earth's crust and constitutes approximately 0.1% of it. Chemical forms of manganese in their natural deposits include oxides, sulfides, carbonates, and

silicates. Anthropogenic sources of manganese are predominantly from the manufacturing of steel, alloys, and iron products. Manganese is widely used as an oxidizing agent, as a component of fertilizers and fungicides, and in dry cell batteries. Methylcyclopentadienyl manganese tricarbonyl (MMT) improves combustion in boilers and motors and can substitute for lead in gasoline as an antiknock agent. Concentrations of manganese in groundwater normally range between 1 and $100\,\mu g\,l^{-1}$, with most values being below $10\,\mu g\,l^{-1}$. Typical airborne levels of manganese (in the absence of excessive pollution) range from 10 to $70\,ng\,m^{-3}$.

Manganese is a transition element located in group VIIA of the periodic table. It occurs in 11 oxidation states ranging from -3 to $+7$, with the physiologically most important valences being $+2$ and $+3$. The $+2$ valence is the predominant form in biological systems and is the form that is thought to be maximally absorbed. The $+3$ valence is the form in which manganese is primarily transported in biological systems.

The solution chemistry of manganese is relatively simple. The aquo-ion is resistant to oxidation in acidic or neutral solutions. It does not begin to hydrolyze until pH 10, and therefore free Mn^{2+} can be present in neutral solutions at relatively high concentrations. Divalent manganese is a $3d^5$ ion and typically forms high-spin complexes lacking crystal field stabilization energies. The previous properties, as well as a large ionic radius and small charge-to-radius ratio, result in manganese tending to form weak complexes compared with other first-row divalent ions, such as Ni^{2+} and Cu^{2+}. Free Mn^{2+} has a strong isotropic electron paramagnetic resonance (EPR) signal that can be used to determine its concentration in the low micromolar range. Mn^{3+} is also critical in biological systems. For example, Mn^{3+} is the oxidative state of manganese in superoxide dismutase, is the form in which transferrin binds manganese, and is probably the form of manganese that interacts with Fe^{3+}. Given its smaller ionic radius, the chelation of Mn^{3+} in biological systems would be predicted to be more avid than that of Mn^{2+}. Cycling between Mn^{3+} and Mn^{2+} has been suggested to be deleterious to biological systems because it can generate free radicals. However, at low concentrations Mn^{2+} can provide protection against free radicals, and it appears to be associated with their clearance rather than their production.

Dietary Sources

Manganese concentrations in typical food products range from $0.4\,\mu g\,g^{-1}$ (meat, poultry, and fish) to $20\,\mu g\,g^{-1}$ (nuts, cereals, and dried fruit). Breast milk is exceptionally low in manganese, containing only $0.004\,\mu g\,g^{-1}$, whereas infant formula can contain up to $0.4\,\mu g\,g^{-1}$. Teas can be particularly rich in manganese, containing up to $900\,\mu g\,g^{-1}$of the element. An important consideration with respect to food sources of manganese is the extent to which the manganese is available for absorption. For example, although tea contains high amounts of the element, the tannin in tea can bind a significant amount of manganese, reducing its absorption from the gastrointestinal tract. Similarly, the high content of phytates and fiber constituents in cereal grains may limit the absorption of manganese. Conversely, although meat products contain low concentrations of manganese, absorption and retention of manganese from them is relatively high. Based on studies utilizing whole body retention curves after dosing with ^{54}Mn, the estimated percentage absorption of 1 mg of manganese from a test meal was 1.35%, whereas that from green leafy vegetables (lettuce and spinach) was closer to 5%. Absorption from wheat and sunflower seed kernels was somewhat lower than that from the leafy greens at 1 or 2%, presumably due to a higher fiber content or to higher amounts of phytates and similar compounds in the wheat and sunflower seeds. The dephytinization of soy formula increased manganese absorption 2.3-fold from 0.7 to 1.6%.

Analysis

Although manganese is widely distributed in the biosphere, it occurs in only trace amounts in animal tissues. Serum concentrations can be as low as 20 nM and typical tissue concentrations are less than $4\,\mu mol\,g^{-1}$ wet weight; tissue concentrations of 4–$8\,\mu mol\,g^{-1}$ wet weight are considered high. Because of the high environmental levels of manganese relative to its concentration in animal tissues, considerable effort must be made to minimize contamination of samples during their collection and handling.

The most common analytical methods that can sensitively measure manganese include neutron activation analysis, X-ray fluorescence, proton-induced X-ray emission, inductively coupled plasma emission, EPR, and flameless atomic absorption spectrophotometry (AAS). Currently, the most common method employed is flameless AAS. All of these methods, with the exception of EPR, measure the total concentration of manganese in the samples. EPR allows selective measurement of bound versus free manganese.

Physiological Role

Tissue Concentrations

The average human body contains between 200 and 400 μmol of manganese, which is fairly uniform in distribution throughout the body. There is relatively little variation among species with regard to tissue manganese concentrations. Manganese tends to be highest in tissues rich in mitochondria; its concentration in mitochondria is higher than in cytoplasm or other cell organelles. Hair can accumulate high concentrations of manganese, and it has been suggested that hair manganese concentrations may reflect manganese status. High concentrations of manganese are normally found in pigmented structures, such as retina, dark skin, and melanin granules. Bone, liver, pancreas, and kidney tend to have higher concentrations of manganese (20–50 nmol g^{-1}) than do other tissues. Concentrations of manganese in brain, heart, lung, and muscle are typically <20 nmol g^{-1}; blood and serum concentrations are approximately 200 and 20 nmol l^{-1}, respectively. Typical concentrations in cow milk are on the order of 800 nmol l^{-1}, whereas human milk contains 80 nmol l^{-1}. Bone can account for up to 25% of total body manganese because of its mass. Bone manganese concentrations can be raised or lowered by substantially varying dietary manganese intake over long periods of time, but bone manganese is not thought to be a readily mobilizable pool. The fetus does not accumulate liver manganese before birth, and fetal concentrations are significantly less than adult concentrations. This lack of fetal storage can be attributed to the apparent lack of storage proteins and the low prenatal expression of most manganese enzymes.

Absorption, Transport, and Storage

Absorption of manganese is thought to occur throughout the small intestine. Manganese absorption is not thought to be under homeostatic control. For adult humans, manganese absorption has been reported to range from 2 to 15% when ^{54}Mn-labeled test meals are used and to be 25% when balance studies are conducted; given the technical problems associated with balance studies, the ^{54}Mn data are probably more reflective of true absorption values. Data from balance studies indicate that manganese retention is very high during infancy, suggesting that neonates may be particularly susceptible to manganese toxicosis.

The higher retention of manganese in young animals relative to adults in part reflects an immaturity of manganese excretory pathways, particularly that of bile secretion, which is very limited in early life. The avid retention of the small amount of manganese from milk and the postnatal changes in its excretory pattern underscore the considerable changes in manganese metabolism that occur during the neonatal period.

In experimental animals, high amounts of dietary calcium, phosphorus, fiber, and phytate increase the requirements for manganese; such interactions presumably occur via the formation of insoluble manganese complexes in the intestinal tract with a concomitant decrease in the soluble fraction available for absorption. The significance of these dietary factors with regard to human manganese requirements remains to be clarified. Studies in avian species have demonstrated that high dietary phosphorus intakes decrease manganese deposition in bone by approximately 50%. Given that the diet of many individuals may be marginal in manganese ($\leq$2 mg per day intake) while high in phosphorus ($\geq$2000 mg per day intake), this antagonism may have important implications for human health. For example, the low fractional absorption of manganese from soy formula has been related to its relatively high phytate content. The mechanism underlying this effect of soy protein on manganese absorption/retention has not been fully delineated. However, dephytinization of soy formula with microbial phytase can markedly enhance manganese absorption.

An interaction between iron and manganese has been demonstrated in experimental animals and humans. Manganese absorption increases under conditions of iron deficiency, whereas high amounts of dietary iron can accelerate the development of manganese deficiency. The chronic consumption of high levels of iron supplements (>60 mg Fe per day) can have a negative effect on manganese balance in adult women. The mechanisms underlying the interactions between iron and manganese have not been fully elucidated; however, they likely involve competition for either a transport site or a ligand. Both iron and manganese can utilize divalent metal transporter 1 (DMT1); however, the expression of DMT1 is regulated by iron status via the IRE/IRP system. Thus, during iron deficiency, DMT1 is upregulated causing an increase in manganese absorption. Rats fed iron-deficient diets accumulate manganese in several brain regions compared to rats fed control diets; the involvement of DMT1 in this accumulation of manganese is an area of active study. It should be noted that the interaction between manganese and iron can also affect the functions of some enzymes. For example, manganese can replace iron in the iron–sulfur center of cytosolic aconitase (IRP-1), resulting in an inhibition of the enzyme and an increase in iron

regulatory protein (IRP) binding activity. Given the central role of IRPs in cellular iron metabolism, elevated cellular manganese concentration could in theory disrupt numerous translational events dependent on IRPs. That this in fact occurs is illustrated by the observation that following the addition of manganese to cells in culture, there can be sharp reductions in ferritin protein abundance, whereas there are increases in transferrin receptor abundance. This results in changes in intracellular iron metabolism, as reflected by decreases in mitochondrial aconitase (m-aconitase) abundance.

Manganese entering the portal blood from the gastrointestinal tract may remain free or become associated with α_2-macroglobulin, which is subsequently taken up by the liver. A small fraction enters the systemic circulation, where it may become oxidized to Mn^{3+} and bound to transferrin. Studies *in vivo* suggest that the Mn^{3+} complex forms very quickly in blood, in contrast to the slow oxidation of the Mn^{2+}–transferrin complex *in vitro*. Manganese uptake by the liver has been reported to occur by a unidirectional, saturable process with the properties of passive mediated transport. After entering the liver, manganese enters one of at least five metabolic pools. One pool represents manganese taken up by the lysosomes, from which it is transferred subsequently to the bile canaliculus. The regulation of manganese is maintained in part through biliary excretion of the element; up to 50% of manganese injected intravenously can be recovered in the feces within 24 h. A second pool of manganese is associated with the mitochondria. Mitochondria have a large capacity for manganese uptake, and the mitochondrial uptake and release of manganese and calcium are thought to be related. A third pool of manganese is found in the nuclear fraction of the cell; the roles of nuclear manganese have not been fully delineated, but one function may be to contribute to the stability of nucleosome structure. A fourth manganese pool is incorporated into newly synthesized manganese proteins; biological half-lives for these proteins have not been agreed upon. The fifth identified intracellular pool of manganese is free Mn^{2+}. Fluctuations in the free manganese pool may be an important regulator of cellular metabolic control in a manner analogous to those for free Ca^{2+} and Mg^{2+}. Consistent with this concept, in pancreatic islets manganese blocks glucose-induced insulin release by altering cellular calcium fluxes, and manganese directly augments contractions in smooth muscle by a mechanism comparable to that of calcium.

The mechanisms by which manganese is transported to, and taken up by, extrahepatic tissues have not been identified. Transferrin is the major manganese binding protein in plasma; however, it is not known to what extent transferrin facilitates the uptake of manganese by extrahepatic tissue. The concentration of manganese citrate in blood can be fairly high, and this complex may be important for manganese movement across the blood–brain barrier. DMT1 may be involved in manganese transport because it is expressed in discrete areas of the brain. Manganese uptake by extrahepatic tissue does not seem to be increased under conditions of manganese deficiency, suggesting that manganese, in marked contrast to iron, does not play a role in the induction (or suppression) of manganese transport proteins.

There is limited information concerning the hormonal regulation of manganese metabolism. Fluxes in the concentrations of adrenal, pancreatic, and pituitary–gonadal axis hormones affect tissue manganese concentrations; however, it is not clear to what extent hormone-induced changes in tissue manganese concentrations are due to alterations in cellular uptake of manganese-activated enzymes or metalloenzymes.

Metabolic Function and Essentiality

Manganese functions as a constituent of metalloenzymes and as an enzyme activator. Manganese-containing enzymes include arginase (EC 3.5.3.1), pyruvate carboxylase (EC 6.4.1.1), and manganese–superoxide dismutase (MnSOD) (EC 1.15.1.1). Arginase, the cytosolic enzyme responsible for urea formation, contains 4 mol Mn^{2+} per mole of enzyme. Reductions in arginase activity resulting from manganese deficiency result in elevated plasma concentrations of ammonia and lowered plasma concentrations of urea. Reductions in arginase activity due to manganese deficiency may affect flux of arginine through the nitric oxide synthase (NOS) pathway, resulting in alterations in NO production. It has been suggested that arginase plays a regulatory role in NO production by competing with NOS for the same substrate, arginine. Rats fed manganese-deficient diets have shown effects indicative of increased NO production, such as increases in plasma and urinary nitrates plus nitrites and decreased blood pressure; however, neither NOS activity nor NO production have been measured directly. In addition, manganese binding by arginase is critical for the pH-sensing function of this enzyme in the ornithine cycle, suggesting that manganese plays a role in the regulation of body pH. With experimental diabetes, liver and kidney manganese concentrations and arginase activity can be markedly elevated.

This manganese effect on arginase has been suggested to be due to an effect of Mn^{2+} on the conformational properties of the enzyme with a resultant modification of arginase activity. Whether this finding implies an increased manganese requirement for people with diabetes has not been determined.

Pyruvate carboxylase, the enzyme that catalyses the first step of carbohydrate synthesis from pyruvate, also contains 4 mol Mn^{2+} per mole enzyme. Although the activity of this enzyme can be lower in manganese-deficient animals than in controls, gluconeogenesis has not been shown to be markedly inhibited in manganese-deficient animals.

MnSOD catalyzes the disproportionation of $\dot{O}_2^-$ to H_2O_2 and $\dot{O}_2$. The essential role of MnSOD in the normal biological function of tissues has been clearly demonstrated by the homozygous inactivation of the *SOD2* gene for MnSOD in mice. Mice with this phenotype die within the first 10 days of life with a dilated cardiomyopathy, accumulation of lipid in liver and skeletal muscle, and metabolic acidosis. The activity of MnSOD in tissues of manganese-deficient rats can be significantly lower than in controls due to downregulation of MnSOD at the (pre)transcriptional level. That this reduction is functionally significant is suggested by the observation of higher than normal levels of hepatic mitochondrial lipid peroxidation in manganese-deficient rats. Tissue MnSOD activity can be increased by several diverse stressors, including alcohol, ozone, irradiation, interleukin-1, and tumor necrosis factor-α, presumably as a consequence of stressor-associated increases in cellular free radical (or oxidized target(s)) concentrations. Stressor-induced increases in MnSOD activity can be attenuated in manganese-deficient animals, potentially increasing their sensitivity to these insults. Transgenic mice have also been produced that overexpress MnSOD; a decreased severity of reperfusion injury has been noted in these animals, further supporting its physiological significance.

Considerable research is focused on the introduction of the human MnSOD gene into research animals utilizing viral vectors or plasmid/liposome delivery. This gene therapy has been shown to decrease radiation-induced injury, extend pancreatic islet transplant function, and slow the growth of malignant tumors in animal models via overexpression of the MnSOD protein. Another field of research that is rapidly advancing utilizes MnSOD mimetics for treatment of a variety of diseases in which the native SOD enzyme has been found to be effective. These mimetics are small manganese-containing synthetic molecules that have catalytic activity equivalent or superior to the native enzyme. They possess the additional beneficial properties of being nonimmunogenic because they are nonpeptides, able to penetrate cells, selective for superoxide (they do not interact with biologically important molecules), stable *in vivo*, and not deactivated by the destructive free radical peroxynitrite, which is capable of deactivating native MnSOD via nitration of tyrosine. These mimetic compounds have been found to be protective in animal models of acute and chronic inflammation, reperfusion injury, shock, and radiation-induced injury. Both of these therapies, MnSOD gene delivery and MnSOD mimetics, hold promise for future treatments in human chronic and acute conditions.

Finally, further evidence for the biological and research relevance of MnSOD is that experiments have been undertaken on the International Space Station to improve three-dimensional growth of MnSOD crystals in order to develop a better understanding of the role of structure in the reaction mechanism of this enzyme.

In contrast to the relatively few manganese metalloenzymes, there are a large number of manganese-activated enzymes, including hydrolases, kinases, decarboxylases, and transferases. Manganese activation of these enzymes can occur as a direct consequence of the metal binding to the protein, causing a subsequent conformation change, or by binding to the substrate, such as ATP. Many of these metal activations are nonspecific in that other metal ions, particularly Mg^{2+}, can replace Mn^{2+}. An exception is the manganese-specific activation of glycosyltransferases. Several manganese deficiency-induced pathologies have been attributed to a low activity of this enzyme class. A second example of an enzyme that may be specifically activated by manganese is phosphoenolpyruvate carboxykinase (PEPCK; EC 4.1.1.49), the enzyme that catalyzes the conversion of oxaloacetate to phosphoenolpyruvate, GDP, and CO_2. Although low activities of PEPCK can occur in manganese-deficient animals, the functional significance of this reduction is not clear.

A third example of a manganese-activated enzyme is glutamine synthetase (EC 6.3.1.2). This enzyme, found in high concentrations in the brain, catalyzes the reaction $NH_3 + \text{glutamate} + \text{ATP} \rightarrow \text{glutamine} + \text{ADP} + P_i$. Brain glutamine synthetase activity can be normal even in severely manganese-deficient animals, suggesting that the enzyme either has a high priority for this element or magnesium can act as a substitute when manganese is lacking. It should be noted that this enzyme can be inactivated by oxygen radicals; therefore, a manganese deficiency-induced reduction in MnSOD activity theoretically could act to depress further the activity of glutamine synthetase.

Manganese Deficiency

Manganese deficiency has been demonstrated in several species, including rats, mice, pigs, and cattle. Signs of manganese deficiency include impaired growth, skeletal abnormalities, impaired reproductive performance, ataxia, and defects in lipid and carbohydrate metabolism.

The effects of manganese deficiency on bone development have been studied extensively. In most species, manganese deficiency can result in shortened and thickened limbs, curvature of the spine, and swollen and enlarged joints. The basic biochemical defect underlying the development of these bone defects is a reduction in the activities of glycosyltransferases; these enzymes are necessary for the synthesis of the chondroitin sulfate side chains of proteoglycan molecules. In addition, manganese deficiency in adult rats can result in an inhibition of both osteoblast and osteoclast activity. This observation is particularly noteworthy, given the reports that women with osteoporosis tend to have low blood manganese concentrations and that the provision of manganese supplements might be associated with an improvement in bone health in postmenopausal women.

One of the most striking effects of manganese deficiency occurs during pregnancy. When pregnant animals (rats, mice, guinea pigs, and mink) are deficient in manganese, their offspring exhibit a congenital, irreversible ataxia characterized by incoordination, lack of equilibrium, and retraction of the head. This condition is the result of impaired development of the otoliths, the calcified structures in the inner ear responsible for normal body-righting reflexes. The block in otolith development is secondary to depressed proteoglycan synthesis due to low activity of manganese-requiring glycosyltransferases.

Defects in carbohydrate metabolism, in addition to those described previously, have been shown in manganese-deficient rats and guinea pigs. In the guinea pig, perinatal manganese deficiency results in pancreatic pathology, with animals exhibiting aplasia or marked hypoplasia of all cellular components. Manganese-deficient guinea pigs and rats given a glucose challenge often respond with a diabetic-type glucose tolerance curve. In addition to its effect on pancreatic tissue integrity, manganese deficiency can directly impair pancreatic insulin synthesis and secretion as well as enhance intracellular insulin degradation. The mechanism(s) underlying the effects of manganese deficiency on pancreatic insulin metabolism have not been fully delineated, but they are thought to be multifactorial. For example, the flux of islet cell manganese from the cell surface to an intracellular pool may be a critical signal for insulin release. It is also known that insulin mRNA levels are reduced in manganese-deficient animals, which is consistent with their depressed insulin synthesis. In addition, insulin sensitivity of adipose tissue is reduced in manganese-deficient rats, a phenomenon that may be related to fewer insulin receptors per adipose cell. Manganese deficiency may also affect glucose metabolism by means of a reduction in the number of glucose transporters in adipose tissue by an unidentified mechanism. Finally, the effect of manganese deficiency on insulin production may also be due to the destruction of pancreatic J3 cells. It is worth noting that constitutive pancreatic MnSOD activity is lower than in most tissues; this, coupled with the observation that most diabetogenic agents function via the production of free radicals with subsequent tissue damage, suggests that an additional mechanism underlying pancreatic dysfunction in manganese-deficient animals may be free radical mediated.

In addition to its effect on endocrine function, manganese deficiency can affect pancreatic exocrine function. For example, manganese-deficient rats can be characterized by an increase in pancreatic amylase content. The mechanism underlying this effect of manganese deficiency has not been delineated; however, it is thought to involve a shift in amylase synthesis or degradation because secretagogue-stimulated acinar secretion is comparable in control and manganese-deficient rats.

Although the majority of studies concerning the influence of manganese deficiency on carbohydrate metabolism have been conducted with experimental animals, there is one report in the literature of an insulin-resistant diabetic patient who responded to oral doses of manganese (doses ranged from 5 to 10 mg) with decreasing blood glucose concentrations. Although this is an intriguing case report, others have reported a lack of an effect of oral manganese supplements (up to 30 mg) in diabetic subjects, and low blood manganese concentrations have not been found to be a characteristic of diabetics.

Abnormal lipid metabolism is also characteristic of manganese deficiency: Specifically, a lipotrophic effect of manganese has been suggested in the literature. Severely manganese-deficient animals can be characterized by high liver fat, hypocholesterolemia, and low high-density lipoprotein (HDL) concentrations. Deficient animals can also be characterized by a shift to smaller plasma HDL particles, lower HDL apolipoprotein (apoE) concentrations, and higher apoC concentrations. As stated previously, tissue lipid peroxidation rates can be increased in

manganese-deficient animals, possibly as a result of low tissue MnSOD activity.

There is considerable debate as to the extent to which manganese deficiency affects humans under free-living conditions. Manganese deficiency can be induced in humans under highly controlled experimental conditions. In one study, manganese deficiency was induced in adult male subjects by feeding a manganese-deficient diet (0.1 mg Mn per day) for 39 days. The subjects developed temporary dermatitis, as well as increased serum calcium and phosphorus concentrations and increased alkaline phosphatase activity, suggestive of bone resorption. Since the late 1980s, several diseases have been reported to be characterized, in part, by low blood manganese concentrations. These diseases include epilepsy, Mseleni disease, maple syrup urine disease and phenylketonuria, Down's syndrome, osteoporosis, and Perthes' disease. The finding of low blood manganese levels in subsets of individuals with the previously mentioned diseases is significant since blood manganese levels can reflect soft tissue manganese concentrations. The reports of low blood manganese concentrations in individuals with epilepsy are particularly intriguing, given the observations that manganese-deficient animals can show an increased susceptibility to drug and electroshock-induced seizures and a genetic model for epilepsy in rats (the GEPR rat) is characterized by low blood manganese concentrations. It is evident that a deficiency of manganese may contribute to the pathology of epilepsy at multiple points, given that Mn^{2+} is implicated in activation of glutamine synthetase, a Mn^{2+}-specific brain ATPase; production of cyclic AMP; altered synaptosomal uptake of noradrenalin and serotonin; glutamate, GABA, and choline metabolism; and biosynthesis of acetylcholine receptors.

Evidence of widespread manganese deficiency in human populations is lacking. Typically, manganese intakes approximate the 2001 US Institute of Medicine's suggested adequate intakes as follows: 3 μg/day for infants 0–6 months old, 0.6 mg/day for infants 7–12 months old, 1.2–1.9 mg/day for children 1–13 years old, 1.6–2.2 mg/day for older children, and 1.8–2.6 mg/day for adults. The Tolerable Upper Intake Level (UL) is the highest level of a daily nutrient intake that is likely to pose no risk of adverse health effects in almost all individuals. The Institute of Medicine's recommended intakes for manganese set ULs at 2, 3, and 6 mg/day for children 1–3, 4–8, and 9–13 years old, respectively. Values were set at 9 mg/day for adolescents 14–18 years old and at 11 mg/day for adults.

Manganese Toxicity

In domestic animals, the major reported lesion associated with chronic manganese toxicity is iron deficiency, resulting from an inhibitory effect of manganese on iron absorption. Additional signs of manganese toxicity in domestic animals include depressed growth, depressed appetite, and altered brain function.

In humans, manganese toxicity represents a serious health hazard, resulting in severe pathologies of the central nervous system. In its most severe form, the toxicosis is manifested by a permanent crippling neurological disorder of the extrapyramidal system, which is similar to Parkinson's disease. In its milder form, the toxicity is expressed by hyperirritability, violent acts, hallucinations, disturbances of libido, and incoordination. The previous symptoms, once established, can persist even after the manganese body burden returns to normal. Although the majority of reported cases of manganese toxicity occur in individuals exposed to high concentrations of airborne manganese ($>5\,mg\,m^{-3}$), subtle signs of manganese toxicity, including delayed reaction time, impaired motor coordination, and impaired memory, have been observed in workers exposed to airborne manganese concentrations less than $1\,mg\,m^{-3}$. Therefore, an inhalation reference concentration range for manganese has been established by the US Environmental Protection Agency to be between 0.09 and $0.2\,\mu g\,m^{-3}$. Manganese toxicity has been reported in individuals who have consumed water containing high levels ($\geq 10\,mg\,Mn\,l^{-1}$) of manganese for long periods of time. Recently, there has been concern that the risk for manganese toxicity may be increasing in some areas because of the use of MMT in gasoline as an antiknock agent, although there is little evidence that air, water, or food manganese concentrations have increased where this fuel is used.

In addition to neural damage, reproductive and immune system dysfunction, nephritis, testicular damage, pancreatitis, lung disease, and hepatic damage can occur with manganese toxicity, but the frequency of these disorders is unknown. Although there is a limited body of epidemiological data that suggests that high levels of manganese can result in an increased risk for colorectal and digestive tract cancers, most investigators do not consider manganese to be a carcinogen. In contrast, both divalent ($MnCl_2$) and heptavalent forms ($KMnO_4$) of manganese are recognized to be strong clastogens both *in vitro* and *in vivo*; exposure to high concentrations of either form results in chromosomal breaks, fragments, and exchanges. High concentrations of manganese can also induce forward and point mutations

in mammalian cells. High levels of dietary manganese have not been reported to be teratogenic in the absence of overt signs of maternal toxicity. However, there are reports that exposure to high levels of manganese during prenatal development can result in behavioral abnormalities. High levels of brain manganese have been reported in subjects with amyotrophic lateral sclerosis, and it has been suggested that this increase may contribute to the progression of the disease. Similar to the cases in humans, chronic manganese toxicity in rhesus monkeys is characterized by muscular weakness, rigidity of the lower limbs, and neuron damage in the substantia nigra. Findings from a recent study suggest that iron and aluminum, which accumulate in the globus pallidus and the substantia nigra of these animals, induce tissue oxidation that may contribute to the damage associated with manganese toxicity. Neural toxicity is a consistent finding in rats exposed to chronic manganese toxicity. Significant manganese accumulation was accompanied by an increase in cholesterol content in the hippocampal region of manganese-treated rats, which was associated with impaired learning; this impairment was corrected by an inhibitor of cholesterol synthesis. The development of manganese toxicity in individuals with compromised liver function, or compromised biliary pathways, is well documented. Significantly, these individuals can have abnormal magnetic resonance imaging (MRI) patterns, which improve following the alleviation of the manganese toxicity. For example, in some cases improvements in brain function have been achieved after liver transplant. The mechanisms underlying the toxicity of manganese have not been agreed upon but may involve multiple etiologies, including endocrinological dysfunction, excessive tissue oxidative damage, manganese-mediated disruptions in intracellular calcium and iron metabolism, and mitochondrial dysfunction caused by manganese inhibition of some pathways of the mitochondrial respiratory chain.

Severe cases of manganese toxicity in humans have been reported for adults, as well as isolated cases in other groups of individuals who are vulnerable, including children on long-term parenteral nutrition and parenteral nutrition patients who have cholestasis or other hepatic disease. In many cases, the previously mentioned groups of individuals have been reported to be characterized by high brain manganese concentrations based on MRI. Although no known cases have been reported, infants may be at a high risk for manganese toxicity due to a high absorptive capacity for the element and/or an immature excretory pathway for it. If manganese is taken up by extrahepatic tissues via the manganese–transferrin complex, the developing brain may be particularly sensitive to manganese toxicity due to the high number of transferrin receptors elaborated by neuronal cells during development, coupled with the putative need by neural cells for transferrin for their differentiation and proliferation. Newborn rats given daily doses of dietary manganese at a level equivalent to that of soy formula exhibited significant neurodevelopmental delays as assessed by several behavioral tests. It should be noted that the concentration of manganese in soy formula is relatively modest but approximately 60–100 times higher than that of breast milk. Brain manganese concentration was increased and striatal dopamine concentrations were significantly decreased even 45 days after the supplementation ended, suggesting that the impact of manganese on the brain and behavior was irreversible. Thus, dietary exposure to high levels of manganese during infancy can be neurotoxic to rat pups and result in developmental deficits. Further studies on human infants fed diets with different levels of manganese are needed to assess whether there are any long-term consequences of early manganese exposure of newborns.

Another group of neuropathological conditions that has been associated with elevated levels of brain manganese is transmissible spongiform encephalopathies. These diseases found in animals and humans are also referred to as prion diseases. There is strong evidence that in their native state, prions are normal brain glycoproteins that bind copper and have an antioxidant function. However, it has been suggested that in the disease process an abnormal isoform of the protein is generated in which manganese is substituted for copper. This isoform is proteinase resistant, no longer has antioxidant activity, and may play a role in the etiology of these diseases. Indeed, elevated levels of brain manganese, along with lower than normal levels of brain copper, have been measured in patients with the prion disease, Creutzfeld–Jakob disease. Whether the elevated levels of brain manganese observed in these patients as well as in animal models of these diseases play an important role in their pathogenesis or are secondary to other factors remains to be determined.

Assessment of Manganese Status

Reliable biomarkers for the assessment of manganese status have not been identified. Whole blood manganese concentrations are reflective of soft tissue manganese levels in rats; however, it is not known whether a similar relationship holds for humans. Plasma manganese concentrations decrease in individuals fed manganese-deficient diets and are

slightly higher than normal in individuals consuming manganese supplements. Lymphocyte MnSOD activity and blood arginase activity are increased in individuals who consume manganese supplements; however, their value as biomarkers for manganese status may be complicated due to the number of cytokines and disease states that may also increase their expression. Urinary manganese excretion has not been found to be sensitive to dietary manganese intake. With respect to the diagnosis of manganese toxicosis, the use of MRI appears to be promising because the images associated with manganese toxicity are relatively specific. Whole blood manganese concentrations can be correlated with MRI intensity and Ti values in the globus pallidus even in the absence of symptoms of neurological damage. Thus, although it is relatively expensive, MRI may be particularly useful as a means of identifying susceptible individuals in, or around, manganese-emitting factories. In addition, the method may be useful in the evaluation of patients with liver failure.

See also: **Carbohydrates**: Regulation of Metabolism. **Cofactors**: Inorganic. **Iron**.

Further Reading

Brown DR (2002) Metal toxicity and therapeutic intervention. *Biochemical Society* **30**: 742–745.

Crossgrove JS, Allen DD, Bukaveckas BL, Rhineheimer SS, and Yokel RA (2003) Manganese distribution across the blood–brain barrier. I. Evidence for carrier-mediated influx of manganese citrate as well as manganese and manganese transferrin. *Neurotoxicology* **24**: 3–13.

Davey CA and Richmond TJ (2002) DNA-dependent divalent cation binding in the nucleosome core particle. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 11169–11174.

Garrick MD, Dolan KG, Horbinski C *et al.* (2003) DMT1: A mammalian transporter for multiple metals. *BioMetals* **16**: 41–54.

Gerber GB, Léonard A, and Hantson PH (2002) Carcinogenicity, mutagenicity and teratogenicity of manganese compounds. *Critical Reviews in Oncology/Hematology* **42**: 25–34.

Guo H, Seixas-Silva JA, Epperly MW *et al.* (2003) Prevention of radiation-induced oral cavity mucositis by plasmid/liposome delivery of the human manganese superoxide dismutase (SOD2) transgene. *Radiation Research* **159**: 361–370.

Kwik-Uribe CL, Reaney S, Zhu Z, and Smith D (2003) Alterations in cellular IRP-dependent iron regulation by *in vitro* manganese exposure in undifferentiated PC12 cells. *Brain Research* **973**: 1–15.

Normandin L and Hazell AS (2002) Manganese neurotoxicity: An update of pathophysiologic mechanisms. *Metabolic Brain Disease* **17**: 375–387.

Sabbatini M, Pisani A, Uccello F *et al.* (2003) Arginase inhibition slows the progression of renal failure in rats with renal ablation. *American Journal of Renal Physiology* **284**: F680–F687.

Salvemini D, Muscoli C, Riley DP, and Cuzzocrea S (2002) Superoxide dismutase mimetics. *Pulmonary Pharmacology and Therapeutics* **15**: 439–447.

Takagi Y, Okada A, Sando K *et al.* (2002) Evaluation of indexes of *in vivo* manganese status and the optimal intravenous dose for adult patients undergoing home parenteral nutrition. *American Journal of Clinical Nutrition* **75**: 112–118.

Takeda A (2003) Manganese action in brain function. *Brain Research Reviews* **41**: 79–87.

Tran T, Chowanadisai W, Crinella FM, Chicz-DeMet A, and Lönnerdal B (2002) Effect of high dietary manganese intake of neonatal rats on tissue mineral accumulation, striatal dopamine levels, and neurodevelopmental status. *Neurotoxicology* **158**: 1–9.

Vahedi-Faridi A, Porta J, and Borgstahl GEO (2002) Improved three-dimensional growth of manganese superoxide dismutase crystals on the International Space Station. *Biological Crystallography* **59**: 385–388.

Yokel RA, Crossgrove JS, and Bukaveckas BL (2003) Manganese distribution across the blood–brain barrier. II. Manganese efflux from the brain does not appear to be carrier mediated. *Neurotoxicology* **24**: 15–22.

MEAL SIZE AND FREQUENCY

F E Leahy, University of Auckland, Auckland, New Zealand

Man's eating habits are changing. Terms such as 'super-sizing,' 'portion distortion,' and 'grazing' have appeared in the contemporary vernacular. Therefore, a better understanding of meal size and frequency is particularly important, especially considering the potential role that these new eating patterns may be playing in the dramatic increase in the incidence of illnesses such as obesity, diabetes, and cardiovascular disease in society.

The principal consequences that changes in meal size and frequency have on the body relate to the absorption and metabolism of food. Several factors in addition to meal size and frequency influence absorption and metabolism, such as the physical characteristics of the food, its macronutrient composition, the energy density of the diet, and the physical volume of the meal. However, the particular contribution that changes in meal size and frequency have made to the dramatic change in

society's eating patterns makes them worthy of special attention.

Effect of Meal Size on Absorption

When a meal of mixed macronutrient composition is consumed, the rate at which the carbohydrate, protein, and fat in that meal is absorbed differs. Carbohydrate in the form of glucose and protein in the form of amino acids enter the portal vein within 30 minutes of meal ingestion and later appear in the general circulation. As the glucose concentration in the portal vein increases, there is an increase in the uptake of glucose into the hepatocytes. Pancreatic islet cells react to the increase in blood glucose and secrete insulin, among other hormones, into the circulation. As a result, there is a decline in the release of nonesterified fatty acids from the adipose tissue. Fatty acid oxidation in the skeletal muscle tissue decreases, and as glucose uptake takes place, the muscle cells increase the rate at which glucose is oxidized. Glycogen synthesis in the muscle and liver cells is increased and the uptake of amino acids by muscle tissue may also occur. Up to 4 h after ingestion of the meal, fat in the form of chylomicron triacylglycerol enters the circulation via the lymphatic system. The action of the hormone lipoprotein lipase in the adipose tissue has by now increased, which promotes the storage of fatty acids as triacylglycerol in adipocytes. This synopsis indicates that following the ingestion of a meal, there is a marked increase in glucose oxidation with a corresponding decrease in fat utilization resulting in the storage of fat.

The larger the meal consumed, the more pronounced are the responses described previously. After a large meal is eaten, the plasma glucose concentration will remain elevated for up to 4 h following ingestion. Conversely, the smaller the meal, the more subtle the effect. This indicates that meal size does indeed influence absorption. However, in order for the relationship between meal size and absorption to be fully understood, the role that absorption plays in determining meal size needs to be considered. The following section focuses on the process of absorption and the systems that control the amount of food eaten.

Regulation of Meal Size by Satiety Peptides and Adiposity Signals

Peptides in the gastrointestinal tract and brain are believed to play an important role in the body's decision to commence and conclude meal consumption. When a meal is being consumed, these peptides are secreted by the gut to indicate the level of satiety. Some of this information can be used by the brain to determine the feeling of fullness, in turn influencing the decision to cease consumption. Cholecystokinin (CCK), a polypeptide located in the peripheral and central nervous systems, is one such satiety signal. It is released in proportion to the amount of food being consumed and helps to determine the amount consumed. Following a meal, CCK is secreted from mucosal epithelial cells in the first segment of the duodenum and stimulates the delivery of digestive enzymes from the pancreas, as well as bile from the gallbladder, into the small intestine. In addition, CCK is produced by neurons in the enteric nervous system and is widely distributed in the brain. The exogenous administration of CCK (and CCK-8, its synthetic analogue) has been shown to influence the amount of food consumed in proportion to the dose given. Although CCK (and other satiety signals) acts to limit meal size, it is important to note that it has little effect on body fat stores, meaning that it does not take into consideration the existing adiposity of the individual when signaling the onset of satiety. Therefore, adiposity signals must be considered in parallel because they also play a part in the process of determining meal size.

Adiposity signals such as leptin act in conjunction with satiety signals in the brain during digestion and their concentration is determined in relation to the degree of adiposity. Like CCK, the effect on meal consumption and body weight of their exogenous administration is dose dependent. Leptin is a peptide hormone produced predominantly by adipocytes, and it is also secreted by the epithelial cells of the stomach. The definitive role of leptin in digestive physiology is still being determined, but it is thought to play a part in limiting food intake in conjunction with CCK. It is when the adiposity signals interact with, and influence, the satiety signals originating from the gut that an attempt at controlling energy intake and meal size is made.

Effect of Meal Size on Metabolism

Energy homeostasis, or the state of balance, achieved by matching energy intake with energy expenditure, is partially dependent on the regulation of meal size consumed. In order for meal size to have an effect on energy metabolism, it must affect either or both components involved in the regulation of energy balance, namely energy intake and energy expenditure. Energy balance is the difference between energy ingested and energy expended over a given period of time. Consequently, energy storage is equal to intake minus expenditure. The following sections examine

the effect of meal size on the two components of the energy balance equation.

Effect of Meal Size on Energy Intake

Meal portion sizes have been increasing steadily since the 1970s, in parallel with the increasing prevalence of obesity in society. It is known that portion and meal sizes vary depending on the food source and location of consumption. Not surprisingly, the largest portions consumed are generally those obtained at fast-food restaurants, although the portion sizes of home-cooked meals have been increasing steadily as well. Meal size may thus be contributing to the problem of obesity by leading to a daily total energy intake that is greater than the daily total energy expenditure, resulting in a positive energy balance.

Effect of Meal Size on Energy Expenditure

Total energy expenditure (EE) can generally be divided into three major components: basal metabolic rate (BMR), thermogenesis, and physical activity (**Table 1**). In order for meal size to have an effect on the EE side of the energy balance equation, it must have an effect on one or more of these components. There is no evidence that meal size has an effect on BMR, which refers to the energy expended to maintain the body on a day-to-day basis. Thermogenesis broadly refers to the body's production of heat, which is divided into three categories: dietary, thermoregulatory, and adaptive. It is the dietary category, commonly known as dietary-induced thermogenesis (DIT), that is of greatest relevance to the discussion of the effect of meal size on energy expenditure. It refers to the heat lost by the body as a result of the absorption and metabolism of a recently ingested meal. DIT represents approximately 10% of energy intake, and therefore the energy expended on DIT increases and decreases in relation to the size of the meal and, more important, the energy value of the meal consumed. The larger the meal, the more energy will be expended to absorb, transport, and metabolize the nutrients consumed during that meal. For example, in the case of a meal containing 2000 kJ (478 kcal) of energy, approximately 200 kJ (48 kcal) will be expended on DIT alone. It is in the physical activity component of energy expenditure that the greatest variation between individuals is observed because physical activity levels (and therefore the energy expended on activity) are contingent on lifestyle choices such as employment and leisure time activities. The effect that meal size may have on physical activity is somewhat difficult to quantify. Meal size is perhaps more important to elite athletes, whose energy expenditure is two or three times greater than that of untrained weight-matched athletes with up to 40% of their energy expenditure being the cost of training.

Table 1 Major components of energy expenditure

Component	*Total energy expenditure (%)*	*Represents*
BMR	60–75	Day-to-day running costs of an individual (e.g., circulation)
Thermogenesis	10–20	Heat produced by the body through dietary, adaptive, and thermoregulatory processes
Physical activity	100 − (BMR + thermogenesis)	The sum of work carried out by an individual

BMR, basal metabolic rate.

Effect of Meal Frequency on Absorption

The perceived health advantages of increased meal frequency (as opposed to eating larger, infrequent meals) have been of interest to researchers since the 1930s. In particular, the benefits of this approach were made apparent by the discovery that insulin requirements in diabetics could be decreased in a frequent meal regime. In a series of case reports on patients taking high insulin doses, it was demonstrated that improved glycemic control and decreased insulin requirements can be achieved when glucose is sipped at hourly intervals throughout the day. Similarly, in healthy individuals a diet composed of many small meals compared with an isoenergetic diet composed of larger meals results in decreased insulin and glucose fluctuations.

Meal frequency not only affects insulin and glucose levels but also influences an individual's circulating lipids. An inverse relationship exists between meal frequency and lipid levels, suggesting that infrequent feeding leads to an increased risk of cardiovascular disease due to large fluctuations in circulating lipids. Increased meal frequency, on the other hand, is associated with several benefits, such as decreased serum cholesterol levels, decreased total:high-density lipoprotein cholesterol ratio, decreased esterified fatty acids, and decreased enzyme levels in adipose tissue associated with fatty acid storage. Paradoxically, individuals who report that they eat more frequently not only have lower total and low-density lipoprotein cholesterol (LDL-C) but also have a greater intake of energy, total fat, and saturated fatty acids. Considering that some of these results were found in a free-living

population, it is possible that dietary misreporting, a common occurrence in overweight populations, may be the cause of this inconsistency.

Mechanisms Underlying the Metabolic Effect of Meal Frequency

The mechanisms underlying beneficial responses to frequent feeding as opposed to an infrequent meal pattern are not fully understood. Frequent feeding has been shown to elicit lower plasma glucose fluctuations than does a more infrequent eating pattern. The absolute amount of carbohydrate eaten at each episode of ingestion in a frequent feeding pattern is simply not great enough to elevate glucose to the same extent as more infrequent eating. Small elevations in plasma insulin seen with frequent feeding are most likely in response to minimal fluctuations in glucose. The mechanisms responsible for the effect of an increased frequency of meal eating on lipid metabolism are not as clear-cut. The lower serum cholesterol levels observed during frequent feeding may be related to lower serum insulin levels. Insulin appears to have a key role in enhancing the hepatic synthesis of cholesterol through its ability to stimulate hydroxymethylglutaryl-coenzyme A reductase (HMG-CoA), the rate-limiting enzyme in hepatic cholesterologenesis. Exogenous insulin quickly increases HMG-CoA reductase activity in rats with diabetes and raises levels of the enzyme in animals without the disorder. It is possible that the reduction of serum cholesterol during a diet of habitual frequent feeding in normal healthy individuals may result from a reduction in hepatic cholesterol synthesis, secondary to the maintenance of euglycemia at lower serum insulin levels. A reduction in cholesterol synthesis would result in an increase in LDL receptors, further lowering total and LDL-C levels.

Alternatively, or in addition, the benefits associated with an increased feeding regimen may reflect unintentional or uncontrolled changes in dietary energy and fat intake that may occur when an individual's meal frequency is altered. It is not clear whether a diet of frequent eating results in any adaptational responses of enzymes or hormones that in turn may be providing additional benefit to the individual.

Much of the research that found these benefits is difficult to interpret due to the variety of methods used, the lack of information available regarding the foods consumed, and the exact nature of the dietary intervention. The majority of measurements are made on fasted blood samples, when in fact most individuals are in a postprandial state for the greater part of every 24-h period. The results of such research must be interpreted with caution for a number of reasons, such as the small sample size used and the interactions with other factors that may prolong absorption time (e.g., soluble fiber, low-glycemic index foods, and the administration of α-glycosides).

As discussed previously, frequent feeding has been demonstrated to lower circulating plasma glucose, insulin, and lipids in both healthy and diabetic subjects in the short term. In addition to the lack of clarity on the mechanisms involved, further research is needed to investigate any medium- and long-term benefits of frequent feeding. It is important that, if deemed desirable in terms of metabolic control, increasing the number of periods of feeding encourages the desired dietary pattern and mix of macro- and micronutrients and is not offset by the failure to decrease meal size.

Effect of Meal Frequency on Metabolism

The maxim that was applied earlier to the study of meal size, namely that it can only influence energy metabolism if it affects energy intake and/or energy expenditure, is applicable to meal frequency. The following sections focus on energy intake and energy expenditure, respectively.

Effect of Meal Frequency on Energy Intake

It has long been argued that the frequency of meal intake may have an effect on body weight regulation. It has been suggested that there is an inverse association between meal frequency and body weight. However, there are a number of flaws in the design of many of the studies from which these data have been derived, and caution is required in the interpretation of the results. Design flaws include (i) dietary underreporting, especially of snacking occasions; (ii) reverse causality, which refers to the possibility that people abstain from eating meals when they become overweight in an attempt to lose weight or to prevent further weight gain; (iii) lack of measurement of physical activity or energy expenditure; and (iv) inclusion of people in a diseased state. These important confounding factors may help to resolve the contradictory results of many research trials. Erroneous conclusions have been drawn from the misinterpretation of such results because these studies are extremely vulnerable to methodological errors that may generate spurious relationships that may not actually exist.

There appears to be very little direct empirical evidence in humans to suggest that frequent feeding per se affects appetite and energy intake. Individuals who eat frequently seem to exhibit a greater capacity to compensate for changes in the energy content

of specific meals relative to individuals who derive most of their energy intake from fewer larger meals. Over very short periods, and under highly controlled experimental conditions, frequent feeding can decrease energy intake at a subsequent meal, which may in turn have an effect on appetite regulation. It remains to be seen, however, whether the same would occur in free-living conditions.

Mechanisms by which meal frequency may influence energy intake Although the evidence is inconclusive, feeding frequency may have an impact on appetite and hence affect energy intake. The control of appetite is very complex and is determined by a number of factors. However, the question remains as to whether the frequency of feeding elicits effects on any of these factors, in turn affecting appetite and possibly body weight.

Frequency of feeding may potentially affect the release of neuroendocrine hormones such as neuropeptide Y, galanin, orexin, and melanocortins from the hypothalamus. The release of such hormones may be either stimulated or suppressed during frequent feeding, leading to either higher or lower than normal hormone levels, which may in turn have knock-on effects on energy intake and/or expenditure. Because it is not feasible to investigate such effects in humans, no studies have been carried out to determine this. The release of gut hormones such as CCK, glucagon-like peptide (GLP), and glucose-dependent insulotropic polypeptide (GIP) may be altered in relation to feeding frequency. In rats, the infusion of the sulfated octapeptide of cholecystokinin (CCK-8) causes a significant reduction in meal size as previously mentioned, whereas meal frequency is increased to compensate for the small meals. However, little is known about the effects of meal pattern on CCK in animals or humans. It is possible that frequent feeding may affect CCK release in one of two ways: It may cause the regular release of the hormone in response to each feed, persistently alerting the brain that the individual is satiated, or CCK may be released into the circulation in such small amounts in response to frequent feeding that it is not recognized by the brain and the individual continues to eat. Similar effects may occur with GLP and GIP.

Effect of Meal Frequency on Total Energy Expenditure

As discussed earlier, the three components of energy expenditure are BMR, thermogenesis, and physical activity. For meal frequency to have an influence on energy expenditure, it must affect one or more of these components. BMR (which represents 60–75% of energy expenditure in sedentary individuals) is not known to be influenced by meal frequency. Much the same can be said for thermogenesis, for which extensive research has failed to demonstrate a link between feeding frequency and DIT. It is reasonable and logical to expect that any difference between frequent and infrequent meal-eating patterns would be seen most clearly during the postprandial period when food has just been eaten, where the rate of ingestion of nutrients may alter EE and fuel storage.

Although much research has been carried out on the effects of meal frequency on total energy expenditure, few studies have isolated the physical activity component per se. Greater attention has been paid to the relationship between meal frequency and physical activity with regard to the performance of elite athletes because the manipulation of the meal pattern can potentially be used as a tool to achieve optimal performance. Because carbohydrate requirements in elite athletes are high and endogenous glycogen reserves are limited, athletes undertaking prolonged strenuous exercise seek to maximize carbohydrate availability at all times.

Irrespective of the above, the key determinant of feeding frequency's overall effect on energy balance is whether it has an impact on 24-h energy expenditure, where energy intake is fixed in content and composition and physical activity is kept constant. Numerous studies have been carried out to investigate this, and all have found that no relationship exists. The majority of these studies used either direct or indirect calorimetry or doubly labeled water in their measurements, both of which are highly reliable energy expenditure measurement techniques.

Conclusion

The contemporary terminology referring to the tendency to increase the amount of food eaten at a meal and the greater frequency at which food is eaten demonstrates the importance of a clear understanding of the consequences of meal size and frequency on health. Satiety peptides and adiposity hormones attempt to control the size of a meal eaten, and increased meal frequency, within the constraints of energy balance, has been found to have beneficial effects attenuating circulating substrates. However, to elucidate the influence that meal size and frequency have on absorption and metabolism, and to clarify whether the increase in the volume of food eaten at a meal and the greater frequency at which food is eaten have a direct affect on health, further research on the free-living population is required.

See also: **Appetite**: Physiological and Neurobiological Aspects; Psychobiological and Behavioral Aspects. **Energy**: Metabolism; Balance; Requirements. **Energy Expenditure**: Indirect Calorimetry. **Weight Management**: Approaches; Weight Maintenance; Weight Cycling.

Further Reading

Bellisle F, McDevitt R, and Prentice AM (1997) Meal frequency and energy balance. *British Journal of Nutrition* **77**: S57–S70.

Blevins JE, Schwartz MW, and Baskin DG (2002) Peptide signals regulating food intake and energy homeostasis. *Canadian Journal of Physiology & Pharmacology* **80**: 396–406.

Blundell JE and King NA (1996) Overconsumption as a cause of weight gain: Behavioral–physiological interactions in the control of food intake (appetite). *Ciba Foundation Symposium* **201**: 138–154; discussion 154–158, 188–193.

Drummond S, Crombie N, and Kirk T (1996) A critique of the effects of snacking on body weight status. *European Journal of Clinical Nutrition* **50**: 779–783.

Frayn KN (1997) In: Keith Snell (ed.) *Metabolic Regulation, A Human Perspective*, 2nd edn. Oxford: Portland Press.

Jenkins DJA (1997) Carbohydrate tolerance and food frequency. *British Journal of Nutrition* **77**: S71–S81.

Mann J (1997) Meal frequency and plasma lipids and lipoproteins. *British Journal of Nutrition* **77**: S83–S90.

Moran TH, Ladenheim EE, and Schwartz GJ (2001) Within-meal gut feedback signalling. *International Journal of Obesity and Related Metabolic Disorders* **25**: S39–S41.

Wilding JP (2002) Neuropeptides and appetite control. *Diabetic Medicine* **19**: 619–627.

Woods SC and Seeley RJ (2000) Adiposity signals and the control of energy homeostasis. *Nutrition* **16**: 894–902.

MEAT, POULTRY AND MEAT PRODUCTS

P A Lofgren, Oak Park, IL, USA

Animal source foods are major contributors to the nutrients in the food supply in many countries. Of these foods, animal muscle (or meat) foods and products are excellent examples of nutrient-dense, or naturally nutrient-rich, foods that provide a relatively large amount of many nutrients per the amount of calories provided in a typical serving. For purposes of this article, discussion is limited to the muscle foods: beef, pork, lamb, veal, poultry, and some of the processed products made from these muscle species.

For meat and meat products there are extensive and comprehensive nutrient databases available for reference for particular products of interest. Thus, this article will provide a sampling of the data available for representative meats and meat products.

One of the best and most comprehensive listings of the nutrient values of all meat, poultry, and other meat products is the nutrient database developed and maintained by the US Department of Agriculture. In this database, complete nutrient profiles are listed for more than 700 beef, 200 pork, 195 lamb, 85 veal, 140 poultry, and 130 turkey products. This database can be accessed and searched on-line at the Web site www.nal.usda.gov/fnic/foodcomp. This database is updated as new data become available for various food products. The most recent version of this database is the *USDA National Nutrient Database for Standard Reference*, Release 17, published in 2004.

For another extensive listing of the nutrient values of many meat and meat products, including some by brand name, the reader is referred to the publication *Bowes & Church's Food Values of Portions Commonly Used* (18th edn.). This reference, although not as extensive in terms of products listed, provides data directly in common serving sizes and provides available data on some additional nutrient and nutrient-related components of meat products (e.g., values for ω-3 and *trans* fatty acids, glutathione, vitamin D activity, and other vitamin-like compounds).

Nutritional Value

The nutritional value of foods, including meat and meat products, can be defined in a number of different ways, from simply listing the quantities of various nutrients contained in the foods to consideration of biological factors that affect the utilization of these nutrients by the body. Some foods may contain nutrients in forms that the body cannot readily utilize. Thus, nutrient bioavailability, or availability, becomes important.

The nutritional value of meat and meat products is related to the quantity and utilization of nutrients and the potential for these products to either enhance or restrict nutrient utilization by the body. There are five major classes of nutrients: protein, lipid, carbohydrate, vitamins, and minerals.

The nutrient content of meat (muscle foods) is fairly similar among the various mammals, birds, and fish. However, differences in the levels of the various nutrients may result from differences in the

carcass composition among species and within species as a result of different fat-to-muscle ratios in the edible portion. As fat percentage increases, nutrient concentration of the muscle portion decreases. Also, to a certain extent, the fat profile/composition and other nutrient content levels may be modified or affected by the animal's diet and/or genetic makeup.

In general, cooking or heat processing has only minimal effects on the nutritional value of muscle foods. In most cases, cooking usually decreases moisture content and concentrates other nutrients, including fat content, especially in lower fat products. This is due to moisture loss. However, in some intensely heated meat products, fat content may also be reduced significantly with negligible loss of other nutrients.

Classes of Nutrients and Meat Products

Protein

Proteins comprise the structural unit of all muscle cells and connective tissue. As such, meat and meat products (muscle foods) are major protein sources. Furthermore, muscle foods, as a group, are excellent sources of high-quality protein that supplies all the essential amino acids in desirable proportions for human consumption. Amino acids are the building blocks of protein, and those provided by meat match or exceed the profile required by humans.

The protein content of most muscle foods, on a wet basis, is between 15 and 35%. This percentage will change due to the moisture and lipid content of the specific product. On a raw weight basis as purchased at a store, the protein content is generally less than 20%. However, people do not eat muscle foods raw, and visible fat in red meat products and skin in poultry products are usually trimmed away. Therefore, muscle foods, as consumed, have a much higher protein content, in the range of 30%.

Lipids

The lipid component of meat and meat products includes a diverse group of substances, such as glycerides (glycerol with fatty acids attached), phospholipids, and sterols. The basic component of most meat lipids is the fatty acids, which can be saturated, monounsaturated, or polyunsaturated.

The relative amount of lipid in muscle foods is probably the most variable aspect of the nutritional profile. Within the lipid component, the relative amount of the different forms of fatty acids present is another variable among meat products. Despite the common reference to animal fats (and especially meat and meat products) as 'saturated,' less than half of all the fatty acids of meats are saturated. The largest proportions of fatty acids in meats are monounsaturated, followed by saturated and then polyunsaturated fatty acids. Among meat products, poultry has a higher proportion of polyunsaturated fatty acids and slightly less saturated fatty acids compared to other meat sources.

The fat in meat products provides much of the flavor associated with these foods and also contributes to the palatability and overall acceptability by consumers.

In addition, the fats in meat and meat products also contain several essential fatty acids (linoleic and linolenic acid), and they contain the fat-soluble vitamins A, D, E, and K.

Carbohydrates

Meat and meat products are not significant sources of dietary carbohydrates. Almost all dietary carbohydrates come from plant sources. The only naturally occurring carbohydrate in muscle foods is glycogen. In some processed meat products, such as those that are 'sugar-cured,' there may be additional sucrose or glucose added.

Vitamins

Meat and meat products are especially good sources of most of the water-soluble vitamins. In general, meat is the major dietary source of vitamin B_{12} and is an excellent source of many of the other B vitamins, such as pyridoxine (B_6), biotin, niacin, pantothenic acid, riboflavin, and thiamin. For vitamin B_{12}, red meat products such as beef and lamb are especially good sources. Pork products are one of the very best sources of thiamin. Although present in muscle foods, the fat-soluble vitamins are less abundant than in plant foods. Vitamins E and K are present, but at lower levels.

Vitamin D activity may be present in some meat products, but at extremely low levels. This is reflected in the USDA nutrient database, in which vitamin D activity is not listed for beef, pork, lamb, veal, and chicken/turkey products; however, it is listed for some processed meat products. In recent years, there has been production research on beef, pork, and lamb to determine if added vitamin D_3 or its metabolites, fed to the animal for a brief period of time prior to slaughter, can result in improved meat tenderness. Although the results are inconsistent, and commercial application is premature, there is some indication that tenderness may be improved with relatively low levels of vitamin D supplementation, which seems to leave very little residual vitamin D_3 or its metabolites in the muscle. Research in Denmark notes that the more biologically active 25-OHD is present at low levels in meat; however, there is no consensus on the conversion factor for 25-OHD to calculate vitamin D

activity. Also, there are very few data on the vitamin D and 25-OHD levels in most meat products. This represents a potential future area of research regarding the nutrient composition of meat.

Minerals

Meat and meat products are good to excellent sources of most minerals. Among the macrominerals, calcium is not high in muscle foods, although phosphorus and potassium are prominent. In natural meat products, sodium is present but not a significant contributor to the diet. However, processed meat products may contain significantly higher levels of sodium (added as part of curing, preserving, or flavor-enhancing ingredients). Some of the microminerals (trace elements) are especially abundant in meat and meat products. Iron is of greatest significance from meat sources because it is present in the 'heme' form, which is more bioavailable than the non-heme form. Of meat products, beef is an especially rich source of iron in this bioavailable form.

Muscle tissue is a very rich source of minerals, such as phosphorus, potassium, magnesium, iron, copper, zinc, and selenium. For instance, pork, poultry, and beef are especially good sources of selenium.

Bioavailability of Nutrients

Muscle foods have been shown to contain 'intrinsic' factors that improve the bioavailability of a variety of nutrients. Moreover, the bioavailability of these nutrients from muscle foods is high, often exceeding the availability for the same nutrients in foods derived from plants. Heme iron is one example. Zinc and copper have been shown to be more available from meat sources than from plant sources. Several of the B vitamins may also be more bioavailable from meat sources than from plant sources.

Another interesting aspect of meat products is their ability to promote the bioavailability of nutrients in nonmuscle foods when the two are eaten together. This has been referred to as the 'meat factor.' Perhaps the best example of this is the positive effect of meat in the diet on non-heme iron sources, also in the diet.

Nutrient Density of Meat and Meat Products

The nutrient density of meat is high. Muscle foods have high levels of essential nutrients per unit of weight and per amount of calories provided. Meat and meat products (muscle foods) provide significant amounts of essential nutrients at levels/concentrations higher than those of most other foods relative to the caloric content provided. The US Food and Drug Administration food labeling guidelines allow a food to be designated a 'good' source of a nutrient if it contributes 10% or more of the Daily Value (DV) and an 'excellent' source if it contributes 20% or more of the DV, for that nutrient, per 3-oz. serving. Most meat products are good or excellent sources of many nutrients. It is generally recognized that in diets that lack muscle foods, greater care is required in diet/menu selection to ensure that adequate levels of essential nutrients are present and bioavailable.

Meat Sources and Nutritional Values

Beef

Beef is an excellent source of high-quality protein, and provides significant contributions of many B vitamins and minerals. In macronutrient terms, the lean-to-fat ratio of the particular beef product influences the calorie and nutrient composition. In general, as the fat content decreases, the concentration of other nutrients (especially protein, B vitamins, and minerals) in beef tends to increase. Most beef products available to the consumer are much leaner than they were 20 or 30 years ago. This is a result of changes in feeding and genetics, producing leaner animals, and also due to closer trim levels on the products that consumers see in the meat case. Whereas in the past, beef cuts with 1/4 in. of fat trim were common, now the same products have only 1/8 in. fat trim or, in some cases, even 0 in. fat trim. In the case of ground beef products, 10 or 20 years ago 17% fat ground beef was considered as 'extra lean.' Today, ground beef is commonly available at fat levels as low as 5 or 10%. Other common fat levels for ground beef are 15, 20, and 25%; however, a large proportion of current ground beef sales are in the 5–15% fat level range.

The fat content of beef contains a varied fatty acid profile, with the largest proportion being contributed by monounsaturated fat, followed by saturated fat and polyunsaturated fatty acids. In addition, because it is a ruminant product, beef is an excellent source of the naturally occurring fatty acid conjugated linoleic acid (CLA), which has been demonstrated to provide anticarcinogenic properties among other health benefits.

Table 1 provides the energy, protein, and lipid profile of beef along with other meat sources. For a comparison of the mineral composition of beef products versus that of other common meat sources, see **Table 2**. For a comparison of the vitamin composition of beef products versus that of other common meat sources, see **Table 3**.

Table 1 Energy, protein, and lipid profile of meats and meat products[a]

Meat species/cut	*Serving size (g)*	*Energy (kcal/kJ)*	*Total protein (g)*	*Total fat (g)*	*Total SFA (g)*	*Total MUFA (g)*	*Total PUFA (g)*	*Total cholesterol (mg)*
Beef								
Composite, Ln 0 in., ckd, all grades	85	179/751	25.4	7.9	3.01	3.32	0.27	73
Top round, Ln 0 in., brld, all grades	85	158/662	27.0	4.8	1.67	2.02	0.18	65
Top loin, Ln 0 in., brld, all grades	85	155/649	24.9	5.4	2.06	2.16	0.20	54
Arm pot roast, Ln 0 in., brsd, all grades	85	173/722	28.4	5.7	2.17	2.44	0.20	57
95% Ln ground beef, brld	85	151/633	22.4	6.2	2.53	2.31	0.28	65
Pork								
Composite, fresh, Ln, ckd	85	180/754	24.9	8.2	2.90	3.70	0.64	73
Tenderloin, fresh, Ln, rstd	85	139/583	23.9	4.1	1.41	1.64	0.35	67
Center loin chop, fresh, Ln, pan-fried	85	197/825	27.4	8.9	3.09	3.78	1.14	78
Shoulder, blade steak, fresh, Ln, brld	85	193/808	22.7	10.7	3.78	4.79	0.92	80
Ham, fresh, Ln, rstd	85	179/751	25.0	8.0	2.80	3.78	0.72	80
Lamb								
Composite, Australian, Ln 1/8 in., ckd	85	171/715	22.7	8.2	3.44	3.28	0.36	74
Loin, Australian, Ln 1/8 in., brld	85	163/683	22.6	7.4	3.13	2.97	0.31	69
Leg, Australian, Ln 1/8 in., rstd	85	162/676	23.2	6.9	2.80	2.81	0.32	76
Foreshank, Australian, Ln 1/8 in., brsd	85	140/586	23.4	4.4	1.60	1.99	0.25	78
Composite, New Zealand, Ln, ckd	85	175/733	25.2	7.5	3.28	2.96	0.44	93
Composite, US domestic, Ln 1/4 in., ckd	85	175/733	24.0	8.1	2.89	3.54	0.53	78
Veal								
Composite, Ln, ckd	85	167/697	27.1	5.6	1.56	2.00	0.50	100
Cutlet, leg top round, Ln, pan-fried	85	156/651	28.2	3.9	1.10	1.40	0.35	91
Loin chops, Ln, rstd	85	149/622	22.4	5.9	2.19	2.12	0.48	90
Shoulder, blade, Ln, brsd	85	168/704	27.8	5.5	1.54	1.96	0.49	134
Chicken/turkey								
Broilers, meat only, rstd	85	162/676	24.6	6.3	1.73	2.26	1.44	76
Broilers, Lt meat only, rstd	85	147/615	26.3	3.8	1.08	1.31	0.83	72
Broilers, Dk meat only, rstd	85	174/729	23.3	8.3	2.26	3.03	1.92	79
Turkey, all classes, meat only, rstd	85	145/604	24.9	4.2	1.39	0.88	1.22	65
Turkey, all classes, Lt meat only, rstd	85	133/558	25.4	2.7	0.88	0.48	0.73	59
Turkey, all classes, Dk meat only, rstd	85	159/665	24.3	6.1	2.06	1.39	1.84	72
Processed meats								
Bacon, pork, cured, pan-fried, 1 slice	7.9	42/176	3.0	3.2	1.05	1.42	0.35	9
Sausage, pork, fresh, ckd, 2 links	48	163/680	9.3	13.6	4.38	5.94	1.79	40
Bologna, beef & pork, low fat, 1 slice	28	64/269	3.2	5.4	2.05	2.56	0.46	11
Salami, beef, ckd, 1 slice	26	67/280	3.3	5.8	2.56	2.77	0.27	18

[a]Amount per 3 oz./85 g, lean only, cooked, except as noted.

Ln, lean and trim level; ckd, cooked; brld, broiled; rstd, roasted; brsd, braised; Lt, light; Dk, dark.

Data from USDA, ARS (2004) *USDA National Nutrient Database for Standard Reference*, Release 17. Nutrient Data Laboratory Web site: www.nal.usda.gov/fnic/foodcomp.

Table 2 Mineral composition of meats and meat products[a]

Meat species/cut	*Serving size (g)*	*Ca (mg)*	*Fe (mg)*	*Mg (mg)*	*P (mg)*	*K (mg)*	*Na (mg)*	*Zn (mg)*	*Cu (mg)*	*Mn (mg)*	*Se (μg)*
Beef											
Composite, Ln 0 in., ckd, all grades	85	7	2.54	22	196	302	56	5.76	0.11	0.02	18.1
Top round, Ln 0 in., brld, all grades	85	6	2.28	18	172	223	36	4.67	0.07	0.01	30.8
Top loin, Ln 0 in., brld, all grades	85	16	1.56	21	195	314	51	4.56	0.07	0.01	28.6
Arm pot roast, Ln 0 in., brsd, all grades	85	14	2.38	19	173	225	46	6.76	0.10	0.01	28.7
95% Ln ground beef, brld	85	6	2.41	19	175	296	55	5.47	0.08	0.01	18.4
Pork											
Composite, fresh, Ln, ckd	85	18	0.94	22	201	319	50	2.52	0.05	0.02	38.2
Tenderloin, fresh, Ln, rstd	85	5	1.25	24	220	371	48	2.24	0.04	0.03	40.9
Center loin chop, fresh, Ln, pan-fried	85	20	0.83	27	230	382	73	2.07	0.07	0.01	40.6
Shoulder, blade steak, fresh, Ln, brld	85	28	1.33	20	187	292	63	4.27	0.05	0.01	33.4
Ham, fresh, Ln, rstd	85	6	0.95	21	239	317	54	2.77	0.09	0.03	42.4
Lamb											
Composite, Australian, Ln 1/8 in., ckd	85	14	1.74	20	176	270	68	4.37	0.13	0.01	9.3
Loin, Australian, Ln 1/8 in., brld	85	18	1.85	22	187	289	68	2.96	0.13	0.01	8.8
Leg, Australian, Ln 1/8 in., rstd	85	8	1.83	21	182	277	61	4.11	0.13	0.01	5.0
Foreshank, Australian, Ln 1/8 in., brsd	85	12	1.62	19	150	217	85	6.74	0.11	0.01	7.7
Composite, New Zealand, Ln, ckd	85	11	2.00	19	209	160	42	3.65	0.10	0.03	1.7
Composite, US domestic, Ln 1/4 in., ckd	85	13	1.74	22	178	292	65	4.48	0.11	0.02	22.2
Veal											
Composite, Ln, ckd	85	20	0.99	24	212	287	76	4.33	0.10	0.03	11.1
Cutlet, leg top round, Ln, pan-fried	85	6	0.74	27	246	376	65	2.87	0.05	0.03	8.8
Loin chops, Ln, rstd	85	18	0.72	22	189	289	82	2.75	0.10	0.03	9.9
Shoulder, blade, Ln, brsd	85	34	1.25	24	214	259	86	6.28	0.15	0.03	12.3
Chicken/turkey											
Broilers, meat only, rstd	85	13	1.03	21	166	207	73	1.78	0.06	0.02	18.7
Broilers, Lt meat only, rstd	85	13	0.90	23	184	210	65	1.05	0.04	0.01	20.7
Broilers, Dk meat only, rstd	85	13	1.13	20	152	204	79	2.38	0.07	0.02	15.3
Turkey, all classes, meat only, rstd	85	21	1.51	22	181	253	60	2.63	0.08	0.02	31.3
Turkey, all classes, Lt meat only, rstd	85	16	1.15	24	186	259	54	1.73	0.04	0.02	27.3
Turkey, all classes, Dk meat only, rstd	85	27	1.98	20	173	247	67	3.79	0.14	0.02	34.8
Processed meats											
Bacon, pork, cured, pan-fried, 1 slice	7.9	1	0.11	3	44	47	192	0.29	0.01	0.00	5.1
Sausage, pork, fresh, ckd, 2 links	48	6	0.65	8	78	141	360	1.00	0.04	0.00	0.0
Bologna, beef & pork, low fat, 1 slice	28	3	0.18	3	51	44	310	0.42	0.02	0.00	3.1
Salami, beef, ckd, 1 slice	26	2	0.57	3	53	49	296	0.46	0.05	0.01	3.8

[a]Amount per 3 oz./85 g, lean only, cooked, except as noted.

Ln, lean and trim level; ckd, cooked; brld, broiled; rstd, roasted; brsd, braised; Lt, light; Dk, dark.

Data from USDA, ARS (2004) *USDA National Nutrient Database for Standard Reference*, Release 17. Nutrient Data Laboratory Web site: www.nal.usda.gov/fnic/foodcomp.

Table 3 Vitamin composition of meats and meat products[a]

Meat species/cut	*Serving size (g)*	*Thiamin (mg)*	*Riboflavin (mg)*	*Niacin (mg)*	*Pantothenic acid (mg)*	*Vitamin B_6 (mg)*	*Folate (μg)*	*Vitamin B_{12} (μg)*	*Vitamin E (mg)*	*Vitamin K (μg)*
Beef										
Composite, Ln 0 in., ckd, all grades	85	0.08	0.20	3.40	0.33	0.30	7	2.64	0.14	1.50
Top round, Ln 0 in., brld, all grades	85	0.06	0.15	4.84	0.53	0.36	9	1.49	0.34	1.30
Top loin, Ln 0 in., brld, all grades	85	0.07	0.13	7.12	0.49	0.53	8	1.39	0.32	1.20
Arm pot roast, Ln 0 in., brsd, all grades	85	0.06	0.18	4.13	0.56	0.27	9	2.07	0.37	1.30
95% Ln ground beef, brld	85	0.04	0.15	5.05	0.55	0.35	6	2.10	0.59	1.10
Pork										
Composite, fresh, Ln, ckd	85	0.72	0.29	4.40	0.58	0.37	1	0.64	0.15	0.00
Tenderloin, fresh, Ln, rstd	85	0.80	0.33	4.00	0.58	0.36	5	0.47	0.17	0.00
Center loin chop, fresh, Ln, pan-fried	85	1.06	0.28	5.10	0.85	0.44	5	0.65	0.19	0.00
Shoulder, blade steak, fresh, Ln, brld	85	0.64	0.37	3.66	0.69	0.26	4	0.96	0.23	0.00
Ham, fresh, Ln, rstd	85	0.59	0.30	4.20	0.57	0.38	10	0.61	0.22	0.00
Lamb										
Composite, Australian, Ln 1/8 in., ckd	85	0.11	0.31	4.94	0.75	0.34	[b]	2.56	[b]	[b]
Loin, Australian, Ln 1/8 in., brld	85	0.15	0.28	6.93	0.71	0.44	[b]	1.71	[b]	[b]
Leg, Australian, Ln 1/8 in., rstd	85	0.12	0.36	4.87	0.84	0.39	[b]	2.71	[b]	[b]
Foreshank, Australian, Ln 1/8 in., brsd	85	0.08	0.24	4.58	0.56	0.22	[b]	2.72	[b]	[b]
Composite, New Zealand, Ln, ckd	85	0.11	0.43	6.53	0.49	0.12	0	2.51	0.16	[b]
Composite, US domestic, Ln 1/4 in., ckd	85	0.09	0.24	5.37	0.59	0.14	20	2.22	0.16	[b]
Veal										
Composite, Ln, ckd	85	0.05	0.29	7.16	1.13	0.28	14	1.40	0.36	5.60
Cutlet, leg top round, Ln, pan-fried	85	0.06	0.32	10.74	1.04	0.43	14	1.28	0.36	4.20
Loin chops, Ln, rstd	85	0.05	0.26	8.04	1.08	0.32	14	1.11	0.42	4.70
Shoulder, blade, Ln, brsd	85	0.05	0.31	4.83	1.35	0.21	13	1.71	0.38	5.80
Chicken/turkey										
Broilers, meat only, rstd	85	0.06	0.15	7.80	0.94	0.40	5	0.28	0.23	2.00
Broilers, Lt meat only, rstd	85	0.06	0.10	10.56	0.83	0.51	3	0.29	0.23	0.30
Broilers, Dk meat only, rstd	85	0.06	0.19	5.57	1.03	0.31	7	0.27	0.23	3.30
Turkey, all classes, meat only, rstd	85	0.05	0.16	4.63	0.80	0.39	6	0.31	0.28	3.10
Turkey, all classes, Lt meat only, rstd	85	0.05	0.11	5.81	0.58	0.46	5	0.31	0.08	0.00
Turkey, all classes, Dk meat only, rstd	85	0.05	0.21	3.10	1.09	0.31	8	0.31	0.54	3.30
Processed meats										
Bacon, pork, cured, pan-fried, 1 slice	7.9	0.04	0.02	0.91	0.10	0.03	0	0.10	0.02	0.00
Sausage, pork, fresh, ckd, 2 links	48	0.14	0.10	3.00	0.35	0.16	1	0.57	0.26	0.20
Bologna, beef & pork, low fat, 1 slice	28	0.05	0.04	0.71		0.05	1	0.37	0.06	0.10
Salami, beef, ckd, 1 slice	26	0.03	0.05	0.84	0.25	0.05	1	0.80	0.05	0.30

[a]Amount per 3 oz./85 g, lean only, cooked, except as noted.
[b]Comparable data not available.
Ln, lean and trim level; ckd, cooked; brld, broiled; rstd, roasted; brsd, braised; Lt, light; Dk, dark.
USDA, ARS (2004) *USDA National Nutrient Database for Standard Reference*, Release 17. Nutrient Data Laboratory Web site: www.nal.usda.gov/fnic/foodcomp.

Pork

Pork, like beef, is an excellent source of high-quality protein and contributes significant amounts of many B vitamins and minerals. As for other muscle foods, pork's nutrient composition is greatly affected by its fat and water content. As fat percentage decreases, the concentration of other nutrients increases. In addition, as pork is cooked and moisture is removed, the concentration of nutrients also increases. Pork is an excellent source of minerals, such as selenium, iron, zinc, phosphorus, and potassium. Compared to other muscle foods, the contribution of pork to selenium in the food supply is especially significant.

Pork is an excellent source of the B vitamins. Pork is an especially good source of thiamin (vitamin B_1), being the single best source of this vitamin among commonly eaten foods. The fat profile of pork can be influenced by feeding regimes such that it is more or less saturated or firm. However, overall the fatty acid profile of pork is largely monounsaturated, followed by saturated and then polyunsaturated fatty acids.

Lamb

Although it represents a smaller portion of overall muscle food consumption, lamb still provides a nutrient profile with significant benefits for the human diet. In addition to being a source of high-quality protein, lamb is also a good source of many minerals and B vitamins. Vitamin B_{12} is especially abundant in lamb. It is also a good source of the minerals iron and zinc.

In addition, as a ruminant, lamb is another naturally occurring dietary source of CLA, a unique fatty acid with anticarcinogenic and other health benefits (from animal model studies).

Veal

Although representing a smaller proportion of overall meat consumption, veal still provides a nutrient profile that is very beneficial. As with all meat sources, veal provides high-quality protein in a product that may be slightly leaner (in terms of fat) than other red meat sources. Compared to other meat sources, veal has a lower iron content.

Poultry

The nutrient composition of poultry (chicken and turkey) is similar to that of red meat animals with a few exceptions. Poultry is lower in iron content, and thus heme iron, than beef. Turkey is slightly higher in several minerals (Ca, Fe, P, K, Zn, and Cu) than chicken. As in red meats, there are significant amounts of several B vitamins (e.g., niacin, B_6, and pantothenic acid) compared to other meat sources, and these are not significantly reduced during cooking.

The fat content of poultry is predominantly monounsaturated fat, followed by saturated fat and polyunsaturated fat. Poultry fat, like pork fat, is somewhat more unsaturated than beef fat. Poultry is significantly higher in polyunsaturated fat compared to beef, pork, lamb, and veal.

Processed Meats

Processed meats represent a diverse array of products that have undergone additional treatment from the fresh meat form to the point of consumption, including curing with other ingredients added and the addition of salt or other flavor or preservative mixtures. Also, these products often represent combined meat sources.

Summary

Muscle foods provide significant amounts of essential nutrients at levels/concentrations higher than those of most other foods relative to the caloric content provided. Almost all of the essential nutrients are present in muscle foods at some level. Furthermore, muscle foods provide nutrients in a form that enhances the bioavailability of nutrients from both the meat and other dietary sources. It is generally recognized that in diets that lack muscle foods, greater care is required in diet/menu selection to ensure that adequate levels of essential nutrients are present and bioavailable.

See also: **Amino Acids**: Chemistry and Classification; Metabolism; Specific Functions. **Bioavailability**. **Biotin**. **Carbohydrates**: Chemistry and Classification; Regulation of Metabolism; Requirements and Dietary Importance; Resistant Starch and Oligosaccharides. **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Copper**. **Dairy Products**. **Dietary Surveys**. **Eggs**. **Energy**: Balance; Requirements; Adaptation. **Fats and Oils**. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Fish**. **Folic Acid**. **Food Composition Data**. **Fruits and Vegetables**. **Iron**. **Lipids**: Chemistry and Classification; Composition and Role of Phospholipids. **Magnesium**. **Manganese**. **Niacin**. **Nuts and Seeds**. **Pantothenic Acid**. **Phosphorus**. **Potassium**. **Protein**: Synthesis and Turnover; Requirements and Role in Diet; Digestion and Bioavailability; Quality and Sources; Deficiency.

Riboflavin. **Selenium**. **Sodium**: Physiology. **Thiamin**: Physiology. **Ultratrace Elements**. **Vegetarian Diets**. **Vitamin A**: Biochemistry and Physiological Role. **Vitamin B_6**. **Vitamin E**: Metabolism and Requirements. **Vitamin K**. **Zinc**: Physiology.

Further Reading

Council for Agricultural Science and Technology (1997) *Contribution of Animal Products to Healthful Diets*, Task Force Report No. 131. Ames, IA: CAST.

Council for Agricultural Science and Technology (1999) *Animal Agriculture and Global Food Supply*, Task Force Report No. 135. Ames, IA: CAST.

Foote MR, Horst RL, Huff-Lonergan EJ *et al.* (2004) The use of vitamin D_3 and its metabolites to improve beef tenderness. *Journal of Animal Science* **82**(1): 242–249.

Godber JS (1994) Nutritional value of muscle foods. In: Kinsman DM, Kotula AW, and Breidenstein BC (eds.) *Muscle Foods—Meat, Poultry and Seafood Technology*, pp. 430–455. New York: Chapman & Hall.

Ovesen L, Brot C, and Jakobsen J (2003) Food contents and biological activity of 25-hydroxyvitamin D: A vitamin D metabolite to be reckoned with? *Annals of Nutrition and Metabolism* **47**(3–4): 107–113.

Pennington JAT and Douglass JS (2004) *Bowes & Church's Food Values of Portions Commonly Used*, 18th ed. Baltimore: Lippincott Williams & Wilkins.

US Department of Agriculture, Agricultural Research Service (2004) *USDA National Nutrient Database for Standard Reference*, Release 17. Nutrient Data Laboratory Web site: http://www.nal.usda.gov/fnic/foodcomp.

Menkes Syndrome *see* **Copper**

MICROBIOTA OF THE INTESTINE

Contents

Prebiotics

J M Saavedra and N Moore, John Hopkins School of Medicine, Baltimore, MD, USA

Introduction

The gastrointestinal (GI) system in humans comprises the largest surface area of any organ in the body. The complexity of this system and its functions provides us with the ability to take in nutrition, selectively process it, assist in maintaining fluid and electrolyte balance, and offers a vehicle for excretion of waste while at the same time offering the first line of defense against toxins, pathogens, and other noxious agents. The indigenous gut microflora make up the complex ecosystem that inhabits the GI lumen, which mediates part of the interaction between the external environment and the host.

The basic development and makeup of the human intestinal microflora, and the metabolic, immune, and functional effects of the host are discussed below. The importance of maintaining a balance in this ecosystem, and the recent use of nutrition for providing beneficial microflora and the clinical effect this offers will be presented.

Normal Microflora

The intestinal microflora of healthy humans is comprised of more than 400 species of bacteria with a population of 10^{12}–10^{14} colony-forming units (CFU) per gram, of which more than 98% are resident in the colon. This bacterial population nearly exceeds the population of cells in the human body. The microflora is composed of both aerobic and

predominantly anaerobic microorganisms that when equilibrium within an individual is maintained confer nutritional and immune benefits. A prime example of the importance of microorganisms in the GI tract was the study of gnotobiotic (germ-free) mice, which suffered persistent enteritis and severe infections with poor survival rate. Through the interaction of the mucosal surface with the GI tract microflora an important system of immune defense is established.

The presence of microorganisms in different segments of the GI tract varies both qualitatively and quantitatively. Bacteria from the mouth are predominantly anaerobes including streptococci, *Bacteroides*, *Lactobacillus*, and some yeasts; these wash down to the stomach with the intake of food and function of swallowing. In the stomach the acid environment destroys most of the oral and food-ingested microorganisms. The microflora of the stomach is comprised of mostly Gram-positive and aerobic microflora at very low levels (10^3 CFU ml^{-1}). *Peptostreptococcus*, *Fusobacterium*, and *Bacteroides* species are present in low numbers while *Clostridium* is uncommon.

The volume of microflora increases exponentially from the small intestine, which is sparsely colonized, to the richly populated colon. The concentrations of bacteria found in the small intestine are between 10^3 and 10^4 CFU ml^{-1}, again both facultative anaerobes and aerobic bacteria with almost complete absence of coliforms and *Bacteriodes*.

The microflora of the colon dramatically increases to a concentration of 10^{11}–10^{12} CFU gm^{-1}. This bacterial load accounts for up to 50% of the volume of colonic content. Although the colonic microflora comprises more than 400 different species it is predominantly anaerobic including *Bacteroides*, *Fusobacterium*, *Bifidobacterium*, *Lactobacillus*, *Enterobacter* and coliforms, and other facultative anaerobes (*Staphlococcus* and *Candida* species).

Development of Microflora

The GI tract is essentially sterile at the time of birth and bacterial colonization begins upon exposure to the environment. Progression of colonization is initially fast, followed by a gradual process of modification over the first few years of life. As the baby passes through the birth canal bifidobacteria and lactobacilli are typically acquired and rapid colonization of mainly enterobacteria occurs. The hospital environment, type of feeding, and type of delivery affect the early colonization of the intestine after birth. Normal vaginal birth permits the transfer of bacteria of the mother as the infant passes through the birth canal. However, with Cesarean delivery this transfer is absent and the hospital or other immediate environment can have a more significant effect on colonization. In these infants, colonization with anaerobic bacteria, especially *Bacteroides*, occurs later than with vaginally delivered infants.

Within the first few days and with introduction of feeding, the newborn intestine (through oxidation–reduction) promotes the establishment of aerobic bacteria, predominantly enterobacteria, *Enterococcus* and staphylococci, and anaerobic bacteria, bifidobacteria, *Bacteroides* and *Clostridia*. As the aerobic bacteria consume oxygen the intestinal milieu becomes more amenable to anaerobic bacteria and aerobic bacteria in turn decline. In breast fed infants bifidobacteria counts increase dramatically and account for 80–90% of the total fecal flora. Lactobacilli and *Bacteroides* also increase but to a lesser extent, while enterobacteria decrease. In formula fed babies *Enterococcus* is the predominant bacteria present with significantly less bifidobacteria and *Bacteroides* than the breast-fed infant. It is the difference in microflora, especially in the greater presence of bifidobacteria, and the presence of oligosaccharide and other bifido-genic factors in breast milk that likely confer a protective effect to the infant against infection, particularly against diarrheal disease.

With the introduction of weaning foods the fecal flora of babies begins to change resembling that of adults by 1 year of age. Concentrations of aerobes decrease and anaerobes (streptococci, *Enterobacter*, *Escherichia coli*, *Bacteroides*, and *Lactobacillus*) increase and predominate by 1–2 years of age. Bifidobacteria concentration also decreases but is generally maintained throughout adulthood (**Figure 1**).

Once well established the microflora is unique to each individual and maintained fairly undisturbed throughout adult life. Changes in general health

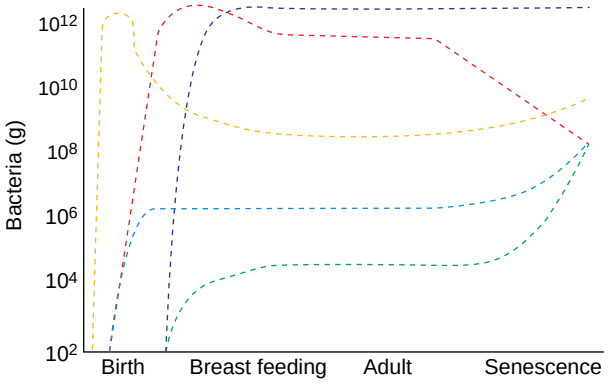


Figure 1 The intestinal flora and its relation to age. (purple), *Bacteroides*, Eubacteriae, Peptococcaceae; (red), Bifidobacteriae; (orange), *E. coli*, streptococci; (blue), lactobacilli; (green), Clostridiae.

and wellbeing, exposure to toxins in the food supply, and utilization of medications, particularly antibiotics, all transiently alter the colonic flora, often profoundly. When the equilibrium of this complex system occurs the host is potentially compromised. However, recovery to the original state of colonization usually occurs upon removal of the altering factors.

Metabolic Activity of Microflora

The structure and function of the gastrointestinal tract is influenced greatly by the presence and make-up of indigenous microflora. In the germ-free murine model, lack of microflora leads to a thinner, less cellular intestinal wall; the villi are thinner, crypts shallower and mucosal surface area is decreased, thus interfering with the gut immune system and nutrient handling processes. The intestinal flora is responsible for production of some micronutrients, particularly vitamins (biotin, folate, and vitamin K), and also fermentation of carbohydrates, which results in the production of short-chain fatty acids (acetate, proprionate, and butyrate). These end products are known to be active in the regeneration and health of the mucosal cellular make up. Glycosylation of complex carbohydrates on the microvillus membrane are specifically related to anaerobic bacteria in the gut lumen. Additionally, the microflora modulates the release of peptides and some proteins from the endocrine cells in the mucosa of the GI tract.

Innate bacteria induce many beneficial responses of the gut immune system. Bacterial interaction with the epithelial cells can enhance local immunity and deter response to antigens. Gram-negative bacteria cause the production of proinflammatory cytokines including interleukin-6 and tumor necrosis factor alpha (TNF-α).

Gut Barrier Effect

Health of intestinal mucosa and the equilibrium of the microflora are essential for the immune response of the GI tract to both ingested and systemic invaders. The colonic microflora plays a key role in maintaining mucosal integrity and the deterrence of pathogenic/toxic bacteria. To maintain the bacterial equilibrium the indigenous bacteria appear to compete with pathogenic bacteria predominantly in the colon for enterocyte receptors, as well as for luminal nutrient. Additionally, the production of short-chain fatty acids from bacterial fermentation of carbohydrates and the transformation of proteins to sterols and bile acids destroy potential mutagens of ingested foods. Dietary antigens are prevented through this system of barrier protection. Alterations will produce an immunoinflammatory response; ultimately entrance of antigens into the body can occur, inducing a systemic allergic response.

Flora-Nutrient Interactions

There is a complex interaction between food and microflora in a feedback-like system. Different types of diets can lead to changes in fecal flora, and its resultant metabolic activity can be altered. When individuals consuming a vegetarian diet were compared to those on a typical Western diet, the latter had microflora that showed greater hydrolyzing ability leading to a more effective metabolism of bile acids and subsequently reduced cholesterol. Similar studies in mice have shown differences with high-fat diets versus low-fat diets.

Disaccharides are broken down in the proximal small intestine by enzymes in the brush border and microvilli of the intestinal epithelium. Glucose, sucrose, lactose, and maltose are the predominant disaccharides hydrolyzed and these rarely reach the colon. When the brush border is unable to produces the enzymes needed for metabolism, the disaccharides are not absorbed in the small bowel and ultimately reach the colon where they interact with the abundant colonic bacteria. Subsequent fermentation causes an osmotic imbalance pulling water into the lumen and causing diarrhea. Significant and rapid production of short-chain fatty acids causes changes in the fecal pH and can irritate the colonic mucosa.

Complex carbohydrates, such as dietary fibers, are predominantly fermented in the colon by colonic bacteria, mostly anaerobic flora. Short-chain fatty acids, acetate, propionate, and butyrate are the predominant by-products and in lesser amounts carbon dioxide, hydrogen, methane, and water. Slow and regular production of short-chain fatty acids provides an energy source that helps the regeneration of colonic mucosa. Fatty acids are also used for hepatic very low density lipoprotein synthesis, which has been reported to have influence in cardiovascular disease.

Dietary protein is only partially digested in the small intestine producing amino acid (NH_2) and carboxyl groups (COOH). These amines enter the colon and are hydrolyzed by enzymes of colonic bacteria. Amines and short-chain fatty acids will enter the systemic system through absorption by the colonic mucosa and portal vein into the bloodstream where they will be appropriately utilized by tissues. These substances then return to the liver through the portal circulation and are excreted as urea in the urine.

Primary bile acids originating from the liver are excreted into the small intestine and conjugate with amino acids, particularly taurine and glycine. Conjugated compounds in general are not well absorbed and cannot re-enter the hepatic circulation without further breakdown. Bacterial action hydrolyzes the conjugated amines, releasing free bile acids 7α- and 7β-dehydroxylation of the bile acid nucleus, and hydroxyl groups C-$_3$, C-$_6$, and C-$_7$. *Bacteroides, Bifidobacterium, Fusobacterium, Clostridium, Lactobacillus,* and *Streptococcus* are the main bacteria that assist in this hydrolysis. These free bile acids can recirculate through the enterohepatic circulation. Bile acids assist in digestion of fats in the intestine. Colonic microflora also transform excess cholesterol found in the large intestine to coprostanol thus reducing available cholesterol and increasing cholesterol to be excreted in stool.

Bacterial microflora of both the small and large bowel synthesizes a number of essential vitamins. Most importantly vitamin K production by the liver is dependent on the metabolic activity of bacteria in the ileum. Prothrombin, a blood-clotting factor, is synthesized in the liver. Glycoprotein arising from the prothrombin complex cannot be synthesized unless the liver contains menaquione. Bacteria in the intestine synthesize menaquione at the terminal ileum where it can be absorbed and reach the liver to promote clotting factors.

Vitamin B_{12} is completely synthesized from microflora in animals. Meats and dairy products from these animals is a primary source of B_{12} for humans, but it is also synthesized in the large bowel. However, the small bowel is the site of optimal absorption of B_{12} so synthesized B_{12} is not well absorbed. Additionally, biotin and other B complex vitamins (folic acid and thiamine) are synthesized by GI microflora.

Microflora and Host Interactions

The immune response within the GI tract is both innate and adaptive. The innate immune system is a pre-existing system that begins to eliminate invading pathogenic microorganisms immediately upon exposure. Natural barriers of the mucosal epithelium begin this immune response. Rapid induction of an immune response occurs with initial inflammation through phagocytosis. Neutrophils and macrophages engulf bacteria in an effort to get rid of them before insult to the epithelium occurs. Phagocytes also release important chemokines and cytokines that increase the inflammatory response activating the adaptive immune mechanisms when necessary.

The adaptive immune mechanisms are able to differentiate indigenous microflora and mount a response to pathogenic microbes. This process involves cells of the gut-associated lymphoid tissue (GALT) resulting in production of IgA. The adaptive branch of the GI immune system is antigen specific allowing a 'memory' of such and responding specifically to re-exposure to offensive bacteria or toxins. It is through this delicate interplay between innate and adaptive immune mechanisms that an adequate immunologic defense response can be maintained while the adaptive immune system is activated in the hope of averting a harmful systemic reaction.

Mucus/Mucin Glycoproteins

Mucus is continuously produced by goblet cells to lubricate and protect the GI epithelium. The primary gene identified that is located in the goblet cell and predominantly responsible for the production and secretion of mucus and its resulting sugar, mucin, is the MUC_2 gene. This is through an elaborate process of encoding a peptide modified by *o*-glycosidic bonds to a variety of carbohydrate residues to amino acids serine or threonine resulting in a glycoprotein with high carbohydrate content that provides the potential to bind sites for both indigenous and pathogenic bacteria. Mucin, the resulting glycoprotein, forms a viscous gel that coats the epithelial surface of the intestine protecting it from chemical and mechanical stress. Coating of the epithelia thus denies pathogenic bacteria the opportunity to adhere preventing an inflammatory response. This is the first line of defense of the intestine against pathogenic microbes. Indigenous bacteria also utilize the carbohydrate component of mucins as fuel, encouraging the growth of health-promoting bacteria (particularly anaerobes).

Colonization Resistance

Varying levels of bacteria throughout the GI tract have inherent benefit to the function of each portion of the GI tract. Thus, selective discouragement of colonization is necessary. For example, the low number of bacteria in the small bowel allows the function of nutrient breakdown and absorption. Intrinsically, the small bowel limits the levels of bacteria through antegrade peristalsis, and bactericidal action of the gastric acid and biliary enzymes of the liver. The ileocecal valve at the terminal end of the small bowel functions as a gate deterring the entrance of colonic bacteria into the small bowel. Presence of higher concentrations of colonic bacteria causes mucosal inflammation and villous atrophy ultimately interfering in its function.

Bacterial overgrowth syndrome is due to anatomical and physiologic alterations of the small bowel

causing proliferation of bacteria in the upper GI tract. Conditions causing hypochlorhydria (decreased secretion of hydrochloric acid) such as gastritis, drug therapy, and dysmotility contribute to bacterial overgrowth. Surgical or anatomical malformations resulting in ineffective peristalsis or absence of the ileocecal valve also contribute to this syndrome. Impaired micelle formation causes fat malabsorption and steatorrhea. Higher levels of free bile acids in the proximal portion of the small bowel bind with vitamin B_{12} thus preventing absorption in the terminal end. Additionally, amino acid and carbohydrate malabsorption occurs leading to increased fecal nitrogen, lower serum proteins, and ultimately protein calorie malnutrition.

The Gut-Associated Lymphoid Tissue (GALT)

The complex function of the gut-associated lymphoid tissue (GALT) is the critical protective immune system in the GI tract. Peyer's patches are cells found in the mucosa and submucosa of the small intestine and contain CD4, CD8 T cells and B cells. M cells that overlay the epithelium transport antigens to the Peyer's patches that initiate the adaptive immune response. Production of secretory IgA occurs and other immune cells then enter systemically through the Peyer's patches and into the mesenteric lymph system. IgA cells prevent pathogens from adhering to the intestinal surface thus preventing gut cell damage.

Altering Gut Flora

The concept of manipulating microflora to enhance the positive aspects of the GI tract has become a more focused endeavor. However, this concept is not new. The early recognition of fermented foods offering health benefits dates back to the early 1900s. Eli Metchnikoff was the first to recognize this benefit when he observed the long lives and good health of Bulgarian peasants and associated this with the large amounts of milk soured with lactic acid bacteria (LAB) they consumed.

Since then much study of the health benefits from introduction of orally supplemented beneficial bacteria has taken place. This concept has been termed probiotics and is defined as the consumption of microbes that confer a positive effect on the host in prevention and treatment of specific pathologic conditions. Bifidobacteria, *Lactobacillus*, and *Streptococcus thermophilus* have been the most recognized and studied probiotics because of their ability to survive the upper GI tract and proliferate, although transiently, in the colon. The purported health benefits of these and other probiotics include prevention and treatment of diarrhea (particularly rotaviral and antibiotic associated), improved lactose digestion, enhanced gut immune function, and, most recently, prevention and treatment of food allergy and its systemic effects (atopic dermatitis and possibly gastrointestinal allergic disease). Use of probiotics to beneficially alter flora composition and its effects will be elaborated on in a separate chapter.

The effects of probiotics on the host are transient and without regular consumption of these products the colon cannot maintain the level of beneficial colonization that connfers the health benefits. Therefore, a key to the probiotic effect and possible enhancement of native colonic flora would be a substrate for gut bacterial growth through fermentation. Certain dietary carbohydrates and fibers that escape digestion in the upper GI tract are ideal for this action. This recent concept involving such carbohydrates is termed prebiotics.

Prebiotics: Definition and Uses

Definition

A prebiotic is generally accepted as a nondigestible food ingredient that selectively stimulates the growth and/or activity of native bacteria in the colon to beneficially affect the host. This generally, but not always, implies:

- a 'natural' food component;
- ability to by-pass the upper GI tract (not digested);
- ability to be selectively fermented by 'beneficial,' nonpathogenic colonic bacteria;
- ability to modify the established microflora; and
- ability to confer an advantageous physiologic activity to the host.

Classifications

Various food components have been recognized to have prebiotic activity, including various fermentable carbohydrates (lactulose, gums, lactilol, soyoligosaccharides, galacto-oligosaccharides Xylo-oligosaccharides. However, the best studied of these have been those classified as dietary fructans. Dietary fructans can either be derived from naturally occurring oligosaccharides or can be artificially synthesized. These carbohydrates contain one or more fructosyl-fructose links that make up the majority of osidic bonds. They are linear or branched fructose (oligo)polymers with either β-2-1 linked inulins or β-2-6 linked levans. These oligosaccharides exist naturally in many plants including onions, garlic, the roots of Jerusalem artichoke,

asparagus root, chicory root, and wheat (**Table 1**). Inulin is extractable from root plants particularly Jerusalem artichoke and chicory, while fructooligosaccharide is hydrolyzed from inulin yielding a shorter chain sugar. It is the degree of polymerization (DP) that distinguishes the fructans. Fructooligosaccharides are β-D-fructans with DP between 2 and 10 while inulin has DP 10–60. Essentially, they are sucrose molecules with 1–3 fructose units linked by a β-(2,1)-glycosidic bond. Most oligosaccharides are synthesized from sucrose and therefore usually have a terminal glucose end. Inulin, derived from chicory, is broken down using an inulase enzyme making a smaller (2–10) chain with lower DP (4). Oligofructose is a form synthesized from sucrose by β-fructofuranosidase linking fructose monomers to sucrose.

Both inulin-derived and synthesized fructooligosaccharides have been shown to resist digestion in the upper GI tract. Ninety per cent of consumed inulin and fructooligosaccharide was excreted at the terminal ileum of adult ileostomy patients. Furthermore, the undigested oligosaccharides are not recovered in the fecal mass indicating they are completely fermented in the colon. In many ways, prebiotics behave as a form of dietary fiber that has specific effects on colonic flora.

Table 1 Fructo-oligosaccharide (FOS) content of common fruits, vegetables, and grains

Food type	*FOS concentration (mg gm^{-1})*
Fruits	
Apples	0.1
Banana	0.1
Banana, ripe	2.0
Blackberry	0.2
Orange, navel	0.3
Peach	0.4
Raspberry, red	0.2
Vegetables	
Acorn squash	0.4
Artichoke, globe	2.4
Artichoke, Jerusalem	58.4
Chicory root, raw	3.9
Garlic	3.9
Onion, red	1.4
Onion, white	3.1
Onion powder	45.0
Peas, snap	1.1
Peas, snow	0.6
Shallot	8.5
Grains	
Barley	1.7
Oats	0.3
Rye	3.8
Wheat	1.3
Wheat bran	3.5
Wheat germ	4.2

Adapted from Campbell J, Bauer L, Fahey G, Hogarth AJCL, Wolf B, and Hunter D (1997) *Journal of Agricultural and Food Chemistry* **45**: 3076–3082.

The mother's milk is a key factor in the early establishment of the infant's colonic environment. Up to 10% of the carbohydrates in human milk are not lactose, and human milk contains high concentrations of other carbohydrates and glycoconjugates that fall under the general category of prebiotic food substances. The monosaccharides of breast milk include D-glucose, D-glactose, sialic acid, L-fructose, and N-acetylglucosamine. Chain lengths range from three to ten with the majority having a lactose end. Combinations of these monosaccharides result in more than 130 varieties of oligosaccharides in human milk. These galacto-oligosaccharides in breast milk have lactose as their reducing end. Many human milk oligosaccharides elongate by enzymatic attachment of *N*-acetylglucosamine linked to a galactose residue. Several of these carbohydrates, including *N*-acetylglucosamine, are considered 'bifidus factors' or 'bifidogenic,' increasing the growth and establishment of bifidobacteria in the intestine of the breast fed infant. Human milk oligosaccharides also appear to prevent attachment of pathogenic microorganisms by competing with epithelial ligands for bacterial binding sites. Several types of human milk oligosaccharides appear to be bacteria specific. For example, sialyated oligosaccharides inhibit attachment of *Pneumococci* and influenza viruses, while galacto-oligosaccharides and fructosylated oligosaccharides can inhibit *E.coli* attachment. The bifidogenic effects, as well as those of direct interaction with the intestinal mucosa, are considered to be some of the mechanisms by which these agents confer a protective effect on the lactating infant.

Oligosaccharide content of the breast milk varies among individuals and within an individual. Levels are highest in the newborn period peaking after 5 days and slowly declining through the first 3 months. The levels of oligosaccharide in the breast milk also are dependent on time of feeding and generally are higher at the beginning of the feed.

Other, less well-studied oligosaccharides including maltose, soya, and xylose-oligosaccharides have some effect on increasing microbe colonization; however, they are weak prebiotics because of the lack of specificity of their fermentation.

Clinical Effects of Prebiotics

Average consumption of dietary fructans as part of a normal diet has been estimated to be 1–4 g day^{-1} in

the US. Europeans tend to have a higher intake ranging from 3 to 10 g day^{-1}. Many products worldwide are produced with supplemental oligosaccharides. Owing to the nondigestible nature of dietary fructans, the nutritional value in terms of calories and energy is negligible. The actual energy produced by these carbohydrates relates to the by-products of fermentation, specifically short-chain fatty acids (SCFA) and lactate.

Effect in the Upper GI Tract

From the dietary point of view, oligosaccharides meet the criteria to be considered a dietary fiber. Fibers are categorized as soluble, insoluble, or mixed. Definition of dietary fiber has focused on biochemical attributes and physiologic effects. Insoluble fibers (nonfermentable) decrease colonic transit time and increase fecal volume thus acting as a bulking agent.

Oligosaccharides because of their fermentable nature are considered a soluble fiber. Their effect on the upper GI tract is to slow down gastric and small bowel transit time, thereby altering glucose metabolism and increasing sensitivity to insulin. Altered fat metabolism by the binding of bile acids thus decreasing serum cholesterol and triglyceride levels has been reported with oligosaccharide supplementation in hypercholestrolemic patients.

There is a strong link between oligosaccharide consumption and the integrity of the GI mucosa. A trophic effect of the mucosa and hyperplasia of the epithelial cells occurs from a hormonal response to dietary fructans although the mechanism it not clear. Adequate or improved trophism of the intestinal wall may increase the absorption capacity for such minerals as calcium, magnesium, iron, copper, and zinc. Of particular interest is the effect of oligosaccharides on calcium absorption. Recent studies have demonstrated increased calcium absorption in teenage girls consuming a prebiotic mixture. Although there are not enough studies as yet to determine what compounds and at what doses connfer this health benefit, it is proposed that the short-chain fatty acids produced from fermentation lower fecal pH and increase colonic absorption of calcium.

Effects in the Colon

The main effect of inulin and fructooligosaccharide in the colon is directly related to fermentation. The process of fermentation from innate bacteria produces short-chain fatty acids and lactate. Increase in biomass contributes to the bulking effect that oligosaccharides have on stool. Additionally, fecal pH is decreased due to suppression of the production of putrefactive substances.

Carbon dioxide and hydrogen are produced in this process contributing to disagreeable side effects when given in high doses. Abdominal cramping, increased flatulence, and bloating have been shown to occur significantly more in studies where adults received 15 g day^{-1} or more of fructooligosaccharide and inulin as compared to a placebo group. However, in a limited number of controlled pediatric studies these symptoms were not seen at doses of up to 3 g day^{-1}.

The greatest value of inulin and fructooligosaccharide is their role in stimulating the growth of innate microbes in the colon. Inulin and fructooligosaccharide selectively promote proliferation of bifidobacteria and *Bacteriodes*. In adult studies fructooligosaccharide and inulin given in doses of 10 g day^{-1} resulted in increased levels of bifidobacteria and decreased levels of enterobacteria and enterococci without GI side effects. In establishing a predominant microbial environment of bifidobacteria, epithelial adherence of pathogenic bacteria is deterred.

It is generally assumed that the immunologic effects seen with probiotic consumption (bifidobacteria, lactobacilli) would apply with the altered microbial balance with prebiotics. However, there is a lack of well-designed trials to support this. Limited animal studies of prebiotics have shown increased lymphocytes in the GALT and peripheral blood, although any impact on the host has not been addressed.

Conclusion

It is clear that the intestinal ecosystem of organisms in humans play a critical role in the development and health maintenance of the human intestine. The intestinal flora can be modified in a positive way via dietary means. Further studies should help define future dietary recommendations in support of improvement in gastrointestinal and immunologic function.

See also: **Biotin**. **Breast Feeding**. **Carbohydrates**: Requirements and Dietary Importance; Resistant Starch and Oligosaccharides. **Colon**: Structure and Function; Disorders. **Dietary Fiber**: Potential Role in Etiology of Disease. **Folic Acid**. **Lactose Intolerance**. **Microbiota of the Intestine**: Probiotics. **Thiamin**: Physiology. **Vitamin K**.

Further Reading

Campbell JM, Bauer LL, Fahey GC Jr, Hogarth AJCI, Wolf BW, and Hunter DE (1997) Selected fructooligosaccharide (1-kestose, nystose and 1 ^{F}-β-Fructofuranosylnystose) composition

of foods and feeds. *Journal of Agricultural and Food Chemistry* **45**: 3076–3082.
Day A and Sherman PM (1998) Normal intestinal flora: pathobiology and clinical relevance. *International Seminars in Pediatric Gastroenterology and Nutrition* **7**(3): 2–7.
Gibson GR (1999) Dietary modulation of the human gut microflora using the prebiotics oligofructose and inulin. *Journal of Nutrition* **129**: 1438S–1441S.
Goldin BR, Lichtenstein AH, and Gorbach SL (1994) Nutritional and metabolic roles of intestinal flora. In: Shills ME, Olson JA, and Shike M (eds.) *Modern Nutrition in Health and Disease*, 8th edn. Malvern, PA: Lea & Febiger.
Hentges D (ed.) (1983) *Human Intestinal Microflora in Health and Disease*. New York: Academic Press.
Kunz C, Rudloff S, Baier W, Klein N, and Strobel S (2000) Oligosaccharides in human milk: structural, functional, and metabolic aspects. *Annual Review of Nutrition* **20**: 699–722.
Mahida YR (ed.) (2001) *Immunological Aspects of Gastroenterology*. Dordrecht: Kluwer Academic Publishers.
Roberfroid MB (1997) Health benefits of non-digestible oligosaccharides. In Kritchevsky D and Bonfield (eds.) *Dietary Fiber in Health and Disease*. New York: Plenum Press.
Roberfroid MB and Delzenne NM (1998) Dietary fructans. *Annual Review of Nutrition* **18**: 117–143.
Simon GL and Gorbach SL (1984) Intestinal flora in health and disease. *Gastroenterology* **86**(1): 174–193.

Probiotics

M Gueimonde and S Salminen, University of Turku, Turku, Finland

Introduction

The human gastrointestinal (GI) tract harbors a complex collection of microorganisms. The individual digestive system contains about 1.5 kg of viable (live) bacteria, made up of more than 500 different identified microbial species. Indeed, the total number of bacteria in the gut amounts for more than 10 times that of eukaryotic cells in the human body, and this bacterial biomass can constitute up to 60% of fecal weight. This complex microbiological community is called the intestinal microflora. While most people are familiar with the side-effects of some members of it (e.g., diarrhea), the beneficial effects in stabilizing gut well-being and general health are less well known. These so-called 'friendly' bacteria are naturally present in the GI tract as part of the normal healthy intestinal microflora and ensure the balance that creates a healthy individual. Such beneficial microbes and a healthy intestinal microflora also constitute the main source of probiotics used to improve intestinal and host health.

Fermented products containing living microorganisms have been used for centuries to restore gut health. Such utilization of live microorganisms to improve host health forms the basis of the probiotic concept.

Usually probiotics are taken in the form of dairy products, drinks, or supplements, but in African countries they have traditionally also been ingested in fermented cereal and in fermented vegetables in Asian countries. The claimed benefits of traditional fermented foods range from treatment of diarrheal diseases to alleviation of the side-effects of antibiotics to the prevention of a number of other health problems. In some countries fermented foods have even been associated with benefits to the skin.

Definition of Probiotics

Probiotics have been defined as 'bacterial preparations that impart clinically verified beneficial effects on the health of the host when consumed orally.' According to this definition the safety and efficacy of probiotics must be scientifically demonstrated. However, as different probiotics may interact with the host in different manners, their properties and characteristics should be well defined. It is understood that probiotic strains, independent of genera and species, are unique and that the properties and human health effects of each strain must be assessed in a case-by-case manner. Most probiotics are currently either lactic acid bacteria or bifidobacteria, but new species and genera are being assessed for future use. The probiotic bacteria in current use have been isolated from the intestinal microflora of healthy human subjects of long-standing good health and thus most of them are also members of the healthy intestinal microflora.

It has been demonstrated that probiotics have specific properties and targets in the human intestinal tract and that they are able to modulate the intestinal microflora.

Intestinal Microflora

Composition of the Intestinal Microflora

The human GI tract hosts a rich and complex microflora that is specific for each person depending on environmental and genetic factors. Different bacterial groups and levels are found throughout the GI tract, as corresponds with the different ecological niches present from mouth to colon. The stomach and the upper bowel are sparsely populated regions (10^3–10^4 CFU per g contents) while the colon is heavily populated (10^{11}–10^{12} CFU g contents). In

the small intestine genera such as *Lactobacillus* and *Bacteriodes* are usually found, whereas those considered predominant in the large bowel include *Bacteriodes*, *Bifidobacterium*, *Eubacterium*, *Clostridium*, *Fusobacterium*, and *Ruminococcus* among others. Several health-promoting properties have been attributed to defined members of the intestinal microflora such as lactobacilli and bifidobacteria. A balanced microflora provides a barrier against harmful food components and pathogenic bacteria and has a direct impact on the morphology of the gut. Hence, the intestinal microflora constitutes an important factor for the health and well-being of the human host and a healthy stable microflora affords a potential source of future probiotics.

Development and Succession of Microflora during Life-Time

The human fetus is sterile and the maternal vaginal microflora comprises the first inoculum of microbes. The indigenous intestinal microflora develops over time, determined by an interplay between genetic factors, mode of delivery, contact with the initial surrounding environment, diet, and disease. As a result, every individual has a unique characteristic microflora. The human intestinal microbiota does not exist as a defined entity; this population comprises a dynamic mixture of microbes in each individual.

The establishment of the gut microflora, a process commencing immediately upon birth, provides an early and massive source of microbial stimuli, and may consequently be a good candidate 'infection.' This step-wise succession begins with facultative anaerobes such as the enterobacteria, coliforms, and lactobacilli first colonizing the intestine, rapidly succeeded by bifidobacteria and lactic acid bacteria. The indigenous gut microflora plays an important role in the generation of an immunophysiological regulation of the gut, providing key signals for the development of the immune system in infancy and also interfering with and actively controlling the gut-associated immunological homeostasis later in life. A healthy microflora can thus be defined as the normal individual microflora of a child that both preserves and promotes well-being and absence of disease, especially in the GI tract, but also beyond it. It provides the first step in long-term well-being for later life and the basis for this development lies in early infancy. Failure in the establishment of a healthy microflora has been linked to the risk of infectious, inflammatory, and allergic diseases later in life. Demonstration of this has stimulated researchers to elucidate the composition and function of the intestinal microflora.

Microflora Research

In spite of the recent development of DNA based methods, microbiota development and characterization in the human host still rests largely on the culture-based assessment pioneered by Japanese researchers. The identification of different microbial species and strains has been dependent on microbial characterization, which is usually based on limited phenotypic properties and the metabolic activity of the microbes, for example, sugar fermentation profiles. There are several bacteria, however, that cannot be cultured and isolated or identified by the traditional methods. The culture technique as used in microbial assessments of feces is also hindered by the fact that microbes in the feces will mainly represent the microflora in the lumen of the sigmoid colon, while the composition of the intestinal microflora differs both along the GI tract and between the lumen and the mucosa. For more accurate information on the population elsewhere in the intestine, samples should be taken by endoscopy or during surgery. Most of our current data on microflora are derived from results obtained from fecal samples and culturing. These data indicate that there are several successive phases in microflora development related to age (**Figure 1**). In early infancy the microflora is scant and simple consisting mainly of bifidobacteria. During breast-feeding it remains so, but following weaning its complexity increases, reaching the state observed in adults where the microflora is specific to each person. Aging is related to further changes and the diversity is again decreased. The microflora becomes more unstable and vulnerable to diseases, for example, diarrheal diseases caused by intestinal pathogens.

Current research efforts focus on revealing genomic data on both probiotic microorganisms and certain important intestinal commensals. This has

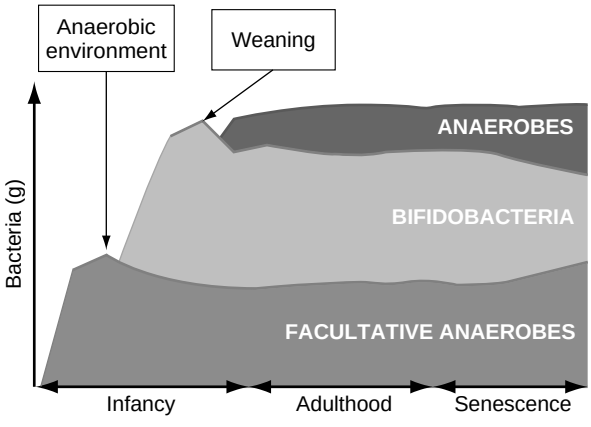


Figure 1 Development of microflora throughout life.

provided information indicating that gut commensals not only derive food and other benefits from the intestinal contents but also have a role in influencing the human host by providing maturational signals for the developing infant and child and providing later signals for alteration to gut barrier mechanisms.

The genomic data on, for instance, *Bifidobacterium longum* and *Bacteroides thetaiotaomicron*, both important members of the human intestinal microflora, give an indication as to how specific bacteria are adapted to the development of the gut by specific genes enabling the use of intestinal mucins and breast milk oligosaccharides as main sources or nutrients.

Genomic information on *B. longum* also gives insight into the adhesive mechanisms that comprise a basis both for populating the infant gut and for communicating developmental signals to specific areas and sites of the gut mucosa. Furthermore, a large part (>8.5%) of the *B. longum* genome is devoted to carbohydrate transport and metabolism, indicating a versatile metabolism well adapted to life in the intestine and making it very different from, for instance, *Lactobacillus johnsonii*.

Bacteroides thetaiotaomicron has also been shown to modulate glycosylation of the intestinal mucus and to induce expression of angiogenins, revealing proposed mechanisms whereby intestinal microbes may influence the gut microecology and shape the immune system. Incorporating such information with host gene expression data from the exposed mucosal sites and beyond them will enable us to understand the role of both microbial transfer and succession and microbe–microbe and host–microbe interactions. Recent information demonstrates that the vast community of indigenous microbes colonizing the human gut also shapes our development and biology.

Role of Microflora in Health and Disease

Major dysfunctions of the GI tract are thought to be related to disturbances or aberrationss of the intestinal microflora. Recent findings confirm that aberrations can be documented and related to disease risk. The microorganisms present in our GI tract thus have a significant influence on our health and well-being.

The development of the intestinal microbiota needs to be characterized to define the composition that helps us to remain healthy. Specific aberrations in the intestinal microflora may predispose to disease. Such aberrations have been identified in allergic disease, including decreased numbers of bifidobacteria and an atypical composition of bifidobacterial microflora. Also, aberrations in *Clostridium* content and composition have been reported to be important. Similar predisposing factors may also exist in the case of microflora and both inflammatory gut diseases and rotavirus diarrhea. Microflora aberrations have also been reported in rheumatoid arthritis, juvenile chronic arthritis, ankylosing spondylitis, and irritable bowel syndrome patients. A thorough knowledge of the intestinal microflora composition will offer a basis for future probiotic development and the search for new strains for human use. Many diseases and their prevention can be linked to the microflora in the gut.

Modulation by Probiotics

In general, probiotic bacteria do not colonize the human intestinal tract permanently, but specific strains are able to transiently colonize or persist for some time in the intestine and may modulate the indigenous microflora. The rationale for modulating the gut microflora by means of probiotics derives from the demonstration that this microflora is important to the health of the host. Specific probiotics have been shown to colonize temporarily the human intestinal tract, thereby modulating the intestinal microflora both locally and at the commensal level. Such modification has not been reported to be permanent; rather it is related to a balancing of aberrant or disturbed microflora to assist it to return to normal metabolic and physiological activities. Such modulation and restoration of the normal state of the microflora activity is a key target for probiotic action. However, the state of the microflora should be well characterized to enable the selection of specific probiotics to counteract the aberration or disturbance in question.

Specific probiotic bacteria can modulate both the intestinal microflora and local and systemic immune responses. Activation of immunological cells and tissues requires close contact of the probiotic with the immune cells and tissue on the intestinal surface. Interestingly, both lactobacilli and bifidobacteria, which colonize mainly the small and large intestine respectively, when given as probiotic supplements were able to modify immunological reactions related to allergic inflammation, whereas lactobacilli were ineffective in protection against cows' milk allergy. In this respect, preferential binding of probiotics on the specific antigen-processing cells (macrophages, dendritic, and epithelial cells) may be even more important than the location of adhesion. It is also known that the cytokine stimulation profiles of

different *Bifidobacterium* strains vary and that strains isolated from healthy infants stimulate mainly noninflammatory cytokines.

Results of an increasing number of clinical and experimental studies demonstrate the importance of constituents within the intestinal lumen, in particular the resident microflora, in regulating inflammatory responses. Probiotic bacteria may counteract inflammatory processes by stabilizing the disturbed gut microbial environment, forming a stable healthy microflora and thus improving the intestine's permeability barrier. Another mode of action comprises enhancing the degradation of enteral antigens and altering their immunogenicity. Yet another mechanism for the gut-stabilizing effect could be improvement of the intestine's immunological barrier, particularly intestinal IgA responses. Probiotic effects may also be mediated via control of the balance between pro- and anti-inflammatory cytokines. Such effects may be mediated through changes in the intestinal microflora, especially by modulation of the bifidobacteria microflora.

Importance of Understanding Intestinal Microflora

It is obvious that an understanding of the cross-talk that occurs between the intestinal microflora and its host promises to expand our conceptions of the relationship between the intestinal microflora and health. There is also an increasing amount of information indicating that specific aberrations in the intestinal microflora may render us more vulnerable to intestinal inflammatory diseases and other diseases beyond the intestinal environment. It is likely that some aberrations may even predispose us to specific diseases. Unfortunately, however, we are still far from knowing the qualitative and quantitative composition of the intestinal microflora and the factors governing its composition in an individual.

Probiotic Effects

Living microorganisms have long been used as supplements to restore gut health at times of dysfunction. It is clear that different strains from a given microbial group may possess different properties. It is thus important to establish which specific microbial strain may have a beneficial effect on the host; even closely related strains can have significantly different or even counteracting effects. Their properties and characteristics should thus be well defined; studies using closely related strains cannot be extrapolated to support each other.

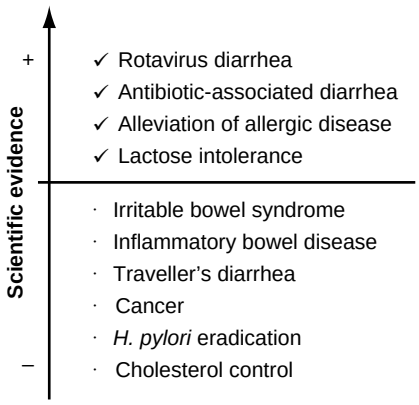


Figure 2 Health benefits of probiotics.

Working hypotheses can be supported by studies carried out *in vitro* using cell culture models or *in vivo* using animal models. However, the studies most important for efficacy assessment are carefully planned and monitored clinical studies in humans.

In summary, well-designed human studies are required to demonstrate health benefits. Using the criteria thus obtained it can be concluded that certain specific probiotics have scientifically proven benefits that can be attributed to specific products (see below). Other reported probiotic health-related effects are only partially established (**Figure 2**), and require more data from larger double-blind placebo controlled studies before firm conclusions can be reached.

Scientifically Documented Effects

Diarrhea The mechanisms by which probiotics prevent or ameliorate diarrhea may involve stimulation of the immune system, competition for binding sites on intestinal epithelial cells (**Figure 3**), or the elaboration of bacteriocins or binding of virus particles in the gut contents. These and other mechanisms are thought to be dependent on the

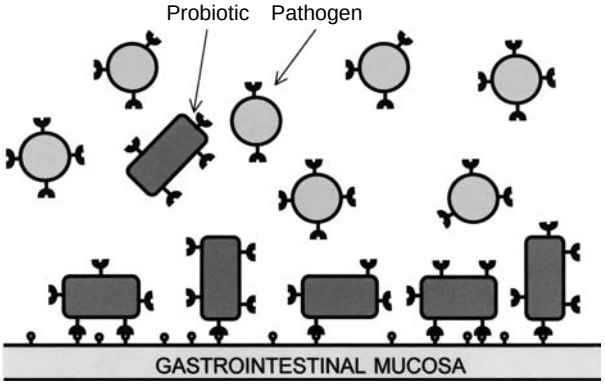


Figure 3 Probiotic adhesion and replacement of pathogenic bacteria.

type of diarrhea being investigated, and may therefore differ between viral diarrhea, antibiotic-associated diarrhea, or traveller's diarrhea.

Viral diarrhea Shortening of the duration of rotavirus diarrhea using *Lactobacillus* GG (LGG) is perhaps the best-documented probiotic effect. A reduction in the duration of diarrhea was first shown in several studies around the world and also in a recent multicenter European study on the use of LGG in acute diarrhea. Other investigators demonstrated that supplementation with a combination of *Bifidobacterium bifidum* and *Streptococcus thermophilus* reduces the incidence of diarrhea and shortens the duration of rotavirus shedding in chronically hospitalized children. On average, the duration of diarrhea was shortened by 1 day in both hospitalized children and those treated at home.

Other investigators have studied the immune modulating effects of probiotics as a means of reducing diarrhea, suggesting that the humoral immune system plays a significant role in the probiotics' effect.

From these numerous studies it is clear that probiotics do indeed play a therapeutic role in viral diarrhea. Even meta-analyses have been conducted in this area, showing that probiotic therapy shortens the duration of acute diarrhea in children. However, the exact mechanism of action involved is not clear and is very likely multifactorial.

Antibiotic-associated diarrhea The incidence of antibiotic-associated diarrhea is between 5 and 30%. The success of probiotics in reducing or preventing this form of diarrhea has been convincing, and includes a number of probiotics as well as various antibiotics.

LGG has been shown to prevent antibiotic-associated diarrhea when consumed in both yogurt form or as a freeze-dried product. Also, *Saccharomyces boulardii* has been found to be effective in preventing antibiotic-associated diarrhea. Other microorganisms such as *Enterococcus faecium* or a combination of *L. acidophilus* and *L. bulgaricus* have also been reported to be effective.

Alleviation of symptoms of allergic disease It has been shown that changes in intestinal microflora composition precede the development of some allergic diseases, indicating a potential area for probiotic application. LGG given prenatally to mothers and during the first months to infants with a high risk of atopic disease has reduced the prevalence of atopic eczema to about half in the infants receiving the strain. Furthermore, extensively hydrolyzed whey formula supplemented with LGG or *Bifidobacterium lactis* Bb12 is more effective than unsupplemented formula in eczema alleviation in infants with atopic eczema.

These results indicate a high potential for probiotic application in the treatment and reduction of risk of allergic diseases.

Lactose intolerance Several studies have shown that lactose-intolerant individuals suffer fewer symptoms if milk in the diet is replaced with fermented dairy products. The mechanisms of action of lactic acid bacteria and fermented dairy products include the following: lower lactose concentration in the fermented product due to lactose hydrolysis during fermentation; high lactase activity of bacterial preparations used in production; and increased active lactase enzyme entering the small intestine with the fermented product or within the viable bacteria.

The bacterial enzyme beta-galactosidase, which can be detected in the duodenum and terminal ileum after consumption of viable yogurt, is thought to be the major factor improving digestibility by the hydrolysis of lactose, mainly in the terminal ileum. Another factor suggested to influence lactose digestion is the slower gastric emptying of semisolid milk products such as yogurt.

In conclusion, there is good scientific evidence to demonstrate the alleviation of lactose intolerance symptoms by specific probiotic lactic acid bacteria. However, the strain-specific lactase activities may vary from nil to very high values. Thus, different products may have varying lactose contents and individual strains, when released into the duodenum, vary in their lactase activity.

Potential Effects Requiring Further Clinical Work

Intestinal microecology and cancer A number of studies have focused on the impact of probiotics on intestinal microecology and cancer. *Lactobacillus acidophilus, L. casei* Shirota strain, and LGG have been shown to have inhibitory effects on chemically induced tumors in animals. Some specific strains of probiotic bacteria are able to bind carcinogens and to downregulate some microbial carcinogenic enzymatic activities. This phenomenon may then reduce carcinogen production and exert a beneficial effect in the colon, the urinary tract, and the bladder.

The most interesting documentation is that concerning *L. casei* Shirota. There have been several mechanistic studies on the effects of the strain reporting decreased mutagen excretion, and some human clinical studies have been conducted using this strain. In clinical and multicenter studies carried

out in Japan, prophylactic effects of oral administration of *L. casei* Shirota on the recurrence of superficial bladder cancer have been reported. Recently, a large Japanese case control study has been conducted on the habitual intake of lactic acid bacteria and risk reduction of bladder cancer. Results suggested that the habitual intake of fermented milk with the strain reduces the risk of bladder cancer in the Japanese population. More studies, and especially human studies in other countries, are needed prior to the establishment of firm conclusions.

Irritable bowel syndrome There is a rationale for investigating the effect of probiotics in the treatment of this common disorder where intestinal motility and dysfunctions in the intestinal microflora are important factors to consider. In a recent study using *L. plantarum* 299v, a reduction of symptoms was reported. *Enterococcus faecium* preparations have also been evaluated for the treatment of patients with irritable bowel syndrome, and although patient-recorded symptoms did not show significant differences, the physician's subjective clinical evaluation revealed an improvement.

Inflammatory bowel disease Inflammatory bowel disease (IBD) comprises a heterogeneous group of diseases of unknown etiology (Crohn's, ulcerative colitis, and pouchitis), but here also factors related to the intestinal microflora seem to be involved, providing a rationale for the application of probiotics. From reviewing studies on the use of probiotics in IBD it can be concluded that, although there are some promising preliminary findings, more well-planned long-term studies are needed before any firm conclusions can be drawn.

Traveller's diarrhea There are a few studies on the prevention of traveller's diarrhea using probiotics and these show a positive outcome for LGG and a combination of *L. acidophilus* LA5 with *B. lactis* Bb-12. The results offer some indication of beneficial effects, even though some studies yielded no reported effects, but information from good and extensive human studies using defined strains for traveller's diarrhea is still largely lacking. The current data on traveller's diarrhea show no scientifically proven effects for any of the strains used. More studies are required for efficacy assessment.

***Helicobacter pylori* eradication** Specific strains of lactic acid bacteria have been reported to inhibit a wide range of intestinal pathogens including *Helicobacter pylori*, which is involved in the process of gastric ulcer development. Lactic acid bacteria are often able to survive acidic gastric conditions and it has therefore been proposed that they may have a beneficial influence during the eradication of *H. pylori*. It has been reported that both the inhibitory substances produced and the specific strains may influence the survival of *Helicobacter*, and studies have been conducted, particularly with a *L. johnsonii* strain. It has been shown that there is good *in vitro* inhibition and that fermented milk containing the strain has a positive effect when consumed during *Helicobacter* eradication therapy. However, more controlled human studies in different populations need be conducted to verify this effect.

Cholesterol control The cholesterol-lowering effects of probiotics have been the subject of two recent reviews with contradictory results. The first, which focused on short-term intervention studies with one yogurt type, reported a 4% decrease in total cholesterol and a 5% decrease in LDL. Contrary to this, the second review concluded that no proven effects could be found. In this context, it is clear that long-term studies are required before the establishment of any conclusion.

Safety

Safety assessment is an essential phase in the development of any new food. Although few probiotic strains or prebiotic compounds have been specifically tested for safety, the long history of safe consumption of some probiotic strains could be considered the best proof of their safety. Although some lactobacilli and bifidobacteria have been associated with rare cases of bacteremia, usually in patients with severe underlying diseases, the safety of members of these genera is generally recognized due to their long history of safe use and their lack of toxicity. Furthermore, the low incidence of infections attributable to these microorganisms, together with a recent study showing that there is no increase in the incidence of bacteremia due to lactobacilli in Finland despite the increased consumption of probiotic lactobacilli, supports this hypothesis. With regard to other bacteria such as enterococci, *S. boulardii*, *Clostridium butyricum,* or some members of the genus *Bacillus* the situation is more complicated, even though they have been used as probiotics for some time.

In addition to the possibility of infection there are other risks that must be taken into account (**Table 1**). These include those risks associated with the metabolic properties of the strain (capacity for deconjugation/dehydroxylation of bile salts,

Table 1 Probiotic action: potential benefits and risks

Action mechanisms	*Potential risks*
Improvement of gut barrier (immunologic, nonimmunologic)	Proinflammatory effects
Modulation of aberrant gut microbiota	Adverse effects on innate immunity
Modulation of inflammatory response	Infection
Degradation of antigens	Production of harmful substances
Binding/inhibition of carcinogens	Antibiotic resistance (Specific risks related to host, strain characteristics, or interactions)

production of enzymes favoring the invasion/translocation through the epithelium, etc.), with the presence of active substances in the probiotic or product (immunoactive substances, toxic compounds, etc.), or with antibiotic resistance. It is clear that strains harboring transferable antibiotic resistance genes should not be used. In this context the specific risks related to each probiotic strain must be carefully identified.

Guidelines are needed to test the safety of probiotics. However, taking into account the great diversity of probiotic microorganisms, it is necessary to identify the specific risks associated with the respective strains, as well as the risk factors associated with the host and the possible interactions between probiotic–host–food components in order to assess the safety of these products. Additional epidemiological surveillance and follow-up of novel strains should be conducted. In this context, the specific risks related to each probiotic strain must be carefully identified. With regard to this, knowledge of mechanisms involved is a key factor not only for the assessment of health effects but also for the safety aspects of probiotics.

Future Challenges

Some of the claimed beneficial effects of probiotics are backed by good clinical studies. However, other possible effects call for further investigation in new, well-planned, long-term human clinical studies prior to any firm conclusions being made. Protocols for human studies need to be developed for probiotics. In some cases, even postmarketing surveillance studies on intakes and long-term effects are useful; such studies have in fact already been used for the safety assessment of current probiotics.

The assessment of potential probiotic strains must be based on a valid scientific hypothesis with realistic studies supporting it. In this respect, knowledge of mechanisms of action is a key factor for hypothesis formulation and for the selection of biomarkers appropriate to the specific state of health and well-being or reduction of risk of disease. It is thus important to improve our knowledge of the mechanisms involved and take into account the fact that probiotic mechanisms of action are multifactorial and that each probiotic may have specific functions affecting the host.

It is also of key interest to increase our knowledge of intestinal microflora composition and to understand its role in health and disease, identifying those microorganisms related to the health status of the host, in order to select probiotic strains able to modulate the intestinal microflora in a beneficial manner.

Knowledge accrued regarding the intestinal microflora, nutrition, immunity, mechanisms of action and specific diseases should be carefully combined with genomic data to allow the development of a second generation of probiotics; strains for both site- and disease-specific action.

See also: **Breast Feeding**. **Cancer**: Epidemiology of Gastrointestinal Cancers Other Than Colorectal Cancers. **Cholesterol**: Sources, Absorption, Function and Metabolism. **Colon**: Disorders. **Diarrheal Diseases**. **Food Allergies**: Etiology; Diagnosis and Management. **Lactose Intolerance**. **Microbiota of the Intestine**: Prebiotics.

Further Reading

Benno Y and Mitsuoka T (1986) Development of intestinal microflora in humans and animals. *Bifidobacteria Microflora* **5**: 13–25.

Dai D and Walker WA (1999) Protective nutrients and bacterial colonization in the immature human gut. *Advances in Pediatrics* **46**: 353–382.

De Roos N and Katan M (2000) Effects of probiotic bacteria on diarrhea, lipid metabolism, and carcinogenesis: a review of papers published between 1988 and 1998. *American Journal of Clinical Nutrition* **71**: 405–411.

Falk PG, Hooper LV, Midvedt T, and Gordon JI (1998) Creating and maintaining the gastrointestinal ecosystem: What we know and need to know from gnotobiology. *Microbiology and Molecular Biology Reviews* **62**: 1157–1170.

Guandalini S, Pensabene L, Zikri M, Dias J, Casali L, Hoekstra H, Kolacek S, Massar K, Micetic-Turk D *et al.* (2000) *Lactobacillus* GG administered in an oral rehydration solution to children with acute diarrhea: a multicenter European trial. *Journal of Pediatric Gastroenterology and Nutrition* **30**: 54–60.

Guarner F and Malagelada JR (2003) Gut flora in health and disease. *Lancet* **381**: 512–519.

Gueimonde M, Ouwehand AC, and Salminen S (2004) Safety of probiotics. *Scandinavian Journal of Nutrition* **48**: 42–48.

He F, Morita H, Hashimoto H, Hosoda M, Kurisaki J, Ouwehand AC, Isolauri E, Benno Y, and Salminen S (2002) Intestinal Bifidobacterium species induce varying cytokine production. *Journal of Allergy and Clinical Immunology* **109**: 1035–1036.

Isolauri E, Kirjavainen PV, and Salminen S (2002) Probiotics: a role in the treatment of intestinal infection and inflammation? *Gut* **50**(Suppl. 3): iii54–iii59.

Isolauri E, Salminen S, and Ouwehand AC (2004) Probiotics. *Best Practice and Research Clinical Gastroenterology* **18**: 299–313.

Jonkers D and Stockbrügger R (2003) Probiotics and inflammatory bowel disease. *Journal of the Royal Society of Medicine* **96**: 167–171.

Kalliomäki M, Kirjavainen P, Eerola E, Kero P, Salminen S, and Isolauri E (2001) Distinct patterns of neonatal gut microflora in infants developing or not developing atopy. *Journal of Allergy and Clinical Immunology* **107**: 129–134.

Kalliomäki M, Salminen S, Arvilommi H, Kero P, Koskinen P, and Isolauri E (2001) Probiotics in the prevention of atopic diseases: a randomised placebo-controlled trial. *Lancet* **357**: 1076–1079.

Ohashi Y, Nakai S, Tsukamoto T, Masumori N, Akaza H, Miyanaga N *et al.* (2002) Habitual intake of lactic acid bacteria and risk reduction of bladder cancer. *Urology International* **68**: 273–280.

Pridmore RD, Berger B, Desiere F, Vilanova D, Barretto C, Pittet A-C, Zwahlen M-C, Rouvet M, Altermann E, Barrangou R, Mollet B, Mercenier A, Klaenhammer T, Arigoni F, and Schell MA (2004) The genome sequence of the probiotic intestinal bacterium Lactobacillus johnsonii NCC 533. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 2512–2517.

Salminen S, Bouley MC, Boutron-Rualt MC, Cummings J, Franck A, Gibson G, Isolauri E, Moreau M-C, Roberfroid M, and Rowland I (1998) Functional food science and gastrointestinal physiology and function. *British Journal of Nutrition* Suppl 1: 147–171.

Salminen SJ, von Wright AJ, Ouwehand AC, and Holzapfel WH (2001) Safety assessment of probiotics and starters. In: Adams MR and Nout MJR (eds.) *Fermentation and Food Safety*, 1st edn, pp. 239–251. Gaithersburg: Aspen Publishers, Inc.

Schell MA, Karmirantzou M, Snel B, Vilanova D, Berger B, Pessi G, Zwahlen M-C, Desiere F, Bork P *et al.* (2002) The genome sequence of *Bifidobacterium longum* reflects its adaptation to the human gastrointestinal tract. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 14422–14427.

Schiffrin EJ, Brassart D, Servin AL, Rochat F, and Donnet-Hughes A (1997) Immune modulation of blood leukocytes in humans by lactic acid bacteria: criteria for strain selection. *American Journal of Clinical Nutrition* **66**: 515S–520S.

Shanahan F (2002) Crohn's disease. *Lancet* **359**: 62–69.

Sudo N, Sawamura S, Tanaka K, Aiba Y, Kubo C, and Koga Y (1997) The requirement of intestinal bacterial flora for the development of an IgE production system fully susceptible to oral tolerance induction. *Journal of Immunology* **159**: 1739–1745.

Tannock GW (2003) Probiotics: time for a dose of realism. *Current Issues in Intestinal Microbiology* **4**: 33–42.

Van Niel CW, Fewudtner C, Garrison MM, and Christakis DA (2002) Lactobacillus therapy for acute infectious diarrhea in children: a meta-analysis. *Pediatrics* **109**: 678–684.

Vaughan E, de Vries M, Zoentendal E, Ben-Amor K, Akkermans A, and de Vos W (2002) The intestinal LABs. *Antonie Van Leeuwenhoek* **82**: 341–352.

Xu J, Chiang HC, Bjursell MK, and Gordon JI (2004) Message from a human gut symbiont: sensitivity is a prerequisite for sharing. *Trends in Microbiology* **12**: 21–28.

Milk *see* **Dairy Products**

Minerals *see* **Calcium**. **Magnesium**. **Phosphorus**. **Potassium**. **Sodium**: Physiology

Molybdenum *see* **Ultratrace Elements**

Monosaturated Fat *see* **Fatty Acids**: Monounsaturated

Mycotoxins *see* **Food Safety**: Mycotoxins

N

NIACIN

C J Bates, MRC Human Nutrition Research, Cambridge, UK

This article is reproduced from the previous edition, pp. 1290–1297, © 1999, Elsevier Ltd.

Absorption, Transport, and Storage

Niacin is a B vitamin that is essential for health in humans and also in most other mammals that have been investigated. Niacin is associated with a characteristic deficiency disease in humans known as pellagra. Pellagra has been described and identified in various communities, notably in Spain and North America, in the last century and the early years of this century. It has persisted in Yugoslavia, Egypt, Mexico, and some African countries. Pellagra is characteristically associated with maize-based diets. The skin lesions found in pellagra are most severe during the summer months because of the effects of the exacerbating sun exposure. However, some countries with a maize diet (e.g., Guatemala) avoid pellagra by means of the niacin present in roasted coffee (**Table 1**). Others avoid it by lime treatment, e.g., in the preparation of tortillas.

Preformed niacin occurs in foods either as nicotinamide (niacinamide) or as the pyridine nucleotide coenzymes derived from it, or as nicotinic acid, without the amide nitrogen, which is the form known as 'niacin' in North America. Both nicotinamide and nicotinic acid are equally effective as the vitamin, but in large doses they exert markedly different pharmacological effects, so it is important, at least in that context, to make and maintain the distinction. In addition to the preformed vitamin, an important *in vivo* precursor is the amino acid L-tryptophan, obtained from dietary protein. Because the human total niacin supply, and hence niacin status, depends on the dietary tryptophan supply as well as on the amount of preformed dietary niacin and its bioavailability, it has become the accepted practice to express niacin intakes as 'niacin equivalents,' which is a combination of mg preformed dietary niacin and mg niacin which can become available by conversion from tryptophan within the body. As discussed later, this calculation involves several assumptions, and is therefore only an approximation to the actual supply to the body for any particular individual; however, it is considered adequate for most practical purposes.

It appears likely that the most important ultimate sources of preformed niacin in most foods, particularly those of animal foods, are the pyridine nucleotides: NAD(H_2) and NADP(H_2). Hydrolases and pyrophosphatases present in biological tissues convert these coenzymes to partly degraded products, which are then available as sources of the vitamin. NAD glycohydrolase and pyrophosphatase enzymes are present in the gut mucosa to assist hydrolysis and absorption of the hydrolyzed products, and these are likely to include both nicotinamide and nicotinamide ribonucleotide, the latter being further degraded to the riboside. Absorption of nicotinamide or nicotinic acid by the mammalian intestine has been shown to consist of a saturable transport component, dominant at low intakes, which is dependent on sodium, energy and pH, and a nonsaturable component, which becomes dominant at high doses or intake levels. Absorption is efficient even at such high discrete doses as 3 g or more: as much as 85% of such a dose is subsequently excreted into the urine. Absorption of test niacin doses introduced directly into the human upper ileum is rapid, with peak levels appearing in blood plasma within 5–10 min.

Transport of niacin between the liver and the intestine can occur *in vivo*, as indicated by radioactive probes in animals, and the liver appears to be a major site of conversion of niacin to its ultimate functional products: the nicotinamide nucleotide coenzymes. Nicotinamide can pass readily between the cerebrospinal fluid and the plasma, thus ensuring a supply also to the brain and spinal cord. Liver contains greater niacin coenzyme concentrations than most other tissues, but all metabolically active tissues contain these essential

Table 1 Niacin equivalents in selected foods[a]

	Niacin equivalents from preformed niacin[b] (mg per 100 g, wet)	*Niacin equivalents from tryptophan[c] (mg per 100 g, wet)*	*Total niacin equivalents (mg per 100 g, wet)*
Milk	0.1	0.8	0.9
Raw beef	5.0	4.7	9.7
Raw white fish	2.4	3.4	5.8
Raw eggs	0.1	3.7	3.8
Raw potatoes	0.6	0.5	1.1
Raw peas	2.5	1.1	3.6
Raw peanuts	13.8	5.5	19.3
White bread	0.8	1.7	2.5
Polished rice	0.2	1.5	1.7
Maize	0.1	0.9	1.0
Cornflakes (fortified)	16.0	0.9	16.9
Coffee[d]	24.1	2.9	27.0

[a]Data adapted from: Paul AA (1969) The calculation of nicotinic acid equivalents and retinol equivalents in the British diet. *Nutrition* (*London*) **23**: 131–136,[a] and supplements to *McCance and Widdowson's The Composition of Foods* (Holland B, Welch AA, Unwin ID, Buss DH, Paul AA, and Southgate DAT (1991), The Royal Society of Chemistry and MAFF),[a] and from Bressani R *et al.* (1961) Effect of processing method and variety on niacin and ether extract content of green and roasted coffee. *Food Technology* **15**: 306–308.
[b]Amount available for absorption. In the case of bread, rice, and maize, the total amounts present are 1.7, 1.5, and 1.2 mg per 100 g, but apart from the niacin added in the fortification of white flour, 90% of this is unavailable for utilization by humans.
[c]Assuming that 60 mg tryptophan yields 1 mg niacin equivalent.
[d]Niacin is released from trigonelline in coffee beans by the roasting process.

metabolic components. Both facilitated diffusion (which is sodium- and energy-dependent and saturable), and passive diffusion (which is nonsaturable) contribute to tissue uptake from the bloodstream. With the exception of muscle, brain and testis, within the body nicotinic acid is a better precursor of the coenzyme form than is nicotinamide. The liver appears to be the most important site of conversion of tryptophan to the nicotinamide coenzymes.

Of the two pyridine nucleotide coenzymes, NAD is present mainly as the oxidized form in the tissues, whereas NADP is principally present in the reduced form, $NADPH_2$. There are important homeostatic regulation mechanisms which ensure and maintain an appropriate ratio of these coenzymes in their respective oxidized or reduced forms in healthy tissues. Once converted to coenzymes within the cells, the niacin therein is effectively trapped, and can only diffuse out again after degradation to smaller molecules. This implies, of course, that the synthesis of the essential coenzyme nucleotides must occur within each tissue and cell type, each of which must possess the enzymatic apparatus for their synthesis from the precursor niacin. Loss of nicotinamide and nicotinic acid into the urine is minimized (except when the intake exceeds requirements) by means of an efficient reabsorption from the glomerular filtrate.

Metabolism and Excretion

The conversion of tryptophan to nicotinic acid *in vivo* is depicted in **Figure 1**. The rate of conversion of tryptophan to niacin and the pyridine nucleotides is controlled by the activities of tryptophan dioxygenase (known alternatively as tryptophan pyrrolase), kynurenine hydroxylase, and kynureninase. These enzymes are, in turn, dependent on factors such as other B vitamins, glucagon, glucocorticoid hormones, and estrogen metabolites, and there are various competing pathways which also affect the rate of conversion. For these reasons, a variety of nutrient deficiencies, toxins, genetic and metabolic abnormalities, etc. can influence niacin status and requirements.

For practical purposes, on the basis of studies performed in the 1950s, 60 mg tryptophan is deemed to give rise to 1 mg nicotinic acid; hence 60 mg tryptophan contributes 1 mg niacin equivalent, for dietary intake calculations and food tables (see **Table 1**).

The two pyridine nucleotide coenzymes, formerly known as 'coenzymes I and II,' then for a period as 'DPN and TPN,' and known nowadays as 'NAD' and 'NADP' (nicotinamide adenine dinucleotide and nicotinamide adenine dinucleotide phosphate), are involved in hundreds of enzyme-catalyzed redox reactions *in vivo*. Although a minority of these

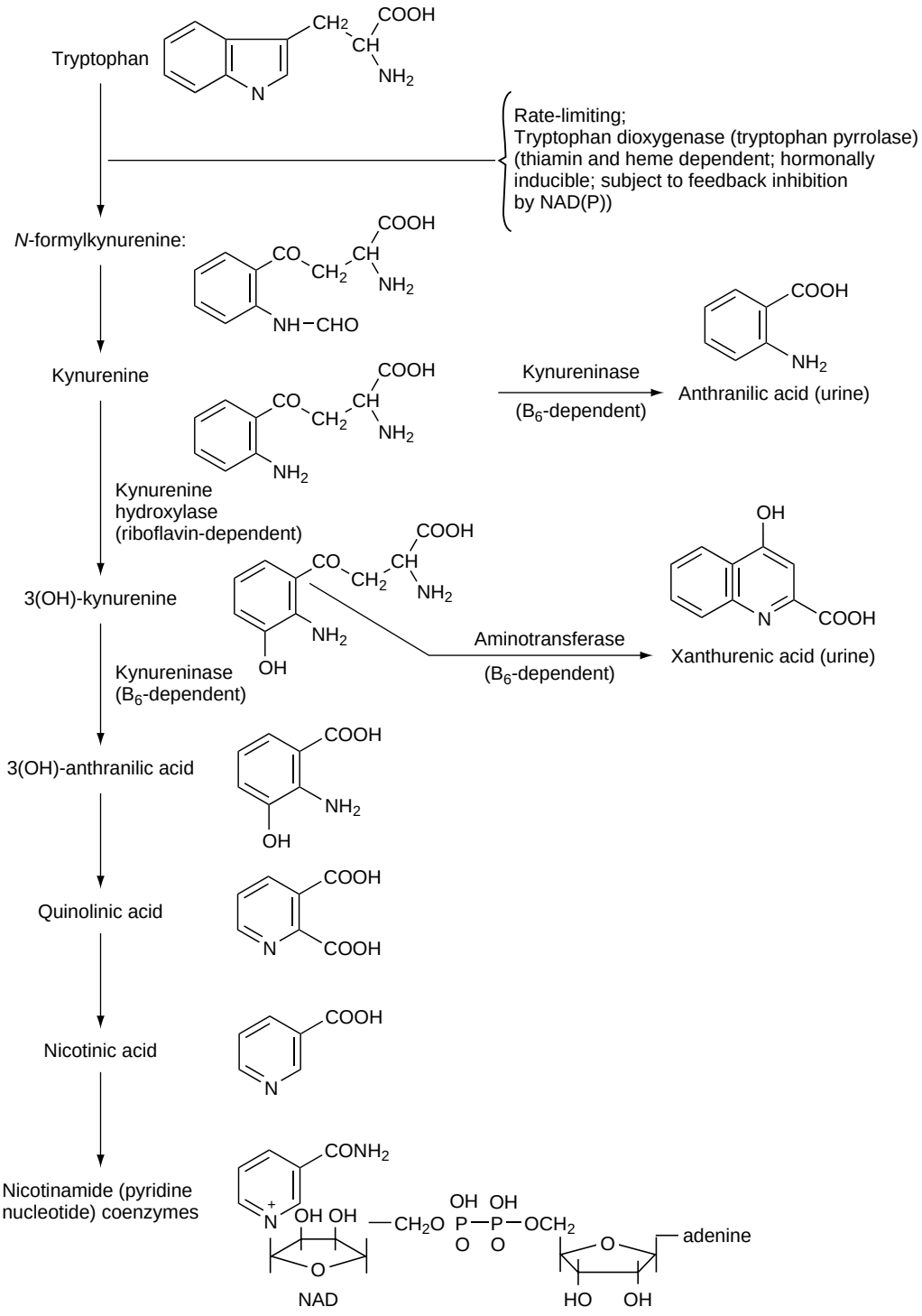


Figure 1 *In vivo* conversion of tryptophan to nicotinic acid and NAD.

diverse reactions can use either of the two niacin-derived cofactors, most are highly specific for one or the other.

Catabolism of the pyridine nucleotide coenzymes *in vivo* is achieved by four classes of enzymes: NAD glycohydrolase, ADP ribosyl transferase, and poly (ADP ribose) synthetase, (all of which liberate nicotinamide), and NAD pyrophosphatase (which liberates nicotinamide mononucleotide which is then further hydrolyzed to nicotinamide). Turnover of nicotinamide then results in the formation of 1-methylnicotinamide (usually described as

N^1-methyl nicotinamide or NMN), an excretory product which is excreted in the kidney and appears in the urine, together with some further oxidation products, typically the 1-methyl-2-pyridone-5-carboxamide and 1-methyl-4-pyridone-3-carboxamide (usually referred to as '2-pyridone' and '4-pyridone', respectively). These excretory turnover products can be used as indicators of whole body niacin status (see below). At high intakes of niacin, as much as 85% of the intake may be excreted unchanged; however the excretion of nicotinamide always predominates over that of nicotinic acid.

Hydrolysis of hepatic NAD to yield nicotinamide allows the release of niacin for utilization by other tissues. Relative protection of the pyridine nucleotide within certain key enzymes such as glyceraldehyde 3-phosphate dehydrogenase confers a protection on certain key metabolic pathways, thus ensuring good homeostatic control. By contrast, there is evidence that the enzymes which catalyze pyridine nucleotide turnover may be hyperactivated within cells that have been damaged by carcinogens, including mycotoxins, thus starving these damaged cells of essential cofactors and causing their death, presumably to protect the rest of the organism. This effect may help to explain the otherwise puzzling observation that moldy grain in the diet can increase the risk of pellagra when niacin and tryptophan intakes are marginal. In normal, healthy cells, the compartmentalization of hydrolytic enzymes prevents unwanted coenzyme turnover, and this compartmentalization seems to become breached in damaged or dying cells.

Other urinary excretion products of niacin include nicotinuric acid (nicotinoyl glycine); nicotinamide *N*-oxide, and trigonelline (N^1-methyl nicotinic acid); the latter may arise from bacterial action in the gut or from the absorption of this substance from foods. The pattern of the different turnover metabolites varies between species, between diets (depending partly on the ratio of nicotinamide to nicotinic acid in the diet), and partly with niacin status; thus there are complex regulatory mechanisms to be considered.

Metabolic Function and Essentiality

The best-known functions of niacin are derived from the functions of its coenzymes: NAD and NADP in the hydrogen/electron transfer redox reactions in living cells. Like most B vitamins, niacin is not extensively stored in forms or in depots that are usually metabolically inactive, but rather those that can become available during dietary deficiency. However, some 'storage' of the coenzymes NAD and NADP in the liver is thought to occur. An inadequate dietary intake leads rapidly to significant tissue depletion within 1–2 months, and then successively to biochemical abnormalities, followed by clinical signs of deficiency, and eventually to death. As with the other B vitamins, rates of turnover and hence the rates of excretion of coenzyme breakdown products decline progressively as dietary deficiency becomes more severe and prolonged, so that the tissue levels are relatively protected and spared. In adult humans a severe deficiency may take many months to develop before it results in the clinical signs of pellagra.

Some of the most important and characteristic functions of NAD manifest in the principal cellular catabolic pathways, responsible for liberation of energy during the oxidation of energy-producing fuels. NADP, however, functions mainly in the reductive reactions of lipid biosynthesis, and the reduced form of this coenzyme is generated via the pentose phosphate cycle. NAD is essential for the synthesis and repair of DNA. NAD has, in addition, a role in supplying ADP ribose moieties to lysine, arginine, and asparagine residues in proteins such as histones, DNA lyase II, and DNA-dependent RNA polymerase, and to polypeptides such as the bacterial diphtheria and cholera toxins. In the nucleus, poly (ADP ribose) synthetase is activated by binding to DNA breakage points and is involved in DNA repair. It is also concerned with condensation and expansion of chromatin during the cell cycle and in DNA replication. Niacin status affects the level of ADP ribolysation of proteins. A high level of poly (ADP ribose) synthetase activity, which is found in some tumors, can result in low levels of NAD. A chromium dinicotinate complex found in yeast extracts may function as a glucose tolerance factor or in detoxification, but this has not yet been proven.

Because the electron transport functions of NAD frequently involve flavin coenzymes, and because both flavin coenzymes and vitamin B_6 coenzymes are involved in the conversion of tryptophan to niacin *in vivo*, there are important metabolic interactions between these B vitamins. A similarity of clinical deficiency signs, making it difficult to distinguish between them, may be encountered in population studies of deficiency.

Because the body's need for niacin can be met completely by dietary tryptophan, it is not, strictly speaking, an essential vitamin. In this respect it resembles carnitine, which can be synthesized entirely from lysine, but for which in some circumstances a dietary requirement exists. Traditionally, however, niacin is classified as an essential vitamin, because some human diets have tended to be lacking

in niacin and its precursor, tryptophan. Some animals such as sheep and cattle appear to be able to synthesize sufficient niacin for their needs from tryptophan, and do not therefore need preformed niacin in their diets.

Assessment of Niacin Status

Whereas the measurement of B vitamin status has, in recent years, tended to focus on blood analysis, perhaps mainly because of the convenience of sample collection, the development of blood-based status analysis for niacin has lagged behind that of the other components of the B complex. Some studies have indeed suggested that the erythrocyte concentration of the niacin-derived coenzyme NAD may provide useful information about the niacin status of human subjects; that a reduction in the ratio of NAD to NADP to below 1.0 in red cells may provide evidence of niacin deficiency; and that a decline in plasma tryptophan levels may indicate a more severe deficiency than a decline in red cell NAD levels. These claims now need to be tested in naturally deficient human populations. The niacin coenzymes can be quantitated either by enzyme-linked reactions or by making use of their natural fluorescence in alkaline solution.

At present, niacin status is most commonly assessed by the assay of some of the breakdown products of niacin coenzymes in the urine. Of these, N^1-methyl nicotinamide (NMN) is the easiest to measure, because of a convenient conversion *in vitro* to a fluorescent product, which can then be quantitated without the need for separation. However, more definitive and reliable information can be obtained by the measurement of urinary NMN in conjunction with one or more of the urinary pyridone turnover products (N^1-methyl-2-pyridone-5-carboxamide and N^1-methyl-4-pyridone-3-carboxamide), which can be detected and quantitated by UV absorption following high-pressure liquid chromatography. The Interdepartmental Committee on Nutrition for National Defense (USA) selected the criterion of niacin deficiency in humans as an NMN excretion rate of $<5.8\,\mu$mol (0.8 mg) NMN per day in 24 h urine samples.

Requirements and Signs of Deficiency

As for most other micronutrients, the requirement of niacin to prevent or reverse the clinical deficiency signs is not known very precisely, and probably depends on ancillary dietary deficiencies or other insults occurring in natural human populations. For the purpose of estimating niacin requirements for dietary reference values, the criterion of restoration of urinary excretion of NMN during controlled human depletion–repletion studies has been selected, and on this basis, the average adult requirement has been estimated as 5.5 mg (45 μmol) of niacin equivalents per 1000 kcal (4200 kJ). Adding a 20% allowance for individual variation this needs to be increased to 6.6 mg (54 μmol) per 1000 kcal, (4200 kJ), which is the current reference nutrient intake (UK). Niacin requirements were, by convention, expressed as a ratio to energy expenditure. For subjects with very low energy intakes, the daily intake of niacin equivalents should not fall below 13 mg, however. If dietary protein levels and quality are high, it is possible for tryptophan alone to provide the daily requirement for niacin equivalents. Dietary niacin deficiency is now rare in most Western countries.

The appearance of severe niacin deficiency as endemic pellagra, especially in North America in the nineteenth and early twentieth centuries, has been ascribed to the very poor availability of bound forms of niacin (in niacytin, a polysaccharide/glycopeptide/polypeptide-bound form, which is 90% indigestible), together with the relatively low content of tryptophan occurring in grains (see **Table 1**). However, the lack of available niacin and tryptophan may not have been the whole story, since coexisting deficiencies or imbalances of other nutrients, including riboflavin, may also have contributed to this endemic disease. It appears also that the choice of cooking methods may have been critical, since the Mexican custom of cooking maize with lime in the preparation of tortillas helps to release the bound niacin from its carbohydrate complex and to increase the bioavailability of tryptophan-containing proteins, and thus to reduce the prevalence of clinical deficiency disease. In parts of India, pellagra has been encountered in communities whose main staple is a form of millet known as 'jowar', which is rich in leucine. It was proposed, and evidence was obtained from animal and *in vitro* studies, that high intakes of leucine can increase the requirements for niacin. However, other evidence is conflicting (this interaction is not fully understood). In parts of South Africa, iron overload has been reported to complicate the metabolic effects of low niacin intakes.

The average content of niacin in human breast milk is 8 mg (65.6 μmol) per 1000 kcal (4200 kJ), and this is the basis for the recommendations (and dietary reference values) for infants up to 6 months. In the UK, the Reference Nutrient Intake niacin increment during pregnancy is nil, and during lactation it is 2 mg per day.

The most characteristic clinical signs of severe niacin deficiency in humans are dermatosis (hyperpigmentation, hyperkeratosis, desquamation – especially where exposed to the sun), anorexia, achlorhydria, diarrhea, angular stomatitis, cheilosis, magenta tongue, anemia, and neuropathy (headache, dizziness, tremor, neurosis, apathy). In addition to the pellagra caused by dietary deficiency or imbalance, there are also reports of disturbed niacin metabolism associated with phenylketonuria, acute intermittent porphyria, diabetes mellitus, some types of cancer (carcinoid syndrome), thyrotoxicosis, fever, stress, tissue repair, renal disease, iron overload, etc. The picture in other species is not radically different; however, deficient dogs and cats typically exhibit 'black tongue' (pustules in the mouth, excessive salivation) and bloody diarrhea, pigs exhibit neurological lesions affecting the ganglion cells, rats exhibit damage to the peripheral nerves (cells and axons), and fowl exhibit inflammation of the upper gastrointestinal tract, dermatitis, diarrhea, and damage to the feathers. All species exhibit reduction of appetite and loss of weight; however, it is of interest that the skin lesions seen in humans are rare in most other species.

Dietary Sources, High Intakes, and Antimetabolites

As can be seen from **Table 1**, different types of foods differ considerably, not only in their total contribution to nicotinic acid equivalents, but also in the ratio of the contribution from preformed niacin and from tryptophan. In a typical Western diet, it has been calculated that if the 60 mg tryptophan = 1 mg niacin formula is applied, then preformed niacin provides about 50% of the niacin supply in the diet. In practice it seems possible for all of the niacin requirement to be provided by dietary tryptophan in Western diets. As is the case for the other B vitamins, meat, poultry, and fish are excellent sources of niacin equivalents, followed by dairy and grain products, but as noted above, certain grains such as maize, and whole highly polished rice, can be very poor sources and may be associated with clinical deficiency if the diets are otherwise poor and monotonous.

In recent years, both nicotinamide and nicotinic acid have been proposed and tested for possibly useful pharmacological properties at high intake levels. This new phase of interest in the vitamin has, in turn, raised concerns about the possible side effects of high intakes, and the definition of maximum safe intakes.

The greatest interest, in pharmacological terms, has been centered around nicotinic acid, which has been shown to have marked antihyperlipidemic properties at daily doses of 2–6 g. Nicotinamide does not share this particular pharmacological activity. Large doses of nicotinic acid reduce the mobilization of fatty acids from adipose tissue by inhibiting the breakdown of triacylglycerols through lipolysis. They also inhibit hepatic triacylglycerol synthesis, thus limiting the assembly and secretion of very low-density lipoproteins from the liver and reducing serum cholesterol levels. Large doses of nicotinic acid ameliorate certain risk factors for cardiovascular disease: for instance they increase circulating high-density lipoprotein levels. The ratio of HDL_2 to HDL_3 is increased by nicotinic acid; there is a reduced rate of synthesis of apolipoprotein A-II and a transfer of some apolipoprotein A-I from HDL_3 to HDL_2. These changes are all considered potentially beneficial in reducing the risk of cardiovascular disease. If given intravenously, large doses of nicotinic acid can, however, produce side effects such as temporary vasodilatation and hypotension. Other side effects can include nausea, vomiting, diarrhea and general gastrointestinal disturbance, headache, fatigue, difficulty in focusing, skin discoloration, dry hair, sore throat, etc. A large trial for secondary prevention of myocardial infarction, with a 15 year period of follow-up, produced convincing evidence for moderate but significant protection against mortality, which was attributed either to the cholesterol-lowering effect or an early effect on nonfatal reinfarction, or both. Nicotinic acid is still the treatment of choice for some classes of high-risk hyperlipidemic patients, although newer drugs may have fewer side effects and therefore be preferred.

The potential benefits of the lipid-lowering effects of nicotinic acid have to be considered in the light of possibly toxic effects, particularly for the liver. These may manifest as jaundice, changes in liver function tests, changes in carbohydrate tolerance, and changes in uric acid metabolism including hyper-uricemia. There may also be accompanying ultrastructural changes. Hyperuricemia may result from effects on intestinal bacteria and enzymes, and from effects on renal tubular function. Such toxic effects are especially severe if sustained release preparations of nicotinic acid are used.

Nicotinamide does not share with nicotinic acid these effects on lipid metabolism or the associated toxicity. However, it has been shown to be an inhibitor of poly (ADP ribose) synthetase in pancreatic β cells in animal studies. A high-risk group of children aged 5–8 years in New Zealand given large doses of nicotinamide daily for up to 4.2 years had

only half the predicted incidence of insulin-dependent diabetes.

Other claims for megadoses of nicotinic acid or nicotinamide, such as the claim that abnormalities associated with schizophrenia, Down's syndrome, hyperactivity in children, etc. can be reduced, have so far failed to win general acceptance. Clearly niacin deficiency or dependency can exacerbate some types of mental illness such as depression or dementia. There have been a number of attempts to treat depression with tryptophan or niacin, or both, on the basis that the correction of depressed brain levels of serotonin would be advantageous. However, these have met with only limited success. Schizophrenics have been treated with nicotinic acid on the basis that their synthesis of NAD is impaired in some parts of the brain, and that the formation of hallucinogenic substances such as methylated indoles may be controlled.

There are various medical conditions and drug interactions that can increase the requirement for niacin. Examples are: Hartnup disease, in which tryptophan transport in the intestine and kidney is impaired; carcinoid syndrome, in which tryptophan turnover is increased; and isoniazid treatment, which causes B_6 depletion and hence interference with niacin formation from tryptophan. Hartnup disease (the name of the first patient being Hartnup) is a rare genetic disease in which the conversion of tryptophan to niacin is reduced, partly as a result of impaired tryptophan absorption. Affected subjects exhibit the classical skin and neurological lesions of pellagra, which can be alleviated by prolonged treatment with niacin. Another genetic disease which may respond to niacin supplements is Fredrikson type I familial hypercholesterolemia; nicotinic acid is effective in reducing the raised blood cholesterol levels associated with this abnormality.

There are several analogs and antimetabolites of niacin that are of potential use or metabolic interest. The closely related isoniazid is commonly used for treatment of tuberculosis; indeed, nicotinamide itself has been used for that purpose. Nicotinic acid diethylamide ('nikethamide') is used as a stimulant in cases of central nervous system depression after poisoning, trauma or collapse. Possible antineoplastic analogs include 6-dimethylaminonicotinamide and 6-aminonicotinamide; however, the latter is also highly teratogenic. These latter compounds inhibit several key enzymes whose substrates are NAD or NADP, by being converted *in vivo* to analogs of these coenzymes. The compound 3-acetyl pyridine, which also forms an analog of NAD, can have either antagonistic or niacin-replacing properties, depending on the dose used. Commonly used drugs such as metronidazole are also niacin antagonists.

See also: **Bioavailability**. **Energy**: Metabolism. **Hyperlipidemia**: Overview; Nutritional Management. **Riboflavin**. **Vitamin B_6**.

Further Reading

Bender DA (1992) Niacin. In: *Nutritional Biochemistry of the Vitamins*, ch. 8, pp. 184–222. Cambridge: Cambridge University Press.

Carpenter KJ (ed.) (1981) *Pellagra: Benchmark Papers in the History of Biochemistry/II*. Stroudsburg, Pennsylvania: Dowden, Hutchinson & Ross Publ. Co.

Di Palma JR and Thayer WS (1991) Use of niacin as a drug. *Annual Review of Nutrition* **2**: 169–187.

Fu CS, Swendseid ME, Jacob RA, and McKee RW (1989) Biochemical markers for assessment of niacin status in young men: levels of erythrocyte niacin coenzymes and plasma tryptophan. *Journal of Nutrition* **119**: 1949–1955.

Hankes LV (1984) Nicotinic acid and nicotinamide. In: Machlin LJ (ed.) *Handbook of Vitamins*, ch. 8, pp. 329–377. New York: Marcel Dekker Inc.

Henderson LM (1983) Niacin. *Annual Review of Biochemistry* **3**: 289–307.

Horwitt MK, Harvey CC, Rothwell WS, Cutler JL, and Haffron D (1956) Tryptophan–niacin relationship in man. *Journal of Nutrition* **60**(supplement 1): 1–43.

Jacob RA, Swendseid ME, McKee RW, Fu CS, and Clemens RA (1989) Biochemical markers for assessment of niacin status in young men: urinary and blood levels of niacin metabolites. *Journal of Nutrition* **119**: 591–598.

Sauberlich HE, Dowdy RP, and Skala JH (1974) *Laboratory Tests for the Assessment of Nutritional Status*, pp. 70–74. Boca Raton: CRC Press.

Swendseid ME and Jacob RA (1984) Niacin. In: Shils ME and Young VR (eds.) *Modern Nutrition in Health and Disease*, 7th edn., ch. 22, pp. 376–382. Philadelphia: Lea & Febiger.

Nitrogen *see* **Amino Acids**: Chemistry and Classification; Metabolism. **Protein**: Digestion and Bioavailability; Quality and Sources; Requirements and Role in Diet; Deficiency

NUCLEIC ACIDS

E A Carrey, Institute of Child Health, London, UK
H A Simmonds, Guy's Hospital, London, UK

The nucleic acids are fundamental to genetics and to metabolism. Deoxyribonucleic acid (DNA) is found in the chromosomes within the nucleus and the mitochondria, and it contains the genetic information. Ribonucleic acid (RNA) is found both in the nucleus and in the surrounding cytoplasm, and its various forms fulfill several tasks associated with the transfer of the genetic message and its eventual translation into proteins. Each nucleic acid is a linear polymer of nucleotides (**Figure 1A**). Nucleosides, the related small molecules, consist of a pentose sugar bound to the N-9 atom of a purine or to the N-1 of the

The nucleic acids, vital constituents of all living cells, were discovered by Miescher in 1868 and isolated from the nuclei of pus cells and spermatozoa of Rhine salmon. The major constituents of nucleic acids were shown to be sugars, phosphate groups, and the characteristic purine and pyrimidine bases, now considered to be some of the first chemicals to emerge from the 'primordial soup' before life began on Earth. Emil Fischer and colleagues established the chemical structure of the purine bases, including uric acid—the end (waste) product of purine metabolism in humans—at the end of the nineteenth century.

This article outlines the biosynthesis of nucleic acids and gives a brief overview of the physiological functions of nucleosides, nucleotides, and nucleic acids. It describes the toxicity that may arise from degradation of both endogenous and dietary (exogenous) nucleic acids in humans and provides a summary of the nucleic acid content of foods.

Physiology

Structure

The nucleic acids are fundamental to genetics and to metabolism. Deoxyribonucleic acid (DNA) is found in the chromosomes within the nucleus and the mitochondria, and it contains the genetic information. Ribonucleic acid (RNA) is found both in the nucleus and in the surrounding cytoplasm, and its various forms fulfill several tasks associated with the transfer of the genetic message and its eventual translation into proteins.

Each nucleic acid is a linear polymer of nucleotides (**Figure 1A**). Nucleosides, the related small molecules, consist of a pentose sugar bound to the N-9 atom of a purine or to the N-1 of the

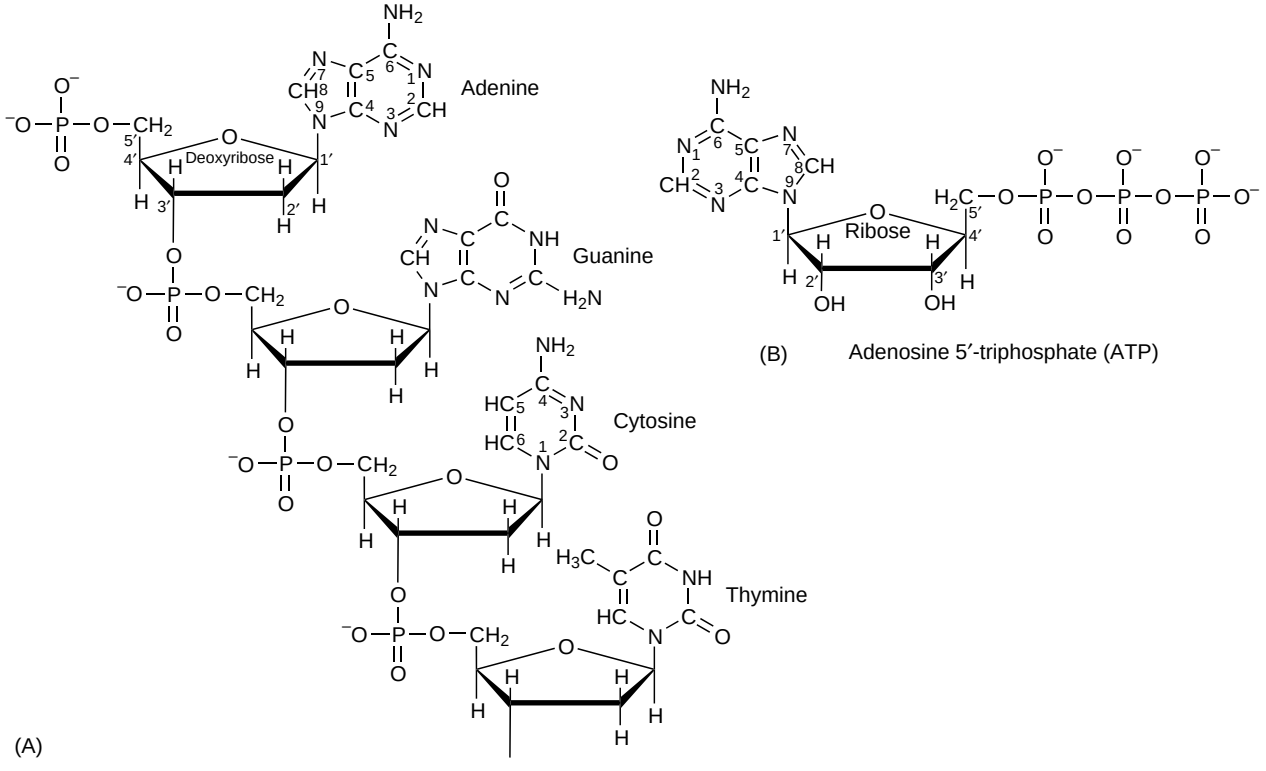


Figure 1 (A) Schematic representation of part of a DNA strand showing the structural formulas of the four constituent bases, adenine, guanine, cytosine, and thymine, linked via the 3′-OH group of the deoxyribose moiety to the 5′-phosphate group of the next nucleotide. Also shown is the numbering of the atoms in the deoxyribose, as well as the pyrimidine and purine rings. The latter consist of a six-membered pyrimidine ring fused to a five-membered imidazole ring. (B) Structural formula of ATP indicating that the ribose, as distinct from deoxyribose, has an OH group at the 2′ position on the pentose ring.

pyrimidine ring: With one or more phosphate groups at the 5′ position of the sugar, the molecule is a nucleotide (**Figure 1B**). When nucleoside triphosphates (NTP) are linked through the 5′ phosphate groups to the 3′ position of the previous residue on the growing chain, the chemical energy for the polymerization is provided by the removal of the second and third phosphate groups.

In DNA the pentose is 2′-deoxyribose (**Figure 1A**) and the bases are adenine (A), guanine (G), cytosine (C), and thymine (T). Two strands are wound in opposing chemical directions in the well-defined double-helix structure of DNA, with each nucleotide of one strand linked by hydrogen bonding to a complementary nucleotide on the other (A-T and G-C). The deoxyribose and phosphate groups form the outer sides of the 'ladder.'

The RNA molecule is chemically single stranded, but double-helical regions arise when stretches of complementary sequences allow hairpin loops to form. In addition, the base uracil (U) is found instead of thymine, and the pentose is ribose.

Nucleic Acid Biosynthesis in Humans

The first step in the synthesis of DNA and RNA in humans involves the formation of the purine and pyrimidine ribonucleotides. They are derived endogenously by two routes: the energetically expensive multistep *de novo* route (**Figure 2**), using small

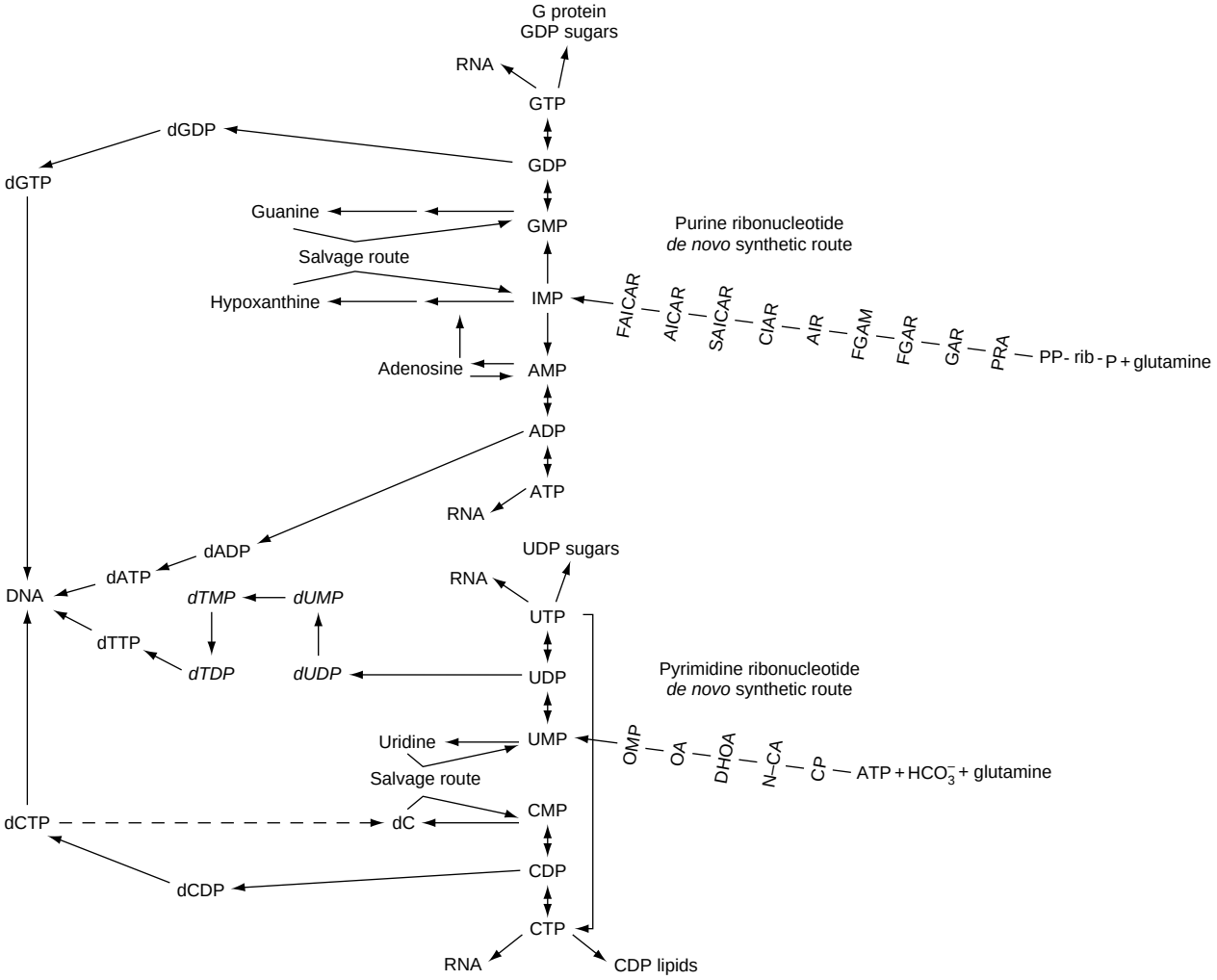


Figure 2 Metabolic pathways for the synthesis of DNA and RNA from their purine and pyrimidine ribo- and deoxyribonucleotide precursors. The dotted line from dCTP indicates the route of breakdown to deoxycytidine (dC), which can be salvaged by dC kinase and incorporated into dCDP lipids or DNA. HCO_3^-, bicarbonate; CP, carbamoyl phosphate; N-CA, carbamoyl aspartate; DHOA, dihydroorotic acid; OA, orotic acid; OMP, orotidylic acid; PP-rib-P, 5-phosphoribosyl-1-pyrophosphate; PRA, 5-phosphoribosylamine; GAR, glycinamide ribotide; FGAR, *N*-formyl glycinamide ribotide; FGAM, *N*-formylglycinamidine ribotide; AIR, 5-aminoimidazole ribotide; CAIR, 5-aminoimidazole-4-carboxy ribotide; SAICAR, *N*-succino-5-aminoimidazole-4-carboxamide ribotide; AICAR, 5-aminoimidazole-4-carboxamide ribotide; FAICAR, 5-formamidoimidazole-4-carboxamide ribotide.

molecules such as carbon dioxide, amino acids, and ribose sugars as precursors; and the energetically less expensive 'salvage' pathway. Purine bases and pyrimidine nucleosides from the breakdown of nucleic acids and nucleotide cofactors are salvaged within the cells, generating nucleotides that can be incorporated into nucleic acids. In most cells in the body, salvage processes are more important than *de novo* synthesis, and the ribonucleotides recycled in this way exert feedback control on the *de novo* routes.

The ribonucleoside 5′-monophosphates generated from either pathway are rapidly phosphorylated within the cell to the triphosphate form for immediate use in the synthesis of RNA or in a variety of cellular processes. The most abundant ribonucleotide in the body is adenosine 5′-triphosphate (ATP), which is the universal energy carrier of all living organisms (**Figure 1B**). In addition, ATP molecules are part of the coenzymes (NAD, NADP, FAD, etc.), which assist in many reactions including the conversion of food into energy. Adenosine and guanosine nucleotides within cells have roles in the transduction of external signals into cellular responses and in the translation and synthesis of proteins. Pyrimidines, present at much lower concentrations in cells, also fulfill diverse functions. UDP-glucose and CDP–lipids are active intermediates in the synthesis of glycogen and membranes, respectively, and sugars linked to UDP or GDP are used in the glycosylation of proteins, many of which are exposed on the outer surface of cells. UDP-glucuronic acid is an essential component of the pathways in liver, gut, and kidney that convert foreign molecules and steroids into soluble forms for disposal from the body.

To make DNA, the ribonucleoside diphosphates (rNDP) are reduced to the corresponding 2′-deoxyribonucleoside diphosphates (dNDP) in a reaction catalyzed by ribonucleotide reductase. This reaction produces dADP, dGDP, and dCDP, which are phosphorylated to the triphosphate form; dUDP is converted via dUMP to dTMP, providing the four substrates essential for DNA synthesis. The DNA polymerase enzymes form double-stranded DNA by sequential addition of monomers complementary to the bases on the opposite strand. Crucially, a 'proofreading' activity in the enzyme ensures the accuracy of the process, and hence double-stranded DNA provides a stable format for genetic information.

Nucleic Acids in the Storage and Transmission of Genetic Information

The role of DNA and RNA in the storage and transmission of genetic information is well established and can be found in standard textbooks. The hereditary material in the nucleus of human cells is packed into 23 chromosomes, and additional DNA is found in the mitochondria. The human genome is known to contain approximately 30 000 coding sequences, or genes, and a substantial proportion of DNA has a regulatory role in transcription of the genes. The sequence of the four bases and the capacity of DNA to be copied into two complementary strands underlie the genetic information of all living organisms. Interactions between DNA and the transcription factors determine the time and place in the body where genes are transcribed, causing development and metabolism to occur.

RNA molecules are synthesized initially on a DNA template by a DNA-dependent RNA polymerase in a process called transcription, in which ribonucleotides complementary to the bases of one strand of DNA are joined by 3′–5′ phosphodiester bonds.

Cells contain three types of RNA, each of which is chemically modified after transcription from the DNA template. The three kinds of RNA together are important in the translation of the genetic message to synthesize proteins in the cell. Most RNA is in the cytoplasm, principally in the form of ribosomal RNA (80% of the total), which performs structural and catalytic roles in ribosomes, the site of the growing polypeptide chain in protein synthesis, whereas messenger RNA (5%) provides the template for protein synthesis. The amino acids are brought to the assembly site covalently bonded to transfer RNA (15%), with their order in the growing protein being specified by the order of the bases in mRNA.

Synthesis of nucleic acids and their precursors in different human cells is related to cellular function

Synthesis of both DNA and RNA is prominent in cells and tissues with a high rate of turnover or metabolism (e.g., liver, gut epithelium, skin, dividing lymphocytes, bone marrow, and hair follicles). In most tissues in the adult, cells differentiate to perform specialized tasks and therefore cell division is used only to replace cells that have been lost. Different complements of enzymes are expressed in each cell type, and therefore tissues have characteristic profiles of internal metabolites, including nucleotides and nucleosides.

For example, in cells that do not continuously divide, such as heart and muscle, nucleotide profiles are relatively simple, relating to the major requirement to sustain levels of cofactors and ATP. In contrast, rapidly dividing cells in liver and intestine show a complex nucleotide pattern, supporting these organs as major sites of nucleic acid

metabolism. The gut is particularly important in this respect. The rate of cell turnover in the luminal villi is high, and it has been calculated in rat that approximately 30 mg of endogenous nucleic acid from dead cells enters the gut lumen daily. This means that the rate of nucleic acid synthesis in liver and intestine is much higher than in tissues such as muscle.

Metabolism of Endogenous Nucleic Acids and Excretion of End Products

There is a considerable turnover of endogenous nucleic acids and ribonucleotides daily during muscle work, wound healing, erythrocyte senescence, mounting an immune response, etc. However, only a small fraction of these vital endogenous compounds are actually degraded and lost from the body. The contents of dead cells are normally used by nearby cells, and degraded RNA or cofactors are recycled within living cells using active 'salvage' routes.

Salvage of the polynucleotides DNA and RNA begins when the molecules are degraded by enzymes—ribonucleases for RNA and deoxyribonucleases for DNA—to liberate nucleotides. The next step, degradation by specific 5′-nucleotidases (removing the phosphate groups) to nucleosides or deoxynucleosides, is essentially irreversible. In turn, removal of amino groups and the sugar residue will give the purine bases hypoxanthine and xanthine or the pyrimidine bases uracil and thymine.

Any pyrimidine nucleosides that are not salvaged are converted first to the bases uracil and thymine, which are further catabolized in a series of steps to β amino acids. All are soluble and readily excreted. There is thus normally no measurable pyrimidine end product and therefore no toxicity from endogenous or dietary pyrimidines.

In contrast, the purine bases are converted to xanthine (an insoluble metabolite) by the enzyme xanthine dehydrogenase (XDH) and then to the equally insoluble end product uric acid. This compound can normally be disposed of in the urine, but high concentrations can lead to the formation of kidney stones or deposits in the joints and under the skin. Some rare genetic disorders of purine biosynthesis can remove feedback regulation, or excessive breakdown of cells may overload the salvage system, resulting in very high endogenous levels of uric acid.

Metabolism of Dietary Nucleic Acids in Humans

The normal human diet is rich in both DNA and RNA since food is derived from once-living organisms. The metabolism of these exogenous nucleic acids follows a similar pattern to the intracellular process described previously, but the bacterial flora of the intestine are the first point of attack. This digestion is rapid. Studies in pigs (confirmed by later studies in humans) demonstrated that up to 50% of radiolabeled dietary purine was degraded and lost as carbon dioxide gas within 30 min, with the remaining 43% being recovered in the urine and 5% in the feces (**Figure 3**). It has been shown that dietary pyrimidine nucleotides, but not purines, are incorporated into RNA.

Humans thus have no apparent requirement for purines from the diet, and the intestinal mucosa provides an effective barrier to their uptake through a battery of enzymes that rapidly degrade purine nucleotides, nucleosides, and bases to the metabolic waste product, uric acid. This phenomenon may represent an important evolutionary development to protect the integrity of the cellular DNA or to ensure that levels of ATP do not fluctuate in concert with the dietary intake of purines.

The potential toxicity of dietary nucleic acids to humans usually arises not from the nucleic acids but

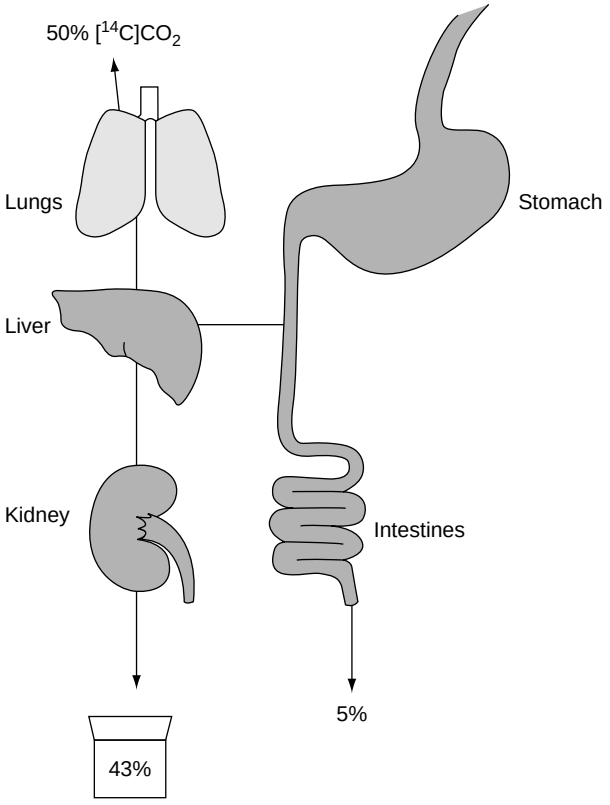


Figure 3 Diagram showing the fate of ^{14}C-labeled exogenous purine (guanine) in an animal model (pig). Radiolabel was recovered only in carbon dioxide gas, urine, or feces. No incorporation into any tissues was found.

from their metabolic end products (principally uric acid). Many investigators have shown that the fate of the dietary purine moiety depends on whether it is administered in the form of DNA, RNA, mononucleotides, nucleosides, or bases, with some being catabolised more readily than others. When normal subjects are fed RNA, the increase in the excretion of uric acid is dramatic alongside a modest increase in plasma urate concentrations. The effect of RNA is also twice that of DNA when the increase in purines in the plasma is measured.

On the other hand, pyrimidine mononucleotides and uridine, but not the base uracil, are absorbed readily from the intestine and utilised for nucleic acid synthesis. This has been demonstrated by studies of humans with hereditary oroticaciduria, a rare defect in *de novo* pyrimidine synthesis at the fourth step involving conversion of orotic acid to UMP. Such patients have a severe megaloblastic anemia that does not respond to the usual forms of treatment. They have been sustained for more than 40 years on oral uridine, indicating that the dietary pyrimidine is absorbed intact and can compensate totally for lack of *de novo* synthesis in humans. Studies using radiolabelled purines and pyrimidines in mice have provided further evidence for the incorporation of dietary pyrimidine nucleosides, but not purine nucleosides, into hepatic RNA.

Nucleosides and Nucleotides in the Diet

In healthy adults, the normal diet is a good source of nucleic acids, nucleotides, and nucleosides, and no supplementation is necessary.

Pharmacological uses for nucleosides and nucleotides Oral uridine, as described earlier, can be used where *de novo* biosynthesis of pyrimidines is defective, and it may be useful in reversing some effects of mitochondrial dysfunction and to minimize the toxic effect of the antitumor drug 5-fluorouracil. Uridine is also a precursor for UDP-glucose, essential for the deposition of glycogen in the liver. Enzymes in the liver, however, rapidly degrade much of each oral dose of uridine. Oral administration of a new prodrug, PN401, inhibits the degradative processes and delivers more uridine into circulation than oral uridine.

Oral CDP–choline is rapidly converted to its components, CDP (which can be recycled to uridine) and choline, an essential component of lipid membranes. Each molecule can then cross the blood–brain barrier, where CDP–choline is used in regeneration of membranes within and around nerve cells. Research in rats and early studies in humans suggest that its pharmacological effects may extend to protection against dementia, memory loss, visual degeneration, and recovery from ischemic strokes.

Beneficial effects of dietary nucleosides and nucleotides There is substantial evidence (principally from research in animal models) that the presence of nucleotides or nucleosides in the diet helps cellular proliferation in the gut, in postoperative trauma, and in the development of the immune response. Although dietary purines are not taken into circulation from the gut, purine nucleotides influence the transcription of several genes in intestinal cells. Nucleotides based on both adenosine and uridine can activate the purinergic receptors on a wide range of cell types. In lymphocytes and other cells, synthesis of nucleotides *de novo* is expanded dramatically when a signal for proliferation is received; the rate of pyrimidine biosynthesis increases more than that of purine biosynthesis. Thus, nucleotides are considered to be 'conditionally essential' since their provision in the diet will provide help through the salvage system where cells are dividing rapidly or where other nutrients, used as precursors, are scarce.

Human milk contains maternal cells, providing nucleic acids, and also nucleosides (particularly cytidine and uridine), and nucleotides equivalent to 10–20 mg/day. Cow's milk contains the *de novo* intermediate orotic acid, which can be taken up by erythrocytes and converted to UMP. Dietary nucleotides have been shown to promote the incorporation of essential fatty acids into membrane lipids in healthy newborn infants and to enhance the integrity and maturation of the intestine and of the immune system. Thus, these components may contribute to the improved immunity seen in breast-fed infants. Many infant formulas are now supplemented with nucleotides/nucleosides, but usually in lower concentrations than in human milk.

Purine ribonucleotides as flavor-enhancing additives The purine 5′-nucleotide monophosphates IMP and GMP, derived from degradation of RNA, have received much attention as the taste-active components in a variety of seafoods and meat. Both IMP and GMP enhance the *umami* flavor generated by monosodium glutamate (MSG). This flavor was generated only by the purine 5′-nucleotides, and not by the pyrimidine nucleotides CMP and UMP, by interaction with receptors on the specific *umami* tastebuds in the mouth. Since ATP is the major free nucleotide in muscle cells, its breakdown into the flavor-enhancing IMP provides a scientific rationale for the improved palatability of meat or game birds that have been hung for several days

after slaughter. Similarly, the distinctive flavors of several cheeses are related to the metabolism, by bacteria, of the characteristic range of nucleotides present in the original milk.

Toxicology

Pathophysiology of Genetic Metabolic Disorders in Nucleotide Metabolism

The important physiological roles played by the nucleic acid precursor rNTP and dNTP molecules in humans has become apparent since the 1970s by the recognition of 28 different inborn errors of purine and pyrimidine metabolism. The spectrum of clinical manifestations ranges from fatal immunodeficiency syndromes to muscle weakness, severe neurological deficits, anemia, renal failure, gout, and urolithiasis (uric acid kidney stones).

Patients in whom the gene encoding a crucial enzyme is absent or defective harbor an 'experiment of nature,' equivalent to the artificial 'knockout' of individual genes in animals. The metabolic consequences of the disorders have highlighted the importance of individual steps in the nucleotide pathways to a particular cell or tissue, particularly the need for intact pathways for nucleic acid synthesis as well as for metabolism, degradation, and recycling of nucleotides.

It is evident from two genetic disorders associated with immunodeficiency that rapid turnover of DNA from cells of the immune system normally produces significant amounts of free deoxyribonucleotides in the bone marrow, which must be degraded to the corresponding base or nucleoside for recycling. The absence of either of two enzymes critical to this pathway, adenosine deaminase and purine nucleoside phosphorylase, results in the accumulation of dATP or dGTP, respectively. Each of these nucleotides inhibits ribonucleotide reductase and may also lead to misincorporation during DNA synthesis. The lymphocytes are particularly sensitive, resulting in a potentially fatal immunodeficiency syndrome with a clinical course similar to that in AIDS. Other rapidly dividing cells, such as those of the skin or the gut, are also affected. These disorders highlight the sensitivity of lymphocytes to the efficient removal of waste from DNA catabolism and, in fact, have provided the basis for development of novel immunosuppressant drugs based on inhibitors of these enzyme activities.

Role of Nucleotide Analogs as Cytotoxic and Antiviral Drugs

DNA in the chromosomes is normally protected by active repair systems against damage by a variety of chemical and physical agents, including ionizing radiation and ultraviolet light. Proofreading activities guard against the incorporation of mismatched nucleotides during DNA replication or transcription. In contrast, the use of certain nucleotide analogs as drugs depends on their incorporation into DNA—the chemical must be recognized and used by the replication enzymes but must prohibit further elongation of the nucleic acid chain. Analogs used in HIV therapy are incorporated by the reverse transcriptase of the virus and bring the reaction to a halt. Toxicity associated with several analogs is known to arise from erroneous incorporation into the patient's mitochondrial DNA because of less stringent proofreading by the mitochondrial DNA polymerase enzyme. Azidothymidine remains one of the most effective and least toxic drugs for AIDS, albeit it is now usually taken in triple therapy.

Nucleotide analogs have been used to inhibit the *de novo* pathways for the synthesis of the precursor nucleosides and nucleotides, leading to depletion of metabolites and imbalance of dNTPs, and hence to misincorporation of nucleotides in RNA or DNA, respectively. Malaria and other parasites rely exclusively on *de novo* pyrimidine biosynthesis; thus, they may be susceptible at drug doses that do not affect the host because the human body can obtain nucleotides from the salvage pathway. Similarly, because of the increased requirement for nucleotides in rapidly proliferating cells, almost all the enzyme reactions (**Figure 2**) have been investigated as potential targets for treatment of cancer, inflammation, or to prevent rejection of transplanted organs. Again, combinations of drugs with different modes of action have often proved most effective.

Toxicity of Exogenous Nucleic Acids to Humans

Dietary nucleic acids are digested fully to their component nucleosides and bases, so nucleic acids are not absorbed *per se* into the body. The potential toxicity to humans of dietary purine bases arises principally because primates lack expression of the gene for the hepatic enzyme uricase. Excess purines are therefore converted to uric acid rather than to the extremely soluble allantoin, as in most other mammalian species. In reptiles, snakes, spiders, and birds, uric acid is the end product of the metabolism of all nitrogenous compounds, analogous to urea in mammals. The main advantage to using this insoluble end product is that there is no obligatory water loss for its excretion, as there is for urea. Consequently, uric acid can be excreted as a slurry by these animals—an evolutionary adaptation enabling survival in arid environments. In contrast, in

humans, excess uric acid may accumulate in the tubules of the kidney as uric acid stones or as crystals in the interstitium, resulting in renal disease. Likewise, crystals can accumulate in the joints or subcutaneously, giving rise to gout symptoms. The most common and effective treatment for gout is the well-known drug allopurinol, which prevents conversion of xanthine to uric acid. The uricase enzyme may be prescribed as short-term therapy to avoid tumor lysis syndrome (the release of massive amounts of nucleic acids when cancerous cells are destroyed by chemotherapy or radiotherapy).

Although the uricase gene appears to be present in human cells, the promoter is not activated, so no enzyme activity is detected in the liver. The biological value of the normal circulating concentration of urate has been debated, but urate ions may provide more than half the antioxidant activity in the plasma of primates.

Not only is uric acid a relatively insoluble form of metabolic waste but also plasma concentrations are kept high because the renal proximal tubule reabsorbs approximately 90% of the filtered uric acid (**Figure 4**) before it can be excreted in the urine. Net reabsorption is higher in healthy males (92%) than in females (88%) and is lower in children of either sex (80–85%). This difference explains the higher plasma uric acid in adult males and means that uric acid is circulating in their plasma at concentrations close to its solubility limit. Reabsorption is much higher (95%) in middleaged males with 'primary gout'—and increasingly in men and elderly women who are treated with thiazide diuretics for high blood pressure. Poor excretion of uric acid is also seen in a familial renal disease associated with gout symptoms in children and young adults of both sexes. Without treatment to lower the uric acid accumulation, the kidney function may be seriously diminished.

Origin of the Body Uric Acid Pool in Humans

The body pool of urate, and hence the plasma urate concentration, is the result of a balance between production, ingestion, and excretion. The main causes of high plasma uric acid concentrations are high intake of exogenous nucleic acid in the diet and overproduction of endogenous purine. Eating less meat, seafood, and other high-purine foods (**Tables 1** and **2**) leads to a lower dietary intake of nucleic acids. In contrast, subjects with genetic defects that remove the usual controls on purine biosynthesis may have overwhelmingly high endogenous levels of the waste product, uric acid.

The contribution of the two sources can be assessed by placing the subject on a purine-free diet for 1 week and measuring the urinary uric acid. In this way, fewer than 5% of patients with gout have been found to excrete abnormally large amounts of urate (>3 mmol/day) derived from endogenous

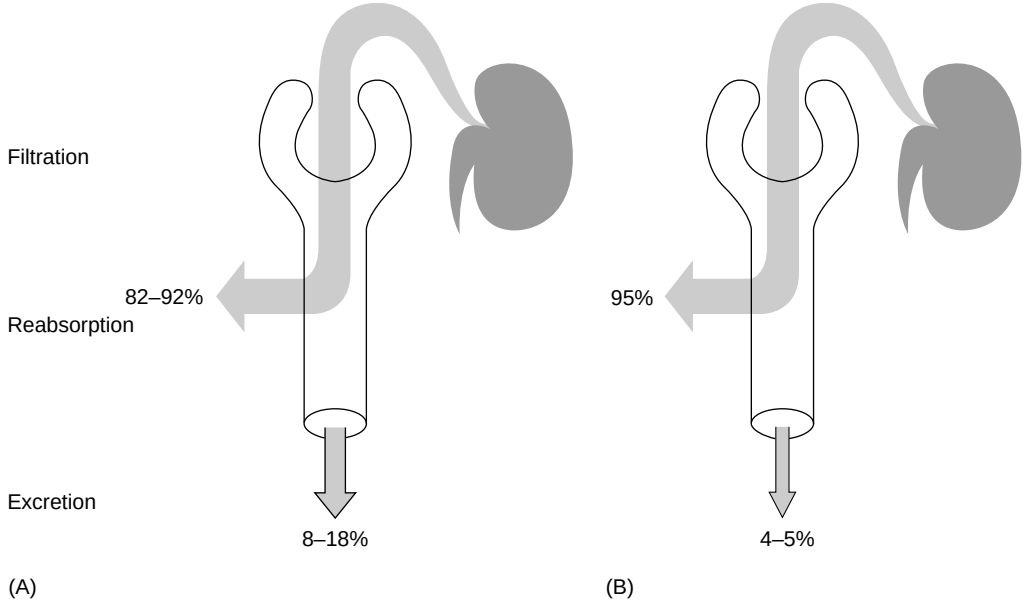


Figure 4 Schematic diagram showing the role of the kidney in influencing plasma uric acid concentration. In the brush border membrane of the proximal tubule, reabsorption by the urate anion exchanger and secretion via a voltage-sensitive pathway both occur. This results in a mean fractional urinary excretion of only 8–18% of the filtered load in healthy subjects and 5% or less in 'primary' gout. The net uric acid reabsorption is (A) higher in healthy men (92%) than in women (88%) and lower in children of either sex (82%), and (B) much higher in middle-aged males with primary gout (95%).

Table 1 Reference guide to the purine content of foods

Foods and beverages rich in nucleic acids/purines

Offal: sweetbreads, liver, kidney, heart, and paté

Wild or farmed game meats (venison, pheasant, rabbit, hare)

Seafoods: sardines, sprats, herring, bloaters, anchovies, fish roe, caviar, taramasalata, trout or salmon, lobster, crab, prawns

Vegetables: asparagus, avocado pears, peas, spinach, mushrooms, broad beans, cauliflower

Pulses and grains: legumes, pulses and soya products such as bean curd, tofu, Quorn

Cereals: all bran, oat, rye, or wheat cereals and products; whole meal, rye, and brown breads

Other: beer and yeast extracts/tablets (Barmene, Tastex); meat or vegetable extracts (Marmite, Vegemite, Bovril, Oxo)

Foods that are moderate or low sources of purine

Beef, lamb, pork (steak or chops), bacon, ham, sausages, some poultry, tongue (all should be eaten in moderation)

Carrots, parsnip, potatoes, lettuce, leeks, cabbage, sprouts, marrow, courgettes

Peanuts, cashew nuts

White bread or flour, cakes, scones, biscuits, cereals

Some fish (see **Table 2**)

Foods and beverages that are purine-free

Milk, cheese, eggs, butter, margarine, cream, ice cream

Sugar, jam, marmalade, honey, sweets

Cucumber, tomato, onions, pumpkin

Fresh, cooked or tinned fruits, nuts

Puddings, custards, yogurt

Fruit juices, soft drinks

Table 2 Concentrations of purines in some common foods and beverages[a]

Food	*Purine (mg/100 g)*	*Protein (mg/100 g)*
Meat		
Beef liver	333	19.7
Beef kidney	285	15.4
Beef heart	285	16.8
Beef tongue	167	16.4
Beef steak	151	19.5
Calf liver	348	19
Sweetbreads	1212	19.6
Veal cutlet	152	19.2
Sheep kidney	312	16.8
Lamb chop	196	14.9
Pork liver	289	22
Pork cutlet	164	16.4
Bacon	85	9.1
Ham	136	19.5
Sausage (beef)	79	13.8
Sausage (pork)	66	11.5
Rabbit	118	20.4
Venison	156	20
Vegetables		
Asparagus	32	2.1
Cauliflower	32	2.1
Celery	20	1.1
Kohlrabi	44	2.1
Mushrooms	72	3.5
Peas	72	6.7
Spinach	96	2.2
Dried legumes		
Split peas	195	21
Red bean	162	20
Lentils	222	28
Haricot beans	230	22
Lima bean	149	21
Other		
Bovril	340	18
Marmite	356	2
Oxo cubes	236	10
Yeast extracts	2257	46
Poultry		
Chicken flesh	181	20.6
Chicken liver	372	22.1
Chicken heart	223	18
Duck	181	16
Goose	177	16.4
Turkey	239	20.1
Fish, seafoods		
Anchovies	411	20
Bass	73	19.5
Bloaters	133	22.6
Bream	72	19.7
Cod	62	18
Crab	61	19.2
Clams	136	17
Eel	108	18.6
Fish cakes	36	12.1
Herring	378	17
Kippers	91	21.2
Lobster	100	20
Lemon sole	54	19.9
Mackerel	246	29
Plaice	53	18.1
Salmon	250	23
Sardines	345	23
Scallops	117	22.3
Sprats	250	25.1
Squid	135	15
Trout	92	19.2
Canned seafoods		
Anchovies	321	30
Herring	378	17
Mackerel	246	26
Oysters	116	6
Salmon	88	26
Sardines	399	24
Shrimp	231	22
Tuna	142	29

[a]Results are recorded relative to 100 g of food for purine and for protein, although serving size for each ingredient may be larger or smaller than 100 g.

purines. In these cases, overproduction of purine nucleotides leading to excess uric acid can be traced to a genetic defect. Two such sex-linked disorders are hypoxanthine–guanine phosphoribosyltransferase (HPRT) deficiency and phosphoribosyltransferase superactivity (PRPS). Boys presenting in infancy usually have severe and eventually fatal neurological deficits. Neurological problems are milder or absent,

and only gout may be evident, in those presenting as adolescents. It is important for clinicians to be aware of these disorders, especially when encountering a young patient with gout or an older male with a history dating back to adolescence. In some families, siblings are also affected, and although the gout symptoms can be alleviated, other aspects are less amenable to treatment and genetic counselling should be given.

Gout

Gout is a painful, acute form of arthritis caused by the accumulation of crystals of uric acid in the joints (typically the great toe is the first site to be affected). The pain may be relieved by antiinflammatory drugs or by colchicine, and the accumulation of urate is halted by the drug allopurinol, which inhibits xanthine dehydrogenase. Nevertheless, adopting a low-purine diet has an important role in alleviating the effects of gout.

Historically, 'primary' gout—affecting predominantly middle-aged males—has long been associated with excessive consumption of 'rich' food and drink (**Tables 1** and **2**). Until World War I, affluent European gentlemen habitually consumed vast nucleic acid-rich meals including many different courses and meats. Alcoholic drinks also played their part, with beer being particularly rich in purines derived from yeast RNA and port a potential cause of lead poisoning from glass bottles and decanters.

Both hyperuricemia and gout are relatively common in Polynesians due to a genetic defect in excretion of urate and in Australasians, traditionally high consumers of nucleic acid-rich seafood, meat, and beer. In such countries, the prevalence of gout is as high as 10% compared with 1–4% in Europe. In Europe, the prevalence is higher in countries such as France where the consumption of seafoods and paté is high. The role of diet in the etiology of primary gout is confirmed by the fact that during and immediately after the two world wars this type of gout was virtually unknown. Gout, a hitherto little known disease, is now more common worldwide where affluence or the consumption of meat have increased.

Urolithiasis (Kidney Stones)

Although modest overindulgence in purine-rich food by normal subjects does not precipitate gout, it can predispose to uric acid lithiasis. Uric acid stones are relatively common in countries where the consumption of nucleic acid-rich beverages and food is high and in hot climates if insufficient fluids are consumed. Health foods such as yeast tablets, Spirulina, or supplements containing nucleotides also contribute to uric acid lithiasis.

A number of compounds, such as vitamin C, increase uric acid clearance and thus can precipitate urolithiasis. Perhaps not so well recognized is the uricosuric effect of a high-protein diet and the fact that purine-rich foods also predispose to renal calcium stones. This may be because many purine-rich foods, such as spinach, are equally rich in calcium oxalate. Approximately 25% of vitamin C intake is also excreted as oxalate, which can compound the problem.

The solubility of uric acid is very sensitive to the pH of the urine, which in turn may be altered by components of the diet. The solubility of uric acid in urine at pH 5.0 is low (approximately 1 mmol l^{-1}), but it can be increased 12-fold at pH 8.0 by alkalinising regimens, such as sodium bicarbonate or potassium citrate.

Exacerbation of kidney stone formation by dietary nucleic acids in inherited purine disorders Excess uric acid from dietary purines can also precipitate symptoms that may draw attention to milder forms in adults of HPRT deficiency or PRPS superactivity. A third genetic defect raises levels of adenine, which is converted by XDH to the even more insoluble uric acid analog, 2,8-dihydroxyadenine (2,8-DHA). Undiagnosed, such subjects have progressed to renal failure and even death. One child presenting in coma had a diet of pulses and grains, which have a particularly high adenine content. Since the accumulation of 2,8-DHA is treatable with allopurinol, such nephropathy can be avoided if the defect is recognized and the consumption of nucleic acid-rich foods reduced to a minimum.

Dietary Sources

Nucleic Acid Content of Foods

The nucleic acid content of different foods is expressed generally in terms of purine equivalents, with the data derived from the hydrolysis of nucleic acids and free nucleotides to the constituent bases. Careful analysis by Robert McCance, Elsie Widdowson, and colleagues since the 1930s forms the basis of tables of the composition of foodstuffs.

Foods may be classified into three groups: high, low, or essentially purine free (**Table 1**). As a general rule, growing organisms such as yeast, or rapidly metabolizing tissues such as liver, will be rich in both DNA and RNA. Seeds, grain, and fish eggs are good sources of the genetic material, DNA. Muscle tissue is an excellent source of nucleotides,

such as the energy source ATP. Extracts of meat and yeast have very high purine contents but are usually eaten in small quantities. Some vegetables may provoke gout attacks by virtue of their oxalic acid content rather than that of purines, but legumes, fast-growing parts of brassicas, and asparagus tips may also have significant nucleic acid content. Fats, white flour, sugar, and fruit juices have been separated from the 'living' part of the food and so they are poor sources of nucleic acids.

Table 2 provides data for specific foodstuffs, obtained from the Documenta Geigy Chemical Composition of Foodstuffs tables. The ideal diet for subjects at risk of gout or of uric acid lithiasis is no more than one meat meal per day, using only the low-purine meat and vegetables indicated.

See also: **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Choline and Phosphatidylcholine**. **Gout**.

Further Reading

Becker MA (2001) Purines and pyrimidines. In: Scriver CR, Beaudet AL, Sly WS, and Valle D (eds.) *The Metabolic and Molecular Basis of Inherited Disease*, 8th edn., pp. 2513–2537. New York: McGraw-Hill.

Carver JD (2003) Advances in nutritional modifications of infant formulas. *American Journal of Clinical Nutrition* **77**(supplement): 1550S–1554S.

Christopherson RI, Lyons SD, and Wilson PK (2002) Inhibitors of de novo nucleotide biosynthesis as drugs. *Accounts of Chemical Research* **35**: 961–971.

Diem K and Lentner C (eds.) (1970) *Scientific Tables—Chemical Composition of Foodstuffs*, 7th edn., pp. 230–243. Basel: Geigy.

Fuke S and Konosu S (1991) Taste-active components in some foods: A review of Japanese research. *Physiology and Behaviour* **49**: 863–868.

Grahame R, Simmonds HA, and Carrey EA (2003) *Gout: The 'At Your Fingertips' Guide* London: Class Publishing.

Lee H, Hanes J, and Johnson KA (2003) Toxicity of nucleoside analogues used to treat AIDS and the selectivity of the mitochondrial DNA polymerase. *Biochemistry* **42**: 14711–14719.

Rolls ET (2000) The representation of umami taste in the taste cortex. *Journal of Nutrition* **130**: 960S–965S.

Secades JJ and Frontera G (1995) CDP-choline: Pharmacological and clinical review. *Methods and Findings in Experimental and Clinical Pharmacology* **17**(supplement B): 1–54.

Uauy R (1989) Dietary nucleotides and requirements in early life. In: Lebenthal E (ed.) *Textbook of Gastroenterology and Nutrition in Infancy*, pp. 265–280. New York: Raven Press.

Zöllner N and Gresser U (eds.) (1991) *Urate Deposition in Man and Its Clinical Consequences*. Berlin: Springer-Verlag.

NUTRIENT–GENE INTERACTIONS

Contents

Molecular Aspects

C D Berdanier, University of Georgia, Athens, GA, USA
H C Freake, University of Connecticut, Storrs, CT, USA

The completion of the sequencing of the human genome has resulted in a broadening of focus to include the investigation of the complex environment in which these genes operate. Although the term 'gene' refers to a specific sequence of DNA, the biological effects of that gene are manifest through its expression as a protein or peptide product. Nutrients affect the expression of genes in a variety of ways. Nutrients are required for the synthesis and packaging of DNA. Some have specific effects on the synthesis of messenger RNA (i.e., either suppress or enhance transcription). Others affect the synthesis of the pyrimidine and purine bases used for DNA and RNA synthesis. Some nutrients have an overall effect on protein synthesis, whereas others influence the translation of the messenger RNA into protein or the post-translational modification of the newly synthesized protein. Still others can affect the outcome of gene expression by influencing the environment in which the gene product functions. This article outlines the process of gene expression, focusing on the ways in which it is influenced and regulated by particular nutrients.

DNA Characteristics

The characteristics of every living creature are dictated by the genetic material, DNA. Nuclear DNA is organized into units called chromosomes, of which there are 46 in the human. The chromosomes are found in pairs and contain the individual units called genes. DNA is a double-stranded helix composed of four bases, two pyrimidines (cytosine and thymine) and two purines (adenine and guanine), that are joined together by ribose and phosphate groups (**Figure 1**). DNA is formed when the bases are joined through phosphodiester bonds using ribose as the common linkage. The phosphodiester linkage is between the 5′ phosphate group of one nucleotide and the 3′ OH group of the adjacent nucleotide. This provides a direction (5′ to 3′) to the chain. The bases are hydrophobic and contain charged polar groups. These features are responsible for the helical shape of the nuclear DNA chain. A double helix forms when the bases of each chain interact through hydrogen bonding.

The DNA base sequence is unique for every protein and peptide that synthesized in the body. The sequence of these bases determines the genotype of the individual for each gene product. Although only four different bases are used for the DNA, it is the sequence of these bases that determines the product being produced. Each gene product is uniquely derived from a specific gene. Although all cells contain the same DNA, not all genes are expressed in every cell; some are particular to specific cell types. Thus, the function of DNA is to determine not only the particular characteristics of the individual but also the properties of each cell through the provision of a multitude of genes, each coding for a particular protein found in that cell. Therefore, it functions to transmit genetic information from one generation to the next in a given species and ensures the identity of specific cell types.

The Human Genome Project has detailed the specific base sequence of nuclear and mitochondrial DNA in the human cell. The identification of each gene and its corresponding controls of expression have not been completely elucidated. Although the nuclear genome has been sequenced, it has not been completely mapped; that is, the location, within the DNA, of each gene (and its promoter region) and the identification of the protein or peptide it encodes

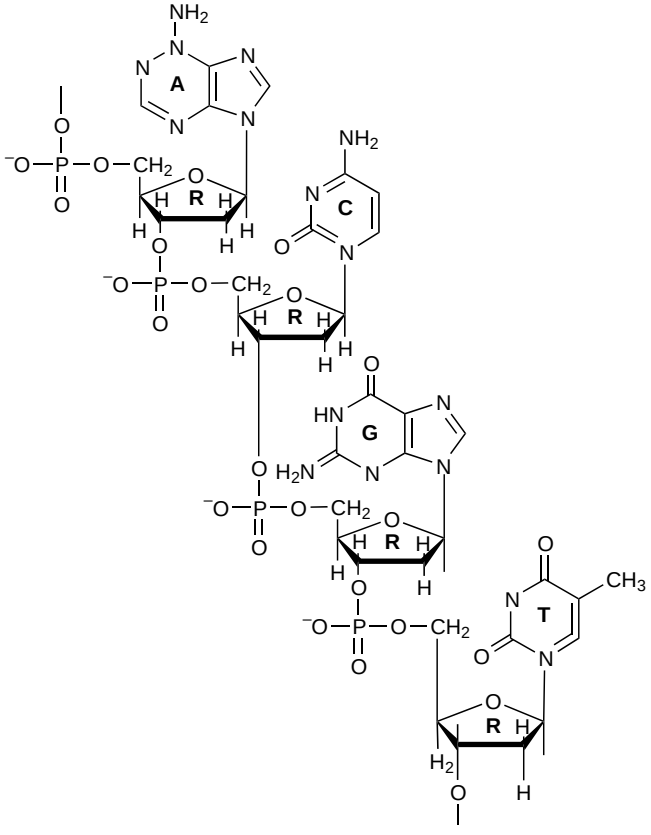


Figure 1 Representation of a segment of DNA showing the phosphodiester bond that uses ribose as the common link between the bases. A, adenine; C, cytosine; G, guanine; T, thymine; R, ribose.

have not been fully determined. In addition, we do not know all the details of the regulation of gene expression. Some genes have been intensely studied, whereas others have yet to be identified. In contrast, the genome in the mitochondria has been completely sequenced and mapped. It is a very small genome encoding only 13 gene products (components of the mitochondrial respiratory chain) under the control of a single promoter sequence, the D-loop. Despite its small size and apparent simplicity, however, we know even less about the regulation of its expression than we know about some of the nuclear-encoded genes.

DNA Synthesis

In the adult, some cell types are extremely long-lived (e.g., neuronal cells), whereas others last only a few days and therefore need constant replacement (epithelial cells, e.g., intestine and skin). This synthesis requires a number of micronutrients, protein, and energy. Should any of these be in short supply, symptoms of malnutrition will be observed, especially in those cell types that have very short half-lives. Typical of niacin deficiency (pellagra), for example, are skin lesions. As epithelial cells die and must be replaced, niacin is needed for this replacement. All the components of the new cells including DNA must be synthesized. The purines and pyrimidines that comprise DNA must be synthesized and this requires energy as well as micronutrients (niacin, riboflavin, pyridoxine, folic acid, vitamin B_{12}, copper, iron, sulfur, zinc, magnesium, and phosphorus). Anemia is another characteristic of malnutrition. Not only must new blood cells be made but also the essential ingredient of these cells, hemoglobin, must be synthesized. Among the nutrients needed for red blood cell synthesis are iron, copper, magnesium, folic acid, vitamin B_{12}, vitamin B_6, and, of course, energy and protein sufficient to support this synthesis.

The nutritional requirements for new cell synthesis are much greater in growing individuals than in adults because growth and cell division are much greater. Thus, energy and protein deficiency can be particularly detrimental. In addition, an adequate supply of specific micronutrients is crucial. For example, zinc deficiency was first described in teenage boys who were stunted and also sexually undeveloped. This report showed that zinc was required for both growth in general and the development of specific organ systems. Folic acid is required for DNA synthesis. Meeting this requirement is crucial during embryonic development. Inadequate folate intake by prepregnant and pregnant women can result in neural tube defects due to insufficient cell division during this time period. Not all women are so affected; there may be genetic differences in the need for folate that in turn determine whether the embryo is affected.

Transcription

Messenger RNA (mRNA) synthesis using DNA as the template is called transcription. The mRNA carries genetic information from the DNA of the chromosomes in the nucleus to the surface of the ribosomes in the cytosol. It is synthesized as a single strand. Chemically, RNA is similar to DNA. It is an unbranched linear polymer in which the monomeric subunits are the ribonucleoside 5′ monophosphates. The bases are the purines (adenine and guanine) and the pyrimidines (uracil and cytosine). Thymine is not used in mRNA. Instead, uracil is used. This base is not present in DNA. Messenger RNA is much smaller than DNA and is far less stable. It has a very short half-life (from seconds to minutes or hours) compared to that of nuclear DNA (years). Because it has a short half-life, the purine and pyrimidine bases that are used to make mRNA must be continually resynthesized. This requires the same array of nutrients noted previously for DNA synthesis.

The synthesis of mRNA from DNA occurs in several stages: initiation, elongation, editing (processing), and termination. Initiation of transcription (the synthesis of mRNA) occurs when factors that serve to stabilize nuclear DNA are perturbed. Perturbation signals pass in to the nucleus and stimulate transcription. A small portion of the DNA (~17,000 bases) is exposed and used as the template for mRNA synthesis. The exposed portion also contains one or more sequences that have control properties with respect to the initiation of transcription. This region is called the promoter region and represents a key site for nutrient interaction. The promoter region precedes the start site of the structural gene and is said to be upstream of the structural gene. Those bases following the start site are downstream. The exposed DNA contains groups of bases called exons and introns. The introns are noncoding and are removed by editing prior to the movement of the mRNA from the nucleus to the cytosol.

Transcription is highly regulated. The DNA in all cell types is identical. However, not all of this DNA is transcribed in all cells all the time. Only certain genes are activated and transcribed into mRNA and subsequently translated into protein or peptides. As mentioned previously, these gene products give

the individual cell type its identity. Central to this regulation are protein:DNA interactions and protein:nutrient interactions. At initiation, basal transcription factors recognize and bind to the start site of the structural gene. They form a complex with RNA polymerase II, an enzyme that catalyzes the formation of mRNA. Transcription factors bind to particular base sequences, called response elements, in the promoter region of the DNA that are upstream of the transcription start site (**Figure 2**). Each gene promoter contains a characteristic array of response elements, and these will determine to which signals the particular gene responds. Transcription factors also bind nutrients, and it is here that some nutrients have their effects on gene expression.

The regulation of transcription often occurs through the regulation of transcription factors. These factors can be regulated by the rates of their synthesis or degradation, by phosphorylation or dephosphorylation, by ligand binding, by cleavage of a pro-transcription factor, or by release of an inhibitor. One class of transcription factors important for nutrition is the nuclear hormone receptor superfamily that is regulated by ligand binding. Ligands for these transcription factors include retinoic acid (the gene active form of vitamin A), fatty acids, vitamin D, thyroid hormone, and steroid hormones. These receptors are proteins with a series of domains. The retinoic acid receptor can serve as an example. Its ligand-binding domain recognizes and binds with high affinity the nutrient signal, retinoic acid. The DNA-binding domain gives gene specificity. It binds to a segment of the gene promoter that contains its corresponding response element, the retinoic acid response element (RARE). A transactivation domain then signals the effective occupation of this response element to the gene as a whole, including RNA polymerase II and its associated proteins. There are additional factors responsible for mediating this interaction between nutrient receptor and the transcription process. They include coactivating proteins, which stimulate transcription, and corepressor proteins, which can cause inhibition of transcription from a particular protein. In general, nutrients can signal the activation of transcription of some genes while at the same time turning off the transcription of others.

An interesting additional feature of this superfamily of nuclear hormone receptors is that they contain two zinc atoms in their DNA-binding domains. Each zinc is bound by four cysteine residues and causes the folding of the protein in a finger-like shape that binds DNA. The zinc ion plays an important role in gene expression because of its central use in the zinc finger of a wide variety of DNA binding proteins. In the case of the receptor superfamily, although zinc is required for receptor function, there is no evidence that it plays a regulatory role. However, there are other transcription factors in which it does play a role. MTF-1 (metal response element (MRE)-binding transcription factor-1) responds to increasing zinc concentrations within the cell by translocating to the nucleus and activating the transcription of genes containing MREs in their promoter region. These genes include

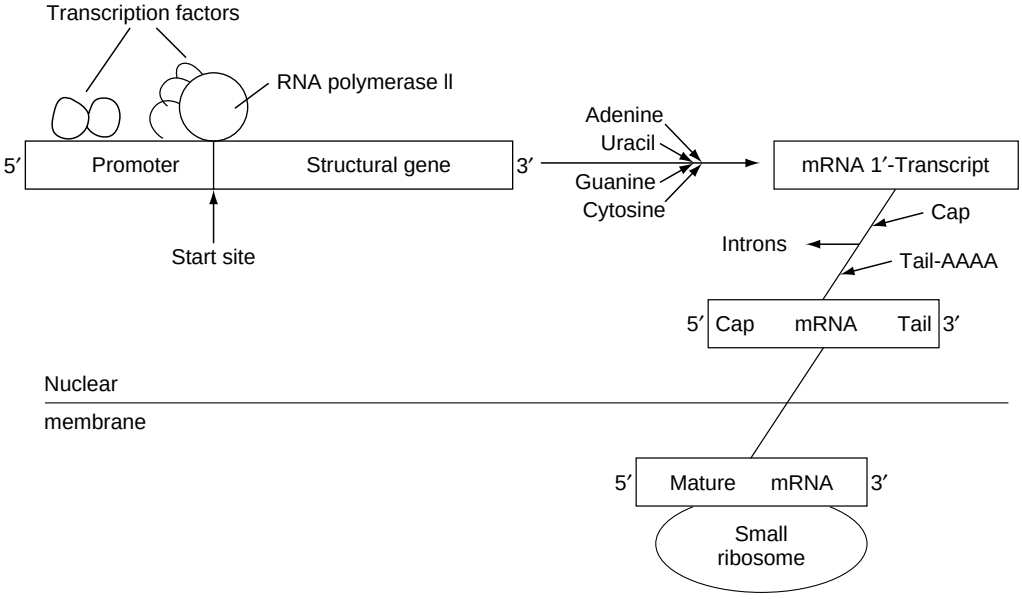


Figure 2 Schematic view of transcription.

metallothionein, which binds zinc and may play a key role in zinc homeostasis.

The direct binding of a nutrient signal to a transcription factor is perhaps one of the simpler ways in which nutrients impact gene transcription. There are other less direct but equally important mechanisms. Genes involved in cholesterol homeostasis are characterized by a sterol response element (SRE) in their promoter regions, which interacts with a sterol response element binding protein (SREBP). This protein is synthesized as a large precursor, incorporated into endoplasmic reticulum membranes, and is unavailable to function in gene regulation until it is cleaved and released. Limited cholesterol availability results in the cleavage and release of SREBP from the membrane compartment and its translocation to the nucleus. There it can perform its gene regulatory function by activating the transcription of genes for cholesterol synthesis as well as the LDL receptor gene. The LDL receptor facilitates cholesterol uptake by the liver. When it is abnormal due to a mutation in its gene, hypercholesterolemia results. The liver is unable to remove cholesterol from the blood and continues to synthesize it since SREBP remains active.

The metabolism and availability of macronutrients also influence gene transcription. Promoter elements have been described that allow a response to glucose (the carbohydrate response element (CHORE)). Although the specifics are unclear, the activity of the protein that binds this element responds to the metabolism of glucose and then stimulates the transcription of relevant genes—for example, those required for glucose metabolism (pyruvate kinase) and fatty acid synthesis (acetyl-coA carboxylase and fatty acid synthase). Fatty acids also influence gene expression. They can affect transcription by binding directly to their own transcription factor (the peroxisome proliferator activated receptor) and also indirectly by reducing the availability of SREBP within the nucleus. The latter mechanism provides a means for linking cholesterol and fatty acid metabolism with the cell.

Nutrients can also affect gene expression indirectly by regulating the release of hormones into the blood. Thus, glucose, in addition to having its own effects on gene expression through the CHORE, also stimulates insulin secretion from the pancreas. Insulin has its own transcriptional effects, often on the same genes that are regulated by glucose. In the postabsorptive state, insulin drops and glucagon is released. This hormone activates an intracellular signaling pathway that results in inhibition of genes involved in glucose metabolism and fatty acid synthesis and stimulation of genes involved in gluconeogenesis (e.g., phosphenolpyruvate carboxykinase). Taken as a whole, macronutrient availability regulates the expression of the complex set of genes responsible for macronutrient metabolism by an aggregate of direct and endocrine-mediated pathways.

Nuclear Processing of mRNA

Once the bases are joined together in the nucleus to form mRNA, it is edited with a reduction in size. Through editing and processing, less than 10% of the original mRNA actually leaves the nucleus. Editing and processing are needed because immature RNA contains all those bases corresponding to the DNA introns. The removal of these introns is a cut-and-splice process whereby the intron is cut at its 5′ end, pulled out of the way, and cut again at its 3′ end. After this group of bases is excised, the bases corresponding to the DNA exons are joined. This cut-and-splice routine is continued until all the introns are removed and the exons joined. Some genes can give rise to multiple protein products since not all exons are necessarily retained in the mature mRNA. Some editing of the RNA also occurs with base substitutions made as appropriate. The mRNA is capped at the 5′ end in a process that adds a guanine base and some methyl groups. Finally, a 3′-terminal poly A tail is added and the mature mRNA is ready to leave the nucleus and move to the cytoplasm for translation. The nucleotides that have been removed during editing and processing are either reused or degraded. Some mRNA is totally degraded, never leaving the nuclear compartment. This serves to control the amount of mRNA. Regulation of the amount of mRNA that leaves the nucleus is a key step in metabolic control.

mRNA Stability

The stability of mRNA can also be regulated within the cytoplasm. Some mRNA have very short half-life (seconds to minutes), whereas others have longer half-lives (hours). This is important because some gene products (i.e., hormones and cell signals) must be short-lived and the body needs to control/counterbalance their synthesis and action. A nutritionally important example of regulation of mRNA stability involves iron and the transferrin receptor. The transferrin receptor is the protein responsible for the uptake of iron into cells. The expression of the transferrin receptor is downregulated by iron in order to limit uptake and potential toxicity of the

mineral at times of high availability. This regulation is achieved through an iron regulatory protein. When iron is limited, this protein is bound to the 3′ untranslated region of the transferrin receptor mRNA. This serves to protect the mRNA from degradative attack and permits its continued translation into active protein. As iron concentrations rise, the binding protein becomes occupied with iron, which results in its dissociation from the transferrin receptor mRNA. The mRNA is then degraded more quickly, concentrations fall, and protein production is limited.

Translation

Following transcription is translation. Translation is the synthesis of the protein or peptide, the gene product. Translation occurs on the ribosomes; some ribosomes are located on the membrane of the endoplasmic reticulum and some are free in the cell matrix. Ribosomes consist of RNA and protein. Ribosomal RNA makes up a large fraction of total cellular RNA. Ribosomal RNA is synthesized via RNA polymerase I in the cell nucleus as a large molecule; there, this RNA molecule is split and leaves the nucleus as a large and a small subunit. The large ribosomal unit serves as the 'docking' point for the activated amino acids bound to the transfer RNA (tRNA). The mRNA is bound to the small ribosomal unit. The two ribosomal units reassociate in the cytosol for the translation step. tRNA is used to bring an amino acid to the large ribosome, the site of protein synthesis. Each amino acid has a specific tRNA. Each tRNA molecule is thought to have a cloverleaf arrangement of nucleotides. This arrangement allows the formation of the maximum number of hydrogen bonds between base pairs. Hydrogen bonding stabilizes the tRNA. tRNA also contains a triplet of bases that pair to a corresponding triplet found in the mRNA. This triplet is not identical to the mRNA triplet and is called the anticodon. The bases pair in a preordained manner: adenine to thymine, guanine to uracil, guanine to guanine, uracil to cytosine, inosine to adenine, and so forth. The amino acid carried by tRNA is identified by the codon of mRNA through its anticodon; the amino acid is not involved in this identification.

Translation takes place in four stages, as illustrated in **Figure 3**. Each stage requires specific cofactors and enzymes. The first stage involves the esterification of the amino acids to specific tRNAs. Each of these esterification reactions requires a molecule of ATP. Here again is an explanation of why the provision of energy is crucial to protein synthesis. If a protein contains several hundred amino acids, this step in translation will require several hundred molecules of ATP. Energy-deficient diets result in a shortfall in ATP and so protein synthesis is compromised.

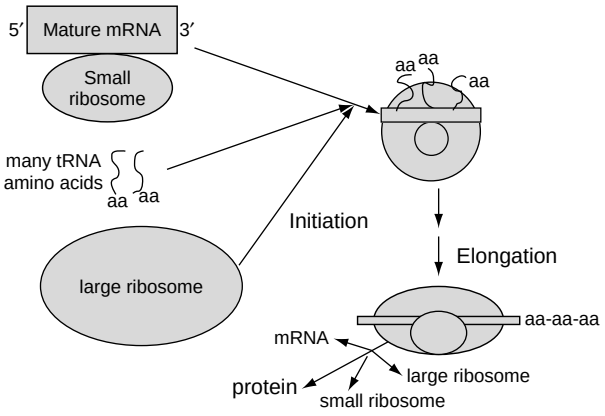


Figure 3 Schematic view of translation.

During the second stage of translation, polypeptide chain synthesis begins. mRNA binds to the small ribosome and an initiation complex is formed. The complex consists of the mRNA cap and the first activated amino acid–tRNA. The ribosome finds the correct reading frame on the mRNA by 'scanning' for an AUG codon. This is the so-called start codon. The large ribosomal unit then attaches and forms a functional ribosome. A number of specific protein initiation factors are involved in this step.

In the third stage of translation, the peptide chain is elongated by the sequential addition of amino acids from the amino acid–tRNA complexes. The amino acid is recognized by base pairing of the codon of mRNA to the bases found in the anticodon of tRNA, and a peptide bond is formed between the peptide chain and the newly arrived amino acid. The ribosome then moves along the mRNA; this brings the next codon into the proper position for attachment to the anticodon of the next activated amino acid–tRNA complex. The mRNA and nascent polypeptide appear to 'track' through a groove between the two ribosomal subunits. This protects the protein being synthesized from attack by enzymes in the surrounding environment.

The final stage of translation is the termination and release of the amino acid chain. The mRNA contains a stop codon that signals termination at the carboxy terminus. The carboxy-terminal amino acid, although attached to the peptide chain, is also esterified to its cognate tRNA–ribosome. A protein release factor promotes the hydrolysis of the ester link between the tRNA and the amino acid. Now the polypeptide is released from the ribosome and is

free to assume its characteristic three-dimensional structure.

Translation is influenced by nutritional status as well as by specific nutrients. Protein synthesis is dependent on the simultaneous presence of all the amino acids necessary for the protein being synthesized and on the provision of energy. If there is an insufficient supply of either, protein biosynthesis will not proceed at its normal pace. This is an example of the consequences of malnutrition with respect to gene expression. Malnourished individuals will not be able to support the full range of *de novo* synthesis of body proteins because their diets are energy poor and/or contain proteins of poor quality. This condition is known as protein-energy malnutrition. It is commonly found in children but may also be observed in adults under severe food deprivation.

An example of an effect of a nutrient on translation of a specific mRNA is that of iron in the synthesis of ferritin. Iron storage in cells occurs through chelation to a protein called ferritin. Ferritin synthesis is highly regulated by iron intake. In iron deficiency, the mRNA start site for ferritin translation is obstructed by an iron regulatory protein. This protein binds to the 5′ untranslated region and inhibits the movement of the 40s ribosome from the cap to the translation start site. When the diet contains sufficient iron and iron status is improved, the iron regulatory protein dissociates from the ferritin mRNA and translation proceeds. When iron availability is limited, the same iron regulatory protein binds to ferritin mRNA (to inhibit its translation) and to the transferrin receptor mRNA, as described previously (to prevent its degradation and ensure its translation). These exquisite mechanisms serve to maintain iron homeostasis.

Post-translational Protein Modification

After translation, the primary amino acid sequence is complete. The secondary and tertiary structure of the protein evolves via numerous interactions between amino acids via hydrogen bonding, disulfide bridges, and ionic bonds. The newly synthesized proteins can be further modified via post-translation reactions. Post-translational protein modification includes the association of various subunits of an enzyme or a carrier or a cell component. For example, the association of the four subunits that make up hemoglobin occurs after the initial synthesis of each of the subunits has occurred. Again, specific nutrients can influence the process. Another example is the post-translational carboxylation of the proteins osteocalcin and prothrombin. Osteocalcin and prothrombin each have glutamic acid-rich regions that, when carboxylated, allow the protein to bind significant amounts of calcium. Calcium binding is an essential feature of the functions of each of these proteins. The post-translational carboxylation of osteocalcin and prothrombin requires vitamin K. Should vitamin K be in short supply, this carboxylation will not occur (or will occur in only a limited way) and these proteins will not be able to bind calcium. Both must bind calcium in order to function. Hence, vitamin K deficiency is characterized by prolonged blood clotting times (inadequate calcium binding by prothrombin) and poorly mineralized bone (inadequate calcium binding by osteocalcin).

Some protein modifications occur in a nutritionally dependent reversible manner. The means whereby macronutrients influence their own metabolism at the level of transcription was outlined previously. In addition, the nutritionally regulated hormones insulin and glucagon influence enzyme activity by phosphorylation/dephosphorylation

Table 1 Examples of nutrient effects on gene expression

Nutrient	*Intermediary protein*	*Gene/gene product*	*Effect*
Cholesterol	Sterol response element binding protein	LDL receptor	Suppresses transcription
Fatty acids	Sterol response element binding protein	Fatty acid synthase	Suppresses transcription in liver
	Peroxisome proliferator activated receptor	Fatty acid binding protein	Increases transcription
Glucose	Carbohydrate responsive factor	Pyruvate kinase, acetyl-coA carboxylase	Increases transcription in liver
Iron	Iron regulatory protein	Ferritin	Increases translation
		Transferrin receptor	Destabilizes mRNA
Vitamin A	Retinoic acid receptor	Retinoic acid receptor	Increases transcription
		Collagenase	Decreases transcription
Vitamin D	Vitamin D receptor	Calcium binding proteins	Increases transcription
Vitamin K		Prothrombin, osteocalcin	Serves as cosubstrate for the post-translational carboxylation of glutamic acid-rich regions of these proteins
Zinc	MTF-1	Metallothionein	Increases transcription

mechanisms. This allows metabolic flux to respond to nutrient availability much more rapidly than would be possible with mechanisms dependent on new protein synthesis.

There are numerous examples of specific nutrients' effects on gene expression. Some of these effects concern the transcription of genes that encode enzymes or receptors or carriers that are important to the use of that nutrient. Examples are listed in **Table 1**. Many nutrients serve more than one function with respect to gene expression. Some influence both transcription and translation, whereas others serve to enhance the transcription of one gene while suppressing the transcription of another. Nutrient–gene interactions can result in either an increase or a decrease in specific mRNA, but there may be no increase in gene product or a measurable increase in gene product function. This speaks to the complicated nature of metabolic control. Simply synthesizing more message units or more enzyme protein does not automatically result in an increase in enzyme activity, an increase in a metabolic pathway, or an increase in a metabolic product. The processes of gene expression and metabolic regulation comprise a complex web of interactions in which nutrients are major and diverse players.

See also: **Carbohydrates**: Regulation of Metabolism. **Cholesterol**: Factors Determining Blood Levels. **Fatty Acids**: Metabolism. **Folic Acid**. **Iron**. **Nutrient–Gene Interactions**: Health Implications. **Vitamin A**: Biochemistry and Physiological Role. **Vitamin K**. **Zinc**: Physiology.

Further Reading

Berdanier CD (1998) In *Advanced Nutrition: Micronutrients*. Boca Raton, FL: CRC Press.

Berdanier CD (2000) *Advanced Nutrition: Macronutrients*, 2nd edn. Boca Raton, FL: CRC Press.

Eisenstein R (2000) Iron regulatory proteins and the molecular control of mammalian iron metabolism. *Annual Review of Nutrition* **20**: 627–662.

Horton JD, Goldstein JL, and Brown MS (2002) SREBPs: Activators of the complete program of cholesterol and fatty acid synthesis in the liver. *Journal of Clinical Investigation* **109**: 1125–1131.

Jump DB and Clark SD (1997) Regulation of gene expression by dietary fat. *Annual Review of Nutrition* **19**: 63–90.

Mangelsdorf DJ, Thummel C, Beato M *et al.* (1995) The nuclear receptor superfamily: The second decade. *Cell* **83**: 835–839.

Matthews JM and Sunde M (2002) Zinc fingers—Folds for many occasions. *IUMBM Life* **6**: 351–355.

Moustaid-Moussa N and Berdanier CD (eds.) (2001) *Nutrient–Gene Interactions in Health and Disease*. Boca Raton, FL: CRC Press.

Health Implications

C D Berdanier, University of Georgia, Athens GA, USA
H C Freake, University of Connecticut, Storrs, CT, USA

Food, and the nutrients it contains, has long been known to influence the health and well-being of humans. Included in the list of nutrition-related diseases that afflict humankind are the specific nutrient deficiency diseases and several of the chronic diseases, including some of the most common ones. Each of these conditions has both a nutrition component and a genetic component. With respect to the nutrient deficiency diseases, there is considerable variation in nutrient requirements since the genetics of the consumer dictates how much of each essential nutrient is needed. If the individual's requirement is met, the deficiency disease is prevented. **Table 1** lists the essential nutrients and the symptoms of the deficient state caused by inadequate intake of the nutrients. Some symptoms (i.e., anemia) characterize several different nutrient deficiency states.

In addition to the diseases that are clearly nutritionally related, many of the chronic diseases have a nutritional component. They are also influenced by genetics, and often these two factors interact so that when certain dietary behaviors are found with a susceptible genotype, the disease results. A third kind of nutrient gene interaction relevant to health occurs when dietary constituents either promote or protect against changes in DNA that result in aging or disease. Nutrients that affect the redox balance within the cell are important in this case.

Some disease states result from alterations in a single gene. Function of the gene product is compromised and the specific pathology develops. Acrodermatitis enteropathica is an example of this, in which the affected individual has impaired zinc absorption due to a mutation in the gene encoding a zinc transporter. Menkes' disease is another example; copper absorption is impaired due to an X-linked mutation in the protein needed to release absorbed copper from the enterocyte into the circulation. People with this disorder develop symptoms of copper deficiency. Single gene mutations have been identified that affect the use of a single nutrient. However, in many genetically determined instances of nutrient malabsorption or abnormal use, the situation is more complex and the disease state may develop as a result of small changes in several genes. These situations may be more common than individual

Table 1 Nutrient deficiency disorders[a]

Nutrient	*Disease: signs of inadequate intake*
Ascorbic acid	**Scurvy**: hyperkeratosis, congestion of the hair follicles, skin hemorrhages, conjunctival lesions, gum swelling and bleeding, peripheral neuropathy with hemorrhages into the nerve sheaths, painful joints, deformed chests in children
Thiamin	**Beriberi**: muscle tenderness and peripheral neuropathy, edema, fast pulse, high blood pressure, decreased urine volume, disorientation, memory loss, ataxia, jerky movements of the eyes
Riboflavin	Ill-defined symptoms that are not necessarily related to inadequate intake: poor growth, poor appetite, cracks in the corners of the mouth, dermatitis of the scrotum
Niacin	**Pellagra**: black, roughened skin especially in areas exposed to sunlight; insomnia; loss of appetite; sores in mouth and tongue; indigestion; diarrhea; mental confusion; nervousness; headache; apprehension; forgetfulness
Vitamin B_6	Ill-defined symptoms: poor growth, muscular weakness, fatty liver, convulsions, anemia, reproductive impairment, edema, neural degeneration, enlarged adrenal glands
Folic acid	**Anemia**; macrocytic anemia. Neural tube defects in infants are associated with inadequate folic acid intakes of the mother during the first trimester of pregnancy
Vitamin B_{12}	**Pernicious anemia**; macrocytic anemia. Also loss of peripheral nerve function
Vitamin A	**Night blindness**, poor growth and reproduction, roughened skin (keratomalacia); **xerophthalmia**, leading to blindness, anemia, reduced immune function
Vitamin D	**Rickets**; inadequate calcification of bones resulting in bone deformities
Vitamin K	Poor blood clotting
Vitamin E	Red cell fragility, increase in blood peroxides
Phosphorus	Anorexia, muscle weakness, rickets, impaired growth, bone pain
Magnesium	Muscle spasms, twitching, tremor, anorexia, nausea
Iron	**Anemia**; microcytic anemia due to low hemoglobin, fatigue, inability to concentrate
Zinc	Growth failure, hypogonadism, impaired immune function, enlarged liver and spleen, mental lethargy
Copper	**Anemia**, poor wound healing
Selenium	**Keshan disease**: fragile red blood cells, enlarged heart, cardiomyopathy, growth retardation, skeletal muscle degeneration, cataract formation
Iodine	**Goiter**: enlarged thyroid gland, poor growth, reduced metabolic rate, mental retardation if deficiency occurs in the perinatal period

[a]Some nutrients have no defined deficiency syndromes.

gene mutations. They are also much more complex and more difficult to identify.

Before discussing the different kinds of nutrient–gene interactions in detail, an outline of the kind of variability found within the human genome is presented.

DNA Variability

The DNA in the nucleus is very stable with respect to the base sequence and content. Humans are more similar than different. Variations in DNA sequence have occurred and continue to occur. DNA replication is not 100% faithful nor is DNA repair 100% accurate. Changes in the sequence of bases that comprise the individual genes and their promoter sequences occur as base substitutions, deletions, or rearrangements. Chemicals that generate free radicals can cause DNA strand breaks and a possible loss of a base. Replacement of that base can occur and the strand can be repaired. However, in some instances, the base used for the repair might not be identical to the one lost, and a base substitution will be made. If any of these changes affect the amino acid sequence of the gene product and this substitution is in a critical area that affects function, then a mutation is said to have occurred. Otherwise, the difference in base sequence is referred to as a polymorphism rather than a mutation.

The significance of a deletion or substitution of bases in the DNA of a particular gene will depend on where it occurs and what change it engenders. It may occur in a noncoding region of the DNA or be a base substitution that does not affect the amino acid sequence of the gene product. Some amino acids have more than one base triplet (codon) that dictate its use in the gene product. If the change results in a substitution of a relatively similar amino acid, the function of the protein may be conserved. Some amino acids can be replaced without affecting the secondary, tertiary, or quaternary structures of the protein (and hence its chemical and physical properties), particularly if the base substitution occurs in an area that encodes a nonactive portion of the gene product. If any of these occur there will be little discernable effect on the gene product. The resultant gene product retains its premutation function but has a slightly different amino acid sequence. Polymorphisms in DNA, particularly mitochondrial DNA, are useful tools because they allow scientists to genetically identify individuals and their relatives and also allow population

geneticists to track mutation and evolutionary events through related family members. Particularly useful in this respect are the polymorphisms in mitochondrial DNA. Anthropologists use this information to track population shifts that have occurred over time.

The amino acid sequence within a given species for a given protein is usually similar. However, some individual variation does occur. Examples of 'acceptable' amino acid substitutions are those that account for the species differences in the amino acid sequence of the hormone insulin. As a hormone, it serves a variety of important functions in the regulation of carbohydrate, lipid, and protein metabolism. However, even though there are species differences in the amino acid sequence of this protein, insulin from one species can be given to another species and be functionally active. Obviously, the species differences in the amino acid sequence of this protein are not at locations in the insulin molecule that determine its biological function in promoting glucose use.

Whether the substitution of one amino acid for another affects the functionality of the protein being generated depends entirely on the amino acid in question. An example of potentially important changes in sequence involves three related proteins important to energy balance regulation—the uncoupling proteins (UCPs) 1, 2, and 3. The UCPs function to uncouple the synthesis of ATP from the synthesis of water in the mitochondrial compartment. If UCPs are present, the cell makes less ATP and releases more energy as heat (thermogenesis), thereby decreasing energetic efficiency. If one or more UCPs are absent or nonfunctional due to a mutation(s) in the codes for these proteins, the reverse occurs. More energy is trapped in the high-energy bond of ATP, and this energy is subsequently transferred to synthetic reactions that produce storage energy products: fats and glycogen. With a decrease in energy wastage by the mitochondria, excess fat accumulates. For whatever reason, the individual is unable to produce or release one or more of the UCPs, and fuel metabolism and energy balance are adversely affected. The individual may not be able to rapidly adjust to changes in the environment, such as a dramatic decrease in environmental temperature. This is an example of a nutrient–gene interaction that is part of a disease process, in this case obesity. However, it is not a single nutrient but all energy-containing nutrients—carbohydrates, fats, and proteins—that play a role in obesity development. If the individual does not have access to a plentiful food supply, the obesity phenotype may not be apparent.

Genetics Affects Nutrient Requirements

There are many examples of nutrient requirements being influenced by genetic background. For example, in the early years of determining the human need for vitamin C, human studies showed that there could be large individual differences in the need for this essential vitamin. In addition to genetic variability, vitamin C need was increased in smokers versus nonsmokers and in people with diabetes compared to people without this disease. Similar observations have been reported for vitamin A. A rare but quite profound example of genetically determined differences in nutrient need is that of vitamin D-resistant rickets. Vitamin D constitutes part of an endocrine system within the body that allows calcium absorption to be adjusted to need. Among other actions, it is responsible for stimulating calcium uptake in the small intestine. It works through a protein, the vitamin D receptor (VDR), that, in association with the activated form of the vitamin, stimulates the transcription of genes associated with calcium transport. The VDR gene is subject to mutation, like all other genes. These can affect function, but since there are two alleles for each gene, one inherited from each parent, even those carrying a mutated gene can usually maintain calcium homeostasis using half the complement of receptors. However, if both parents carry an allele for the mutation, their child may inherit two copies of the mutated VDR genes and vitamin D would not therefore stimulate the transcription of the calcium transport proteins. These children would develop a severe form of rickets that would be vitamin D resistant. This is a rare event. These children can be treated with calcium, but they have a requirement for vitamin D that can never be satisfied.

There is another genetic condition, vitamin D-dependent rickets, in which the affected gene is not the VDR but rather an enzyme required for the activation of vitamin D. The receptor is fully functional, but affected individuals are unable to metabolize vitamin D to its active form. In this case, treatment with vitamin D is possible, giving the active form of the vitamin rather than the precursor form normally found in food. Both kinds of rickets are rare but provide a clear example of the influence of genetic background on nutrient function.

A more common example may be found with folate. Folate deficiency in some prepregnant and pregnant women can result in an infant with a condition known as spina bifida. Hydrocephaly can also result. This is a neural tube defect in which the bony covering of the spinal column is incomplete. The defect occurs early in embryonic development when the cells are differentiating into specific cell types. Folate plays an important role in this

differentiation, and in some women (not all) an insufficient intake of folate just before pregnancy and in the early weeks of pregnancy can result in these neural tube defects. It was estimated that 2500 infants per year were born with this problem. As a prevention measure, foods are now fortified with the vitamin. However, many women with very low intakes of folate give birth to healthy infants. It appears likely that some women have an enhanced requirement for folate, and it is the infants of these women who are at risk. Although the specific gene mutations are not clear, a number of candidate genes involved in folate metabolism and transport have been identified.

Another example is hemochromatosis (HH). This is a disorder resulting from unregulated absorption of iron. Usually, iron absorption is downregulated when stores are adequate, but this does not occur with HH and toxic levels build up. The condition is caused by a mutation in the *HFE* gene. Although mutation of both alleles of the gene is rare, it has been estimated that 10% of some population groups may be heterozygous carriers. This is important because enhanced iron absorption is found in heterozygotes and results in liver disease, diabetes, and other chronic conditions. For many individuals and populations, a lack of dietary iron and its association with a host of iron deficiency disorders are concerns. However, individuals with HH need to limit iron intake; thus, information about genotype is clearly useful.

It is interesting to note that for some genes involved in nutrient function, no clinical syndromes linked to their mutation or deletion have been described. A possible explanation for this comes from experimental work with mice, in which gene function is typically investigated using gene deletion studies. MTF-1 is a transcription factor that regulates genes in response to zinc availability. Knock-out of the *MTF-1* gene is embryonically lethal (i.e., no offspring develop in the absence of this gene). Thus, mutations or deletions of genes that play a key role in embryonic development will not be seen in adults.

The examples given previously demonstrate a general principle that holds true for all nutrients: There is a variation in requirement that depends on genetic background. This may result from alterations in a single gene, but perhaps more likely it is due to the aggregate effect of differences in many genes that encode products relevant to that nutrient's function. As the human genome becomes better annotated and the significance of sequence differences better understood, it will be possible to make more precise recommendations for nutrient intakes for both individuals and populations.

Nutrient–Gene Interactions in Chronic Disease

A further refinement for recommended nutrient intakes is to include consideration of the relationship of one or more nutrients to the development of chronic disease. Many chronic diseases are the result of an interaction between the genetic heritage or genotype of the individual and the lifestyle choices that individual makes. Conditions such as heart disease, diabetes mellitus, and obesity are in this category. There are also a number of genetic conditions that can be managed by diet. One of the most common of these is lactose intolerance. Approximately 75–80% of the adult population in the world today is lactose intolerant. That is, they cannot consume quantities of milk and some milk products without experiencing gastrointestinal distress. **Table 2** lists some genetic disorders that are amenable to dietary management. There are also some relatively rare genetic diseases that affect genes involved in key nutritionally relevant biological processes. For example, mutations in the gene encoding the low-density lipoprotein receptor can impair the ability of the liver to clear cholesterol from the circulation. If cholesterol in the circulation cannot be removed by the liver, it can accumulate in the blood and perhaps lead to cardiovascular disease.

Many of the major chronic diseases (i.e., heart disease, cancer, stroke, diabetes, and obesity) have identifiable genetic linkages that are nutritionally responsive. That is, if an individual carries one or more gene messages that predispose that individual to one of these diseases, nutrient intake can affect the time course and appearance of the disease. For example, more than 150 mutations have been identified that associate with the development of

Table 2 Genetic disorders amenable to dietary management

Disorder	*Nutrition strategy*
Acrodermatitis enteropathica	Increase zinc intake
Fructosemia	Avoid fructose-containing foods
Galactosemia	Avoid lactose-containing foods
Hereditary hemochromatosis	Limit iron intake
Lactase deficiency	Avoid lactose-containing foods (milk and milk products)
Methylmalonuria	Vitamin B_{12} injections
Obesity (some forms)	Consume only enough energy to meet energy need; increase energy output (exercise)
Phenylketonuria	Control phenylalanine intake such that the need for this amino acid is met but that no surplus is consumed
Sucrase deficiency	Avoid sucrose-containing foods

diabetes. The phenotypic expression (the development of diabetes) of some of these genotypes can be influenced by diet. Numerous mutations, especially in the genes for the lipid-carrying proteins, have been identified as being associated with heart disease. These too may be nutrient responsive, but the details of this responsiveness are not known. Still other mutations have been found that associate with the development of obesity or with one or more of the diseases generically referred to as cancer. Again, the details of nutrient–gene expression in these diseases are lacking.

Although many genetic signatures have been associated with specific diseases, not all people who have these genetic characteristics develop the associated disease. This suggests that not only must one have the genetic characteristic but also one must provide the environment for the disease to flourish. An example of this was reported in the early 1960s. Newly arrived Yemenite Jews and Yemenite Jews who had resided in Israel for at least 20 years were compared with respect to diet, lifestyle, and the prevalence of type 2 diabetes. The newly arrived immigrants had very little diabetes, whereas the established Yemenite Jews had as much diabetes in their population as in the Israeli Jewish populations from other areas of the world. The diets and lifestyles of these population groups were compared to a matched group of Arabs living in the same locations in Israel. The diets were not greatly different among the groups, but the disease was far more prevalent in the Jews than in the Arabs. Studies of the diet consumed by the Jews in Yemen versus that in Israel revealed that there were very few differences with one exception: In Yemen very little refined carbohydrate was consumed. Sugar was not readily available, and what was available was very expensive. Once the Yemenites settled in Israel and adopted the Israeli diet with its abundance of refined carbohydrates, type 2 diabetes began to appear. It was suggested that the change in diabetes prevalence in the Yemenite group was due to an interaction between their genetic heritage and their increased consumption of refined carbohydrate. This report was the first to suggest such an interaction.

As mentioned previously, more than 150 mutations associate with diabetes mellitus, but the presence of one or more of these mutations does not necessarily mean that the person will become a diabetic. Diabetologists have acknowledged that there are far more people with a diabetes genotype than with a diabetes phenotype. That many of the diabetes phenotypes take so many years to develop suggests that given the appropriate lifestyle choices, the phenotype may never develop; however, it may develop very rapidly if poor lifestyle choices are made. In support of this argument, one has only to examine the numbers of new cases of diabetes in times of abundant food supplies and in times of food restriction. During World War II when food was rationed (as was gasoline for automobiles), people ate less and were more active. During this period, the number of new cases of type 2 diabetes declined. The number of new cases of type 1 diabetes (autoimmune diabetes or insulin-dependent diabetes) remained fairly constant. Because food was rationed and activity was increased, fewer people had excess fat stores, and this was probably a contributing factor to the decrease in diabetes development. When food became abundant after the war, food intake again was unrestricted, and over time the prevalence of both diabetes and obesity increased.

Some forms of diabetes and obesity share a genotype that phenotypes as obesity/diabetes, called 'diabesity.' As with the group of diseases called diabetes, obesity has a number of mutations that associate with it. The expression of these genotypes depends largely on whether sufficient food is available and consumed to make possible the phenotypic expression of the obesity genotype. Several of these mutations affect food intake regulation and thus energy balance. If the brain does not receive an appropriate appetite-suppressing signal, then excess energy is consumed, with the result of excess body fat stores. Excess fat stores, particularly in the adipocyte, interfere with the action of insulin in facilitating the entry of glucose into the fat cell. When this occurs, abnormal glucose metabolism (type 2 diabetes) develops. Individuals with excess fat stores can normalize their glucose metabolism if these stores are significantly reduced through food intake restriction and increased physical activity. However, not all instances of diabesity can be resolved in this way.

Technological developments have made it easier to routinely determine the presence of polymorphisms in genes associated with the development of diseases. Given the significance of cardiovascular disease and its association with lipoprotein metabolism, much effort has been focused on this area. Currently, dietary recommendations on fat intake are made for the whole population, but it appears reasonable to suppose that individual responses to a diet designed to lower plasma lipid concentrations will depend on genotype. Evidence for this comes from a G/A polymorphism in the promoter region of the gene encoding the ApoA1 lipoprotein, a major constituent of the high-density lipoprotein (HDL). The A polymorphism is less common and some studies have found an association with the possession of this form of the gene and higher HDL concentrations.

HDL is thought to be protective against heart disease. However, the association between the A polymorphism and elevated HDL is quite inconsistent. This is explained by considering diet. Women with the A polymorphism who consumed >6% of energy as polyunsaturated fatty acids (PUFA) had higher HDL cholesterol concentrations than women consuming <6% dietary PUFA. In women lacking this polymorphism, no such effect of diet was seen. Thus, although women with the A polymorphism would clearly benefit from the standard recommendation for increasing dietary PUFA, those who lack it may not, at least with respect to HDL.

The extent to which individual polymorphisms determine disease risk is likely to be limited. These chronic diseases are complex and outcome is likely to depend on polymorphisms in a number of genes and their interactions with dietary as well as other lifestyle factors. However, as technology improves and more polymorphisms are identified, an aggregate picture of risk will develop that will allow much more refined and specific dietary recommendations to be made. Knowing that these conditions are influenced by lifestyle choices, the identification of susceptible individuals should enable the design of effective strategies to delay disease development. It may not be possible to eliminate the problem, but it may be possible to appreciably delay its onset.

Nutrition Influences Mutation Risk

Although parents largely determine one's genotype, DNA is always subject to mutation. A few or many bases can be destroyed by free radicals, for example, leaving the cell unable to produce a given protein. In turn, this affects the function of the cell. It should be noted that a single affected cell does not represent a lethal event, except for that particular cell. It becomes a problem when many cells have their DNA damaged in the same way and the loss of cell function is significant. In most instances, DNA repair will occur so well that there is little noticeable effect of the initial insult. However, over time, mismatch repair or cumulative assaults on the DNA can have cumulative effects on DNA and cellular function. This cumulative effect of assault has been suggested to explain aging. Aging as a result of cumulative effects of free radical attack on DNA as well as on the vulnerable membranes within and around the cell has been used to explain the gradual loss in cellular function that occurs with age.

Nutrients that protect cells against free radical attack are additional examples of nutrient–gene interactions in health and disease. Such nutrients as vitamin E, ascorbic acid, carotene (vitamin A), selenium, and others serve to suppress free radical formation or to promote the synthesis of enzymes that function in the free radical suppression system. **Table 3** lists nutrients and their roles in free radical protection.

Table 3 Nutrients that have a role in free radical protection

Nutrient	*Role*
Vitamin E	Quenches free radicals as they form via the conversion of tocopherol to tocopheroxyl radical, which is then converted to a quinone
Vitamin K	Serves as a H^+/e^- donor/acceptor
Carotene	Serves as a H^+/e^- donor acceptor (precursor of vitamin A)
Ascorbic acid	Serves as a H^+/e^- donor acceptor; copper is used as well
Selenium	Incorporated as selenocysteine into glutathione peroxidase
Copper, zinc	Essential cofactors for cytosolic superoxide dismutase
Manganese	Essential cofactor for mitochondrial superoxide dismutase

Whereas the nucleus has a very efficient DNA repair process, the mitochondrion does not. However, there is only one nucleus in each cell, whereas there are many mitochondria in that same cell. If one or two are damaged, there are many in the cell to compensate. Disease develops only when damage occurs to a large majority of the mitochondria. A certain threshold of damage must be reached for such damage to have a physiological effect. Again, nutrients that function as free radical suppressants or that enhance the synthesis of enzymes of the free radical suppression system function to protect mitochondria from free radical damage. In the nucleus, the DNA is protected from free radical attack by histone and nonhistone proteins. Histones are highly basic proteins varying in molecular weight from ~11 000 to ~21 000. The histones keep the DNA in a very compact form. In contrast, the mitochondrial DNA does not have this protective histone coat. It is 'naked' and much more vulnerable to damage. In addition, ~90% of oxygen free radicals are generated in the mitochondria, providing the means for such damage should the enzyme superoxide dismutase, a manganese-dependent enzyme found in this compartment, not suppress these radicals. The damage can be quite severe, but because each mitochondrion contains 8–10 copies of its genome and there are many mitochondria in each cell (up to 2000), the effects of this damage may not be apparent. There is another superoxide dismutase found in the cytosol that has a similar function. It is a copper/zinc-dependent enzyme. Again, note the dependence of function on particular nutrients

(manganese, zinc, and copper) that in turn have effects on gene expression in health and disease.

Another way that changing DNA in a single cell can profoundly affect the health of the whole organism involves cancer. In this case, the DNA changes occur in particular genes related to the growth regulatory properties of the cell. Normal homeostatic mechanisms fail and the individual cell multiplies rapidly and therefore has a widespread influence. Nutrition interacts with cancer in a number of ways. It can promote or prevent the initiating mutation. It also will influence the progression of cancer by providing the nutrients required for its growth. The cancer often ultimately influences nutrition by limiting food intake. Another level of complexity is added with chemo- and radiotherapies and their interactions with nutrition.

Throughout this article, examples have been given that illustrate the interactions that occur between nutrients and genes. The ultimate goal of understanding such interactions is to use our knowledge to enhance the expression of genes that sustain good health while suppressing the expression of genes associated with disease. Although it is currently not possible to identify individuals with genetic dispositions to chronic diseases, there is no doubt that such screening tests will be developed and will be used as a basis for recommending nutrient (food) intakes. Optimizing health, after all, is the ultimate goal of good nutrition.

See also: **Aging**. **Antioxidants**: Diet and Antioxidant Defense. **Cancer**: Epidemiology and Associations Between Diet and Cancer; Effects of Nutritional Status. **Children**: Nutritional Problems. **Coronary Heart Disease**: Hemostatic Factors; Lipid Theory; Prevention. **Diabetes Mellitus**: Etiology and Epidemiology. **Folic Acid**. **Hyperlipidemia**: Overview. **Inborn Errors of Metabolism**: Classification and Biochemical Aspects. **Iron**. **Lactose Intolerance**. **Nutrient–Gene Interactions**: Molecular Aspects. **Obesity**: Definition, Etiology and Assessment. **Vitamin D**: Rickets and Osteomalacia. **Zinc**: Physiology; Deficiency in Developing Countries, Intervention Studies.

Further Reading

Acworth IN and Bailey B (1995) *Handbook of Oxidative Metabolism*. Chelmsford, MA: ESA.

Berdanier CD (1998) *Advanced Nutrition: Micronutrients*. Boca Raton, FL: CRC Press.

Krauss RM (2001) Dietary and genetic effects on low-density lipoprotein heterogeneity. *Annual Review of Nutrition* **21**: 283–295.

Malloy PJ and Feldman D (1999) Vitamin D resistance. *American Journal of Medicine* **106**: 355–370.

Moustaid-Moussa N and Berdanier CD (eds.) (2001) *Nutrient–Gene Interactions in Health and Disease*. Boca Raton, FL: CRC Press.

Moyers S and Bailey LB (2001) Fetal malformations and folate metabolism: Review of recent evidence. *Nutrition Reviews* **59**(7): 215–224.

Ordovas JM (2002) Gene–diet interaction and plasma lipid responses to dietary intervention. *Biochemical Society Transactions* **30**: 68–73.

Pietrangelo A (2002) Physiology of iron transport and the hemochromatosis gene. *American Journal of Physiology: Gastrointestinal and Liver Physiology* **282**: G403–G414.

Strachen T and Read AP (1996) *Human Molecular Genetics*. New York: Wiley–Liss.

Tolstoi LG (2000) Adult-type lactase deficiency. *Nutrition Today* **35**: 134–141.

Wei Y-H and Lee H-C (2002) Oxidative stress, mitochondrial DNA mutation, and impairment of antioxidant enzymes in aging. *Experimental Biology and Medicine* **227**: 671–682.

Zeisel SH, Allen LH, Coburn SP *et al.* (2001) Nutrition: A reservoir for integrative science. *Journal of Nutrition* **131**: 1319–1321.

NUTRIENT REQUIREMENTS, INTERNATIONAL PERSPECTIVES

A A Yates, ENVIRON Health Sciences, Arlington, VA, USA

Determining human requirements for nutrients has been a major activity for nutritionists, biochemists, and physiologists for the past 100 years since the advent of methods that have allowed for their isolation, quantification in food, and determination of their function in cell and whole body metabolism. Whereas initial efforts focused on identifying constituents in food required to maintain life and promote growth and thus were considered essential or indispensable, research during the past 60 years has become increasingly focused on elucidating the

specific roles each nutrient plays in health and quantifying, through experimentation and study of healthy populations, the amounts needed on a daily basis to provide for optimal health and prevent disease. This process of estimating requirements for an individual with any level of precision is still in the early stages of development. Nevertheless, many facets of maintaining and improving the health of the public hinge on knowing how much is needed of which nutrients or chemical components of food, and how this differs at different stages of growth and development.

Multiple terms have been adopted to define nutrient requirements, allowances, or standards (**Table 1**). They have been established or adopted by various countries and then used for the major functions of planning food programs or assessing diets for adequacy or excess (**Figure 1**). Major efforts during the past two decades by nutrition scientists throughout the world have resulted in a shift from establishing and periodically revising nutrient allowances or recommendations based on general consensus of adequate levels (e.g., the Recommended Dietary Allowances (RDAs) of the Food and Nutrition Board in the United States, the Recommended Nutrient Intakes (RNIs) of Canada, or the Safe Levels of Intake derived by the expert groups convened by the World Health Organization and Food and Agriculture Organization of the United Nations) to more definitively anchoring the reference values to specific, well-described scientific studies so that when new information becomes available from research, it is clear that new evaluations need to be undertaken. For example, in the past, the RDAs in the United States have been used as the reference values in many situations, from setting the standards for nutrient content in programs that provide single meals, such as in school lunch programs, to the basis for government reimbursement for costs of care in skilled nursing homes (**Table 2**). It is not surprising that one reference value or number, even when adjusted for age or body size and based on scientific studies, is at times not appropriate for the situation in which it is used.

What Is a Nutrient?

The traditional approach to establishing the human essentiality of a nutrient is to show that it can be chemically isolated from foods and can improve or remove a deficiency sign resulting from its lack in the diet. The number of required nutrients defined in this way has increased over the years (**Table 3**).

During the past two decades, as a result of scientific inquiry and experimentation, the line between nutrients that might be considered essential versus nonessential has blurred. There are few new chemicals in foods or food components that, when identified, have been shown to cause severe dysfunction or death when removed from the diet in a similar manner to many of those listed in **Table 3**. However, many chemical constituents of food do contribute to health; current controversy focuses on whether such substances should be considered nutrients. The major difference with the use of modern scientific techniques is the ability now to detect finer gradations of inadequacy so that with some newer constituents the end result is not necessarily death or severe organ dysfunction but decline in health status or ability to function optimally. It could be said that there is merely a longer latency period than with typical nutrient deficiencies or excesses before the effect becomes manifest; such a situation may well characterize the typical diet-related chronic disease. An example of this is the role of vitamin E in decreasing onset of cardiovascular disease: Demonstrated to be effective in animal studies, large-scale studies in humans have so far not documented the expected positive effects on primary prevention of the specific chronic disease.

Groups throughout the world have come to define health as not just the absence of overt disease, such as nutrient deficiency diseases like pellagra (inadequate vitamin B_6) or goiter (inadequate iodine), but also a level of reserve to protect against stress, either environmental or self-induced, and preventive in nature rather than therapeutic. In 1946, the World Health Organization defined health as follows: "Health is a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity."

Scientific Basis for Establishing Recommended Intakes

Since the initial development of quantitative recommended intakes of nutrients in the 1930s and 1940s, new approaches have provided a stronger science base to the reference values so established. Early development of recommended intakes usually involved convening a group of scientists who considered the available literature and, based on their expert judgment, developed quantitative estimates of requirements for specific subpopulation groups, including by age and gender. Newer statistically supported methods allow for a more science-based approach to such deliberations and consensus.

Table 1 Definitions of reference nutrient values used by selected countries and groups

AI: Adequate Intake

Canada and the United States (1997–present): A value based on experimentally derived intake levels or approximations of observed mean nutrient intakes by a group (or groups) of healthy people. The AI for children and adults is expected to meet or exceed the amount needed to maintain a defined nutrition state or criterion of adequacy in essentially all members of a specific apparently healthy population.

The Netherlands (2000–present): An amount of the nutrient that provides for the needs of almost all those in the group.

DRI: Dietary Reference Intake

United States and Canada: A set of nutrient-based reference values, each of which has special uses.

DRV: Dietary Reference Value

United Kingdom: A term used to cover LRNI, EAR, RNI, and safe intake.

EAR: Estimated Average Requirement

United Kingdom: The required intake of a group of people for energy, protein, a vitamin, or a mineral. About half will usually need more than the EAR and half less.

United States: The daily intake value that is estimated to meet the requirement, as defined by the specified indicator or criterion of adequacy, of half of the apparently healthy individuals in a life stage or gender group.

LRNI: Lower Reference Nutrient Intake

United Kingdom: An amount of the nutrient that is enough for only the few people in a group who have low needs.

RDNI: Recommended Daily Nutrient Intake

Nordic countries: The average nutrient intake that meets the requirement needs of 50% of a group. The remaining 50% of the group will have requirements above the RDNI.

RNI: Recommended Nutrient Intake (formerly in Canada); Reference Nutrient Intake (United Kingdom)

Canada (prior to 1997): The recommended intakes of essential nutrients.

United Kingdom: An amount of the nutrient that is enough, or more than enough, for about 97% of people in a group. If average intake of a group is at the RNI, then the risk of deficiency in the group is small.

PRI: Population Reference Intake

Belgium and European Community: The intake that is enough for virtually all healthy people within a group.

RDA: Recommended Dietary Allowance

United States (prior to 1997): The intake that meets the nutrient needs of 97 to 98% of a group.

Canada and the United States (1997–present): The average daily dietary intake level that is sufficient to meet the nutrient requirements of nearly all (97–98%) healthy individuals in a particular life stage and gender group.

The Netherlands (since 2000): The mean requirement plus twice the standard deviation of the requirement (defined as the smallest intake of a nutrient that both prevents symptoms of deficiency and at which, at the same time, the risk of chronic diseases—to the extent that this is influenced by the nutrient concerned—is minimal, and is thus sufficient for almost all people in a group).

SUL: Safe Upper Level

United Kingdom: An intake level that can be consumed daily over a lifetime without significant risk to health on the basis of available evidence.

UL: Tolerable Upper Intake Level

Canada and the United States (1997–present): Highest level of daily nutrient intake that is likely to pose no risk of adverse health effects for almost all apparently healthy individuals in the specified life stage group. As intake increases above the UL, the potential risk of adverse effects may increase.

European Community (2000–present): The maximum level of total chronic intake of a nutrient (from all sources) judged to be unlikely to pose a risk of adverse health effects to humans.

The Netherlands (2000–present): Intake level above which there is a risk of adverse effects.

Sources: Committee on Medical Aspects of Food Policy (1991) *Dietary Reference Values for Food Energy and Nutrients in the United Kingdom*, Report on Health and Social Subjects No. 41. London: HMSO.

Food Standards Agency, Expert Group on Vitamins and Minerals (2003) *Safe Upper Levels for Vitamins and Minerals*. London: HMSO.

Health Council of the Netherlands (2001) *Health Council of the Netherlands; Reports 2000*, Publication No. A2001/01, pp. 53–54. The Hague: Health Council of the Netherlands.

Institute of Medicine, Standing Committee on the Scientific Evaluation of Dietary Reference Intakes, Food and Nutrition Board (1997) *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride*. Washington, DC: National Academy Press.

Institute of Medicine, Subcommittee on Upper Reference Levels of Nutrients and the Standing Committee on the Scientific Evaluation of Dietary Reference Intakes, Food and Nutrition Board (1998) *Dietary Reference Intakes: A Risk Assessment Model for Establishing Upper Intake Levels for Nutrients*. Washington, DC: National Academy Press.

Institute of Medicine, Subcommittee on Interpretation and Uses of Dietary Reference Intakes and the Standing Committee on the Scientific Evaluation of Dietary Reference Intakes, Food and Nutrition Board (2003) *Dietary Reference Intakes: Applications in Dietary Planning*. Washington, DC: National Academies Press.

European Commission, Scientific Committee on Food, Health and Consumer Protection Directorate-General (2000) *Guidelines of the Scientific Committee on Food for the Development of Tolerable Upper Intake Levels for Vitamins and Minerals*, SCF/CS/NUT/UPPLEV/11 Final, 28 November.

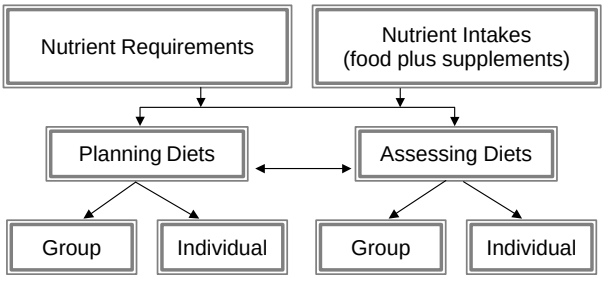


Figure 1 Uses of reference intakes in planning and assessing diets.

Table 2 Pre-1997 uses of RDAs in the United States

Planning for feeding groups of healthy people (school lunch, elderly feeding programs)
Nutrient goals for healthy individuals
Basics for foods provided in supplemental feeding programs (e.g., WIC)
Procurement of and purchasing food supplies for groups of healthy people
Reference point for evaluating the dietary intake of population subgroups
Nutrient intake targets in intervention programs
Basis of food groups in food and nutrition education programs
Reference point for the nutrition labeling of food and dietary supplements
Basis for fortification of food products
Basis for formulating dietary supplements and special dietary foods
Standards for menu planning for hospitals, correctional facilities, military operations, and other institutional feeding settings

A number of factors must be present before quantitative requirements for nutrients can be made most useful to those who use such estimates for program planning and evaluation:

- There must be some understanding of the chemical. For example, in early work on vitamins, an isolated fraction of cod liver oil was determined to be required for normal eye growth and bone development and was named 'vitamin A.' Subsequent isolation and characterization allowed the isolated mixture to be further separated into what was called the fat-soluble factor for bone growth compared to another required for sight. Thus, vitamin A was differentiated from vitamin D in the lipid-soluble fraction.
- There must be data on how much is present in the diet. In order to obtain these data, the content of the nutrient or food component in multiple typical foods must be analyzed, which thus allows the data to be used to estimate intake or exposure.
- There should be some idea of intake among the population groups of interest. Studies in which known amounts of a nutrient are consumed at varying levels and evidence of inadequacy detected should be conducted. This is typically done first with animal models, followed by human clinical trials or metabolic studies, which include at least one level of intake at which effects of inadequacy are observed and can be linked directly to the nutrient under study. Frequently, it is not possible to remove or add some nutrients to a diet without altering the content of other nutrients; this is particularly true for energy-yielding nutrients, such as omega-3 fatty acids, or substances such as fiber. This makes the interpretation of the resulting data less clear.

Table 3 Nutrients for which RDAs and recommended intakes or ranges (in parentheses) have been established since 1941

Nutrient	*1941*	*1989*	*1997–2004*
Calories	X	X	X
Protein	X	X	X
Calcium	X	X	X
Iron	X	X	X
Vitamin A	X	X	X
Thiamin	X	X	X
Riboflavin	X	X	X
Niacin	X	X	X
Vitamin C	X	X	X
Vitamin D	X	X	X
Vitamin E		X	X
Vitamin K		X	X
Vitamin B_6		X	X
Vitamin B_1		X	X
Folate		X	X
Pantothenic acid		(X)	X
Biotin		(X)	X
Choline			X
Chromium		(X)	X
Copper		(X)	X
Fluoride		(X)	X
Iodine		X	X
Magnesium		X	X
Manganese		(X)	X
Molybdenum		(X)	X
Phosphorus		X	X
Selenium		X	X
Zinc		X	X
Potassium		(X)	X
Sodium			X
Chloride			X
Total water			X
Carbohydrate			X
Total fiber			X
Linoleic acid (n-6)			X
α-Linolenic acid (n-3)			X

From the Food and Nutrition Board, US National Research Council, Institute of Medicine.

Adequate for What?

Usually, once these data are known or have been estimated, it becomes possible to establish an intake recommendation, initially based on observations of

how much appears to prevent the deficiency and how much is in the diet of those not demonstrating the symptoms or signs (indicators) of inadequacy. Many of the earlier recommended intakes were established on this basis, which is why, in many cases, the values may vary greatly across expert groups and countries. As additional data derived from experiments, observations of intake, and consequences of inadequacy of a nutrient in the diet are generated, there is a need for periodic updates of nutrient requirements and recommended intakes (**Table 4**). Changing recommendations may result in the need to make changes in programs and activities, such as food labeling, and thus frequently represent new costs. Of great importance from a scientific perspective is an overt statement of the goal of the derived reference value: Will the reference value provide guidance for minimizing overt deficiencies, usually by providing enough to prevent a known deficiency sign or symptom, or is it set at a dietary level required to maintain a blood concentration or function that might represent storage or a reserve and thus be available in times of stress?

For example, the prevention of scorbutic gums, one of the signs of overt vitamin C deficiency, requires far less vitamin C on a daily basis than the amount needed to maintain 70% saturation of white blood cell ascorbate (vitamin C) levels to counteract potential oxidative stress and damage at the cellular level. Generally, for a nutrient there exists a growing list of possible indicators or outcomes that could be used to estimate requirements (**Table 5**), and for each, a different amount may be needed daily for the specific indicator to meet the body's need and thus demonstrate adequacy.

There is usually a continuum of benefits that occur as the level of intake increases. It becomes very important to define what the criterion(ia) is that has been used to establish the quantitative level of intake recommended. **Figure 2** shows data relating iron intake to three possible criteria or indicators that could be used to determine adequate intakes for women in a national survey in The Netherlands and analyzed by George Beaton. The data show that as the level of iron intake decreases, the number of individuals (or percentage of the population group of women in this age group) who would have their needs met as documented by a given indicator of adequacy decreases. Thus, if prevention of anemia is used as the criterion (in this case, hemoglobin value <110 g/l), an individual whose intake averaged 6 mg/day would have a 40% probability that she would be inadequate (i.e., her hemoglobin value would be below the cutoff). However, if a biochemical marker of function of iron (e.g., total iron binding capacity) were used, the level of intake needed for a 40% probability of being inadequate using that criterion would be approximately 9 mg/day. Finally, if the goal were to maintain a level of storage, such as ferritin concentration, the dietary level would need to approximate 18 or 19 mg/day. Thus, when comparing recommended

Table 4 Changing US recommendations for nutrients: RDAs for vitamins (adult males, moderately active)

Vitamin	*1941*	*1943*	*1945*	*1948*	*1953*	*1958*	*1968*	*1976*	*1980*	*1989*	*1997–2001*
Vitamin A (mg RE)	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	900[c,d]
Vitamin D	400 IU[a]	400 IU[a]	[b]	[b]	[b]	[b]	400 IU	400 IU	5 μg	5 μg	5 μg[d]
Vitamin E							30 IU	15 IU	10 IU	10 mg	15 mg[e]
Vitamin K (μg)										80	120[d]
Vitamin C (mg)	75	75	75	75	75	75	60	45	60	60	90
Thiamin (mg)	1.8	1.8	1.5	1.5	1.5	1.6	1.3	1.4	1.4	1.5	1.2
Riboflavin (mg)	2.7	2.7	2.0	1.8	1.6	1.8	1.7	1.6	1.6	1.7	1.3
Niacin (mg)	18	18	15	15	15	21	17	18	18	19	16
Vitamin B_6 (mg)						1–2[f]	2.0	2.0	2.2	2.0	1.3
Pantothenic acid (mg)										4–7[f]	5[d]
Biotin (mg)										0.03–0.1[f]	0.03[d]
Folate (μg)						500[f]	400	400	400	200	400[g]
Vitamin B_{12} (μg)							3.0	3.0	5.0	2.0	2.4

[a]When not exposed to sunshine (400 IU $\approx$ 10 μg).
[b]Small amount needed when not exposed to sunshine.
[c]Unit changed from RE (Retinol Equivalent) to RAE (Retinol Activity Equivalent).
[d]Adequate Intake (AI), not RDA.
[e]As α-tocopherol only.
[f]Estimate or range, no recommendation made.
[g]As Dietary Folate Equivalents (DFE).
From the Food and Nutrition Board, US National Research Council/Institute of Medicine.

Table 5 Possible indicators or criteria to evaluate adequacy of iron intakes

Erythrocyte indexes
Erythrocyte protoporphyrin levels
Factorial modeling
Hemoglobin concentration and hematocrit
Iron balance studies[a]
Plasma total iron binding capacity
Serum ferritin concentration
Soluble serum transferrin receptor levels
Serum transferrin saturation

[a]Balance studies measure or estimate total excretion of a nutrient at different levels of intake and determine the lowest level of intake at which intake = excretion.

intakes, it is critical to know specifically the criterion or criteria used in setting the recommended intake and evaluating adequacy.

Role of Estimates of Average (Median) Requirements

For many of the uses given for reference values, it becomes important statistically to not depend on an allowance that would cover the needs of everyone and thus might include a safety factor added to some adequate level of intake but, rather, to apply estimates of the average requirement for the group of interest. For most nutrients, with iron a notable exception, it can be assumed that nutrient requirements are symmetrically distributed in a population of similar people (**Figure 3**), which means that some will have higher requirements than other similar individuals due to genetics and other factors, and that a median requirement intake level can be determined, such that consumption of a nutrient at that level would be adequate for half of the individuals in the group but inadequate for the other half. If this

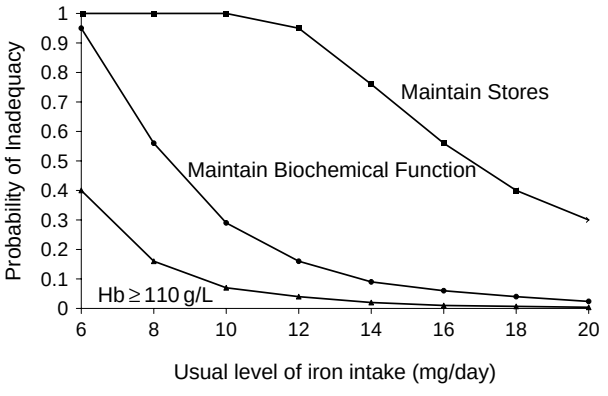


Figure 2 Probability that specified usual iron intake would be inadequate to meet the needs of a randomly selected menstruating woman. (Used with permission, G. Beaton, 1994)

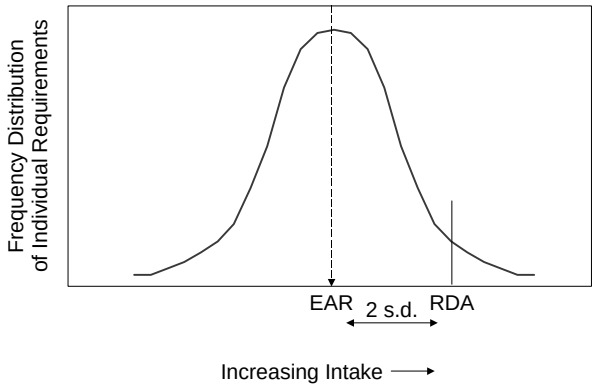


Figure 3 Probability distribution of individual nutrient requirements.

distribution of requirements is symmetrical, then the median and the mean requirement are the same.

Why have an Estimated Average Requirement? There are two main reasons to have an Estimated Average Requirement (EAR): to use as the basis for establishing the recommended intake for an individual and to assess the adequacy of intakes of similar population groups. The concept of establishing an average requirement, and assuming that the requirements of individuals in a population of similar people are symmetrically (or normally) distributed, is not new. Conceptually, it has served as the ideal basis for recommended intakes during the past few decades. However, it was rigorously used on only rare occasions. The RDA has been conceptually defined in the United States during the past few decades as the lowest amount of a nutrient that, in the judgment of the Food and Nutrition Board, meets the known nutritional needs of almost all of the population (subgroup), and it was also more mathematically defined as the mean requirement plus two standard deviations (SD), which would equal an amount required by 97 or 98% of the population to whom it is applied.

The Dietary Reference Intake (DRI) process—a joint effort of the United States and Canada—retained the term RDA, limiting its use to serving as the goal for intake when planning diets for individuals and standardizing the method by which it is established. It is defined as follows: $RDA = EAR + 2SD_{EAR}$. When data on variation in requirements of a specific nutrient are lacking, it is assumed that the standard deviation (variation) in requirements is approximately 10%. This variation in requirements is derived from the variation seen in basal metabolic rate in individuals and the variation seen in protein requirements, with protein being the nutrient whose variability has been most studied.

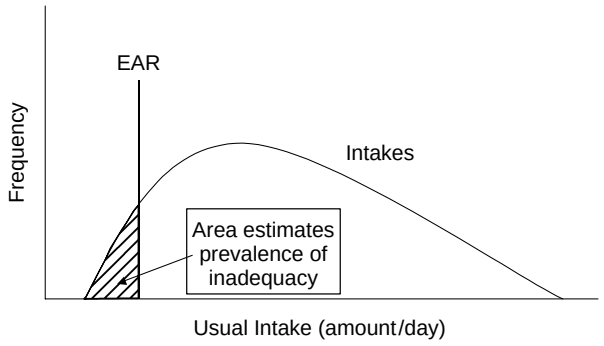


Figure 4 Using the EAR to estimate the prevalence of inadequacy in a population from the distribution of nutrient intakes.

It has been demonstrated statistically that the prevalence of inadequacy in a population whose requirements are symmetrically distributed can be estimated by comparing its intake to the EAR for that nutrient in the same (or a similar) population (**Figure 4**). Thus, in the DRI process, when evaluating vitamin C requirements, experimental data from a clinical study indicated that the average intake for men needed to achieve 70% white blood cell ascorbate saturation (the chosen indicator) was ~75 mg/day, and the EAR was set at 75 mg/day. This is a value that can be applied to the intakes of other similar populations of men who have similar characteristics to determine the percentage of the population who may be inadequate based on this criterion of adequacy (**Figure 5**).

To use this method to assess adequacy of population groups, there are other basic statistical assumptions that should be met. First, an individual's requirement for a nutrient must be statistically independent of the intake for that nutrient (this does not hold for nutrients such as total energy or water—people eat or drink because they know they need energy or water). Second, the amount of variation (the distribution) in the nutrient intake levels in the population group must be greater than the variation in the group of the requirements for the nutrient (this is almost always the case, except when everyone in the group consumes the same food in the same amounts—thus, there is little variability in intake). If these two assumptions are met, along with the symmetry mentioned previously, then the EAR can be used as the cut point for adequacy in other similar populations (as shown in **Figure 4**). This is called the EAR cut point method.

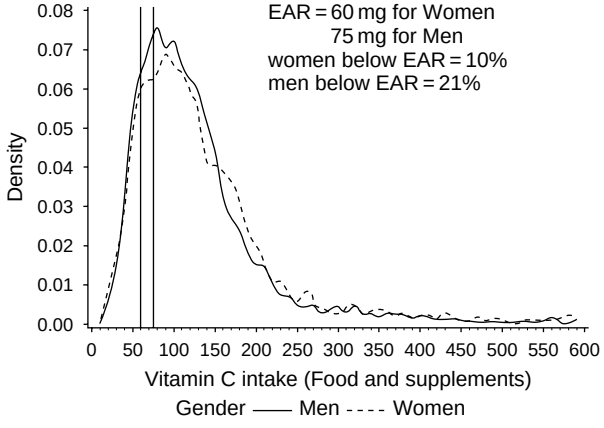


Figure 5 Vitamin C intake data from NHANES III for men and women; using the EAR to determine the expected prevalence of inadequacy.

Because the RDA has been misused as a tool to assess adequacy of intakes of groups in the past by policymakers and scientists alike, it has been argued by some that it is better for scientific panels of experts not to provide, in addition to EARs, any recommended intakes since their only use is to provide guidance to the individual, and health professionals can easily develop recommended intakes from reference values that are average requirements. However, the concept of RDAs in the United States and RNIs in Canada has been accepted in the general population to the extent that to not provide RDAs (and, as an extension, recommended intakes such as Adequate Intakes (AIs) where data are lacking) would result in more misguided actions than would result from providing them along with instructions for their specific and only use: to plan diets for the individual.

Adequate Intake: Used when an EAR cannot be determined Whereas for many nutrients enough data exist to be able to establish levels of nutrient intake at which half of the individuals in a group would be inadequate based on the criterion chosen, for some nutrients the necessary data may be conflicting or lacking. In order to give some guidance to users of nutrient reference values, it is still necessary to provide quantitative numbers. To further differentiate the appropriate uses of the RDA, the DRI framework provides an additional category of a recommended intake for use with individuals to plan diets—termed the AI. This is a level that is considered adequate for all members of the group and thus may overestimate the needs of many, if not all. Statistically, it cannot be used as if it were an EAR to assess adequacy. It does, however, provide guidance for how much an individual should consume. In some cases, it is derived from the average intake of a population in which inadequacy appears to be nonexistent based on review of available indicators or criteria (such as is the case for vitamin K).

Reference Values: Which to Use and When

As mentioned previously, there are two main uses of reference values: to assess diets for adequacy or excess and to plan diets (**Figure 1**). Although these may seem to be the same, in many ways the best reference values to use in these situations may be quite different from each other on a quantitative basis. In addition, each of these major functions is frequently applied in two different situations: to a group's intake (i.e., the intake of a population or subpopulation) or to an individual's intake.

Using DRIs to plan diets If the goal is to plan a diet or menu for a specific group so that the nutrient intake of all but a small number (e.g., 2 or 3%) in the group will have their needs met, it is not necessary for each person to consume at least the RDA; this actually overstates the need of almost all individuals. It is only necessary that the nutrient be consumed such that the intake of only 2 or 3% would be below the EAR. Thus, the goal would be to have a very low percentage of intakes below the EAR (**Figure 4**).

On the other hand, if one is planning a diet for the individual, and there is little knowledge about the individual other than his or her gender and age, then one would want to provide what is thought to be adequate for almost everyone in the group, which is the RDA—by definition set at 2 SD above the median or average requirement (EAR)—or the AI.

Using DRIs to assess diets Frequently, such as when considering whether to fortify the food supply with a specific nutrient or when evaluating the nutritional status of a subgroup in the population, it is necessary to assess the diets of groups through surveys of food intake and from such surveys determine which nutrients may be consumed at inadequate levels. If data on intakes for the group of interest are available, and the group possesses similar characteristics to the individuals studied when deriving the EARs, it is possible to estimate the prevalence of inadequacy in the group of interest from their intake data without information on their requirements or variation in intake.

This is a key reason for establishing EARs for nutrients, and it replaces the questionable past practice of comparing intakes to the RDA. Frequently when this was done, a group might appear to be at low risk of inadequacy because the mean intake of the group as a whole for a nutrient might be at or above the RDA, despite a sizable portion of the group being below their individual requirements, if they had been determined (**Figure 6**).

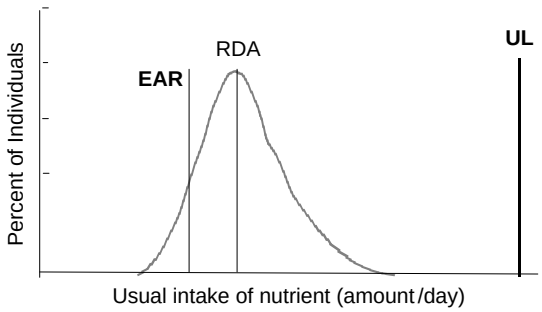


Figure 6 Example of assuming that when the mean intake of a population group is equal to the RDA, there will be a low prevalence of inadequacy. As shown, there may still be a substantial proportion of the population with intakes below the EAR, which would be a better estimate of the prevalence of inadequacy.

Whether this occurred or not would depend on whether the RDA was based on the mean intake of a population in which no one was inadequate or whether the RDA came from data for which some members of the population had inadequate intakes and thus demonstrated one or more possible criteria of inadequacy, which are usually not possible to determine. By using the EAR as the cutoff to determine the prevalence of inadequacy (this applies to those nutrients for which requirements are symmetrically or normally distributed), it is possible to set an acceptable level of inadequacy in situations of scarce resources in which it is not possible to assume that all have an adequate intake.

DRIs for Other Nutrients and Food Constituents

As indicated previously, assumptions regarding variability and independence are involved in using EARs to estimate adequacy and to plan diets. When these cannot be followed, the Food and Nutrition Board's DRI framework included other categories of reference values to provide guidance for program planning and nutrition policy: the AI, including the Estimated Energy Requirement (EER) and the Acceptable Macronutrient Distribution Range (AMDR). In the United Kingdom, population averages along with minima and maxima for some energy-yielding nutrients have been established.

The EERs for use in the United States and Canada are derived from regression equations for adults and for children based on pooled data obtained from a group of international investigators. They represent the first time that energy recommendations have been based on quantitative estimates of energy expenditure (made by the technique of measuring doubly labeled water metabolism) directly in individuals over 2 or 3 weeks for a

large number of people rather than estimating the amount of time spent in various energy-requiring activities over a 24-h period and then multiplying each type of activity by indirect estimates of energy expended.

Reference values for macronutrients such as starch, fiber, and other carbohydrates, various fatty acids, and other lipids such as cholesterol are primarily related to the role that each macronutrient plays in chronic disease development and risk factor reduction. As such, the data that support such reference values are usually less definitive, and definitely more complex, than those for single nutrients that can be easily isolated and manipulated in the diet. This additional set of reference values is given as ranges to provide guidance to federal agencies and others related to nutrient intakes. The ability to identify and quantitate the relationship of accepted risk factors for diseases is also important in reviewing literature to develop macronutrient ranges compatible with low risk of disease and maintenance of health.

Finally, physical activity has been included in the recent DRI series to highlight the very important role it plays in decreasing risk of chronic disease in terms of both maintaining sufficient energy expenditure to allow for maintenance of body weight and maintaining cardiovascular fitness to decrease the risk of heart disease.

Application of Risk Assessment Methodology to Nutrients

One of the many needs for reference values is to provide guidance about when intake of a nutrient may be too much, where the level of intake has the potential for an increased risk through excess consumption. In the past, this was rarely a concern because it was difficult to consume, on a chronic basis, large enough amounts of a specific nutrient from foods to result in serious adverse effects.

Most adverse effects of overconsumption are self-limiting because they usually involve gastrointestinal disturbances (as is the case for dietary fiber) or involve objectionable and readily reversible effects (e.g., turning orange when consuming very high amounts of carotenoids from carotene-rich foods). However, instances of serious adverse effects have been reported in the past few decades due to over-ingestion of isolated nutrients or food constituents, typically in pill form and given in therapeutic doses, or through mistakes in fortification and enrichment of the food supply, but rarely from overconsumption of foods in their natural state.

Recent increases in demand for nutrients as a result of consumer interest in self-management of health, and provocative findings relating specific dietary constituents to possible health benefits, have provided incentives for industry to increase the availability and use of nutrients and food components in dietary supplements and for the voluntary fortification of foods. Thus, the need for science-based reviews of data on the potential for increased risk of serious adverse effects that may result from chronic consumption of individual nutrients in higher amounts than typically encountered with foods has grown in importance. Such reviews have been conducted by Canadian and US scientists through the Food and Nutrition Board, by the United Kingdom's Expert Group on Vitamins and Minerals, and by the Scientific Committee on Food of the European Commission, among others. Each has worked on developing approaches to evaluating reports of adverse effects and establishing, if possible, upper levels of intake for which little concern about risks of serious adverse effects may be expected. Although somewhat differing in the review of specific studies and in defining what might be considered serious, these efforts are all aimed at incorporating the basic components of toxicological risk assessment (**Figure 7**) in the review of nutrients, primarily from a qualitative perspective and on an individual (nutrient-by-nutrient) basis. In all cases, attempts are made to quantitate no-observed-adverse-effect levels as well as lowest-observed-adverse-effect levels of exposure and then divide by an uncertainty factor to obtain the upper reference level or limit (**Figure 8**).

Issues in Establishing Reference Intakes

Extrapolating Data to Other Life-Stage and Gender Groups

Invariably, there is not enough information on studied populations to establish reference values directly for each subgroup. Knowledge of nutritional needs as well as response to higher levels of intake and exposure for such groups, such as during pregnancy or preadolescence, would be very useful. In order to provide adequate guidance when data are lacking, reference intakes are routinely provided by extrapolating the available primary data to these important age or life-stage groups from those subgroups for whom data are available. Consensus on the best methods to use for extrapolation when data are lacking, with modeling and consideration of more sophisticated approaches than just body size or caloric expenditure, is needed to enhance the utility of the derived reference values.

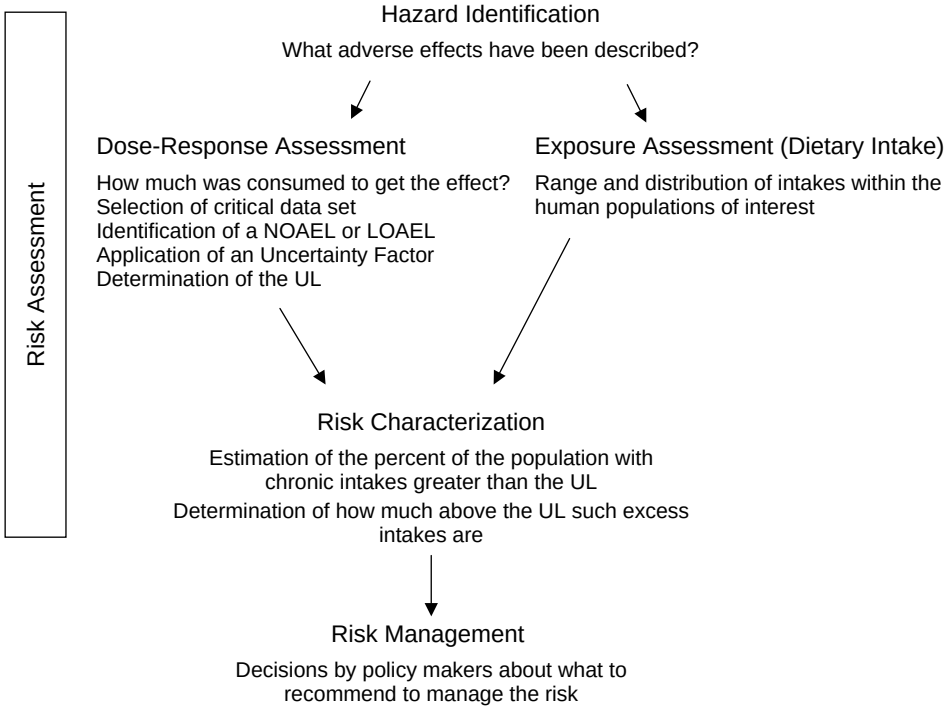


Figure 7 Steps in a model of risk assessment for nutrients.

Role of Nutrient Intake Surveys and Food Composition Databases

Surveys such as the National Health and Nutrition Examination Surveys and the What We Eat in America Survey in the United States, the Dutch National Food Consumption Survey in The Netherlands, and the National Diet and Nutrition Surveys in the United Kingdom serve as the underpinning for tracking changes in consumption and eating behavior of specific vulnerable population groups, such as young children or the elderly, in order to evaluate the potential for targeted intervention programs, either through programs aimed at changing eating behavior (e.g., the 5 A Day Program to enhance fruit and vegetable consumption in the United States) or through fortification of specific foods (e.g., calcium with bread in Canada) or changes in food product formulation (e.g., decreasing *trans* fat in high-fat processed foods). The absence of surveys that link intake with health or quantifiable and validated disease indicators makes it almost impossible to determine risk of inadequacy as well as risk of excess, particularly in vulnerable groups, without very expensive laboratory tests and clinical observation.

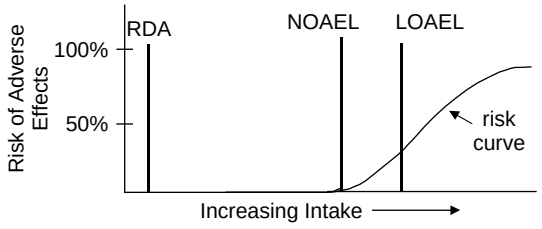


Figure 8 Identifying the hazard: dose–response.

Lack of data or nutrient content in a variety of foodstuffs, as well as lack of valid intake data, decreases the utility of subsequent estimates of inadequacy or exposure. An issue that continues to hamper reliable estimates of intake is selective underreporting and overreporting of intakes of specific foods or portion sizes by responders in surveys, usually related to foods known to be associated with causation of disease in the first case (underreporting) or considered more healthy in the second (overreporting). Although conducting large-scale surveys is costly and highly labor-intensive, poor collection of intake data and lack of replicate food composition information available to estimate intakes continue to hamper attempts to improve accuracy of the estimates. Much work is currently under way to increase the ability for such surveys to estimate intakes.

Approaches to Evaluating Bioactive Food Components

As new technologies, such as metabolomics, develop that allow better understanding of cell

metabolism and interaction among nutrients within cell systems, food constituents that have previously gone unnoticed are gaining recognition for their potential roles in maintaining health and decreasing risk of chronic disease. Some food components appear to work in concert with other nutrients and chemicals and are highly active at nanogram concentration levels in cellular systems involved in decreasing inflammatory responses or cell death. These bioactive substances may be difficult to analyze in food stuffs when they rapidly convert or oxidize into other less active compounds, making traditional methods of determining potential roles in health very difficult to apply. However, such new technologies offer the opportunity to study not pathways but, integrated circuits of multiple systems and bioactive food components simultaneously, modeling from multiple perspectives rather than the typical linear relationships diagrammed in the metabolic pathways identified by the mid-twentieth century. Using these tools, the integrated nature and role of known and unknown chemical constituents of foods will form the basis for evaluating human nutritional requirements in the future.

Steps Toward International Consensus

An issue that is obvious in any consideration of how best to approach estimating human requirements is the need to achieve consensus on the best science-based approaches to determine them. Internationally, the diversity of requirement estimates might mislead one to assume there was significant variability in nutrient needs based on geographic location or genetic makeup. As more information regarding the role that genetic factors play in disease becomes available, the variability seen in actual requirements will diminish. There will continue to be a need to recognize and use information about nutrient bioavailability, which may well be different for diets based on different foods and staples and thus require different reference values for such varied situations, but human physiology is remarkably similar.

Harmonizing approaches to reviewing data and achieving consensus among scientists is an important first step to deriving truly borderless reference values that represent differences that are physiologically and genotypically related rather than culturally related.

Efforts to harmonize are ongoing in a number of settings. Germanic language countries now have joint reference values; Australia and New Zealand are working on joint reference intakes, as are nutritionists in Southeast Asia; countries in the European Union have plans for increasing such joint deliberations beyond the activities involved in developing upper levels; and the United Nations, through the coordinating efforts of the United Nations University, is initiating extragovernmental discussion of basic issues involved in evaluating the human data that serve as the basis for establishing requirements and reference values. All these activities are in the beginning stages. With the enhanced level of communication due to computers and the Internet, such efforts are feasible as well as critical to undertake.

See also: **Antioxidants**: Diet and Antioxidant Defense. **Bioavailability**. **Dietary Guidelines, International Perspectives**. **Dietary Intake Measurement**: Methodology; Validation. **Dietary Surveys**. **Energy**: Balance; Requirements; Adaptation. **Food Composition Data**. **Food Fortification**: Developed Countries. **Functional Foods**: Regulatory Aspects. **Nutritional Surveillance**: Developed Countries; Developing Countries. **Phytochemicals**: Epidemiological Factors. **World Health Organization**.

Further Reading

Institute of Medicine (1997) Dietary Reference Intakes. In *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride*, pp. 21–37. Washington, DC: National Academy Press.

Institute of Medicine (1998) *Dietary Reference Intakes: A Risk Assessment Model for Establishing Upper Intake Levels for Nutrients*. Washington, DC: National Academy Press.

Institute of Medicine (2000) Using the estimated average requirement for nutrient assessment of groups. In *Dietary Reference Intakes: Applications in Dietary Assessment*, pp. 73–105. Washington, DC: National Academy Press.

Trumbo P, Schlicker S, Yates A, and Poos M (2002) Dietary Reference Intakes for energy, carbohydrate, fiber, fat, fatty acids, cholesterol, protein, and amino acids. *Journal of the American Dietetic Association* **102**: 1621–1630.

NUTRITION POLICIES IN DEVELOPING AND DEVELOPED COUNTRIES

C Geissler, Kings College London, London, UK

This article reviews the definition of nutrition policy and aspects specific to developing and developed countries; components of typical policies; government structures for their formulation and implementation; the types of programs used to implement policy; historical trends of emphasis; the international promotion of nutrition policies; constraints imposed by major development organisations; the effectiveness and characteristics of successful policies and programs; and emerging issues.

What Is a Nutrition Policy?

Nutrition is the process whereby living organisms use food for maintenance of life, growth, the normal functioning of organs and tissues, and the production of energy. Human nutrition therefore encompasses food composition, food consumption, food habits, the nutritive value of foods, nutritional requirements, the relationship between diet and health, and research in all these fields. Diet in this context means the total solid and liquid foods consumed by an individual or a population group. Nutrition is therefore at the center of a web of a number of disciplines and so policy affecting nutrition involves many government sectors.

In the process of national policy formulation, various ministries and departments of the government (sectors) prepare programs for implementation during a specific plan period. Those aspects of the national policy that are specifically designed to improve the state of nutrition in a country are together defined as 'nutrition policy' or 'food and nutrition policy.'

In developing countries, national policies are published for each sector in periodic national development plans, usually every 5 years. In developed countries, they are formulated at irregular intervals within the term of office of the elected government. Nutrition does not usually constitute a separate sector and so aspects of nutrition policy appear under the policies of specific sectors, such as agricultural, food, health, education, and social welfare. These aspects are by no means comprehensive and during the stage of implementation are generally not coordinated through any official mechanism.

Policy Differences in Developing Countries

Nutrition policy preparation and implementation in developing countries differs from that in developed countries in two main aspects, as discussed in the following sections.

Types of Nutrition Problems Addressed

In developing countries, these are mainly undernutrition, labeled as protein energy malnutrition, and specific deficiency conditions, most commonly vitamin A, anemia, and goiter. Although the so-called 'diseases of affluence' often affect the richer urban sections of the population, they have not been policy priorities, but they have recently become so in some Asian and Latin American and other developing countries. However, countries such as China that are in nutritional transition between the predominance of diseases of poverty and of affluence have to consider how to reduce remaining nutritional deficiencies but avoid the nutrition-related problems afflicting developed countries. In developed countries, chronic diseases related to poor nutrition, such as obesity, coronary heart disease, diabetes, and osteoporosis (sometimes referred to as 'overnutrition'), are the main problems addressed because most micronutrient deficiencies have been contained, although anemia remains prevalent, as does goitre in some areas, and pockets of undernutrition also exist in developed countries.

Influence of External Aid Agencies

In many developing countries, governments are assisted in the formulation of nutrition policies by agencies such as the World Bank, the United Nations Children's Fund (UNICEF), the World Health Organisation (WHO), and the United Nations Food and Agricultural Organisation (FAO), and specific projects are often resourced by external funding, technical assistance, and food aid.

National Nutrition Policies and Government Structures

Since the 1970s, nutrition has been recognized as an important objective of national development and an indicator of such development. In developing countries, this objective determines the goals that form the major ingredients of a national nutrition policy. In both developed and developing countries, a typical policy would aim to ensure a biologically safe and physically clean food supply sufficient to amply meet people's physiological, social, and cultural requirements of a variety of foodstuffs at commonly affordable prices. The specific programs to implement such policies usually include a mixture of analyses of the situation and individual interventions of the types outlined in the following section. Analyses for the purpose of monitoring the nutritional situation and providing public information and recommendations may include the periodic assessment of consumption patterns and energy and nutrient intakes, the identification of populations at risk of deprivation and excessive or imbalanced food consumption through specific studies and surveys, analyzing and composing dietary patterns in terms of food groups and nutrients, and defining minimum and desirable standards of the requirement of food energy and nutrients for various age groups and specific groups of the population with special needs.

Because nutrition is generally not a sector *per se*, it usually does not have a direct budget and in many countries the ministry of health is charged with improving the state of nutrition of the people, whereas aspects of food come under the ministry of agriculture. In most developing countries, implementation of the nutrition policy is carried out by these ministries or a ministry of planning through autonomous or semiautonomous councils, commissions, or committees, which may or may not be intersectoral in their compositions. These may report to the relevant ministry, cabinet, or, rarely, directly to the president. In some countries, nutrition units also exist in the provinces to provide regional planning information and actions. Some nutrition planning bodies receive advice from ad hoc technical committees as needs arise. The weakness or strength of such bodies can be judged from the change in nutritional status of the people since their establishment.

In developed countries, the ministries or departments of health and agriculture also share the main responsibility for nutrition. However, there is often conflict between the interests of producers and consumers within ministries of agriculture. Many of the nutrition and health policies in both types of countries incorporate aspects such as education and modifying activity levels, and so other ministries become involved to provide the facilities for these.

Types of Programs and Interventions

The types of interventions that form part of national nutrition policies in both developing and developed countries tend to be limited to palliative measures such as vitamin supplementation, nutrition education, and child feeding programs because many of the underlying factors that lead to malnutrition, such as unemployment, low wages, and land tenure arrangements, involve fundamental economic and political interests that are much more difficult and contentious to address. In developed countries, which are by definition richer, the governments generally provide economic safety nets for the unemployed, disabled, and other disadvantaged sections of the population. These people have to be cared for by extended family or other means in developing countries. The pattern of programs is therefore different between developed and developing countries because of differences in the nutritional problems and the wealth of the population and government. However, the types of programs are similar. The types of interventions that affect nutrition can be divided into general categories summarized in **Table 1**.

Historical Trends

There have been changes in emphasis in the type of programs advocated throughout the decades to improve nutrition as knowledge of nutrition has grown and as governments and development agencies have experienced success or failure in various approaches.

During World War II and the postwar period of the 1940s, the emphasis in developed countries was on institutional feeding, such as school meals, school milk, and the distribution of concentrated vitamin sources to children and mothers. These approaches were continued in developing countries by international agencies, such as FAO and WHO, after their establishment postwar.

As decolonization progressed, a growing interest in the process of economic development and nutrition led to the recognition that individual interventions had little impact on malnutrition and that a more integrated approach was needed to improve the use of available resources. In the 1950s, the international agencies therefore promoted 'applied nutrition programs,' which are village-based

Table 1 Types of nutrition programs

Explicitly nutritional
Programs directly related to food, including those aiming to improve food availability, accessibility, quality, safety, consumption, and knowledge.
Nutrition-oriented food policies
Agricultural production, kitchen gardens, marketing, storage, processing, safety
Food price and distribution control, food price subsidies, taxation, food stamps, rationing
Feeding programs
Mother and child: nutrition rehabilitation centres; on site, take home
Schools: lunch, breakfast, snack, milk
Workers: canteens, Food for Work
Elderly: community center; Meals on Wheels
Weaning foods
Formulated, fermented, amylase rich
Fortification, supplementation
Iron, B vitamins, iodine, vitamin A, iron, vitamin D, vitamin C, amino acids
Nutrition education
Implicitly nutritional
Programs with indirect nutritional impact through improvement of effective food demand, food utilization, and energy balance
Health
Primary health care: immunization, antiparasites, rehydration, basic medicines, prenatal care, health education, first aid
Sanitation: water supply, water treatment, water storage, waste disposal, drainage and spraying, hygiene education
Economic
Income generation
Income maintenance: welfare benefits, unemployment benefits, child allowances, etc.
Income substitution by subsidized basic needs
Activity moderation
Cereal mills, water storage and transport, child care crèches
Sports facilities, cycle paths, etc.
Integrated
Combining explicit and implicit nutritional interventions; For example, targeted 'applied nutrition programs', and 'community development programs'

programmes with components addressing several of the multiple factors of malnutrition, such as income generation activities, horticulture, health care, and nutrition education.

During the 1960s, attention focused on the world food supply and concern that the population could outstrip production. Thus, food and nutrition programs were centered around the production and dissemination of high-yielding varieties of cereals, wheat, rice, and maize—the 'green revolution' package. In the same decade, the idea developed that a specific 'protein gap' existed between the amount of protein available in national food supplies and population needs. A second focus was therefore on means to increase the production and consumption of protein from a variety of novel sources. During this same period, in the United States awareness increased that pockets of food poverty still existed in this affluent country. The resulting concern about 'Hunger in America' led to new welfare programs, such as the Women, Infants, and Children program.

By the 1970s, after 25 years of experience in nutrition interventions within economic development strategies in developing countries, it was recognized that increased national wealth did not always result in improved welfare and nutrition as predicted by the 'trickle-down' theory of development. Nutrition was therefore proposed as a specific goal for national development because a better nourished population would achieve more effective development. Government nutrition policy should be integrated and coordinated by a nutrition planning unit in an umbrella organization such as a ministry of planning or a prime minister's office so that the underlying causes would be simultaneously addressed by the appropriate government sector. This approach was fostered by development agencies such as FAO and the US Agency for International Development and was adopted by several countries, particularly after it was endorsed at the World Food Conference in 1974.

In developed countries, an increased interest in nutrition policy emerged as it was realized that existing legislation, based on food purity and the prevention of adulteration, and also the control of deficiencies, was not adequate to deal with the changing nature of nutrition problems of chronic nutrition-related diseases. Several government

advisory, professional, and consumer bodies in the United States, United Kingdom, and other countries recommended appropriate dietary goals with the common theme of reducing fat, sugar, and salt intake and increasing the intake of dietary fiber, fruit, and vegetables. The recommendation to reduce the intake of certain nutrients appeared to be a threat to some sectors of the food industry, resulting in considerable opposition to the recommendation and arguments about the validity of the evidence on which the recommendation was based. Other constraints to updating food and nutrition policy included legislation designed to prevent adulteration and maintain quality as previously perceived, such as minimum fat levels in milk and premiums on animals with high fat content. However, the proposals were gradually accepted and incorporated into government policies, while industry recognized new opportunities in the production of high-fiber, low-fat, low-salt and -sugar food products. Norway was the first developed country to have an integrated food and nutrition policy in 1975. Other developed countries subsequently formulated food and nutrition policies within their health and agriculture sectors.

By the 1980s, attempts in developing countries to apply the rational procedures advocated by the national nutrition planning approach for the selection of appropriate interventions had demonstrated the paucity of data on which to decide nutritional priorities and the effectiveness of various interventions, the difficulties of placing a policy priority on nutrition, and the problems of effective intersectoral coordination. This approach was subsequently abandoned by the development agencies. These hurdles, however, led to better evaluation of interventions and to measurements of the functional impact of malnutrition. In the 1980s, the promotion of intersectoral planning gave way to ensuring that existing sectoral interventions such as agricultural development programs included nutrition considerations. The other main theme was the targeting of nutrition interventions to those most in need and the involvement of local communities in self-sustaining development programs. This was brought about by the structural adjustment programs described later.

The 1980s also saw greater recognition of the role that women play in child nutrition through their economic as well as reproductive roles; the income that they control empowers them to make decisions beneficial to their own health and that of their children. There was also a renewed recognition of the role of diet quality in the promotion of nutrition status by the understanding that micronutrients have a function in child survival beyond deficiency diseases. This led to the promotion in developing countries of small-scale home gardening, capsule distribution, and fortification programs. Such programs had been in use for several decades in developed countries, providing land for kitchen garden allotments, the provision of supplements to children and mothers during the world wars, and the fortification of white flour with vitamins and minerals from that period to the present day as well as later compulsory fortification of margarine with vitamins A and D.

The main theme of the 1990s in developing countries was subsequently micronutrient intervention, including particularly vitamin A, iron, iodine, and, to a lesser extent, folic acid. There developed a research interest in population trials with several micronutrients that may lead to changes in nutrition policy and interventions. For example, the importance of vitamin A was investigated not only for eye lesions and blindness but also for resistance to respiratory and diarrheal infections; antioxidants began to be tested for their possible role in protection against cancer, heart disease, and other conditions; and zinc and other micronutrients were explored as a means to address the issue of restricted growth, which is widespread in developing countries.

In the 1990s, more developed countries produced explicit nutrition policies and also integrated measures to increase physical activity. For example, in the United Kingdom explicit nutritional goals were set for the first time in the 1992 government health policy, *The Health of the Nation*, which focused on five key areas for action—coronary heart disease and stroke, cancers, mental illness, HIV/AIDS and sexual health, and accidents—the first two of which are diet related. The diet and nutrition targets were as follows:

Reduce the average percentage food energy from saturated fats by at least 35% (to no more than 11% food energy)

Reduce the average percentage food energy from total fat by at least 12% (to no more than approximately 35% food energy)

Reduce the percentage of men and women aged 16–64 years who are obese by at least 25 and 35%, respectively (to no more than 6% of men and 8% of women)

Reduce the percentage of men drinking more than 21 units of alcohol per week and women drinking more than 14 units per week by 30% (to 18% of men and 7% of women)

By concentrating on these targets, it was expected that the associated dietary changes and reduction in

obesity would have beneficial consequences on such diseases as cancer, osteoarthritis, diabetes, etc. A nutrition task force was set up to oversee implementation and a physical activity task force was also set up to develop physical activity targets and detailed strategies. On a regional basis, the WHO European Region prepared the First Action Plan for Food and Nutrition Policy 2000–2005, which includes a food and nutrition task force. Food-based dietary guidelines were produced in the United States and subsequently in other countries and by FAO/WHO.

Food safety became a major concern in the 1990s, particularly in the developed countries. European consumers in particular lost faith in the science establishment due to initial assurances that BSE (mad cow disease) was no danger to human health, and they became extremely cautious about the safety of the food they purchased. In the United Kingdom, this distrust also led to a revision in government structure via the Food Standards Act 1999 so that agricultural and consumer food interests that had been combined within the Ministry of Agriculture Food and Fisheries were separated into the Food Standards Agency to champion consumer interests and the Department for the Environment, Food and Rural Affairs to oversee agriculture. A similar body was established within the European Union, the European Food Safety Authority, in 2002. In the United States, food and nutrition are regulated by the Food and Drug Administration, whereas other aspects of food and nutrition policy are regulated by the US Department of Agriculture. Due to the increased requirements to adhere to the new food safety expectations, it is more difficult for developing country exporters to gain market share in the developed world, affecting their own food security through constrained export opportunities. Along with these new food safety considerations have been concerns about genetically modified (GM) foods.

Also in the 1990s, there was increased awareness that pockets of food poverty still existed in developed countries, following the increased economic inequality that occurred in the 1980s in both developed and developing countries, partly due to government cutbacks in welfare programs (see International Constraints). This led to actions to relieve the constraints of the poor. This echoed a similar period in the United States in the 1960s involving 'Hunger in America,' which resulted in new welfare programs.

In the twenty-first century, many of the interventions emphasized in previous decades continue to form the tools of nutrition policy, and community trials continue. The experiences of these efforts in developing countries have been drawn together in a United Nations Administrative Committee on Coordination/Standing Committee on Nutrition review of 'what works.' No such review has been carried out on the effectiveness of various interventions in reducing the chronic nutrition-related diseases, possibly because the history of interventions is shorter. However, some systematic reviews on interventions for specific diseases have been conducted.

The main new emerging intervention is the development of genetic modification to provide crops with higher levels of the micronutrients that are commonly deficient in developing countries, such as iron and vitamin A, and with resistance to poor environmental conditions, such as drought and soil salinity. To date, GM foods have realized benefits largely for producers in developed countries in terms of higher productivity and lower costs. Despite no obvious benefits to consumers other than perhaps lower prices, GM soybean products have been consumed in the United States since the 1990s. However, European consumers and many in the United States are concerned that the food safety and environmental safety issues related to GM foods have not been adequately researched. The mandatory labelling of foods as 'containing GM organisms' is proposed as one solution, allowing consumers to make informed choices, but this has been opposed by GM producers as being too expensive to keep the GM and non-GM crops separate throughout the food distribution chain. The public sector has a role to play, and some new institutional arrangements, including the Global Alliance for Improved Nutrition concerning food fortification, are seeking to create incentives for the private sector to develop fortified foods for the benefit of the poor.

International Context

International Promotion

Hunger and malnutrition were put on the international agenda by the League of Nations in the 1930s, and the first conference of the United Nations in 1943 was devoted to food and agriculture. It remained an important focus of the United Nations technical agencies, FAO, WHO, and UNICEF, which were created immediately after World War II. Other international organizations have since been established, including the World Food Programme, World Food Council, International Fund for Agriculture Development, United Nations Fund for Population Activities, the World Bank, and the Consultative Group on International Agriculture. All these organizations and other international

supporting bodies have explicit objectives to eradicate human suffering due to hunger and malnutrition and to promote well-being and sound standards of health for all peoples of the world. The focus of these groups has been mainly on developing countries, but developed countries have recently been considered. These organizations have played an important role in relation to nutrition policies in developing countries by (i) providing technical assistance in the formulation and implementation of policies, programs, and activities; (ii) providing program and project funding; (iii) collecting and disseminating data, such as the World Food Surveys conducted by FAO every decade since 1946, which have greatly influenced the ideas of nutritionists and development policymakers in estimating the extent and defining the causes of malnutrition and have shaped the technical assistance deemed to be appropriate; (iv) organizing fora for debate on topics relevant to food and nutrition policy, such as the World Food Conference (1974); Alma Ata Conference of Primary Health Care (1978); World Conference on Agrarian Reform and Rural Development (1979); Convention on the Elimination of All Forms of Discrimination Against Women (1979); Fourth UN Development Decade (1990); World Summit for Children (1990); Innocente Declaration on Protection, Promotion and Support of Breast-Feeding (1990); Montreal Policy Conference on Micronutrient Malnutrition (1991); Rio Declaration on Environment and Development (1992); and International Conference on Nutrition (1992). The World Food Summit convened in November 1996, two decades after the influential World Food Conference of 1974, with the objective "to renew the commitment of the world leaders at the highest level to the eradication of hunger and malnutrition and the achievement of lasting food security for all." The UN Millennium Summit in 2000 produced the Millennium Development Goals, espoused by each of the UN agencies. These were to (i) eradicate extreme poverty and hunger; (ii) achieve universal primary education; (iii) promote gender equality and empower women; (iv) reduce child mortality; (v) improve maternal health; (vi) combat HIV/AIDS, malaria, and other diseases; (vii) ensure environmental sustainability; and (viii) develop global partnership for development. The agencies have set specific targets for each of these goals.

Some of the resolutions of these fora are very broad and clearly unachievable, such as the nutrition goals of the Fourth United Nations Development Decade (1990s), which were to (i) eliminate starvation and death caused by famine, (ii) reduce malnutrition and mortality among children substantially, (iii) reduce chronic hunger tangibly, and (iv) eliminate major nutritional diseases.

The more specific targets of the World Summit for Children (1990) to be reached by the year 2000, included (i) reduction in severe as well as moderate malnutrition among children younger than 5 years old by half of 1990 levels, (ii) reduction in the rate of low birth weights (2.5 kg or less) to less than 10%, (iii) reduction of iron deficiency anemia in women by one-third of the 1990 level, (iv) virtual elimination of iodine deficiency disorders, and (v) virtual elimination of vitamin A deficiency and its consequences, including blindness. These have clearly not been reached, and the setting of such unobtainable targets has been criticized on the grounds that they divert the attention of nutrition planners away from local priorities to global issues.

International Constraints

In the early 1980s, many developing countries experienced severe economic crises and had to implement a variety of 'structural adjustment policies,' enforced by the international finance agencies, the International Monetary Fund and the World Bank, to reduce government spending and improve balance of payments. Reduced spending resulted in cutting a variety of welfare programs that had been effective in controlling malnutrition, such as food price subsidies. Structural adjustment conditions have been rigidly imposed by the agencies for countries to obtain new financial loans. These institutions are funded by quotas from members who have voting rights in proportion to their contribution, assessed according to economic status, so that decisions are effectively in the hands of the major industrialized countries, especially the United States. This banking structure means that the policies of borrowing countries are dictated by the richer industrialized nations.

Structural adjustment has frequently resulted in changes of particular concern to the poor, such as increased food prices and decreased expenditure on social programs. The effects of these policies on health care, food consumption, incomes, and prices appear to have led to a serious deterioration in indicators of nutrition, health status, and school achievement in several countries, although it is difficult to distinguish policy effects from those of general economic decline. Efforts were subsequently made by UNICEF and other bodies to buffer vulnerable groups from these effects. During the same period, there were also cutbacks in welfare programs in developed countries, such as the provision of school meals.

International Trends in Malnutrition

To what extent have nutrition policies in developing and developed countries been effective in reducing malnutrition? During the 20 years between the World Food Conference in 1974 and the International Nutrition Conference in 1992, there have been considerable changes in the extent of malnutrition. The percentage of underweight children has declined in all areas of the world except sub-Saharan Africa and South America, but the numbers have declined only in China and have increased markedly in Southeast Asia and sub-Saharan Africa. Most data on nutritional status relate to preschool children because these are considered the most vulnerable, but other age groups are certainly not immune to malnutrition. Since 1992, international assessments of nutritional status have included women, but no information is available on trends. On the other hand, the prevalence of obesity and associated diseases has increased alarmingly in developed countries and also in several developing countries. Since the 1990s, international nutrition reports have moved from an almost exclusive focus on developing countries to include the nutrition concerns of developed countries.

What Is the Secret of Success?

Although undernutrition is clearly related to poverty, some countries are better nourished than others at similar levels of national wealth. Some countries are much better than others with a similar gross national product (GNP) in terms of indicators of nutrition and health, such as food available for consumption and infant mortality. Countries that have done best to improve undernutrition in recent years are those in which there is greater equity or in which policies have concentrated on ensuring the satisfaction of basic needs, including adequate food. Their political ideologies range from communist China to capitalist South Korea and Taiwan. China is the classic example of a country that is still poor but has largely dominated malnutrition and famine through effective organization of food production and distribution. Other examples are Costa Rica, Chile, Cuba, Kerala state in India, Sri Lanka, and Thailand, which have better nutrition conditions than other counties with similar GNPs. In contrast, some countries have extensive chronic malnutrition despite massive aid (e.g., Bangladesh) and rapid economic growth (e.g., Brazil).

These improvements cannot all be ascribed to specific nutrition policies. What are the lessons that can be learned about the effectiveness of the nutrition interventions commonly used to implement nutrition policies? This is not an easy question to answer because the evaluation of effectiveness of specific interventions is theoretically simple but practically difficult since evaluation has to take into account general economic change. An important function of international development and research agencies such as the World Bank, the International Food Policy Research Institute, FAO, UNICEF, and WHO since the 1970s has been to draw together research on the impact of policies and programs on the economic, health, and nutritional status of beneficiaries to distinguish the characteristics of success. Most of these have concentrated on developing countries. Several features of successful large-scale nutrition interventions in relation to undernutrition have been extracted and are summarized next.

The objectives must be based on a careful analysis of the real problem and be achievable in a timescale set within the program design. Community and local nongovernment organization involvement is essential in the design and implementation so that there is a sense of joint ownership for self-sustaining success. The overall effectiveness depends on coverage, and if interventions are targeted at specific groups there has to be a trade-off between the cost-effectiveness of targeting and wider coverage of the population. Charismatic leadership and good management are essential, and the appropriate mix of components must be accompanied by effective administration with a balance between bottom-up and top-down actions. Most successful programs include strong training and supervision. Effective implementation is helped by setting clear targets and by monitoring and evaluating the process, with flexibility to modify the program where necessary. The attitude of the workers is crucial in determining the potential for scaling up from a pilot project with selected staff to a large-scale operational program that has to use existing staff. Awareness of the consequences and causes of malnutrition and a political commitment at all levels are important. These common characteristics are a useful basis for the planning of future programs to maximize their success.

Emerging Issues in the Twenty-First Century

Some of the main emerging and reemerging nutrition issues of the new millennium for developing countries are those that reflect changing economic,

demographic, and disease patterns and include HIV/AIDS, the nutrition transition, refugees, adolescents, and aging.

In developed countries, aging is also one of the main emerging issues, along with the continued increase in obesity in both adults and children, with concomitant increases in related diseases such as diabetes. It has been recognised that this cannot be dealt with only on the nutrition front, and nutrition policies and recommendations are now including measures to increase activity in the population. Research continues to refine the association of various food factors with aspects of health and so determines policy. For example, the United States has already undertaken folic acid fortification of flour and the United Kingdom is considering doing so.

Millions have died of AIDS, especially in sub-Saharan Africa, with devastating effects on people's livelihoods. For the individual, the disease raises nutrient requirements and reduces the immune system, increasing vulnerability to other diseases. A major issue is the transmission of HIV from mother to child during pregnancy, at birth, or with breastfeeding. For the household, HIV/AIDS reduces the capacity to care for young children and infected household members and to work to ensure food security, resulting in deteriorating nutritional status. Women feel the impact most severely. Nutrition policy has to relate to prevention and nutritional care, which can significantly postpone illness and prolong life.

More developing countries will have to modify their nutrition policies to address the shift from problems of nutritional deficiency and infectious diseases to problems of chronic diet-related diseases, including obesity, diabetes, cardiovascular disease, hypertension, and various forms of cancer. These shifts are associated with changes in diet and lifestyle patterns that accompany industrialization, urbanization, economic development, and market globalization and that result in the increased consumption of energy-dense diets and sedentary work and leisure occupations.

National and international conflict has resulted in millions of refugees and internally displaced persons, estimated by the United Nations to be 35 million, of which 80% are women and children, and acute malnutrition is frequently reported. Nutritional support of these displaced populations is a concern for national governments and international agencies.

Most attention in the past has been focused on the nutrition of young children and on pregnant and lactating women, and other groups have been relatively neglected. Recently, the special needs of adolescents have begun to be addressed. Adolescents comprise approximately 20% of the world's population, and adolescence is a period of intense physical, psychosocial, and cognitive development, during which they gain up to 50% of their adult weight, height, and skeletal mass, caloric requirements are maximal, and poor eating habits and pregnancy are additional concerns. More attention will be paid to this group and to ways to avoid the consequences of poor nutrition during this period.

Another neglected group is the elderly. Currently, there are 580 million people older than 60 years (61% in developing countries), and this number is projected to increase to 1 billion by 2020 (71% in developing countries). The majority are women because they live longer than men. Special problems associated with nutrition include osteoporosis and fractures, vulnerability to malnutrition, and degenerative diseases.

Conclusions

A nutrition policy is easy to draw up on paper but is useless unless implemented. Many countries have adopted nutrition policies that were ineffective because they were not or could not be implemented. Some policies could not be implemented even if the political will existed because they were too complex, such as National Nutrition Planning, or could not be scaled up successfully from pilot projects to operational programs because they did not have funding for an equivalent level of training and supervision, such as the Applied Nutrition Program. Successful implementation depends on many economic and technical factors but most important on political will. Success in improving nutrition has been achieved in countries with a wide range of political ideologies but with a common theme of government commitment to promoting equity and to satisfying basic needs. Some types of specific interventions can be successful without such commitment, such as nutrient supplement programmes, but the criteria of success have to be clearly defined in terms of population coverage, sustainability, and to what extent the program addresses the main nutritional problems.

See also: **Dietary Guidelines, International Perspectives**. **Food Fortification**: Developing Countries. **Malnutrition**: Primary, Causes Epidemiology and Prevention. **Nutrient Requirements, International Perspectives**. **Nutritional Surveillance**: Developing Countries. **United Nations Children's Fund**. **World Health Organization**.

Further Reading

de Onis M, Frongillo EA, and Blossner M (2000) Is malnutrition declining? An analysis of changes in levels of child malnutrition since 1980. *Bulletin of the World Health Organisation* **78**(10): 1222–1233.

FAO/WHO (2003) *Diet, Nutrition and the Prevention of Chronic Diseases. Report of the Joint FAO/WHO Expert Consultation*, Technical Report Series No. 916. Geneva: FAO/WHO.

Flores R and Gillespie S (eds.) (2001) *Health and Nutrition: Emerging and Re-emerging Issues in Developing Countries*, 2020 Focus 5. Washington, DC: International Food Policy Research Institute.

Geissler C (1995) Nutrition intervention. In: Ulijaszek SJ (ed.) *Health Intervention in Less Developed Countries*. Oxford: Oxford University Press.

Grantham-McGregor SM, Pollit E, Wachs TD, Meisels SJ, and Scott KG (1999) Summary of the scientific evidence on the nature and determinants of child development and their implications for programmatic interventions with young children. *Food and Nutrition Bulletin* **20**(1): 4–6.

Heaver R (2002) *Improving Nutrition: Issues in Management and Capacity Development. Health, Nutrition and Population Discussion Paper, Human Development Network*. Washington, DC: World Bank.

Jolly R and Cornia GA (eds.) (1984) *The Impact of World Recession on Children. UNICEF Report*. New York: Pergamon Press.

Kennedy E (1999) Public policy in nutrition: The US nutrition safety net—Past, present and future. *Food Policy* **24**: 325–333.

Khush GS (2002) The promise of biotechnology in addressing current nutritional problems in developing countries. *Food and Nutrition Bulletin* **23**(4): 354–357.

Pinstrup-Andersen P (ed.) (1993) *The Political Economy of Food and Nutrition Policies*. Baltimore: Johns Hopkins University Press.

Popkin BM, Horton S, and Kim S (2001) The nutrition transition and prevention of diet-related chronic diseases in Asia and the Pacific. *Food and Nutrition Bulletin (United Nations University)* **22**(4 supplement).

Riches G (ed.) (1997) *First World Hunger: Food Security and Welfare Politics*. Basingstoke, UK: Macmillan.

Standing Committee on Nutrition (2002) *Nutrition in the Context of Conflict and Crisis*, SCN News No. 24. Geneva: ACC/SCN.

United Nations Administrative Committee on Coordination, Standing Committee on Nutrition (2000) *Fourth Report on the World Nutrition Situation*. Geneva: ACC/SCN and IFPRI.

NUTRITION TRANSITION, DIET CHANGE AND ITS IMPLICATIONS

B M Popkin, University of North Carolina, Chapel Hill, NC, USA

The world is experiencing rapid shifts in structures of diet and body composition with resultant important changes in health profiles. In many ways, these shifts are a continuation of large-scale changes that have occurred repeatedly over time; however, the changes facing low- and moderate-income countries appear to be very rapid. Broad shifts continue to occur throughout the world in population size and age composition, disease patterns, and dietary and physical activity patterns. The former two sets of dynamic shifts are termed the demographic and epidemiological transitions. The latter, whose changes are reflected in nutritional outcomes, such as changes in average stature and body composition, is termed the nutrition transition. These three relationships are presented in **Figure 1**.

Human diet and activity patterns, and nutritional status, have undergone a sequence of major shifts, defined as broad patterns of food use and their corresponding nutrition-related diseases. During the past three centuries, the pace of dietary and activity change appears to have accelerated to varying degrees in different regions of the world. Furthermore, dietary and activity changes are paralleled by major changes in health status as well as by major demographic and socioeconomic changes. Obesity emerges early in these shifting conditions, as does the level and age composition of morbidity and mortality. Although there are five broad nutrition patterns dating back to the origins of modern man, the focus of this article is on the three most recent periods (**Figure 2**). For convenience, the patterns are outlined as historical developments; however, 'earlier' patterns are not restricted to the periods in which they first arose but, rather, they continue to characterize certain geographic and socioeconomic subpopulations. The first two patterns relate to earlier periods in the evolution of man—the first pattern of collecting food and the second

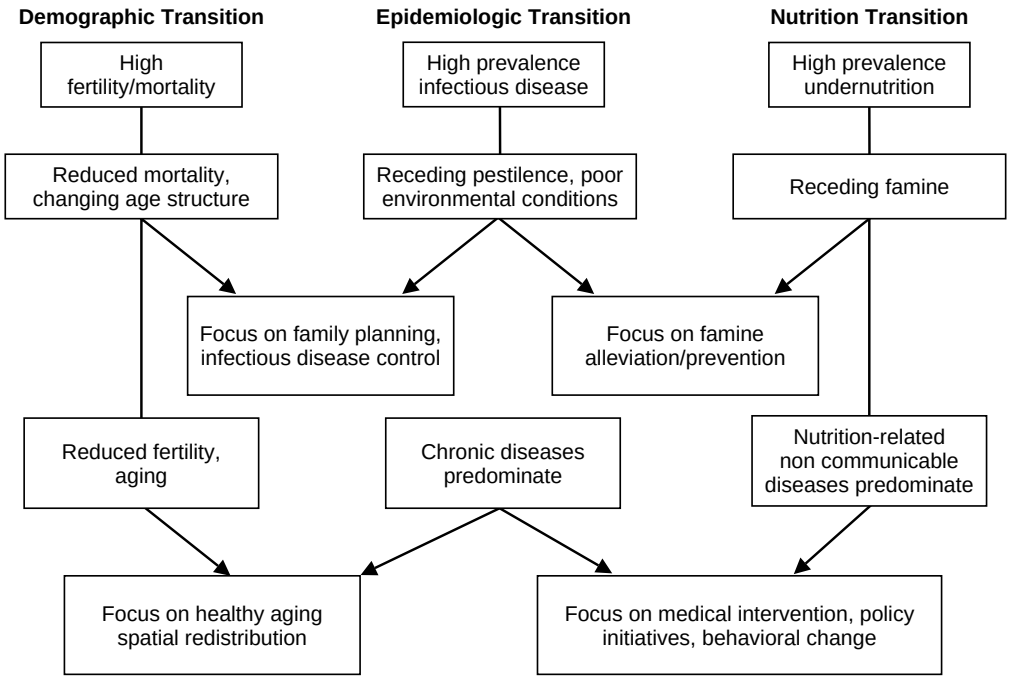


Figure 1 Stages of health, nutritional, and demographic change. (From Popkin BM (2002) The shift in stages of the nutrition transition in the developing world differs from past experiences! *Public Health Nutrition* **5**(1A): 205–214.)

pattern of famine. The following are the three later periods:

Pattern 3: Receding famine: The consumption of starchy staples had predominated and continues to do so, but these items become less important in this low-fat diet as limited amounts of fruits, vegetables, and animal protein are increasingly added to the low-fat and high-fiber diet. Many earlier civilizations made great progress in reducing chronic hunger and famines, but only in the last third of the past millennium have these

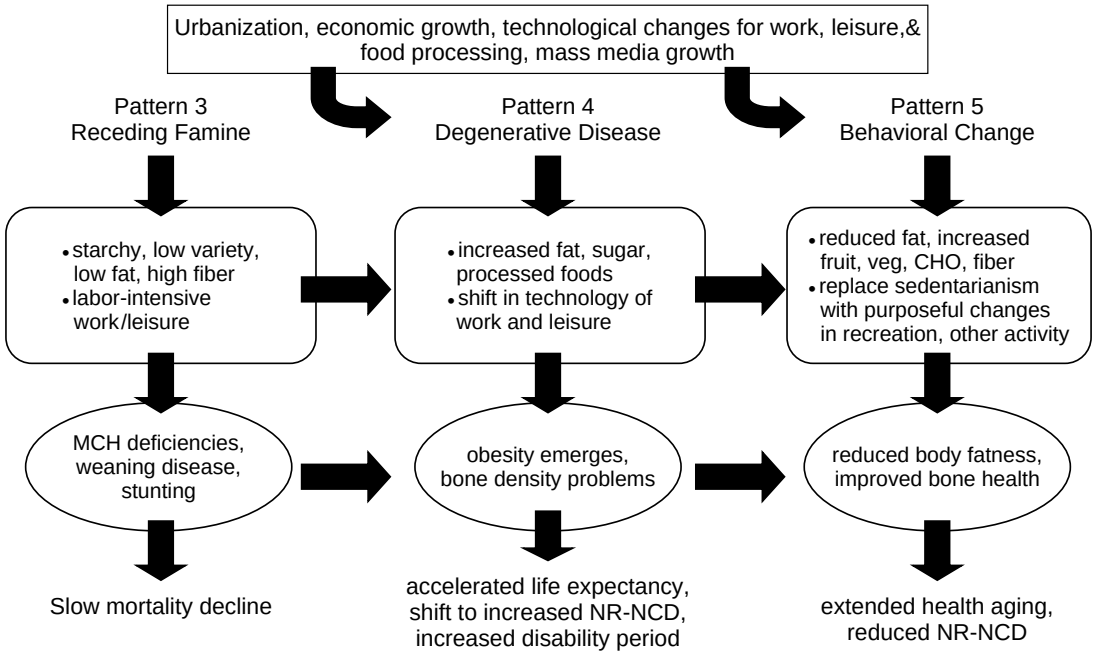


Figure 2 Stages of the nutrition transition. (From Popkin BM (2002) The shift in stages of the nutrition transition in the developing world differs from past experiences! *Public Health Nutrition* **5**(1A): 205–214.)

changes become widespread, leading to marked shifts in diet. However, famines continued well into the eighteenth century in portions of Europe and remain common in some regions of the world. Activity patterns start to shift and inactivity and leisure become a part of the lives of more people.

Pattern 4: Nutrition-related noncommunicable disease (NR-NCD): A diet high in total fat, cholesterol, sugar, and other refined carbohydrates, low in polyunsaturated fatty acids and fiber, and often accompanied by an increasingly sedentary life is characteristic of most high-income societies (and increasing proportions of the population in low-income societies), resulting in increased prevalence of obesity and contributing to the degenerative diseases that characterize the final epidemiologic transition stage.

Pattern 5: Behavioral change: A new dietary pattern appears to be emerging, evidently associated with the desire to prevent or delay degenerative diseases and prolong health. Whether these dietary changes, instituted in some countries by consumers and in others also prodded by government policy, will create a large-scale transition in dietary structure and body composition remains to be seen.

Our focus is increasingly on patterns 3–5, particularly the rapid shift in much of the world's low- and moderate-income countries from the stage of receding famine to NR-NCD. **Figure 2** presents this focus. The concern about this period is so great for many that the term 'nutrition transition' is synonymous with this shift from pattern 3 to 4.

Shifts in Dietary and Activity Patterns and Body Composition Seem to Be Occurring More Rapidly

The pace of the rapid nutrition transition shifts in diet and activity patterns from the period termed the receding famine pattern to one dominated by NR-NCDs seems to be accelerating in the lower and middle-income transitional countries. We use the word 'nutrition' rather than 'diet' so that the term NR-NCDs incorporates the effects of diet, physical activity, and body composition rather than solely focusing on dietary patterns and their effects. This is based partially on incomplete information that seems to indicate that the prevalence of obesity and a number of NR-NCDs is increasing more rapidly in the lower and middle-income world than it has in the West. Another element is that the rapid changes in urban populations are much greater than those experienced a century ago or less in the West; yet another is the shift in occupation structure and the rapid introduction of the modern mass media. Underlying such changes is a general concern for rapid globalization as the root cause.

Clearly, there are quantitative and qualitative dimensions to these changes. On the one hand, changes toward a high-density diet, reduced complex carbohydrates, increased added sugar and other caloric sweeteners, and inactivity may be proceeding faster than in the past. The shift from labor-intensive occupations and leisure activities toward more capital-intensive, less strenuous work and leisure is also occurring faster. On the other hand, qualitative dimensions related to multidimensional aspects of the diet, activity, body composition, and disease shifts may exist. The social and economic stresses that people face and feel as these changes occur may also be included.

Scholars often note that the pace and complexity of life, reflected in all aspects of work and play, are increasing exponentially. There are also unanticipated developments, new technologies, and the impact of a very modern, high-powered communications system. It is this sense of rapid change that makes it so important to understand what is happening and anticipate the way in which changes in patterns of diet, activity, and body composition are occurring. Although the penetration and influence of modern communications, technology, and economic systems related to 'globalization' have been a dominant theme of the past few decades, there seem to be some unique issues that have led to a rapid increase in globalization and its impact.

Stating that globalization is the cause results in a focus on broad and vaguely measured sets of forces; this ignores the need to be focused and specific, which would allow us to develop potentially viable policy options. It is difficult to measure each element of this globalization equation and its impact. These processes certainly have been expanded, as indicated by enhanced free trade, a push toward reduction of trade barriers in the developing world, and the increasing penetration of international corporations into the commerce in each country (measured by share of gross national product (GNP) or manufacturing). Similarly, other economic issues related to enhanced value given to market forces and international capital markets are important. Equally, the increasing access to Western media, the removal of communication barriers enhanced by the World Wide Web, cable TV, mobile telephone systems, etc. are important. The accelerated introduction of Western technology into manufacturing and the basic sectors of agriculture, mining, and services is also a key element.

Another way to understand the types of changes the developing world is facing is to consider an urban squatter's life and a rural villager's life in China approximately 20 years ago and today. During the 1970s, food supply concerns still existed; there was no television, limited bus and mass transportation, little food trade, minimal processed food, and most rural and urban occupations were very labor intensive. Today, work and life activities have changed: Small gas-powered tractors are available, modern industrial techniques are multiplying, offices are automated, soft drinks and many processed foods are found everywhere, TVs are in approximately 89% of households (at least one-fifth of which are linked to Hong Kong Star and Western advertizing and programing), younger children do not ride bicycles, and mass transit has become heavily used. Considering that such changes are also occurring in much of Asia, North Africa, the Middle East, Latin America, and many areas (particularly cities) in sub-Saharan Africa, it is evident that the shift from a subsistence economy to a modern, industrialized one occurred in a span of 10–20 years, whereas in Europe and other industrialized high-income societies, this occurred over many decades or centuries.

To truly measure and examine these issues, we would need to compare changes in the 1980–2000 period for countries that are low and middle income to changes that occurred a half century earlier for the developing world. However, data on diet and activity patterns are not available, and there are only minimal data on NR-NCDs and obesity.

The elements of the nutrition transition known to be negatively linked with NR-NCDs are obesity, adverse dietary changes (e.g., shifts in the structure of diet toward a greater role for higher fat and added caloric sweeteners in food, reduced fruit and vegetable intake, reduced fiber intake, greater energy density, and greater saturated fat intake), and reduced physical activity in work and leisure. The causes of these elements of the nutrition transition are not as well understood as the trends in each of them. In fact, few studies have attempted to research the causes of such changes, and there are only a few data sets equipped to allow such crucial policy analyses to be undertaken.

Obesity Trends

The most commonly measured health outcome of the shifts in the structure of diet is obesity. The shifts in adult overweight and obesity in the developing world in the past 10–30 years are far faster than in the higher income countries. We examined the shifts in body composition among Chinese adults aged 20–45 years during an 8-year period. Not only did mean body mass index (BMI) increase but also the shape of the BMI distribution curve changed during the 8-year period. From 1989 to 1997, the proportion of underweight men and women declined considerably and the prevalence of both overweight and obesity increased greatly. In fact, the proportion of overweight or obese men more than doubled from 6.4.0 to 14.5% and the proportion of overweight or obese women increased 50%, from 11.5 to 16.2%.

China is not unique; here, data from a few low- and middle-income countries are presented to compare their increase in the annual prevalence of overweight and obese adults with that of the United States. **Figure 3** presents the annualized increase in the percentage points of prevalence for data from high-income countries with comparable data. **Figure 4** shows how quickly overweight and obesity have emerged in Mexico as a major public health problem. Compared to the United States and European countries, where the annual prevalence increase in overweight and obesity is approximately 0.25 each, the rates of change are very high in Latin America. Cuba's data only represent Havana. Similar shifts in the prevalence of obesity are presented for North Africa and the Middle East and Asia in **Figures 5** and **6**, respectively.

What is important to note is that the increase in the proportion of the adult population that is overweight is far greater in all the lower income countries than in the United States or most European countries. Only Spain, with its large shift in overweight in the past decade, is close to the speed of change of these countries.

Dietary Changes: Shift in the Overall Structure over Time

The diets of the developing world are shifting equally rapidly. There are no good data for most countries on total energy intake, but there are reasonable data to examine shifts in the structure of the diet. Food balance data were used to examine the shift over time in the proportion of energy from fat.

The dramatic changes in the aggregate income–fat relationship from 1962 to 1990 are displayed in **Figure 7** by the estimated regression lines based on cubic polynomial regressions. Most significantly, even the poor nations had access to a relatively high-fat diet by 1990, when a diet deriving 20% of energy (kcal) from fat was associated with countries having a GNP of only $750 *per capita*,

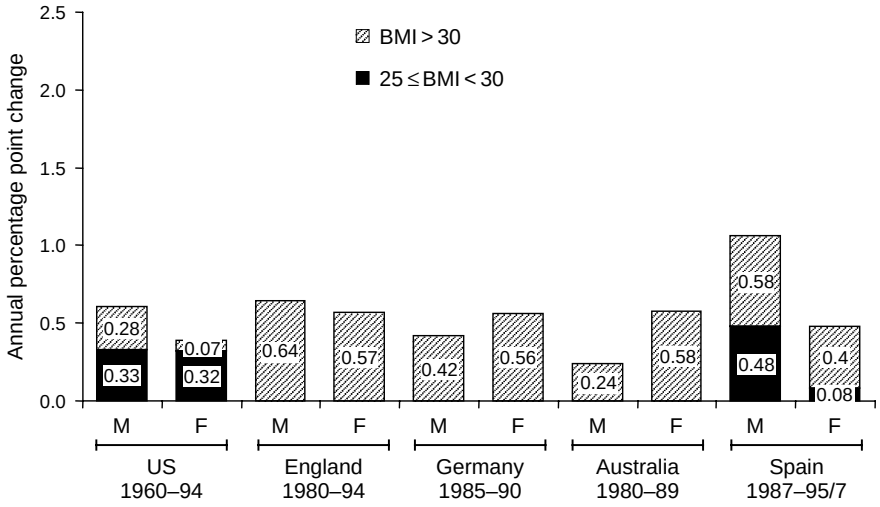


Figure 3 Obesity trends among adults in the United States and Europe (the annual percentage point increase in prevalence). BMI, body mass index; F, female; M, male. (Popkin BM (2002) The shift in stages of the nutrition transition in the developing world differs from past experiences! *Public Health Nutrition* **5**(1A): 205–214.)

whereas in 1962 the same energy diet (20% from fat) was associated with countries having a GNP of $1475 (both GNP values in 1993 dollars). This dramatic change arose from a major increase (10–13%) in the consumption of vegetable fats by poor and rich nations; similar increases (3–6%) also occurred in mid- and high-income nations.

At the same time, there were decreases in the consumption of fat from animal sources for all except the low-income countries. The availability of animal fats continued to be linked to income, though less strongly in 1990 than in 1962. These decreases, combined with the increase in vegetable fat intake for all income countries, resulted in an overall decrease in fat intake for moderate-income countries of approximately 3% but an increase of approximately 4 or 5% for low- and high-income countries. **Figure 7** shows these substantial shifts in the relationships between GNP and the composition of diets over time.

Vegetable fats in 1990 accounted for a greater proportion of dietary energy than animal fats for

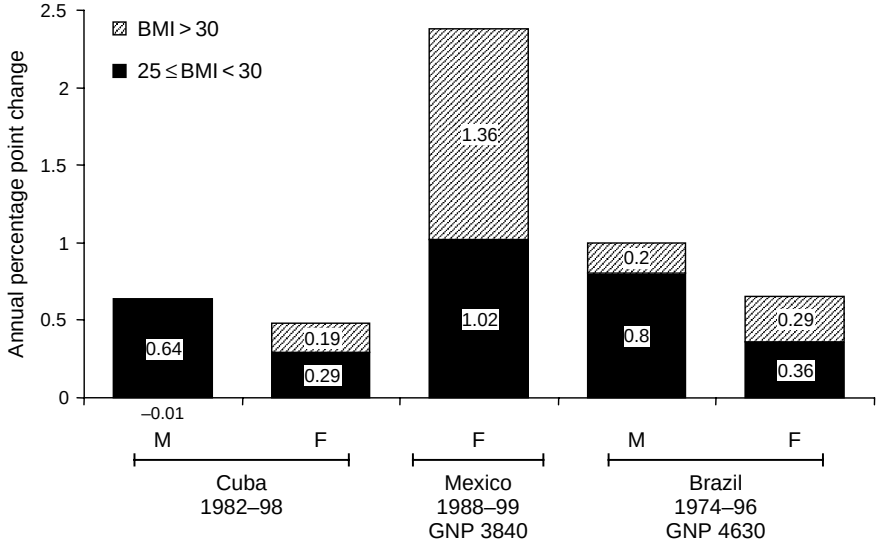


Figure 4 Obesity trends among adults in Latin America (the annual percentage point increase in prevalence). BMI, body mass index; F, female; GNP, gross national product; M, male. (Data from Rodriguez-Ojea A, Jimenez, Berdasco A and Esquivel M. (2002) The nutrition transition in Cuba in the nineties: an overview. *Public Health Nutrition* **5**(1A): 129–33. Rivera (2002) Reference: Rivera JA, Barquera S, Campirano F, Campos I, Safdie M and Tovar V. (2002) Epidemiological and nutritional transition in Mexico: rapid increase of non-communicable chronic diseases and obesity. *Public Health Nutrition* **5**(1A): 113–22.)

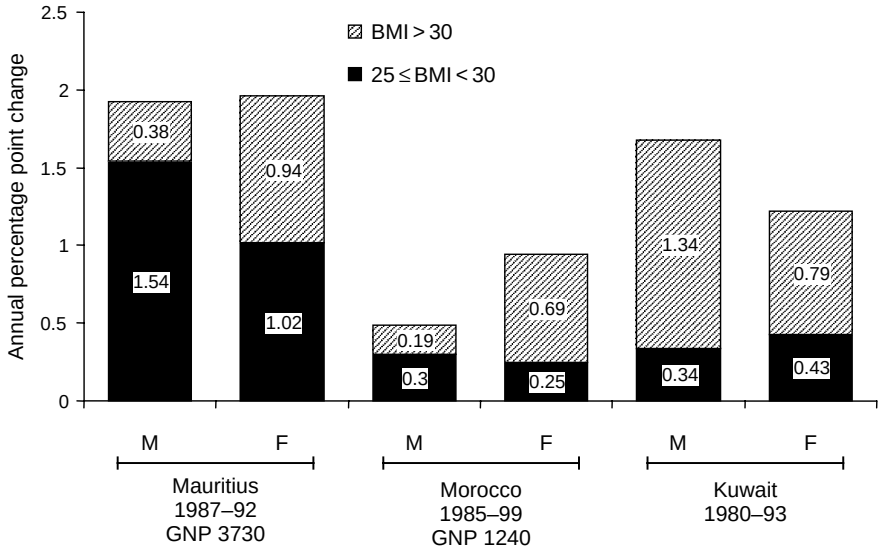


Figure 5 Obesity trends among adults in North Africa/Middle East (the annual percentage point increase in prevalence). BMI, body mass index; F, female; GNP, gross national product; M, male. (Data from Benjelloun S. (2002) Nutrition transition in Morocco. *Publlic Health Nutrition* **5**(1A): 135–40. Hodge (1996) Reference: Hodge AM, Dowse GK, Gareeboo H, Tuomilehto J, Alberti KG, Zimmet PZ. (1996) Incidence, increasing prevalence, and predictors of change in obesity and fat distribution over 5 years in the rapidly developing population of Mauritius. *International Journal of Obesity* **20**: 137–46. Al-Isa (1995,1997) References: 1-Isa AN. (1995) Prevalance of obesity among adult Kuwaitis: a cross-sectional study. *International Journal of Obesity and Related Metabolic Disorders*. **19**(6):431–3. A1-Isa AN. (1997) Changes in bidy mass index (BMI) and prevalance of obesity among Kuwaitis 1980–1994. *International Journal of Obesity* **21**: 1093–9.)

countries in the lowest 75% of countries (all of which have incomes less than $5800 per capita) of the per capita income distribution. The absolute level of vegetable fat consumption increased, but there remained at most a weak association between GNP and vegetable fat intake in these aggregate data. The change in edible vegetable fat prices, supply, and consumption is unique because it equally affected rich and poor countries, but the net impact is relatively much greater on low-

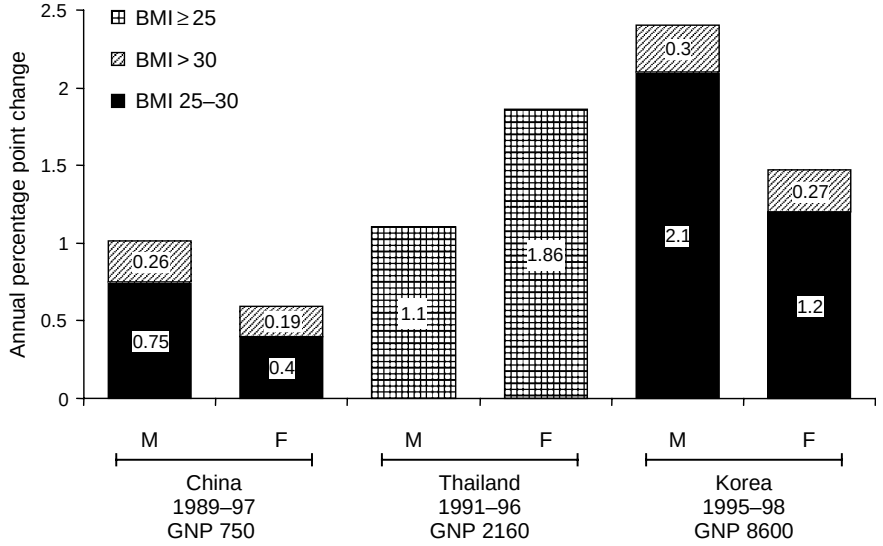


Figure 6 Obesity trends among adults in Asia (the annual percentage point increase in prevalence). BMI, body mass index; F, female; GNP, gross national product; M, male. (Data from Kosulwat V. (2002) The nutrition and health transition in Thailand. *Public Health Nutrition* **5**(1A): 183–89. Du (2002) Reference: Du S,Lu B, Zhai F and Popkin BM. (2002) A new stage of the nutrition in China. *Public Health Nutrition* **5**(1A): 169–74. Lee (2002) Reference: Lee M-J, Popkin BM and Kim S. (2002) The unique aspects of the nutritioin transition in South Korea: the retention of healthful elements in their traditional diet. *Public Health Nutrition* **5**(1A): 197–203.)

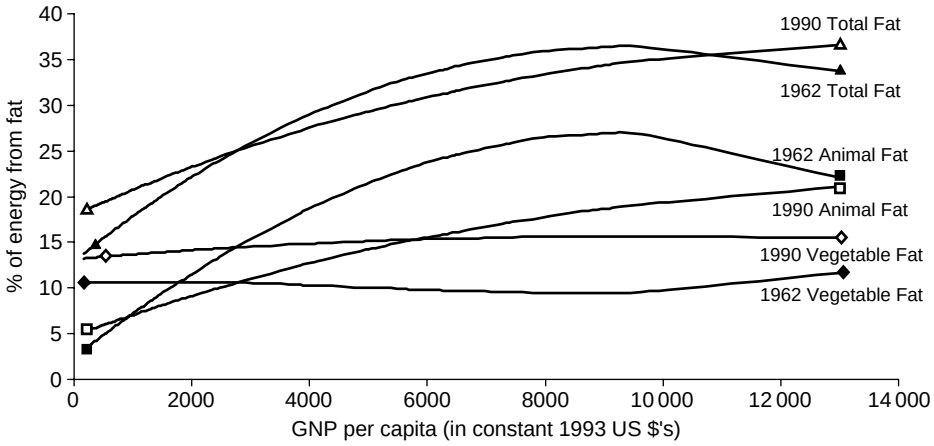


Figure 7 Relationship between the percentage of energy from fat and gross national product (GNP) *per capita*, 1962 and 1990. (Source: Nonparametric regressions run with food balance data from FAOUN and GNP data from the World Bank for 134 countries; Guo X, Mroz TA, Popkin BM and Zhai F (2000) Structural changes in the impact of income on food consumption in China, 1989–93. *Econoomic Development and Cultural Change* **48**: 737–60.)

income countries. Recent analysis in China shows that the pace of change for increased energy density and animal source foods in the diet has accelerated.

There is also an equally large and important shift in the proportion of energy from added caloric sweeteners in the diets of lower income countries. In fact, an additional 100–200 kcal per day was available for daily consumption from added caloric sweeteners in the diet in 2000 compared to 1962 in the developing world. In the United States, this added caloric sweetener increase derives mainly from soft drinks and fruit drinks, but in many other countries the source of this increase is other foods, even basic processed foods that have sweeteners added to them. Increasingly, high-fructose corn syrup is used as the sweetener of choice. This is unfortunate because there are mechanisms by which glucose may limit intake, but not fructose.

When we specifically examine the combined effect of these various shifts in the structure of rural and urban Chinese diets, we find an upward shift in the energy density of the foods consumed. In this study, the kilocalories of energy intake from foods and alcohol per 100 grams of food in both urban and rural Chinese adult diets increased by more than 10% (to 2.42) between 1989 and 1997. These are very rapid shifts in energy density. It is important to note that the value of 2.42 is not comparable with the normal measure of energy density of the diet. The normal method includes full measures of all beverages, whereas the Chinese Food Composition Table, from which this data was extracted, measures only a few beverages (milk, coconut juice, sugarcane juice, spirits, beer, wine, champagne, and brandy) and excludes many beverages, particularly tea and coffee. A number of clinical investigations have varied the energy density of the diet in ad libitum studies. Each study shows that increases in energy density, often as small as from 1 to 1.3 kcal/g, can increase total energy intake. For these reasons, energy density changes in China, and most likely in other developing countries, are critical components of dietary change to be monitored.

Rapid Social Change Is Important: Urbanization, Rapid Demographic Change, and Other Behavioral Changes Are Occurring Simultaneously

Diets have shifted in urban areas in a far more dramatic fashion than in rural areas. We do not focus on many of the complex issues related to the type of urban change that has occurred. Nevertheless, critical sociodemographic issues include the following:

- Rapid reductions in fertility have enhanced the shift in the age distribution.
- Urbanization continues unabated in Asia and Africa. More poor will reside in urban than rural areas in future decades.
- Economic changes, particularly increased income and income inequality, appear to define changes in many regions of the developing world.
- Globalization of mass media is occurring at an earlier stage of economic development than occurred in higher income countries in the past.

Urbanization

In other published work, we have shown how the structure of diet has shifted markedly as populations have urbanized. This relationship will, by itself, shift the structure of diet significantly at the national level as urbanization continues and as the proportion of the population in urban areas grows.

Structural Shifts in Income–Diet Relationships Are Occurring

Changes in dietary behavior can be caused either by shifts in the composition of society regarding the plurality of the educated, rich, or urban residents or by changes in actual behavior of those with specific characteristics. This latter type can include a change in consumption behavior such that for the same level of education or income, a person would buy different amounts or types of commodities at different points in time. Research conducted in China shows that there have been profound behavioral shifts of this type during the past decade (i.e., for each extra dollar of income, additional high-fat foods are purchased vs. what would have been purchased in previous years for the equivalent extra dollar). Economists speak of this effect as one that shows how the decision-making demand pattern for food has changed, so for the same income level the patterns of demand have changed significantly from earlier periods. The explosion in access and exposure to mass media may very well have created this situation.

Mass Media

There is no doubt that access to modern mass media has increased very rapidly, particularly in the past decade. Elsewhere, we have shown worldwide trends. It is most useful to examine the proportion of households in a country that have TV sets to gain insight into this topic. Again, we use China Health and Nutrition Survey (CHNS) data to demonstrate the types of changes in one setting. Overall, 88.5% of Chinese households in the CHNS sample had TVs in 1997. It is important to note that not only the proportion of people with access to TV was shifting but also the types of programs and access to Western influences were shifting. In the 1980s, cable systems in China did not provide outside programming, but by 1997 approximately one-fourth of Chinese provinces provided access to Phoenix Star TV, a Hong Kong TV system that relies heavily on US and British programming and provides modern TV advertising.

Again, although there are no extensive data on the proportions of Chinese households with access to mass media 30–70 years ago, it is certain that the penetration into Chinese households in 1997 was far greater than it was into US households 50 years ago, when TV was in its infancy.

Health Effects: Is the Biology Different? Rather, Do We Have Different Social Structures and Body Composition Patterns That Affect BMI–Disease Relationships? Are There Genetic Variants That Are Important?

There are a number of different ways these questions could be answered in the affirmative. One is if the body composition and other unmeasured race/ethnic factors affect susceptibility to NR-NCDs. Another might be if previous disease patterns (e.g., the presence of malaria or other tropical diseases) led to disease patterns that predisposed the population to certain problems. One component of this may be the fetal insult syndrome developed and popularized by Barker.

There is a growing body of research that shows the international standards, used to delineate who is overweight and obese, are not appropriate for many large subpopulations in the world. For instance, a BMI of 25 in an Asian adult appears to have a far greater adverse metabolic effect than in a Caucasian adult. In fact, the World Health Organization and the International Obesity Task Force formed a group of scientists and agencies in Asia to review this topic. This group held international meetings and has proposed a lower BMI cutoff for Asians of 23 for overweight and 25 for obesity. In one paper comparing China, the Philippines, and US Hispanics, blacks, and whites, the odds of being hypertensive were higher for Chinese men and women compared to other subpopulation groups at lower BMIs in the 23–25 range. Ethnic differences in the strength of the association between BMI and disease outcomes warrant further consideration.

Zimmet and others who have focused on this issue as it relates to lower income countries believe that the highest genetic susceptibility for adult-onset diabetes is for Pacific Islanders, American Indians, Mexican Americans and other Hispanics, and Asian Indians. Those with modest genetic susceptibility include Africans, Japanese, and Chinese. The age of onset (usually after age 50 years) of non-insulin-dependent diabetes mellitus is much lower for these susceptible populations, and it appears that the prevalence is higher for a given level of obesity and waist:hip ratio. Zimmet

summarizes a large selection of literature that has explored these issues relating to diabetes among susceptible populations.

It is not clear how much of this difference between subpopulations regarding BMI–diabetes or other BMI–morbidity relationships is a function of differences of body composition, metabolic or genetic factors, or social causes. We have shown that part of the apparent race–hypertension relationship may also be explained by socioeconomic status.

Another dimension relates to the issue of inflammatory burden. Evidence that inflammation plays a central role in cardiovascular disease (CVD), particularly at all stages of atherosclerosis, is persuasive. This position is supported by basic science and epidemiology. As reviewed in a meta-analysis, the magnitude of the associations between CVD outcomes and levels of inflammatory factors, such as C-reactive protein, albumin, white blood cell count, and fibrinogen, is surprisingly consistent across studies, despite differing designs, populations, duration of follow-up, and case definitions.

There is another pathway related to the role of previous health problems for which there is less understanding and no real documentation of its impact (e.g., malnutrition that caused a virus to mutate, parasitic infections that affected long-term absorption patterns, or a parasite that is linked to an unknown genotype—comparable to sickle cell anemia and its evolutionary linkage with malaria). We have no basis for speculation about this potential pathway.

However, the final pathway—the effect of fetal and infant insults on subsequent metabolic function—appears to be a critical area. If the rapid shifts toward positive energy imbalance are occurring concurrently with higher levels of low birth weight in a population, then this becomes a much more salient aspect of this argument. For the developing world, where intrauterine malnutrition rates are high and there is a high prevalence of nutrition insults during infancy, the work of Barker and many others portends important potential effects on the prevalence of NR-NCDs in the coming decades. Not only is there an emerging consensus that fetal insults, particularly with regard to thin, low-birth-weight infants who subsequently face a shift in the stage of the transition and become overweight, are linked with increased risk of NR-NCDs but also infancy may equally be a period of high vulnerability. Three studies by Hoffman suggest that fat metabolism of stunted infants is impaired to the extent that this may lead to increased obesity and other metabolic shifts. Other work on the role of stunting on obesity suggested such an effect, but Hoffman's work suggests the mechanism and fits with the correlational work.

The CVD Epidemic Is Beginning

Evidence from many developing countries shows that nutrition-related chronic diseases prematurely disable and kill a large proportion of economically productive people, a preventable loss of precious human capital. This includes countries in which HIV/AIDS is a dominant problem. Four out of five deaths from nutrition-related chronic diseases occur in middle- and low-income countries. Reddy reviewed these data and noted that

> the current high burden of NCDs is highlighted by the estimates for 1998 that indicate these disorders contributed to 58.8% of global mortality and 43% of the global burden of disease, measured as disability adjusted life-years lost. The contribution of low- and middle-income countries to this burden is large; about 77% of the total mortality and 85% of the total burden of disease attributable to NCDs arises from these countries. (page 231)

The burden of cardiovascular disease alone is now far greater in India, and also in China, than in all economically developed countries in the world combined. Low-income communities are especially vulnerable to nutrition-related chronic diseases, which are not just diseases of affluence. CVD, cancer, diabetes, neuropsychiatric ailments, and other chronic diseases are becoming major contributors to the burden of disease, even as infections and nutritional deficiencies are receding as leading contributors to death and disability.

Furthermore, CVD in the developing world emerges at an earlier age. As Reddy notes,

> Thus in 1990, 46.7% of CVD-related deaths in developing countries occurred below the age of 70 years, in contrast to only 22.8% in the high-income industrial countries. The Global Burden of Disease Study projected 6.4 million deaths would occur due to CVD in the developing countries in 2020, in the age group of 30–69 years. (page 233)

A World Health Report updates this analysis and focuses on the important role of obesity and CVD and cancer deaths in the developing world.

There are major differences in the profiles of the CVD epidemic across the developing world. For instance, hypertension and stroke are more likely to emerge in east Asia, whereas diabetes occurs earlier in south Asia.

As would be expected from the dietary and obesity data noted previously, CVD levels are far greater in urban areas of the developing world, but

often the opposite is true in the developed higher income countries.

Social Burden of Changes in Diet, Body Composition, and Health

In higher income countries, increasingly higher income groups follow a more healthful lifestyle, whereas the poor do not. Thus, higher income Americans consume a more healthful diet pattern, exercise more, and smoke less, and similar patterns are found in other high-income countries. In contrast, the prevailing opinion has been that the opposite is found in the developing world, namely that the poor are less likely to have a heavy burden of NR-NCDs compared to the rich. This is changing rapidly. It has been shown that obesity has declined among the better educated and increased among the lower educated in southeastern Brazil. It has also been shown that not only are less healthful dietary patterns consumed by higher income Chinese but also other dimensions of lifestyle (inactivity, smoking, and drinking) are poorer among the higher socioeconomic status (SES) Chinese. In other research, scholars of China have shown a rapid shift in food consumption patterns among different income groups in China that seems to indicate a shift in the burden of poor diets toward the poor in China. It has been shown that for countries with a GNP per capita of more than $2500, the likelihood is very great that there will be more obesity among the lower SES groups compared to higher SES groups.

The Future

Consuming a more tasteful and richer diet is a goal of most of the world's population. As shown here, dietary change is universal. In particular, rapid change is seen in the poorest areas of the world. The challenge is to learn how to continue to improve the palatability and quality of our diet while doing so in a more healthful manner.

See also: **Diabetes Mellitus**: Etiology and Epidemiology. **Dietary Intake Measurement**: Methodology; Validation. **Famine**. **Fats and Oils**. **Nutrient Requirements, International Perspectives**. **Obesity**: Definition, Etiology and Assessment.

Further Reading

Adair LS, Kuzawa CW, and Borja J (2001) Maternal energy stores and diet composition during pregnancy program adolescent blood pressure. *Circulation* **104**: 1034–1039.

Barker DJP (2001) *Fetal Origins of Cardiovascular and Lung Disease* New York: Marcel Dekker.

Bell AC, Adair LS, and Popkin BM (2002) Ethnic differences in the association between body mass index and hypertension. *American Journal of Epidemiology* **155**: 346–353.

Bell C, Ge K, and Popkin BM (2001) Weight gain and its predictors in Chinese adults. *International Journal of Obesity* **25**: 1079–1086.

Bell EA, Castellanos VH, Pelkman CL *et al.* (1998) Energy density of foods affects energy intake in normal-weight women. *American Journal of Clinical Nutrition* **67**: 412–420.

Bray GA, Nielsen SJ, and Popkin BM (2004) Consumption of high-fructose corn syrup in beverages may play a role in the epidemic of obesity. *American Journal of Clinical Nutrition* **79**: 537.

Drewnowski A and Popkin BM (1997) The nutrition transition: New trends in the global diet. *Nutrition Reviews* **55**: 31–43.

Eaton SB and Konner M (1985) Paleolithic nutrition: A consideration on its nature and current implications. *New England Journal of Medicine* **312**: 283–289.

Hoffman DJ, Sawaya AL, Coward WA *et al.* (2000) Energy expenditure of stunted and nonstunted boys and girls living in the shantytowns of Sao Paulo, Brazil. *American Journal of Clinical Nutrition* **72**: 1025–1031.

International Diabetes Institute (2000) *The Asia-Pacific Perspective: Redefining Obesity and Its Treatment.* Victoria Australia: Health Communications Australia.

Monteiro CA, Conde WL, and Popkin BM (2002) Is obesity replacing or adding to undernutrition? Evidence from different social classes in Brazil. *Public Health Nutrition* **5**(1A): 105–112.

Omran AR (1971) The epidemiologic transition: A theory of the epidemiology of population change. *Milbank Quarterly* **49**: 509–538.

Popkin BM (2002) The shift in stages of the nutrition transition in the developing world differs from past experiences! *Public Health Nutrition* **5**(1A): 205–214

Popkin BM and Nielsen SJ (2003) The sweetening of the world's diet. *Obesity Research* **11**: 1325–1332.

Reddy KS (2002) Cardiovascular disease in the developing countries: dimensions, determinants, dynamics and directions for public health action. *Public Health Nutrition* **5**(1A): 231–237.

Zimmet PZ, McCarty DJ, and de Courten MP (1997) The global epidemiology of non-insulin-dependent diabetes mellitus and the metabolic syndrome. *Journal of Diabetes Complications* **11**(2): 60–68.

NUTRITIONAL ASSESSMENT

Contents

Anthropometry

J Eaton–Evans, University of Ulster, Coleraine, UK

This article is reproduced from the first edition, pp. 1357–1363, © 1999, Elsevier Ltd. with revisions made by the Editor.

Anthropometric measurements include weight, height, length, selected skinfold thicknesses, and head, waist, hip, and arm circumferences. When compared with reference values, these measurements or combinations of these measurements can provide information on body size and the proportion and distribution of body fat and lean body mass in adults; they can also be used to assess growth in children. Anthropometric measurements indirectly indicate present or past nutrition and may be markers of future ill-health.

This article reviews the uses, advantages, and limitations of anthropometric measurements; discusses the technical errors of the measurements; describes the most frequently used measurements, derived nutritional indices, and reference values; and summarizes the laboratory methods that may validate the assessments.

Uses of Anthropometric Measurements

In adults and children, anthropometric measurements can be used to estimate body fat and lean body mass and assess their distribution and change over time. Body fat includes storage fat, found inter- and intra-muscularly, around the organs and gastrointestinal tract and subcutaneously, as well as lipids in bone marrow, central nervous tissue, mammary glands, and other organs. Normal-weight men and women have about 10 and 20% body fat, respectively. Lean body or fat-free mass is mostly water and protein with relatively small amounts of glycogen and minerals. Inadequate diets are associated with low body fat stores and reduced lean body mass in adults and growth failure of children. Consumption of food greater than requirements results in excessive body fat stores in adults and children. Body fat stores that are too low or too high are associated with increased risk of morbidity and mortality. The proportion and distribution of fat and fat-free mass varies with age, sex, genetics, disease, some hormones, and some drug treatments. Extensive physical exercise may be associated with increased muscle mass.

Different anthropometric measurements and combinations of measurements provide information on body composition and fat distribution and, therefore, nutritional status. The choice of measurements depends on the purpose of the assessment, the equipment available, the subjects being measured, and the skills of the observer making the measurements. Measurements can be made in laboratories, clinics, and hospitals using fixed, precision equipment with a high degree of accuracy, or in the field, including peoples' homes or rural centers, with lighter, robust, and portable equipment.

Advantages and Limitations of Anthropometric Measurements

Anthropometric measurements are noninvasive. Compared with other methods of assessing nutritional status, the measurements are quick and easy to make using relatively cheap and simple equipment. They can be made by relatively unskilled people.

Anthropometric measurements cannot identify protein and micronutrient deficiencies, detect small disturbances in nutritional status, nor identify small changes in the proportions of body fat to lean body mass. Some anthropometric measurements may not be socially or culturally acceptable, such as the measurement by men of womens' subscapular and supra-iliac skinfold thicknesses; some measurements may be impractical to make, such as the height in people who are unable to stand straight. Observers with limited literacy skills may not be able to read and therefore record some measurements. A single anthropometric measurement, such as weight, does not normally in itself assess growth and/or body composition and, therefore, indicate nutritional

status. To interpret anthropometric measurements, single measurements or combinations of measurements must be compared with reference values, by age and sex. Such reference values are not available for all population groups nor for all ages.

Errors of Anthropometric Measurements

All anthropometric measurements should be made as accurately as possible. Measurement errors may result in the misclassification of subjects' nutritional status or may lead to changes in nutritional status over time being over- or underestimated. Very precise and accurate measurements are needed for nutrition research and in some clinical situations. The same degree of precision may not be possible in nutritional screening and surveillance programs in field studies. Errors in making measurements arise from the equipment, the physical state and age of the subjects, the time of day when the measurements are made, misreading of measurements by the observer, and as a result of rounding up or down to the nearest half or whole integer. These technical errors of measurement (TEM) vary with the age of the subjects, the measurements being made, and between (inter-) and within (intra-) observers. Values for a particular anthropometric measurement of a group of people by age and sex can be considered accurate if the inter- and intraobserver error is close to a reference value for TEM in a series of repeated measurements and if there are no biases in the measurement. For measurements of subjects outside the age range, the coefficient of variability (R) can be calculated as $R = 1 - [(TEM)^2/(SD)^2]$, where SD is the total intersubject variance including measurement error. It has been recommended that an R of 0.90, that is a measurement 90% error-free, is an acceptable lower limit of accuracy, although an intraobserver R of 0.95 might be more realistic in some circumstances.

TEM can be minimized by careful training of all observers and by making measurements using appropriate equipment in triplicate and then calculating the mean. If measurements for a research study are to be made by more than one person, the interobserver measurements made must be comparable. R can be calculated for interobserver variability by making a series of measurements.

Anthropometric Measurements

Height

Height, or stature, is measured in adults and children over the age of 2 years using a stadiometer, a portable anthropometer, or a moveable headboard on a vertical measuring rod. The measuring device should be checked for accuracy using a standard 2-m steel tape. Subjects should be measured to the nearest 0.1 cm. Subjects, in minimal clothing with bare heads and feet, should stand straight, arms hanging loosely to the side, feet together and with heels, buttocks and shoulder blades in contact with the vertical surface of the stadiometer. Errors occur if subjects do not stand straight, do not keep heels on the ground, or overstretch. Diurnal variation results in people being 0.5–1 cm shorter in the evening than in the morning.

Height cannot be measured accurately in adults with severe kyphosis of the spine and in those who are bed- or chair-ridden. Since knee height is highly correlated with stature, height in such adults can be estimated from the measurement of knee height, using a sliding calliper. The regression equations, derived from a nonrandom sample of American people over the age of 60 years, are:

$$\text{Height (cm) for men} = (2.02 \times \text{knee height, cm}) - (0.04 \times \text{age, years}) + 64.19$$

$$\text{Height (cm) for women} = (1.83 \times \text{knee height, cm}) - (0.24 \times \text{age, years}) + 84.88$$

Variations in the proportion of limb length to trunk length can lead to a standard error in the estimate (SEE) of height from knee height of ± 8 cm. Demispan, which is the distance between the sternal notch of the right collar bone and the left finger root of the middle and ring finger when the subject's arm is horizontal and in line with the shoulders, can also be used to estimate height.

Length, rather than height, is measured in infants and children under the age of 3 years. Length is measured by laying a child face upwards on a measuring board with the head against the fixed headboard, and moving another board up to and resting against the child's heels with the legs straight (**Figure 1**). Small changes in length (± 0.5 cm) may not be significant as it is a difficult measurement to make. Children wriggle and will not stretch out their legs. Length measurements are 1–2 cm longer than height.

Height (stature) or length indicates attained size or growth of adults and children. Long periods of inadequate food intake or increased morbidity result in a slowing of skeletal growth and individuals being short for their age, or stunted. Consecutive measurements of height every 3–6 months can be used to assess growth velocity in children and to indicate the timing of the adolescent growth spurt.

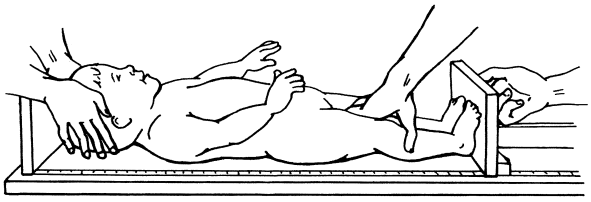


Figure 1 Measurement of recumbent length in children younger than 3 years of age. The head should be in contact with the fixed headboard, with child facing straight up. With legs fully extended, the mobile footboard should be placed firmly against the infant's heels. (Reproduced with permission from Frisancho AR (1990) *Anthropometric Standards for the Assessment of Growth and Nutritional Status*. Ann Arbor: University of Michigan Press.)

Weight

Weight is measured with digital weighing scales, using a pan, basket, sling, standing platform or chair, depending on the age and mobility of the people being measured. Weighing scales must be set on a hard, level, and even surface. Scales should be accurate, sensitive, and robust. They must be carefully maintained, calibrated, regularly checked for accuracy using known weights, and always set at zero before use. Weight is usually measured to the nearest 0.1 kg for adults and 0.01 kg for infants.

Weight measures total body mass but does not provide information on the proportions of fat, water, protein, and minerals. Weight and fat are only synonymous in very heavy people. Adults can be heavy for height if very muscular, overfat, and/or big framed. With accurate scales, small changes in weight are detectable but may not necessarily reflect change in body fat or lean body mass. In healthy persons, day-to-day variation in body weight is usually small (± 0.5 kg). Consecutive measurements of weight can be used to monitor the effects of treatment such as weight loss on reduction diets or weight gain with nutritional interventions and supplementation. Weight changes are assumed to reflect changes in the amount of body fat. However, changes in body weight may also result from differences in hydration, oedema, tumour growth, and trauma, as well as from factors such as the amount of food in the gastrointestinal tract and the fullness of the bladder. Weight may remain constant if the loss of muscle mass is masked by increased fat as seen in sarcopenia, the age-related loss of muscle, or by increased fluid retention.

Weight-for-height (or length) can be used to indicate body composition in adults and is an age-independent measure of body composition in children. Growth can be measured in children by consecutive measurements of weight over time (growth velocity) or by weight-for-age if the children's ages are known.

Head Circumference

Head circumference is measured in infants and young children, to the nearest 0.1 cm, with a narrow flexible nonstretch tape laid over the supraorbital ridges and the part of the occiput which gives the maximum circumference. The head circumference of infants increases rapidly in the first 2 years of life. Increase in head circumference in the first 2 years of life is affected by nutritional status and nonnutritional problems, including some diseases, genetic variation, and cultural practices.

Mid-Upper Arm Circumference

Mid-upper arm circumference (MUAC) is measured in adults and children, to the nearest 0.1 cm, using a flexible nonstretch tape laid at the midpoint between the acromion and olecranon processes on the shoulder blade and the ulna, respectively, of the arm (**Figure 2**). MUAC is a measure of the sum of the muscle and subcutaneous fat in the upper arm. In severe malnutrition both fat and muscle are reduced in the upper arm. Oedema may increase a limb's circumference but it is not usually a problem of the upper arm. MUAC can be used as a indicator of body composition in adults and children. Since MUAC increases little between the age of 6 months and 5 years, it can be used in preschool children as

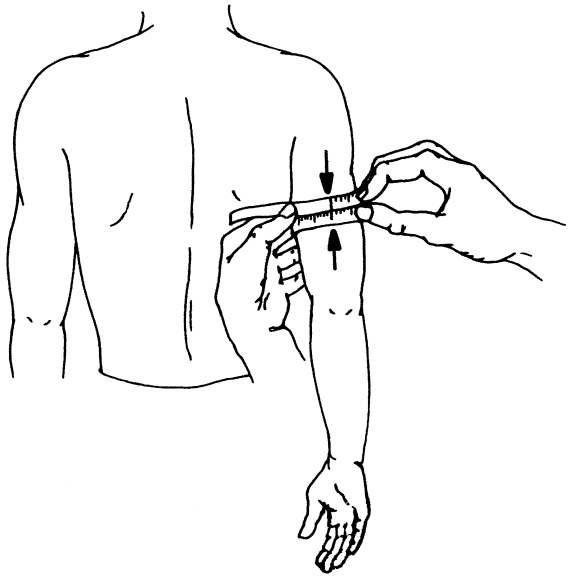


Figure 2 Measurement of upper arm circumference at the mid-point of the upper arm. (Reproduced with permission from Frisancho AR (1990) *Anthropometric Standards for the Assessment of Growth and Nutritional Status*. Ann Arbor: University of Michigan Press.)

an age-independent screening tool for severe malnutrition. A MUAC less than 12.5 cm suggests malnutrition. A MUAC greater than 13.5 cm is normal.

Skinfold Thickness

Precision skinfold thickness callipers are used to measure the double fold of skin and subcutaneous fat to the nearest millimeter. The usual sites of measurement are at the triceps (TSFT), the midpoint of the back of the upper arm (**Figure 3**); the biceps (BSFT) at the same level as the TSFT but to the front of the upper left arm; the subscapular (SSFT) just below and laterally to the left shoulder blade (**Figure 4**); and the suprailiac (SISFT) obliquely just above the left iliac crest. Skinfold thicknesses can also be measured at the mid-thigh, mid-calf, and abdomen.

Skinfold thicknesses are difficult measurements to make with precision and accuracy: It is difficult to pick up a consistent fold of skin and subcutaneous fat; in the very obese, the skinfold may be bigger than the callipers can measure; the fold of skin and fat compresses with repeated measurements; and the careless use of the callipers causes pain, bruising, and skin damage to subjects. There is, therefore, likely to be considerable inter- and intraobserver error in the measurements.

Skinfold thicknesses measure subcutaneous body fat and, therefore, indicate body composition. TSFT and SSFT indicate subcutaneous fat on the limbs and body trunk, respectively. Skinfold thickness measurements mistakenly assume that subcutaneous fat, measured at one or more selected sites, measures total body fat stores. However, subcutaneous fat at one site may not reflect fat stores at another site, and may not be positively correlated with the amount of visceral fat deposited around the internal organs of the body. Subcutaneous fat, and therefore skinfold thicknesses at the different sites, changes at varying rates with age, weight change, with diseases such as diabetes, and in women during pregnancy, postpartum, and at the menopause. Skinfold thicknesses are not useful for monitoring short-term change in fat stores. If only one skinfold thickness measurement is made, TSFT is most commonly selected. TSFT correlates with estimates of total body fat in women and children. SSFT is better than TSFT as an indicator of total body fat in men. SSFT has been shown to be a predictor of blood pressure in adults independently of age and racial group.

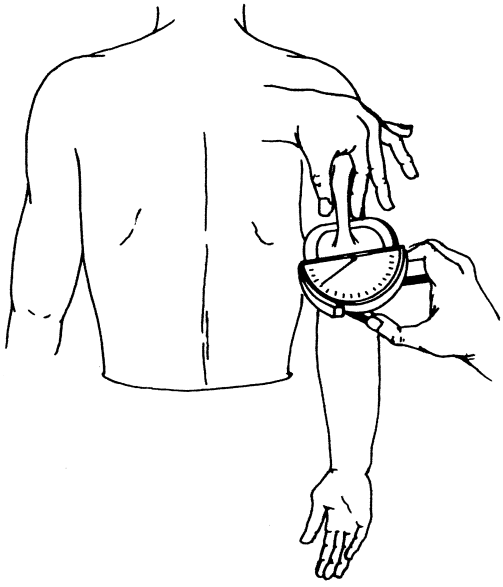


Figure 3 Measurement of triceps skinfold using a Lange caliper. With the subject's arm in a relaxed position, the skinfold is picked with thumb and index fingers at the mid-point of the arm. (Reproduced with permission from Frisancho AR (1990) *Anthropometric Standards for the Assessment of Growth and Nutritional Status*. Ann Arbor: University of Michigan Press.)

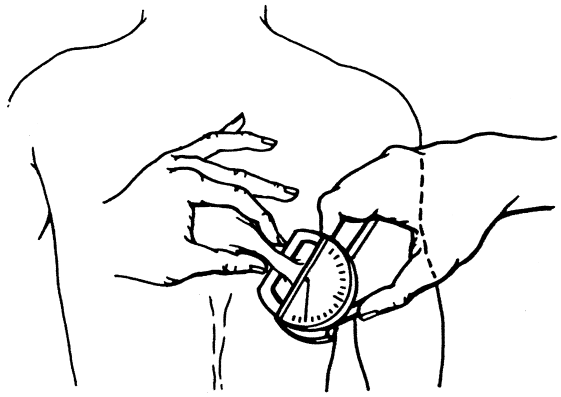


Figure 4 Measurement of subscapular skinfold using a Lange caliper. With subject's arm and shoulder relaxed, a horizontal skinfold is picked approximately 1 cm below the tip of the scapula with thumb and index fingers. The caliper is applied 1 cm from fingers. (Reproduced with permission from Frisancho AR (1990) *Anthropometric Standards for the Assessment of Growth and Nutritional Status*. Ann Arbor: University of Michigan Press.)

Waist and Hip Circumferences

Waist and hip circumferences are measured to the nearest 0.1 cm using a flexible narrow nonstretch tape in adults wearing minimal clothing, standing straight but not pulling in their stomachs. Waist circumference is measured halfway between the lower ribs and the iliac crest, while hip circumference is measured at the largest circumference around the buttocks. Measurement error occurs if the tape is pulled too tight or loose, or if subjects wear clothes with belts and/or full pockets.

With increase in waist circumference there is an increase in insulin sensitivity, while a waist circumference greater than 94 cm in men and 80 cm in women has been associated with increased risk factors for cardiovascular disease.

Elbow Width

Elbow width is the width of the epicondyles of the humerus with the elbow flexed at 90°. Sliding callipers are used to measure elbow width in adults to the nearest 0.1 cm. Elbow width is a measure of bone size. Frame size can be determined by comparison with reference values either by age or by height and sex.

Nutritional Indices

Most single anthropometric measurements do not in themselves assess nutritional status. Nutritional indices are derived either by combining two or more anthropometric measurements, shown in laboratory studies to be predictive of body composition, or by comparison of the anthropometric measurements with reference values of healthy, well-fed populations. A combination of these methods can also be used.

Body Mass Index

Body mass index relates weight (kg) with height (m) by a simple calculation to indicate body composition ($BMI = weight/height^2$). It is the most commonly used screening measurement for both obesity and underweight as very low and high BMI are associated with increased mortality and morbidity. BMI classifies adults as underweight, normal, overweight, or obese (**Table 1**). BMI and percentage body fat is only highly correlated at extremes of the distribution. BMI cannot distinguish between adults who are heavy because of fat or heavy because of muscle, takes no account of frame size, and provides no information on body fat distribution. These limitations may result in heavy, muscled sports people being classified as overweight or obese.

Table 1 Classifications of nutritional status as a percentage of ideal body weight and body mass index

% of ideal body weight for height	*Body mass index*	*Nutritional status*
>120	>30	Obese
110–120	25–29.9	Overweight
90–109	20–24.9	Normal
<90	<20	Underweight

In children, BMI is age-dependent. BMI increases rapidly in the first year of life and then more slowly. Children, at extremes of the distribution of reference values by age and sex, can be identified as being abnormally thin or fat for height. The BMI classification used for adults should not be used for children.

Weight-for-Height

Weight-for-height is an indicator of body composition in adults. Reference values by age, sex, and frame size can be used to estimate desirable body weights (%desirable body weight = (actual weight ÷ ideal body weight) × 100), which are categorized by cutoff points (**Table 1**).

Weight-for-height by sex is a sensitive indicator of body composition in children. It appears to be relatively independent of ethnic group in children aged 1–5 years and age-independent in children aged 1–10 years. Children with weights less than 85% of the median reference weight-for-height are considered wasted. It is a useful screening tool for current malnutrition, especially if used with height-for-age. Oedema and obesity, however, may confound the index.

Weight-for-Age

Weight-for-age can be used to monitor growth in children of known age when a series of measurements are made and compared with reference values by sex. A single measurement of weight-for-age does not discriminate between a child who is light for age because of stunting and/or wasting owing to malnutrition, and one who is small for age but healthy and well fed. Children should gain weight as a percentile of the reference values. Failure to gain weight as expected or a weight loss indicates an inadequate diet, infection, and/or lack of care and should be investigated. Maintenance of weight or weight gain may mask the loss of lean body fat and the increased oedema of kwashiorkor.

Growth Velocity

Growth velocity, or change in weight or height over time, can be used to assess growth in children when compared with reference values by age and sex. Growth rates decline in the first few years of life and then increase with the pubertal growth spurt. Premature and small-for-dates children and those recovering from malnutrition and severe infections tend to have higher growth velocities (catch-up growth). Growth velocities are useful to monitor growth and assess the response to therapy including nutritional supplementation.

Head Circumference-for-Age

Head circumference-for-age by sex is used by paediatricians to identify children up to 2 years of age with severe chronic malnutrition pre- and postpartum and the need for further medical investigations. It is not a good indicator of children's nutritional status.

Mid-Upper Arm Circumference-for-Age

Mid-upper arm circumference (MUAC)-for-age indicates body composition (upper arm fat and muscle) in adults and children when used with measurements of weight and height. MUAC measurements are compared with reference values by age and sex. Since the rate of change of arm circumference is slow, it cannot be used to assess growth or monitor the response to therapy.

Mid-Upper Arm Circumference-for-Height

Mid-upper arm circumference-for-height (the QUAC stick) is a cheap, quick, age-independent screening tool for children with malnutrition. It is a vertical stick on which are inscribed the 80 and 85% median reference values for MUAC and height, respectively. A child is considered malnourished if the MUAC is less than 80% of the MUAC expected for height.

Skinfold Thickness-for-Age

Skinfold thickness-for-age and sex indicates subcutaneous body fat stores in adults and children. Reference values for TSFT, SSFT, and the sum of the TSFT and SSFT are available by age and sex.

Measurement of BMI with skinfold thicknesses can identify people who are heavy owing to excess fat or muscle mass. A high BMI and low TSFT and/or SSFT indicate a large muscle mass; a high BMI and high TSFT and/or SSFT indicate a high subcutaneous body fat.

Mid-Upper Arm Muscle Circumference and Upper Arm Muscle Area

Mid-upper arm muscle circumference (MUAMC) and upper arm muscle area (AMA) are estimates of upper arm muscle and, therefore, body composition. They can be used as indicators of muscle mass and protein stores. Both MUAMC and AMA are calculated from measurements of MUAC and TSFT on the mistaken assumption that the arm is cylindrical, the subcutaneous fat is equally distributed, the bone atrophies in proportion to muscle wastage in malnutrition, and the cross-sections of neurovascular tissue and bone are small. The formula, with MUAC and TSFT in mm, is:

$$\text{MUAMC} = \text{MUAC} - (\pi \times \text{TSFT})$$

AMA can be calculated from revised formulae which take account of errors resulting from the noncircular nature of muscle and the inclusion of nonskeletal muscle with MUAC and TSFT in cm:

$$\text{For men: AMA} = \left[\frac{\text{MUAC} - (\pi \times \text{TSFT})}{4\pi}\right]^2 - 10.0$$

$$\text{For women: AMA} = \left[\frac{\text{MUAC} - (\pi \times \text{TSFT})}{4\pi}\right]^2 - 6.5$$

MUAMC and AMA can be compared with reference values by age and sex. AMA cannot be used to monitor change in muscle stores because of the problems in making this measurement. The ratio of AMA to total body muscle mass changes with age and certain diseases.

Arm Fat Area

Arm fat area (AFA) can be derived from measurements of MUAC and TSFT. AMA is a better indicator of total body fat but not percentage body fat, than TSFT alone. The formula used to calculate AFA (with MUAC and TSFT in mm) is:

$$\text{AMA} = \frac{\text{TSFT} \times \text{MUAC}}{2} - \frac{\pi \times (\text{TSFT})^2}{4}$$

AMA can be compared with reference values by age and sex. Theoretically, limb fat area can be calculated for other limbs and the body trunk, but there are no reference values available.

Total Body Fat

Total body fat can be estimated as a percentage of body fat by comparing the sum of TSFT, BSFT, SSFT, and SISFT with reference values derived from laboratory studies by age and sex, or via the estimation of body density from regression equations with skinfold thickness measurements. There are no specific empirical equations which can be used for specific population groups. Lean body mass is calculated by difference. These calculations may overestimate body fat in lean individuals and underestimate body fat in fat adults. They should not be used in undernourished individuals or in those with diseases where the total body water content may be markedly increased.

Waist-to-Hip Ratio

The waist-to-hip ratio (WHR) in adults discriminates between those with upper body or intraabdominal obesity (WHR greater than 1 in men and 0.8 in women) and those with lower body or

peripheral obesity. Genetics, sex, and age partly determine body fat distribution. A high WHR is associated with an increased risk of premature mortality and morbidity.

Reference Values

Anthropometric assessments are interpreted by comparison with reference values by age and sex. Ideally, reference values should represent the range of 'optimum' measurements for health and longevity of a population of the same ethnic origin. Individuals should be considered at nutritional risk when above or below predetermined reference limits or cutoff points, based on functional impairment, clinical signs of deficiency, or increased risk of mortality and morbidity.

In practice, reference values are derived from large sets of cross-sectional anthropometric measurements of representative samples of populations of the same ethnic origin, who are assumed to be well nourished and free from infection, parasitic disease, and any other environmental factors which may affect growth and nutritional status. These data can be supplemented by data derived from direct laboratory studies of body composition. International reference values are used if local reference data are too difficult, time-consuming, and expensive to obtain.

Normal, healthy, well-fed people vary in size. Therefore, reference values are usually presented as percentiles, with values less than the 5th centile or greater than the 95th centile considered outside the normal range. If international reference values are used, it may be necessary to modify the cutoff points used for identifying those at risk for particular populations. Since the rate of growth of children is age-dependent, growth charts of the most commonly used anthropometric measurements and derived indices have been constructed. Children's growth is best monitored by plotting their sequential measurements on growth charts. A well-nourished healthy child should progress along a centile between the 5th and 95th centile for each measurement. When a child's measurements cross centiles of the growth charts, whether owing to growth faltering, failure-to-thrive, or excessive growth, the cause needs to be investigated. In this way anthropometric measurements can indicate the adequacy of a child's diet, the timing of the introduction of weaning foods, the impact of illness, and the response to treatment. To use growth charts such as weight-for-age, height-for-age, and head circumference-for-age, it is essential to know the age of a child accurately. A child's age cannot be estimated accurately by examination. A malnourished child is smaller and looks younger and less mature than a well-fed, healthy child.

To analyse data for screening, surveillance, or research programmes, measurements can be compared with the median (50th centile) of the reference data and expressed as either a percentage of the median value or a modified statistical z-score transformation, where

$$\text{SD score} = \frac{\text{individuals' measurement} - \text{median value of reference value}}{\text{standard deviation value of reference value}}$$

SD scores are appropriate for use in areas with a high incidence of malnutrition. A high proportion of the population have measurements less than the 5th centile of reference values in these areas. A child with a SD score less than -2, irrespective of the nutritional indices used, is considered malnourished. A SD score greater than 2 suggests obesity.

The most commonly internationally used anthropometric reference values for noninstitutionalized adults aged 25–74 years have been derived from the American NHANES I and NHANES II studies, which were undertaken during 1971–74 and 1976–80, respectively. Reference values by sex, height, and frame size for desirable weights-for-height have also been derived from data on the longevity of holders of life insurance policies. These reference values are not taken from a random sample of the population, as only the more affluent in society are likely to hold life insurance; in addition, these values take no account of body composition and make no reference to the incidence of disease.

Many countries have developed their own reference standards for weight and height of children and adolescents. In the United States, the Centers for Disease Control and Prevention (CDC) published in the year 2000 growth charts for children 0–20 years of age, which for the first time included age-adjusted BMI data. The CDC dataset excluded anthropometric information from the most recent surveys fro certain age groups because of the marked secular trend to higher body weights in the US population.

The United Kingdom, France, and several other countries have developed reference growth curves based on national datasets. The World Health Organization is also leading an international effort to develop a truly global reference standard for growth, including reference values that reflect growth rates of exclusively breast-fed infants, who are known to grow at lower rates than formula-fed

infants during the first year of life. Although growth rates differ across countries, there is general consensus that in healthy populations these differences are encompassed within the boundaries of acceptable percentile ranges, usually the 5th and 95th. Furthermore, environmental rather than genetic factors appear to be the main determinants of differences in growth across populations since migration of individuals from a poor to an adequate nutritional environment is usually accompanied by marked gains in height after only a few generations.

Children's growth is influenced by many factors, including sex, ethnic group, breast or bottle milk feeding, their birth order, gestational age of premature children, as well as the size (height) of their parents. Growth charts do not allow for these factors. With the secular changes in height and weight, growth charts derived from anthropometric measurements of children made over 30 years ago may no longer be appropriate to monitor growth of all children today. However, they may still be relevant to disadvantaged groups in the population. Similarly, growth charts of children derived from reference data of today may not be relevant to the children of the future.

A new reference data set, derived from mostly cross-sectional studies undertaken between 1978 and 1990 of 23 000 British children aged 0–20 years, has been developed by the Human Measurements Anthropometry and Growth Research Group. Growth charts from these data have been produced by sex for weight-for-age, height-for-age, length-for-age, head circumference-for-age, and BMI-for-age.

See also: **Dietary Intake Measurement**: Methodology; Validation. **Growth Monitoring**. **World Health Organization**.

Further Reading

Centers for Disease Control and Prevention (2005) *Growth Charts Dataset*. Available at www.cdc.gov/nchs/about/major/nhanes/growthcharts/datafiles.htm.

Frisancho AR (1990) *Anthropometric Standards for the Assessment of Growth and Nutritional Status*. Ann Arbor: University of Michigan Press.

Gibson RS (1990) *Principles of Nutritional Assessment*. New York: Oxford University Press.

Jelliffe DB (1966) *The Assessment of the Nutritional Status of the Community*. Geneva: World Health Organization.

World Health Organization (1995) *Physical Status: The Use and Interpretation of Anthropometry*, WHO Technical Report Series No. 854. Geneva: World Health Organization.

Biochemical Indices

F Fidanza, University of Rome Tor Vergata, Rome, Italy

Biochemical methods are considered to be the most objective measures for the assessment of nutritional status of the individual. The method employed should cover a range of cutoff points specific and sensitive to depletion of the nutrient body pool or tissue store.

The evolution of deficiency for most nutrients, particularly vitamins, progresses in successive stages. The first stage of deficiency is when nutrient body stores begin to be depleted; in this stage, nutrient urine excretion decreases, whereas homeostatic regulation ensures that the level of nutrient in the blood does not change. In the next stage, depletion is more marked; nutrient urinary excretion continues to decrease and its blood and other tissue concentrations are reduced.

A lowering of nutrient metabolites and/or dependent enzymes often characterizes the following stage. Sometimes, lower hormone concentrations and some physiological alterations are observed. In the last stages, morphological and/or functional disturbances are present; at first they are reversible, and then they become irreversible. Nonspecific signs and symptoms can be present; without therapeutic intervention, death can be expected.

Within the framework of the evolution of nutrient deficiencies, the biochemical static and functional tests most commonly used in nutritional status assessment in humans are discussed here.

Static Biochemical Tests

Static tests measure chemically the content of nutrients, their active or inactive metabolites, or other related components in tissues and urine. The choice of tissue or fluid depends on the information required (short-term or long-term status, body pool or tissue store) and on the condition of the subject.

Various confounding factors affect static biochemical tests. Some are of a general kind, such as age, sex, ethnic group, physiological and hormonal status, seasonality, elevation, and thus cannot be eliminated; others are of a technical nature and can be reduced or eliminated by standardization; and others are biological or environmental (e.g., alcohol intake, smoking habits, and use of medicines). The most relevant confounding factors are considered for

each method; those that occur during infection are examined separately.

Protein Nutritional Status

Total serum protein determination is very seldom used because it is no longer considered a sensitive index of status.

Plasma proteins are albumin, transport proteins (transthyretin (TTR) involved in thyroid hormone transport and formerly called prealbumin, retinol binding protein (RBP), and transferrin (TF)) and fibronectin (FB; an apsonic glycoprotein). Serum albumin, measured by an automated dye-binding method, has a rather large body pool and a long half-life and so it is a less sensitive index of immediate nutritional status. TTR, complexed with RBP in the carriage of vitamin A, TF, and FB have a smaller pool size and a shorter half-life than serum albumin and so their concentrations can change more rapidly. Therefore, they are immediate indicators of protein status. Plasma transport proteins are usually measured on radial immunodiffusion plates or alternatively with laser nephelometry. Useful commercial kits are available. Plasma fibronectin is measured only with laser nephelometry. Albumin and transport proteins are negative acute phase reactants. Other confounding effects of protein-losing diseases, such as reduced protein synthesis diseases, conditions involving an increase in plasma volume, or hemodilution and zinc depletion, have been reported. In addition, RBP is sensitive to deficiencies of vitamin A, and TF is affected by iron status. Insulin has also been demonstrated to interfere with plasma transport protein levels.

Urinary creatinine, usually measured with a colorimetric method (also automated), is used as a biochemical marker of muscle mass. In fact, urinary creatinine is a nonenzymatic product of creatine and cannot be reutilized. Various assumptions are required for correct urinary creatinine determination, and various confounding effects are reported (age, diet, intensive exercise, pregnancy, injury, fever, and renal diseases with impaired creatinine clearance). In a clinical setting, the creatinine/height index is preferred, but because of some limitations, it is not very useful.

Urinary 3-methyl-histidine (3-MH) can be measured by ion exchange chromatography or high-performance liquid chromatography (HPLC). 3-MH is present in myofibrillar proteins, and during breakdown it is excreted quantitatively because it cannot be reused or oxidized. Accordingly, it is used as an indicator of muscle protein turnover. Various confounding effects are reported (sex, age, diet, intensive exercise, stress, hormonal and catabolic states, etc.) and so the use of the urinary 3-MH test is considered to be rather problematic.

Insulin-like growth factor-1 (IGF-1), or somatomedin C, is a regulator of anabolic properties. It has been proposed as a sensitive indicator of protein deficiency. It is assayed in serum by a radioimmunoassay method available also in a kit. IGF-1 can also be used as a nutritional marker in adults receiving total parenteral nutrition. Confounding effects of stress, some hormonal diseases, and obesity have been reported.

Plasma amino acid levels have been used in the past to diagnose protein–energy malnutrition. The ratio of free nonessential amino acid levels (glycine, serine, glutamine, and taurine) to the essential amino acid levels (leucine, isoleucine, valine, and methionine) was proposed for the diagnosis of kwashiorkor. In children with this disease, this ratio can be much higher than the normal value of 2. Plasma amino acids were previously assessed by paper chromatographic methods; automated ion exchange or HPLC techniques are now preferred. However, in recent years there has been much less interest in this test.

Essential Fatty Acid Status

A number of measures can be used to assess deficiency of essential fatty acids. In serum cholesterol esters, fatty acids determination is related to recent intake, in erythrocyte membranes it is related to intake during the previous 2 or 3 months, and in subcutaneous fat tissue it is related to intake of fatty acids for more than 1 year. Essential fatty acids are measured by gas–liquid chromatography.

Vitamin Nutritional Status

Vitamin A (retinol) status can be assessed in the liver and plasma/serum. The best method is determination in the liver, but hepatic biopsy is very invasive and unsuitable in population studies. Plasma retinol is usually measured by HPLC after separation from its carrier (RBP), but its marginal values do not always reflect status because of homeostatic control and confounding effects (e.g., protein–energy malnutrition, infection, parasitic diseases, zinc deficiency, liver disorders, and chronic alcoholism). In the case of inflammation, the degree of depression of serum retinol can be quantified by assessing the concentration of certain acute phase proteins (CRP and AGP).

Because serum retinol is closely correlated with serum RBP, the measurement of this transport

protein by the immunodiffusion technique or a portable apparatus has been proposed to assess vitamin A status.

The RBP:TTR molar ratio has been introduced to detect vitamin A deficiency (VAD) in the presence of inflammation. This test was based on the observation that VAD and inflammation were independent causes of low plasma RBP, whereas plasma TTR concentration was reduced only by inflammation. Nonsatisfactory results were reported from two African population groups.

The deuterated-retinol-dilution (DRD) technique is used to indirectly assess total body vitamin A reserves. A dose of deuterium-labeled retinyl acetate is given orally. After allowing time to reach equilibration (3–20 days), deuterated and nondeuterated retinol is measured by gas chromatography-mass spectrometry. A mathematical formula is used to estimate total body stores of vitamin A. Because of a set of assumptions and technical difficulties, this method is used mostly in research projects. In inflammation, the release of RBP is inhibited, so the test is probably unreliable.

In connection with vitamin A, its provitamins and non-provitamins, the carotenoids, need some consideration. β-Carotene and a few other carotenoids play an independent and specific role in preventing oxidation, genotoxicity, and malignancy. The serum level of carotenoids is correlated with vegetable and fruit intake. Lutein is the best indicator of green leafy vegetable consumption. Lycopene is a good measure of tomato-based product consumption. α-Carotene in industrialized countries is probably a biomarker of carrot consumption and in West Africa a good marker of red palm oil consumption. The plasma level of carotenoids is measured by the HPLC system. However, there are difficulties with peak identification and quantification. Confounding effects of diet and season, sex and age, infection, smoking, and drinking habits are reported. With an appropriate HPLC system, it is possible to measure in a single assay vitamins A and E and individual carotenoids.

Vitamin D status is generally assessed by measurement of serum 25-hydroxyvitamin D (25-OHD) and in some circumstances 1,25-dihydroxyvitamin D ($1,25(OH)_2D$). The current test for 25-OHD and $1,25(OH)_2D$ determination in serum is by radioimmunoassay, also available as commercial kits. The HPLC method with ultraviolet detection can be used as an alternative. Confounding effects of seasons, age, sex, drugs, and liver and renal diseases are reported.

Vitamin E status can be assessed in plasma, erythrocytes, platelets, and adipose tissue. The most common and practical measure is α-tocopherol in plasma by HPLC. Because α-tocopherol is bound to lipoproteins, it is preferred to express plasma α-tocopherol relative to serum cholesterol. The determination of α-tocopherol in adipose tissue biopsy provides information on long-term nutritional status, but this test is too invasive. Confounding effects of chronic enteropathies, protein–energy malnutrition, hemolitic anemia, cholestatic liver disease, and some drugs and heavy metals are reported.

Vitamin K status requires a multiple approach including a functional test. Plasma phylloquinone is measured by reversed-phase HPLC using postcolumn chemical reduction followed by fluorometric detection. Determination of the serum undercarboxylated form of prothrombin (PIVKA-II) by enzyme-linked immunosorbent assay (ELISA) and urinary γ-carboxyglutamic acid by HPLC with fluorometric detection has been proposed. Confounding effects of age, sex, season, malfunction of gastrointestinal tract, osteoporosis, liver diseases, antibiotics, and other drugs are reported.

Thiamin status can be assessed by urinary excretion and erythrocyte thiamin pyrophosphate (TPP) tests. Thiamin urinary excretion is indicative of recent dietary intake; thiamin is detected fluorometrically after conversion to thiochrome. If 24-h urine cannot be collected, thiamin should be determined in the fasting morning urine and expressed in relation to creatinine concentration. Erythrocyte TPP is indicative of long-term nutritional status and is assessed by HPLC using fluorometric detection after precolumn derivatization to thiochrome pyrophosphate. The only limitation is TPP instability, and determination should be carried out within 2 h of blood drawing. Erythrocyte TPP levels present large interindividual variation probably as a results of confounding factors (age and sex, alcohol intake, smoking habits, physical activity, and drugs).

Riboflavin status can be assessed by urinary excretion and whole blood flavinadeninedinucleotide (FAD) tests. Riboflavin urinary excretion is indicative of recent dietary intake; riboflavin is measured by HPLC using fluorometric detection. As for thiamin, if fasting morning urine is collected, riboflavin value is expressed in relation to millimoles of creatinine. Confounding effects of physical activity, bed rest, chronic alcoholism, antibiotics, and other drugs are reported. Whole blood FAD is considered a reliable indicator of long-term nutritional status and is assessed by reversed-phase HPLC using fluorometric detection. This test presents some advantages over the functional test

erythrocyte glutathion reductase activation coefficient (EGR-AC).

Vitamin B_6 status is generally assessed by urinary 4-pyridoxic acid (4-PA) and whole blood or plasma pyridoxal-5′-phosphate (PLP) tests. The 4-PA test is indicative of recent intake but also of a deep compartment with slow elimination rate. 4-PA is measured by reversed-phase HPLC using fluorescence detection. When the completeness of 24-h collection is impossible, 4-PA is expressed in relation to millimoles of creatinine. PLP in whole blood or plasma is considered to be an indicator of depletion of vitamin B_6 reserves. In whole blood, PLP can be measured by reversed-phase HPLC using fluorometric detection. A HPLC system with fluorescence detector for determination of vitamin B_6 vitamers and pyridoxic acid in plasma is available. Plasma PLP can also be measured by radioenzymatic assay using tyrosine decarboxylase apoenzyme, which is more sensitive than other methods of analysis. Confounding effects of age and sex, acute phase status, tissue injury, catabolic state, smoking habits, alcoholism, pregnancy, drugs, physical exercise, organic diseases, and some inborn errors of metabolism are reported.

Niacin status can be assessed by measuring the two end products N'-methylnicotinamide (N'MN) and N'-methyl-2-pyridone-5-carboxamine (2-Py) in urine by HPLC. The ratio of these two urinary products is considered to be the best index of niacin nutritional status. With a single HPLC assay, the previously mentioned two nicotinamide metabolites and N^1-methyl-4-pyridone-3-carboxamide (4-Py) can be measured. The ratio (2-Py + 4-Py)/N'MN is proposed; it has a diurnal variation and decreases with cold. However, for nutritional status assessment further investigation is needed.

Folate status can be assessed by serum/plasma folate, which provides information on recent intake, and erythrocyte folate, indicative of body folate stores and long-term nutritional status. Folate is measured by radioassay kits, sometimes simultaneously with vitamin B_{12}. Less practical, although more accurate, are microbiological assays. In a EC-Flair programme intercomparison study, it was observed that radioassay tends to overestimate serum folate and presents considerable between-kit variability; improved standardization of diagnostic kits and the provision of suitable reference material are still of paramount importance. HPLC, liquid chromatography–mass spectrometry (LC–MS), and LC–MSMS methods are now available. Confounding effects of starvation, dietary folate intake and alcohol abuse, pregnancy, smoking habits, and drugs are reported for serum folate; iron deficiency, age, and other disease states are reported for erythrocyte folate.

Vitamin B_{12} status can be assessed by measuring serum or plasma total cobalamins and serum holotranscobalamin II. Serum or plasma cobalamins are determined by competitive protein-binding assay. Kits are available to measure folate simultaneously. Microbiological assays tend to give lower results. Confounding effects of age, sex, impaired absorption by some diseases or drugs, myeloproliferative disorders, worm infestations, and severe liver disease are reported. Holotranscobalamin II is the transport protein of absorbed cobalamin and has been considered as an early indicator of vitamin B_{12} deficiency and possibly a marker of cobalamin malabsorption. Plasma holotranscobalamina II is measured by microparticle enzyme intrinsic factor assay (together with total vitamin B_{12}) or by indirect immunoadsorption method.

Biotin status can be assessed in whole blood by microbiological assay. Radioimmunoassay tests are also available not only for plasma but also for urine. These tests give slightly higher values than microbiological assay.

Vitamin C status can be assessed by ascorbic acid in plasma, buffy-coat, and leucocytes. Ascorbic acid in plasma is considered an index of the circulating vitamin available to tissues, in buffy-coat it is indicative of the intracellular content, and in leucocytes (particularly polymorphonuclear) it is believed to be a good indicator of tissue stores. Whole blood and erythrocyte ascorbic acid determinations are considered of lesser value than plasma for ascorbic acid status assessment. Ascorbic acid in the previously mentioned blood components is measured with a dinitrophenylhydrazine assay and with a more practical HPLC method coupled with electrochemical or amperometric detectors. Also, a HPLC with fluorometric detection method is available. Confounding effects of acute stress, infection, surgery, smoking habits, chronic alcoholism, sex, and drugs are reported. The urinary excretion of ascorbic acid is an index of recent intake; because of instability of the collected sample, the determination is limited to special cases.

Essential Mineral and Trace Element Nutritional Status

Sodium and potassium in plasma/serum have little meaning in nutritional terms; total body Na or K are measured by radioisotope dilution.

Calcium status can be assessed measuring serum or plasma ionized calcium or indirectly by

measuring bone mass and bone density. Plasma ionized calcium provides information on physiological function and is measured by a calcium-selective electrode; bone calcium content is an index of body calcium stores and is measured by neutron activation analysis or dual-photon absorptiometry. Confounding effects of venous stasis, cardiac arrest, large volumes of citrated blood infusion, and high or low pH are reported for plasma ionized calcium.

Magnesium status can be assessed by measuring magnesium in serum, erythrocyte, leucocyte, and urine. Serum magnesium is the method most commonly used. Confounding effects of haemolysis, energetic exercise, and pregnancy are reported. Erythrocyte magnesium is considered indicative of a long-term status. Confounding effects of age, thyroid disease, and premenstrual tension are reported. Leucocyte magnesium is considered indicative of intracellular status. Urinary magnesium is used as an indicator of magnesium deficiency after a load test. Some precautions are necessary for this test. Magnesium is measured by flame atomic absorption spectroscopy (AAS) or automated colorimetric methods. The serum/plasma free ionized magnesium determination by selective electrode has been considered a better indicator of status. Further studies are required.

Iron status is assessed in relation to three stages of development of iron-deficiency anemia. In the first stage, to evaluate the size of body iron stores, serum or plasma ferritin can be measured by radiometric methods or using ELISA. Commercial kits are available. Confounding effects of infection, liver and malignant diseases, acute leukemia, Hodgkin's disease, rheumatoid arthritis, thalassemia major, alcohol consumption, age, and sex are reported. In the second stage, to determine the adequacy of iron supply to the erythroid marrow, serum iron (measured by the colorimetric method, available as commercial kits; AAS is not recommended because it gives higher values), plasma or serum total iron binding capacity (TIBC; by colorimetric or radioactive methods available as commercial kits), erythrocyte protoporphyrin (by specific hematofluorometer), and serum transferrin receptor (by ELISA using developed monoclonal antibodies) are measured. The percentage of transferrin saturation is computed as follows: serum iron/TIBC $\times$ 100. Confounding effects of infection, chronic alcoholism, folate and vitamins B_6, B_{12}, and C deficiencies, acute viral hepatitis, malignancy, shock, physical trauma, pregnancy, and altitude are reported for serum iron; infection, protein–energy malnutrition, alcoholic cirrhosis, malignancy, nephrotic syndromes, entheropathy, pregnancy, viral hepatitis, and oral contraceptive intake are reported for TIBC; and infection, lead poisoning, and porphyrin disorders are reported for erythrocyte protoporphyrin. In the third stage, as indicators of iron-deficiency anemia, hemoglobin (by spectrophotometry or automatically with an electronic counter), hematocrit or packed cell volume (by specially designed centrifuge or an electronic counter), and red cell indices (mean cell volume and mean corpuscolar hemoglobin, both by electronic counter) are measured. Confounding effects of chronic infection, deficiencies of folate and vitamin B_{12}, chronic diseases, hemoglobinopathies, parasitosis, sex, altitude, and smoking habits are reported. All tests of the third stage present low sensitivity and, for the confounding factors, low specificity. The measure of serum transferrin receptor seems to be a promising technique for the evaluation of iron deficiency or toxicity because it is not influenced by infection, inflammation, and chronic diseases. The assessment of serum ferritin and transferrin receptors is considered valuable in screening iron deficiency. Because the measurement of only one variable is not sufficient for the assessment of mild iron deficiency, and also to avoid other limitations, it is recommended to combine two or more independent variables.

Zinc status can be assessed by using AAS to measure zinc in plasma or serum, leucocyte and leucocyte subsets, urine, hair, nails, and saliva. Plasma or serum zinc is the method most commonly used. Many precautions are required during sample collection to avoid the influence of time of day, proximity of meal, stress, hemolysis, and contamination. There are also many pathophysiological conditions that can negatively influence specificity and sensitivity of serum zinc (e.g., infection, stress, chronic disease, exercise, oral contraceptive use, pregnancy, hypoalbuminemia, diabetes, starvation, severe malnutrition, and other catabolic conditions). Therefore, plasma zinc levels are generally considered a poor measure of marginal zinc deficiency. Leucocyte subset zinc, particularly monocyte zinc, is considered a useful indicator of zinc deficiency, but monocyte separation is difficult and a large blood sample must be collected. Zinc in other fluids or tissues is not considered a useful or reliable indicator of zinc deficiency.

Copper status is most frequently assessed in serum or plasma by AAS, even though this measure is of low sensitivity or specificity in the general population. Levels of copper in other tissues or fluids are difficult to assess or are not considered valid indices of copper status. Confounding effects of infection,

inflammation, pregnancy, leukemia, Hodgkin's disease, some anemias, myocardial infarction, malabsorption, ulcerative colitis, Wilson's disease, hepatitis, high-level physical activity, cigarette smoking, age, and sex are reported.

Selenium status is usually assessed measuring plasma or serum selenium by AAS with a Zeeman background correction and also by the fluorometric technique. Although plasma Se determination provides information on short-term Se status, the determination in whole blood or erythrocytes is indicative of long-term status. Confounding effects of some inborn errors of metabolism, congestive cardiomyopathy, age, and some physiological conditions are reported.

The determination of Se in urine presents some limitations and Se levels in hair and nails display some drawbacks.

Iodine status is generally assessed by measuring urinary iodine using the colorimetric method, which reflects iodine intake within the past few days. If 24-h urine cannot be collected, iodine excretion can be expressed per gram of creatinine but only in areas with very low inter-and intraindividual variation in urinary creatinine. In clinical settings, the measurement of uptake of radioactive iodine is used.

Functional Tests

Functional tests are defined by Solomons and Allen as tests that measure behavioral, physiological, or biochemical functions of the organism dependent on the adequate availability of a nutrient or responses to the regularity process to maintain body stores and harmonic internal distribution for those many nutrients that are homeostatically regulated by the organism.

There are few reliable and specific functional tests; other simpler tests, not commonly used, lack specificity. Further studies are required on these tests because they are important for a correct assessment of nutritional status in humans. Some of the confounding factors reported for static biochemical tests also apply to functional tests.

Vitamin Nutritional Status

Vitamin A functional tests are the relative dose–response (RDR) and the modified relative dose–response (MRDR). The RDR test can provide information on liver store and is indicative of marginal status of vitamin A. It consists of the determination of plasma retinol level at baseline (A0), administration of a small dose of retinyl acetate or retinyl palmitate, and a second determination of plasma retinol 5 h later (A5). The response is related to the release from the liver of holo-RBP, and in deficient subjects the plasma retinol will increase after 5 h. RDR is calculated as follows:

$$\mathrm{RDR} = (\mathrm{A5} - \mathrm{A0}) \times 100/\mathrm{A5}$$

The MRDR uses a metabolite of vitamin A (3,4-didehydroretinol (DR)). After a test dose, DR binds to RBP and after 5 h appears in the serum if vitamin A reserves are low. Serum retinol (SR) and DR (SDR) are measured by HPLC. The ratio is calculated as follows:

$$\mathrm{MRDR} = \mathrm{SDR}/\mathrm{SR}$$

This ratio is abnormal when greater than 0.06. Confounding effects of protein–energy malnutrition, malabsorption, inflammation (due to inhibition of release of RBP), and liver disease are reported for both tests.

The vitamin D functional test can be the measure of serum alkaline phosphatase activity. For determination of serum alkaline phosphatase, automated procedures and commercial kits are available. The specificity is not very high and confounding factors are age, sex, pregnancy, and unrelated pathologies.

Vitamin E functional tests consist of the following assays: erythrocyte hemolysis, erythrocyte malondialdehyde, breath pentane, susceptibility of low-density lipoprotein to oxidation, and diene conjugate second derivatives. For the first two assays, there are methodological limitations; for the other assays, further experimentation is needed.

A vitamin K functional test that has been recently proposed is the determination of serum underdecarboxylated osteocalcin by radioimmunoassay. This test is well correlated with static indices. Commercial kits are available, as is a semiautomated bead-based enzyme immunoassay that is less time-consuming.

A thiamin functional test that is commonly used is the erythrocyte transketolase activation coefficient (ETK-AC) test. Transketolase is a thiamin-dependent enzyme with a specific role in the glucose oxidative pathway. Transketolase activity in haemolysed erythrocytes (ETK) is measured either by the disappearance of pentose or by the appearance of hexose by spectrophotometry. In the case of thiamin deficiency, the quantity of hexose is reduced. When TPP is added to the reaction mixture, the enzyme activity is enhanced in thiamin-deficient hemolysates only. The activation coefficient is given by the ratio of enhanced (with TPP addition) to basal (without TPP addition) activity.

Because of limitations, to obtain a correct thiamin nutritional status, basal activity should be carried out together with the activation test. Automation of this test is available. Confounding effects of chronic ethanol exposure, conditions that reduce thiamin intake or absorption, uncontrolled diabetes, hyperparathyroidism, age, stress, and infections are reported. Because various methods of measurement have been proposed, in order to obtain a better interpretation and comparison of results, the standardization of the procedure and the use of quality control samples at various time points have been recommended.

A riboflavin functional test that is commonly used is the erythrocyte glutathione reductase activation coefficient (EGR-AC) test. Glutathione reductase is a flavoenzyme with FAD as a prosthetic group. By measuring the EGR activity by spectrophotometry in erythrocyte hemolysate without FAD addition (basal) and with FAD addition (stimulated), the activation coefficient (the ratio of stimulated to basal activity) can be calculated. The higher the coefficient, the lower the coenzyme content. Automation of this test is available. Confounding effects of glucose-6-phosphate dehydrogenase deficiency, severe uremia, liver cirrhosis, biliary disorders, diabetes, thyroid diseases, congenital heart disease, chronic alcoholism, pyridoxine deficiency, stress, and drugs are reported.

Vitamin B_6 functional tests are the erythrocyte aspartate aminotransferase activation coefficient test and the tryptophan load test. Erythrocyte aspartate aminotransferase activity is measured spectrophotometrically in erythrocyte hemolysate without PLP addition (basal) and with PLP addition (stimulated). The activation coefficient is given by the ratio of stimulated to basal activity. Automation of this test is available. Confounding effects of renal and liver diseases, cancer, celiac disease, high protein diet, thiamin status, alcohol intake, stress, and drugs are reported. The tryptophan load test was used in the past because vitamin B_6-dependent enzymes are involved in the conversion of tryptophan to niacin. After an appropriate loading dose of tryptophan and under controlled conditions, vitamin B_6-deficient subjects excrete tryptophan metabolites (kynurenine, kynurenic acid, and xanturenic acid) in urine measured spectrophotometrically after thin-layer or ion exchange chromatography separation. Confounding effects of protein intake, exercise, pregnancy, some hormones, and acute phase status in young people are reported. Plasma total homocysteine (tHcy) in the absence of folate and vitamin B_{12} deficiencies can be considered indicative of vitamin B_6 status. For its determination, see the following discussion of folate functional tests.

Folate functional tests are the plasma homocysteine, urinary formiminoglutamic acid (FIGLU), lymphocyte deoxyuridine (dU) suppression, and hypersegmentation of neutrophilic granulocytes assays. Folate and, to a lesser extent, vitamins B_{12} and B_6 are involved in tHcy metabolism. Plasma homocysteine concentration, in the absence of vitamin B_{12} and B_6 deficiencies, is considered a test of folate status. Because an elevated plasma tHcy concentration is associated with an increased risk of cardiovascular diseases, the determination of this amino acid in plasma has become very common. Various methods are available for tHcy determination. The most commonly used is the HPLC method with fluorescence or ultraviolet (UV) detection, which presents some problems for standardization. Capillary electrophoresis methods with laser fluorescent or UV detection have several advantages. Immunoassay methods are all automated, not time-consuming, and easy to use because of the availability of commercial kits. The chromatographic method coupled to mass spectrometry and with isotopic dilution is considered a reference method due to its high level of accuracy and precision. In a Dutch population study, it was observed that after adjustment for confounders (age, intake of other B vitamins and methionine, smoking, and alcohol consumption) folate was independently inversely associated with plasma tHcy concentration. Other confounding factors are renal failure, inborn errors affecting enzymes involved in lowering tHcy level, lack of exercise, hypothyroidism, psoriasis, and a few drugs. FIGLU acid is eliminated due to the inhibition of the conversion of histidine to glutamic acid in folate deficiency. However, the specificity of this test is low and its use limited. The other two tests are too complex, not very specific, and require further investigation.

Vitamin B_{12} functional tests are the urinary/serum methylmalonic acid (MMA), plasma/serum tHcy, and dU suppression assays. MMA increases in vitamin B_{12} deficiency; the loading with valine or isoleucine produces a marked increase in both urine and serum. MMA is measured by gas chromatography–mass spectrometry. In vitamin B_{12} deficiency in the absence of folate and vitamin B_6 deficiencies, tHcy in plasma increases and decreases with B_{12} administration. For tHcy determination, see the discussion of folate functional tests. The dU suppression test is rather complex and not specific for assessment of vitamin B_{12} status.

Vitamin C functional tests are the lingual vitamin C and intradermal 2,6-dichlorphenolindophenol

solution assays. The time to decolorize this solution is inversely correlated with plasma vitamin C levels. However, in humans both methods have low precision.

Zinc functional tests are serum or plasma alkaline phosphatase, erythrocyte metallothionein (MT), monocyte metallothionein mRNA (MTmRNA), and serum thymulin assays. Alkaline phosphatase is a zinc metalloenzyme; rather than being indicative of zinc deficiency, it is considered to be of value after zinc supplementation but with contrasting results. A commercial kits is available for plasma alkaline phosphatase determination. Alkaline phosphatase activity has low specificity and is subject to pathophysiological conditions. Erythrocyte MT decreases in moderate and severe zinc depletion and changes in response to elevated dietary zinc intake. Erythrocyte MT is measured by sandwich ELISA assay. MTmRNA is a new approach to zinc status assessment. It responds more rapidly to zinc supplements than erythrocyte MT. MTmRNA is measured in monocytes by competitive reserve transcriptase-polymerase chain reaction. An improvement of the this method is the determination of MTmRNA on blood samples spotted onto filter paper. Confounding effects are limited to infection. MT and MTmRNA assays are very promising; further studies are needed because of the difficulty in their determination. Serum thymulin activity is decreased in zinc deficiency because it requires zinc to maintain its structure. This test needs further investigation.

Copper functional tests are serum caeruloplasmin, erythrocyte superoxide dismutase (SOD), and leucocyte/platelet cytochrome *c* oxidase assays. Serum caeruloplasmin, an acute phase reactant protein, can be measured for its oxidase activity on various substrates or by radial immunodiffusion (a commercial kit is available). Serum ceruloplasmin levels are increased with exercise, stress, pregnancy, trauma, cigarette smoking, infection, and malignancy and decreased with nephrosis, advanced liver disease, malnutrition, protein-losing enteropathies, and drugs. Cu, Zn SOD is a cytosolic metalloprotein that catalyzes the reduction of superoxide to hydrogen peroxide and oxygen. It is considered to be a better indicator of reduced copper status than serum copper or caeruloplasmin. The major disadvantage of this test is the lack of a standard assay. Reference values depend on the analytical method. For determination, commercial kits are available. Confounding effects of Down's syndrome, uraemia, various anaemies, Duchenne muscular dystrophy, glutathione reductase deficiency, and porphyria are reported. Tests on cytochrome *c* oxidase seem to be more reliable than erythrocyte SOD as indicators of copper stores. They are not affected by sex and hormone use, but the enzyme is rather labile and presents large intersubject variations. For technical improvement, further studies are warranted.

Selenium functional tests are plasma, erythrocyte, and platelet glutathione peroxidase activity (GSF-px) assays. The plasma GSH-px is a useful index only in populations with low Se intake; it responds rapidly to supplementation. Erythrocyte GSH-px presents a plateau at $1.77\,\mu mol\,l^{-1}$, above which it is independent of Se status. In addition, erythrocyte GSH-px responds slowly to depletion and supplementation. Platelet GSH-px responds rapidly to Se dietary changes and presents the maximum activity at Se levels of $1.25–1.45\,\mu mol\,l^{-1}$. Accordingly, it is considered to be a sensitive indicator of changing Se status. GSH-px can be measured with coupled enzyme assay or ELISA; commercial kits are available. Confounding effects of age and sex, physical activity, essential fatty acid deficiency, vitamin B_{12}, and iron deficiencies, and stress from antioxidants are reported.

Iodine functional tests are the determination of the thyroid hormones, thyroxine (T_4) and 3,5,3′-triiodothyroxine (T_3), and pituitary thyroid-stimulating hormone (TSH) in serum by specific competitive radioimmunoassay methods (available in kits).

Choice of Laboratory Tests

The choice of laboratory tests depends on the type of study to be carried out. In field nutritional epidemiology studies, particularly in developing countries, the number of tests will be limited by the sample size, the suspected prevalence of deficiencies, the local laboratory conditions, and the availability of skilled personnel and economic resources. In general, the following common tests can be suggested: hemoglobin, hematocrit, serum iron, TIBC, serum ferritin, blood protoporphyrin, serum albumin, plasma transport proteins, and serum zinc. In specific cases, serum retinol, other vitamins in blood or urine, and some hormones (thyroxine and TSH) and minerals (urinary iodine) can be added.

In population studies to be carried out in developed countries with high-level laboratory facilities, the selection of laboratory tests depends on the purpose of the study, sample size, and financial resources. The assessment of protein status is in general limited to plasma transport proteins, unless there are other specific reasons for using other variables. Essential fatty acid status is assessed in lipid pattern studies; the choice of test is determined by

Table 1 Summary of Flair Concerted Action No. 10 recommended methods

Micronutrient	*Recommended method*	'*Best available*' *method*	*Additional methods of use in some circumstances*
Vitamin A		Serum retinol	RDR test MRDR test Isotope dilution technique RBP:TTR ratio
Carotenoids		Serum carotenoid profile	Serum lutein Serum lycopene Serum α-carotene
Vitamin E	Lipid standardized serum α-tocopherol		
Vitamin D	Serum 25-OH vitamin D		Serum 1,25-dihydroxyvitamin D Serum calcium Serum phosphate Serum alkaline phosphatase
Thiamin	ETK stimulation test		RBC TPP
Riboflavin	EGR stimulation test		
Vitamin B_6	Plasma PLP		EAST stimulation test
Vitamin B_{12}	Serum cobalamins		Serum MMA
Folate	RBC folate		Serum folate
Vitamin C	Plasma vitamin C Leucocyte ascorbate		
Selenium		Plasma selenium	RBC GSHPx
Iron	Serum ferritin	Transferrin receptors	
Copper		Serum copper (?)	RBC SOD
Zinc		Serum zinc	RBC metallothionein

RDR, relative dose–response; MRDR, modified relative dose–response; RBP, retinol binding protein; TTR, transthyretin; ETK, erythrocyte transketolase; RBC, red blood cell; TPP, thiamin pyrophosphate; EGR, erythrocyte glutathione reductase; PLP, pyridoxal-5′-phosphate; EAST, erythrocyte aspartate aminotransferase; MMA, methylmalonic acid; GSHPx, platelet glutathione peroxidase; SOD, superoxide dismutase.Adapted with permission from van den Berg H, Heseker H, Lamand M, Sandstrom B and Thurnham D (1993) Flair Concerted Action No 10 Status Papers—Introduction, Conclusions and Recommandations. *International Journal for Vitamin and Nutrition Research* **63**: 247–251, with changes suggested by D. Thurnham (personal communication).

the interest in recent intake or long-term status. In association with this test, serum cholesterol and triacylglycerols and also lipoprotein fractions are measured. The selection of micronutrient tests can be determined by the suspected deficiencies from previous dietary surveys; in the absence of dietary data, several tests should be measured because preclinical deficiencies are common in developed societies. A sensible selection can be found in **Table 1**. In the US Third Nutritional Health and Nutrition Examination Survey, most of the recommended and best available methods were used. This is also the case for vitamin status analysis in the recent UK government diet and nutrition surveys of specific population groups.

In a hospital setting, the selection of laboratory tests depends on the clinical conditions of patients on admission and during the subsequent course of injury or illness. Because protein–energy malnutrition can be present in some cases, protein status should be assessed using laboratory tests for serum albumin, plasma transport proteins, and urinary creatinine and 3-methylhistidine, and also for acute phase proteins. Using some of the previous values associated with other variables (immunological functions and anthropometric measurements), indices relating nutritional status to clinical outcome can be computed. Among hospital patients, vitamin and trace element deficiencies are also common; the determination of deficient variables suspected on the basis of history and physical examination is suggested.

Because on various occasions major differences in interlaboratory comparisons and ring tests have been observed, it is essential in the selection of laboratory tests to favor definitive reference methods or, in their absence, standardized and validated methods for which careful collection and handling of samples is compulsory and also appropriate quality control. Commercial-quality control samples or external quality assurance schemes are available only in some cases. For quality control of in-house samples, it is suggested to prepare one sample with low or deficient content and one with normal or

high content. Interlaboratory cross-comparison is highly recommended.

Evaluation of Laboratory Indices

In general, reference values are population specific; accordingly, each major laboratory in homogeneous areas has to derive them from a clinically healthy reference population selected with very specific criteria. These values should preferably be given in percentiles.

In general, cutoff points for an appropriate interpretation of results have been derived statistically from reference values. A current procedure for constructing cutoff point consists of determining the biochemical values that correspond to the earliest determinable physiological, metabolic, functional, and morphological alterations. Such an approach has been followed only in a very few cases, and consequently most available cutoff points should be considered as tentative.

For albumin, the guidelines for interpretation suggested in 1974 are still in use. For children and adults, values $<28\,g\,l^{-1}$ are indicative of a deficient (high-risk) status, and for pregnant women this value is $<30\,g\,l^{-1}$. A marginal (moderate-risk) status is indicated by the following values: infants, $<25\,g\,l^{-1}$; children 1–5 years, $<30\,g\,l^{-1}$; children 6–17 years, $<35\,g\,l^{-1}$; adults, 28–34 $g\,l^{-1}$; pregnant women at first trimester, 30–39 g/l^{-1}; and pregnant women at second and third trimester, 30–34 $g\,l^{-1}$. All values above the moderate risk are indicative of an acceptable (low-risk) status.

Reference values for transport proteins are provided by plate producers. Tentatively, $0.10\,g\,l^{-1}$ for prealbumin and $25\,mg\,l^{-1}$ for retinol binding protein are considered indicative of protein deficiency. For transferrin, values less than $1\,g\,l^{-1}$ are considered indicative of severe protein depletion; marginal status values are between 1 and $2\,g\,l^{-1}$.

For the interpretation of serum phospholipid essential fatty acid values, the ratio triene–tetraene (C 20:3 n-9/C 20:4 n-6) above 0.2 was considered by Holman to be the upper limit of normalcy. The ratio C 22:5 n-6/C 22:6 n-3 can be a sensitive index of n-3 fatty acid deficiency.

The cutoff points for the most widely used micronutrient tests in adults are reported in **Table 2**. Cutoff points are different for children, pregnant and lactating women, and the elderly. These values can be found in reference texts. For antioxidant vitamins and provitamins to prevent chronic diseases, the following optimal plasma levels have been proposed: retinol, $>2.5\,\mu mol\,l^{-1}$; β-carotene, $>0.40\,\mu mol\,l^{-1}$; α-tocopherol, $>30\,\mu mol\,l^{-1}$; and ascorbic acid, $>50\,\mu mol\,l^{-1}$.

Confounding Effects of Infection on Laboratory Assessment

As already indicated, many confounding effects of infection have been observed in many laboratory tests for nutritional status. For protein status, confounding effects of infection are reported for almost all laboratory tests, excluding that for total serum protein. In particular, serum albumin, plasma transport protein, and fibronectin levels decrease because of the increase of acute phase proteins.

For vitamin A, severe systemic infections (e.g., pneumonia, bronchitis, diarrhoea, septicaemia, rheumatic and scarlet fever, malaria, and measles) cause a marked decrease in serum retinol level. This decrease may be due to various factors (e.g., increased retinol excretion in urine and reduced liver release of retinol and RBP to plasma). A reduction of vitamin A liver reserves assessed by the RDR test has been observed in children with chickenpox.

Plasma vitamin E is reduced in malaria-infected patients. This influence is retained via the lipoproteins and not directly. Tests for thiamin status can be confounded by infections that prevent normal absorption (diarrhea and dysentery) or increase the requirement (fever).

For vitamin B_{12}, fish tapeworm or hookworm infestations give a low level of serum vitamin B_{12} because of their preferential consumption of this vitamin. For vitamin C, acute and chronic infections can depress markedly the serum ascorbic acid level due to a decrease in vitamin C reserves.

For iron status tests, infection induces an increase in serum ferritin and blood protoporphyrin levels and a decrease in serum iron binding capacity, serum iron, and hemoglobin. Zinc status tests are influenced by acute and chronic infections. A decrease in plasma zinc has been reported, due initially to redistribution of zinc within the body tissues and then to a negative body balance. This is due to anorexia, which reduces dietary intake, and also to increased losses via the faeces (diarrhea), sweat, and urine.

Regarding copper status tests, infection results in an increase in serum copper level because the leucocytic endogenous mediator induces an increase in serum ceruloplasmin. Jodine status can be influenced by infection because the synthesis of TTR is markedly suppressed.

In nutrition surveys, to correct misclassification of laboratory values due to positive acute phase proteins, the concurrent serum determination of these proteins has been suggested.

Table 2 Tentative cutoff points for interpretation of results of micronutrient tests in adults

	Severe deficiency	*Marginal deficiency*	*Physiological level or range*
Lv retinol (μmol g^{-1})	<0.07		
P retinol (μmol l^{-1})	<0.35	0.35–1.05	>1.05
RBP:TTR ratio	≤0.37		
Relative dose response (%)		>20	<20
S 25-OHD (nmol l^{-1})	<12.5	12.5–25.0	>25
P/S α-tocopherol (μmol l^{-1})	<11.6	11.6–16.2	>16.2
P α-tocopherol (μmol l^{-1})	<9.25	9.25–13.9	>13.9
E TPP (nmol l^{-1})	<120	120–150	>150
U thiamin (μg/24 h)	<27	27–65	>66
ETK-AC[a]	>1.25	1.15–1.25	1.00–1.15
E FAD (nmol l^{-1})	<200		
U riboflavin (μg/g creatinine)	<27	27–79	>80
*E*GR-AC[a]	>1.4	1.2–1.4	<1.2
*E*GR-AC[b]	>1.30	1.20–1.30	<1.20
P PLP (nmol l^{-1})	<20		20–86
U 4-PA (nmol/nmol creatinine)			128–680
*E*AST-AC[a,b]	>1.80	1.70–1.80	<1.70
P vitamin B_{12} (pmol l^{-1})[b]	<258		>260
S TCII (pmol l^{-1})	<15		
S methylmalonic acid (μmol l^{-1})		0.5–1.0	<0.5
P homocysteine (μmol l^{-1})	>100	100–12	<12
P folate (nmol l^{-1})	>7.0		
P folate (nmol l^{-1})[b]	<5.7	5.7–11.4	>11.4
RBC folate (nmol l^{-1})	<317	317–354	>354
L lobe average	>3.6	3.6–3.2	<3.2
P biotin (nmol l^{-1})[b]	<0.5	0.5–1.0	>1.0
P ascorbic acid (μmol l^{-1})	<11.4	11.4–17	>17
B ascorbic acid (μmol l^{-1})	<17	17–27	>28
L ascorbic acid (nmol/10^8 cells)		53–95	114–301
S/P ferritin (μg)	<12	20	100
S iron (μmol l^{-1})	<10.7	20	
S TIBC (μmol l^{-1})	<71.6		
Transferrin saturation (%)	<15%		
E PP (μmol l^{-1})	<1.24		
Haemoglobin (g l^{-1})	M <130		
	F <120		
Haematocrit (%)	M <40		
	F <36		
MCV	<80		
P Zn (μmol l^{-1})	<10.7		
S caeruloplasmin (μmol l^{-1})			2–4
P Se (μmol l^{-1})	<0.38	0.38–0.76	0.76–1.52
E Se (μmol l^{-1})	~0.45		1.13–2.41

[a]The percentage stimulation is now very seldom used. It can be calculated as follows: (AC × 100) − 100.
[b]From Benton D, Haller J and Fordy J (1997) The vitamin status of young british adults. *International Journal for Vitamin and Nutrition Research* **67**: 34–40, with permission.
Lv, liver; *P*, plasma; *S*, serum; *E*, erythrocyte; *U*, urine; *L*, leucocytes; *B*, whole blood; RBP, retinol binding protein; TTR, transthyretin; TPP, thiamin pyrophosphate; ETK-AC, erythrocyte transketolase activation coefficient; FAD, flavinadeninedinucleotide; EGR-AC, erythrocyte glutathione reductase activation coefficient; PLP, pyridoxal-5′-phosphate; 4-PA, 4-pyroxic acid; EAST-AC, erythrocyte aspartate aminotransferase activation coefficient; TCII, transcobalamin II, RBC, red blood cell; TIBC, total iron-binding capacity; PP, protoporphyrin; MCV, mean cell volume.
Adapted with permission from Flair Concerted Action No. 10 Status Papers (1993) *International Journal for Vitamin and Nutrition Research* **63**: 252–316, with changes suggested by C.J. Bates (personal communication).

See also: **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Carotenoids**: Chemistry, Sources and Physiology. **Cobalamins**. **Copper**. **Fatty Acids**: Omega-3 Polyunsaturated; Omega-6 Polyunsaturated. **Folic Acid**. **Iron**. **Magnesium**. **Niacin**. **Nutritional Assessment**: Anthropometry; Clinical Examination. **Potassium**. **Riboflavin**. **Selenium**. **Sodium**: Physiology. **Thiamin**: Physiology. **Vitamin A**: Physiology. **Vitamin B_6**. **Vitamin E**: Metabolism and Requirements. **Vitamin K**. **Zinc**: Physiology.

Further Reading

Bates CJ (1997) Vitamin analysis. *Annals of Clinical Biochemistry* **34**: 599–626.

Bates CJ (1999) Diagnosis and detection of vitamin deficiencies. *British Medical Bulletin* **55**: 643–655.

Brody T (1999) *Nutritional Biochemistry*, 2nd edn. San Diego: Academic Press.

De Leenheer AP, Lambert WE, and Van Boexlaer JF (2000) In *Modern chromatographic analysis of vitamins*. New York: Decker.

Fidanza F (1991) *Nutritional Status Assessment—A Manual for Population Studies*. London: Chapman & Hall.

Gibson RS (1990) *Principles of Nutritional Assessment*. New York: Oxford University Press.

Gregory J, Foster K, Tyler H, and Wiseman M (1990) *The Dietary and Nutritional Survey of British Adults (Office of Population Censuses and Surveys, Social Survey Division)*. London: HMSO.

Gunter EW, Lewis BG, and Koncikowski SM (1996) *Laboratory Procedures Used for the Third National Health and Nutrition Examination Survey*. Atlanta: Centers for Disease Control and Prevention.

Iyengar GV (1989) *Elemental Analysis of Biological System*. Boca Raton, FL: CRC Press.

Jelliffe DB and Jelliffe EFP (1989) *Community Nutritional Assessment—With Special Reference to Less Technically Developed Countries*. Oxford: Oxford University Press.

McLaren DS and Frigg M (2001) *Sight and Life Manual on Vitamin A Deficiency Disorder (VADD)*, 2nd edn. Basel: Task Force SIGHT AND LIFE.

Report of the International Nutritional Anemia Consultative Group (1985) *Measurements of Iron Status*. Washington, DC: Nutrition Foundation.

Sauberlich HE (1999) *Laboratory Tests for the Assessment of Nutritional Status*, 2nd edn. Boca Raton, FL: CRC Press.

van den Berg H, Heseker H, Lamand M, Sandstrom B, and Thurnham D (1993) Flair Concerted Action No. 10 Status Papers. *International Journal for Vitamin and Nutrition Research* **63**: 247–316.

Wright R and Heymsfield S (1984) In *Nutritional Assessment*. Boston: Blackwell Scientific.

Clinical Examination

B Caballero, Johns Hopkins Bloomberg School of Public Health and Johns Hopkins University, Baltimore, MD, USA

The clinical evaluation of nutritional status is a fundamental component of health assessment at any age. Along with anthropometry, dietary assessment, and laboratory tests, the physical examination is one of the key tools to evaluate nutritional status.

The two most common settings for a clinical examination are the hospital (inpatient or outpatient) and the field health care unit. In the first situation, the physician or examiner may have access to resources that are usually not available in the field. Because of this and other constraints, the assessment of nutritional status in the field is frequently more narrowly aimed at identifying a specific clinical condition or set of signs and symptoms. In either case, it is essential that information on history and physical findings be collected in a standardized manner in terms of both format and procedures. The former is usually best achieved by the use of preprinted or computerized forms. Electronic forms can be programmed to perform immediate range checking as values are entered, thus alerting the operator when values out of range are entered. Procedures for examination must be clearly defined in writing, and any health worker should be able to follow the instructions and perform an acceptable measurement. Although many components of the examination are subjective, it is important to standardize as much as possible terms such as 'minor,' 'average,' and 'large' within the group of examining persons, attributing a numeric value whenever possible. If data entry requires selecting from a numeric scale, they should be also standardized by cross-validation with experienced personnel or by means of photographs or models.

The two components of the clinical assessment are the medical history and the physical examination (**Table 1**).

Table 1 Major components of a nutrition-oriented medical history

Medical history
History of weight loss or gain
Gastrointestinal symptoms (nausea, diarrhea, flatulence, pain, etc.)
History of changes in color or texture of skin, hair, conjunctiva, buccal mucosa
Use of medications
Physical activity level (work-related, leisure)
History of fatigue, shortness of breath, muscle cramps
Other lifestyle practices
Places of residence, travel (exposure to toxins, sunlight, food contaminants)
In children and adolescents
– Growth history
– Neurodevelopmental history
– General school performance
– Parental and siblings' body size (body mass index)
– Pubertal stage
– Food preferences, fads
Dietary history
Habitual dietary intake and preferences
Past diet history
Alcohol consumption
Food allergies and intolerances
Assessment of dietary intake
– 24-hr recall
– Food frequency questionnaire

Medical History

The medical history for nutritional assessment is no different from a general medical history, in which familial and past and present environmental factors and their possible association with specific diseases or disease risk are considered. For the purpose of nutritional assessment, this information will be used to determine if any nutritional finding or complaint may be caused by an underlying medical condition, particularly one that remains unrecognized at the time of the examination. Additionally, specific medical conditions and their current status are important factors altering nutrient requirements and dietary prescriptions.

One specific focus of medical history in a nutritional assessment context is the exploration of gastrointestinal function. Conditions such as chronic diarrhea, gastroesophageal reflux, and colonic disorders may be associated with reduced nutrient absorption or food avoidance that result in impaired nutritional status. Past history of gastrointestinal problems and/or surgery may also point to current alterations in nutrient digestion or absorption. Other important components of the medical history are history of weight loss or gain, past and present use of medications, use of special foods or formulas, changes in taste or smell, and food allergies and intolerances.

In children and adolescents, the medical history must also obtain information on neurodevelopmental stages, history of behavioral problems, and overall school performance. Food preferences must be noted, particularly in adolescence, when adoption of unconventional dietary practices is more likely to occur.

Physical Examination

As noted previously, anthropometric measurements are a key component of the physical examination. Measurement of weight and height is perhaps one of the most frequently performed nutritional measurements. Although its value is limited with regard to identifying specific nutrient deficiencies, it is invaluable to evaluate growth and adequacy of past and present diet in infants, children, and adolescents and to identify undernutrition and obesity in adults. Measurements should be done by trained personnel and following standard protocols. In addition to anthropometry, the physical examination focuses on signs of nutrient deficiency or excess. These signs usually appear only when the deficiency is advanced and are not to be expected in marginal deficiencies. Furthermore, the time that it takes for a deficient intake of a given nutrient to cause clinical manifestation of deficiency varies considerably, depending on whether the nutrient is stored in the body and on the initial status of the reserves. Typical signs for selected nutritional deficiencies are presented in **Table 2**. Virtually none of these signs, with the exception of Bitot's spots, are pathognomonic for one specific deficiency. However, they are useful in indicating a specific nutrient impairment and prompting further evaluation.

The physical examination should start with a general visual assessment of the patient. In children, state of alertness, willingness to engage in play, or

Table 2 Typical clinical signs of selected nutritional deficiencies

Deficiency	*Signs*
Protein–energy malnutrition	Hair: depigmentation, thinning, pluckability
	Edema in lower extremities (generalized in severe cases)
	Muscle wasting
	Decreased subcutaneous fat
	Skin: diffuse depigmentation, flaky dermatosis
	Liver enlargement
Vitamin A	Bitot's spot
	Conjunctival xerosis
	Corneal xerosis
	Keratomalacia
	Night blindness
Riboflavin	Angular stomatitis
	Cheilosis
	Scrotal (vulvar) dermatosis
	Red tongue
	Corneal vascularization
Thiamin	Edema
	Hyporeflexia
	Muscle tenderness
	Cardiac enlargement
	Tachycardia
Niacin	Pellagroid dermatosis
	Scarlet, raw, fissured tongue
	Malar and supraorbital pigmentation
Vitamin C	Bleeding, spongy gums
	Petechiae
	Ecchymoses
	Epiphyseal enlargement
	Atrophy of lingual papillae
	Follicular hyperkeratosis
Vitamin D	Active rickets: rib beading, epiphyseal enlargement, persistently open fontanelle, craniotabes, hypotonia
	Residual rickets: frontal or parietal bossing, bowlegs, knock-knees, thorax deformities
Iron	Pale conjunctiva
	Atrophy of lingual papillae
	Koilonychia
Folic acid, B_{12}	Usually associated with pallor of anemia
	Peripheral neuropathy (B_{12})
Iodine	Thyroid enlargement

resisting examination are important clues to energy level and physical strength. A generalized loss of fat depots, or excess adiposity as in the obese, is readily identifiable in most circumstances. A general overview can also identify pallor, loss of muscle mass, and skin changes.

Numerous signs of nutritional deficiencies can be identified in the skin and hair. Because skin exhibits a relatively rapid turnover, impairments in protein synthesis can result in fragile, flaky, and discolored skin. Vitamin A deficiency typically causes a dry, hyperkeratotic skin. The dermatitis of pellagra consists of patchy areas of hypo- or hyperpigmentations, usually in sun-exposed body regions, eventually progressing to hardened, broken surfaces. In protein–energy malnutrition, hair may become brittle, thin, and easily pluckable. Fluctuations in the rate of synthesis of hair protein may result in band discoloration, where pale and normal colors alternate, resulting in the 'banner sign,' typical of kwashiorkor. Petechiae or hematomas may result from protein–energy malnutrition or vitamin K or vitamin E (in the newborn) deficiencies.

One of the most specific signs of nutritional deficiency can be identified in the eye. Vitamin A deficiency produces a series of alterations in the conjunctiva and the cornea that not only indicate a deficiency of this nutrient but also help grade its severity. The most commonly used classification of vitamin A deficiency is primarily based on eye findings, from Bitot's spots to perforated keratomalacia. Conjunctival pallor has been a classic sign of anemia, but its sensitivity varies substantially depending on ethnicity, ambient lighting, and experience of the observer.

The mouth and tongue are also areas where typical manifestations of deficiency can be detected. A red tongue is a classic sign of riboflavin deficiency but has also been associated with niacin deficiency; the latter may also include fissures. Conversely, a pale tongue may indicate iron deficiency. Glossitis, with or without color changes, has been linked to pyridoxine deficiency. A similar condition, including pain and intense red color, has been associated with biotin deficiency. Angular stomatitis and ulcerations and other lip lesions are associated with riboflavin or ascorbic acid deficiencies. In the latter, extensive involvement of the gums (swelling and bleeding) is also typical. Atrophy of the papillae occurs in vitamin B_{12}, niacin, and folate deficiencies. Excess vitamin A intake may result in discoloration of the gingival mucosa.

Rib beading (also known as rickets rosary) is a typical sign of vitamin D deficiency in children, but a similar manifestation may appear in vitamin C deficiency (scurvy). Ephyphiseal enlargement and bowlegs are other classic signs of rickets. A distended abdomen is characteristic of protein–energy malnutrition in children. In the lower limbs, inspection must ascertain the presence of edema, which is also associated with protein–energy malnutrition.

Peripheral neuropathies such as those associated with beriberi or vitamin B_{12} deficiencies may result in visible impairment of limb movements, such as the 'foot drop' of dry beriberi.

In preadolescents and adolescents, assessment of sexual maturation (usually following the Tanner staging) is an important component of the physical examination, although it is not always feasible due to cultural and practical reasons. Alternatively, more limited information may be obtained in girls by self-reported menarcheal status. Self-assessment of Tanner stage by comparison with photographs is another useful alternative, but use of these photographs with children may not be acceptable in some communities.

In order to obtain a unified rating of a person's nutritional status, it is desirable to integrate clinical, laboratory, and functional data into a single scoring system. Several approaches to achieve this have been proposed, and their use will depend primarily on the target population and the intended use of the score. The Subjective Global Assessment is an approach that relies primarily on data from the physical examination and thus can be readily performed after this examination has been completed. Other scoring systems, such as the Prognostic Nutritional Index or the Instant Nutritional Index, rely to variable degrees on combinations of clinical and laboratory data.

See also: **Dietary Intake Measurement**: Methodology; Validation. **Energy Expenditure**: Indirect Calorimetry; Doubly Labeled Water. **Nutritional Assessment**: Anthropometry; Biochemical Indices.

Further Reading

McLaren DS (1992) *A Colour Atlas and Text of Diet-Related Disorders*. London: Wolfe.

Morrison G and Hark L (1996) *Medical Nutrition* Cambridge, MA: Blackwell Science.

Newton JM and Halsted CH (1998) Clinical and functional assessment of adults. In: *Modern Nutrition in Health and Disease*, 9th edn., pp. 895–902. Philadelphia: Lippincott.

Sardesai VM (2003) *Introduction to Clinical Nutrition*, 2nd edn. New York: Marcel Dekker.

Stallings VA and Fung EB (1998) Clinical nutrition assessment of infants and children. In: *Modern Nutrition in Health and Disease*, 9th edn., pp. 885–893. Philadelphia: Lippincott.

NUTRITIONAL SUPPORT

Contents

In the Home Setting

M Elia and R J Stratton, University of Southampton, Southampton, UK

The prevalence of nutritional problems in developed societies is a cause of growing concern. At one end of the nutritional spectrum, the obesity 'epidemic' is spreading at an alarming rate. At the other end of the spectrum, protein–energy malnutrition and nutrient deficiencies are also common, especially in the elderly and in those with disease. **Table 1** shows the frequency of specific vitamin deficiencies and underweight (body mass index <20 kg/m^2) in people aged 65 years or older resident in the United Kingdom. Complimentary information on protein–energy status can be obtained by considering simple criteria, such as those used by the 'Malnutrition Universal Screening Tool' (MUST) (**Figure 1**). This tool, which depends on weight loss and body mass index (and an acute disease effect, which does not normally apply to community patients), has been used to estimate that 10–15% of older people in the United Kingdom are at medium to high risk of malnutrition. The prevalence of malnutrition increases with age, and it is more common in the presence of disease and in institutions, where about one in five people are at risk. With the growing number of older people, especially those living in nursing homes and alternative care facilities, the overall prevalence of malnutrition may increase. It is disturbing that malnutrition is underrecognized and undertreated, despite its adverse effects on the individual and society.

The first important step in the management of malnutrition is identifying it using one of a number of validated nutritional screening tools. MUST was developed specifically for all types of patients in all health care settings. The potentially broad application of the same tool encourages consistency of thought and continuity of care through different health care settings. The care plan linked to this tool varies from dietary restriction in the case of obesity to supplementation and other forms of nutritional support in the case of malnutrition. For special situations, enteral tube feeding (e.g., in some patients with swallowing problems) and parenteral (intravenous) nutrition are required.

This article focuses on the treatment of malnutrition (rather than obesity) in the home setting. This treatment includes dietary counselling and fortification, oral nutritional supplementation (mixed macro- and micronutrient supplements), and artificial nutritional support (enteral tube feeding (ETF) and parenteral

Table 1 Proportion of subjects 65 years or older with selected vitamin deficiencies and body mass index <20 kg/m^2

	Free living (%)	*Institutions (%)*[a]	*Criteria*	
Vitamin deficiencies				
Folate deficiency	29	35	Red blood cell concentration	<345 μmol/l
– Severe deficiency	8	16		<230 μmol/l
Thiamine deficiency	9	14	Erythrocyte transketolase activation coefficient (ratio)	>1.25
Vitamin B_{12} deficiency	6	9	Plasma concentration	<118 pmol/l
Vitamin D deficiency	1–2	1–5		<12 μmol/l
Vitamin C deficiency	14	40	Plasma concentration	<11 μmol/l
– Severe deficiency	5	16		<5 μmol/l
Underweight	3	16	Body mass index	<20 kg/m^2

[a]Registered residential homes (57%), nursing homes (30%), dual-registration homes (9%), and other facilities (4%)
Based on the National Dietary and Nutrition Survey (1998) in the United Kingdom.

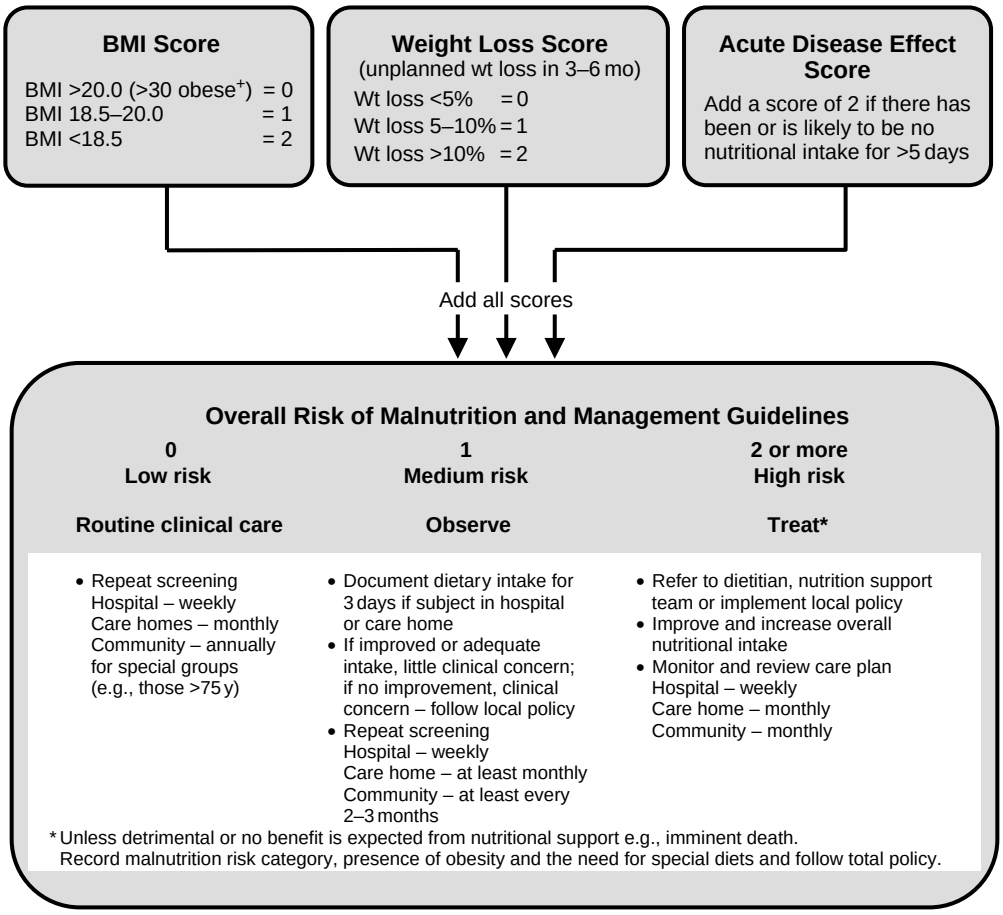


Figure 1 'Malnutrition Universal Screening Tool' (MUST). A copy of MUST and further details on taking alternative measurements, special circumstances, and subjective criteria can be downloaded at www.bapen.org.uk.

nutrition (PN)). The simplest and most commonly used treatment involves oral nutritional support, which is considered before home enteral tube feeding (HETF) and home parenteral nutrition (HPN).

Oral Nutritional Support

Dietary Counselling and Fortification

Dietary counselling, usually provided by a dietitian, is an integral part of oral nutritional support. It includes advice on dietary fortification, which is often the first-line treatment of malnutrition in the home and other care settings. Counselling may involve advice on eating patterns (e.g., eating certain types of snacks at particular times of day) or addition of energy- and protein-rich food ingredients (e.g., cream, milk, oil, butter, sugar, and skimmed milk powder) to meals. Commercial energy- and protein-containing supplements can also be used to improve intake without substantially altering the taste of food and drink. The use of nutritionally fortified food snacks as part of the diet may improve both the intake and the status of micronutrients. However, the success of these dietary strategies is limited in patients with severe anorexia, those living in poverty and due to other social factors, and in those with inadequate motivation. Thus, patients may find it difficult to purchase, manipulate, or prepare their meals. Financial or other forms of social support, such as help with shopping, cooking (or provision of 'meals on wheels'), and help with eating, may do much to improve intake in some individuals. Although dietary counselling, with or without dietary fortification, is widely used in clinical practice, there is little research supporting its clinical efficacy in patients at risk of malnutrition in developed countries.

Oral Nutritional Supplements

Mixed macro- and micronutrient liquid sip feeds and other oral nutritional supplements (bars, powders, and puddings) are widely used in the treatment

Table 2 Summary of significant functional and clinical outcome improvements following oral nutritional supplementation in community patients from randomised controlled trials

Patient group	*Functional/clinical outcome*
Chronic obstructive pulmonary disease	Respiratory muscle function Hand grip strength Walking distances
Elderly	Reduced number of falls Increased activities of daily living Muscle power
HIV/AIDS	Cognitive function
Liver disease	Lower incidence of severe infections Lower frequency of hospitalisation
Malignancy	Immunological benefits
Osteoarthritis	Increased activities of daily living[a] Improved osteoarthritis index[a]

[a]Nutritional supplement also containing immunoglobulin G (90 mg).
Based on Stratton RJ, Green CJ, and Elia M (2003) *Disease-Related Malnutrition: An Evidence Based Approach to Treatment*. Oxford: CABI Publishing.

of malnutrition in the community setting. A systematic review of 78 randomized controlled trials (RCTs) (including 44 RCTs from the community setting) suggests oral nutritional supplements can improve energy and nutrient intakes, improve body weight (or attenuate weight loss), and improve a number of functional and clinical outcomes in various patient groups (**Table 2**). Meta-analysis of RCTs from both hospital and community settings suggests significantly lower mortality (odds ratio, 0.62; 95% confidence interval, 0.49–0.78) and complication rates (infections and postoperative complications) (odds ratio, 0.29; 95% confidence interval, 0.18–0.47) in patients given oral nutritional supplements (typically 1.05–2.5 MJ (250–600 kcal) daily).

For some patients, nutrition via the oral route is either unable to meet the nutritional requirements (e.g., patients with a poor appetite) or contraindicated (e.g., a cerebrovascular accident patient with aspiration and intestinal failure). For such patients, HETF and HPN may be required, although the treatment is usually initiated in hospital.

Artificial Nutrition Support: Home Parenteral Nutrition and Home Enteral Tube Feeding

Patients suffering from chronic conditions often prefer to be treated in the familiar surroundings of their home rather than in hospital. When the treatment involves sophisticated techniques, it is essential that either the patient or the caregiver is adequately trained to distinguish between problems that can be easily remedied at home and those that need expert advice and treatment in hospital. With the increasing pressure for hospital beds and the increasing cost of hospital care, many forms of treatment that were previously restricted to the hospital environment have extended to the community, including renal dialysis, cytotoxic drug therapy, HETF, and HPN. HETF has grown rapidly so that its prevalence in several developed countries is now several times greater than in hospital. In contrast, PN is still practiced less commonly outside hospital than in hospital and is likely to remain so in the foreseeable future. Both forms of treatment have led to the development of professional teams specialising in nutritional support in both the hospital and the community. These teams deal with problems ranging from simple day-to-day management issues to difficult ethical problems, such as concerning withholding or withdrawing nutritional support.

Origins and Development

The first report of HPN appeared in 1970 in North America, and in Europe the first reports appeared in the late 1970s. The number of people receiving HPN has increased considerably since then but remains substantially lower than for HETF (**Figure 2**).

HETF is a much older technique than HPN, with the first reports appearing centuries ago. Accurate information on the numbers of people receiving HETF is difficult to obtain because HETF tends to be initiated from many centres and centralized reporting and record keeping in most countries are not fully established. There has been rapid growth in HETF attributable to developments in tube technology (flexible fine bore tubes) and endoscopic procedures for placement of gastrostomy tubes (facilitating easier initiation and management of long-term feeding), as well as the development of home care services provided by commercial enteral

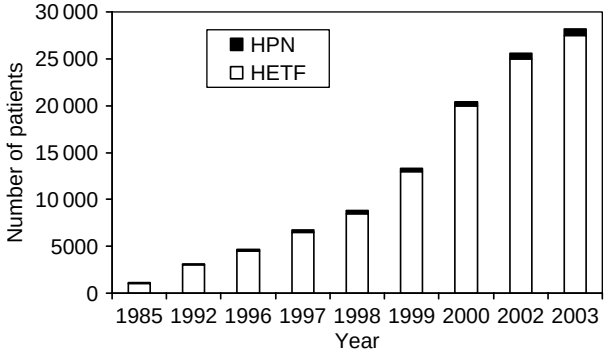


Figure 2 Estimated growth in point prevalence (amount of feeding taking place at a given point in time) in home enteral tube feeding (HETF) and home parenteral nutrition (HPN) in the United Kingdom.

feeding companies. In many developed countries there is considerably more ETF taking place in the community than in hospital. In Britain, there continues to be steady growth (10–20% per year) in the numbers of people receiving HETF, and in 2003, 21 527 people were registered with the British Artificial Nutrition Survey (BANS), with an estimated total number receiving HETF in excess of 25 000. As with HPN, HETF is less common in Europe than in North America and is practised much less in Eastern Europe, India, and China than in industrialized Western countries.

In addition to the differences in prevalence of HETF and HPN between countries, there may also be marked variations within countries. Even within one region of the United Kingdom (south and west regions) the number of individuals receiving HETF in 2002 within different primary care trusts varied from 82 to 632 per 1 million people. Similarly, considerable variation in the point prevalence of HPN was found to exist in different regions of the United Kingdom in 1999 (0 to 36 per 1 million). This large variation, which is unlikely to be due to chance, can be explained by variations in the availability of expertise and support staff, resources to fund such treatment, or local differences in attitudes/policies toward the use of artificial nutrition.

The wide variations in the prevalence of home artificial feeding throughout the world are related to health care economies. There is a relationship between expenditure on health care, as a percentage of gross domestic product (GDP), and the incidence of HPN and HETF. In India, Pakistan, and Africa, where spending on health is low, home artificial nutrition is less common. In Western Europe, where health care accounts for a greater proportion of GDP, home artificial nutritional support is more common. In the United States, with an even greater expenditure on health care, the prevalence of HPN and HETF is higher than anywhere else in the world.

Indications

Home enteral tube feeding The indications for HETF are different for adults and children. In adults, the most common indications are neurological disorders of swallowing resulting from cerebrovascular accidents, Parkinson's disease, and obstructive lesions of the upper gastrointestinal tract. These mainly affect older individuals so that in various countries approximately half of HETF is administered to individuals aged 65 years or older. In children, HETF is usually used in conditions that lead to failure to thrive, such as cerebral palsy, cystic fibrosis, congenital malformation, and metabolic disorders.

As with HPN, one of the main differences between countries in the indications for HETF concerns malignant disease. In North America, ~40% of people receiving HETF have been reported to have malignant disease, and up to ~70% in Italy. In the United Kingdom, the number of patients receiving HETF because of malignancy has steadily increased in both adults and children so that by 2000, 1 in 4 adults and 1 in 12 children who started HETF had cancer, usually of the upper gastrointestinal tract (mainly oesophageal, head and neck, and oropharyngeal). The age distribution of people receiving HETF is influenced by the indications. Because disorders of swallowing (strokes, motor neurone disease, and other neurological conditions) and cancer of the upper gastrointestinal tract tend to occur in older age groups, adults receiving HETF tend to be elderly (with more than 60% of those in the UK being older than 60 years and 46% older than 70 years). In recent years, there has been a trend to provide HETF to an older and more disabled population. Recent surveys in the UK suggest that approximately 50% of patients are house or bed bound and require total help to manage their tube feeding. Because the majority of these patients with high levels of disability are at home (spending <1% of their time in hospital), there are resource implications associated with the provision of health care by the underrecognized and underappreciated voluntary caregivers. Approximately 20% of those receiving HETF are children, and many children who started HETF because of cerebral palsy or congenital handicap continue tube feeding into adulthood.

Home parenteral nutrition The main indications for HPN are Crohn's disease, ischaemic bowel disease, motility disorders, or bowel and malignant disease. Patients receiving HPN are usually younger than those who receive HETF, although there is an overlap. There are also differences between the practice of HPN in different countries. One of the main differences concerns malignant disease. In the United States, 40–50% of patients receiving HPN have been reported to have cancer, and similar if not higher percentages have been reported in some European countries, such as Italy. Early reports from the United Kingdom and Denmark suggested that only a small proportion of HPN (~5%) involved patients with cancer, although this has increased with time. For example, in the United Kingdom it has steadily increased so that by 2003, one in seven patients starting HPN had cancer.

Table 3 Standards of practice for home enteral tube feeding (HETF)

Structure	*Process*	*Outcome*
There will be a training program for the health care professionals involved in the care of patients receiving HETF.	Discharge planning will be performed only by professionals who have the necessary experience or who have undertaken a course of training in the topic.	The patient has confidence in the hospital team planning his/her discharge.
There will be a model of care for patients needing HETF.	The members of the multidisciplinary team will be involved in writing the 'mission statement' on which the model is based.	The patient will know the benefits, aims and objectives of the HETF team.
There will be a relaxed, quiet area suitable for private discussion.	There will be a caring and compassionate atmosphere with adequate time for discussion.	The patient will feel able to express his/her fears and expectations.
The discharge planning documentation will include sections on domestic, family and social circumstances.	The nutrition team will evaluate, with the patient and family, how HETF will alter his/her way of life.	The patient will believe that the feeding system can be integrated into an acceptable way of life.
There will be written patient/carer learning goals for HETF.	A designated nurse or dietitian will be responsible for teaching the patient according to his/her individual capacity for learning.	The patient will be able to demonstrate the necessary skills and achieve all the learning goals.
There will be an instruction manual for HETF.	Information and procedures will be regularly updated in order to reflect developments and innovations in tube feeding, access, nutrients and delivery systems.	The patient will perform therapy based on current practice.
A relative, friend or appropriately trained health care professional will be available to deliver therapy if the patient is unable to do so.	The nurse/dietitian will help the patient identify the most appropriate carer. A community nurse will be given the opportunity to visit the patient in hospital and observe therapy before the patient is discharged.	The patient has confidence that safe care will be available at home.
Access to the gastrointestinal tract will be achieved by a tube suitable for long-term use.	The patient, nurse and doctor will choose the most appropriate tube and access site.	The patient will use a feeding tube which is acceptable and accessible.
There will be a policy for sharing care with the patient's General Practitioner (GP).	The GP will be contacted and a shared care protocol agreed.	The patient will know the responsibility of each health care professional.
Written information describing HETF will be available for the GP.	The hospital team will provide the GP with the information before the patient is discharged, together with the discharge date and on-call telephone numbers.	The patient will have confidence in his/her GP's knowledge of HETF.
There will be written procedures for the management of feeding tubes.	The nurse/dietitian will adapt the procedures according to the patient's physical skills and domestic circumstances.	The patient's daily life will not be restricted by prolonged inappropriate procedures.
There will be a written prescription for the enteral feed (and other prescribable items).	The patient's GP will be contacted and advised on how to prescribe the feed.	The patient will have the enteral feed available at home on the day of discharge.
There will be a list of the required equipment, e.g., syringes, connectors, administration sets, pump, drip stand, telephone.	Before discharge the patient's home health authority will be provided with the list and asked to arrange supply by making local arrangements or establishing a contract with a commercial supplier.	The patient will have all the necessary supplies in his/her home on the day of discharge.
There will be an on-call system for providing expert advice to the patient by telephone day and night.	The nurse/dietitian/doctor will explain the system to the patient and identify the professions involved.	The patient will know the names and telephone numbers of health care professionals to contact in case of emergency day or night.
Information will be available describing how the nutrient solutions and supplies will be provided following discharge.	The nurse/dietitian will explain the ordering system and discuss storage, depending on the patient's home circumstances.	The patient will know how to obtain supplies and store and dispose of unwanted material.
There will be a post-discharge monitoring protocol, established by the nutrition team.	Monitoring will be performed by a designated health professional as defined by the protocol.	The patient will know what the follow-up arrangements are.

There has also been an increase in age (due at least partly to the increasing use of HPN in patients with cancer and ischemic bowel disease) so that by 2003 nearly one-fourth of all patients receiving HPN in the United Kingdom were older than 60 years of age.

Organization

The organization and management of HETF and HPN has evolved over time. For example, delivery of feeds and equipment to the first patients who received HPN or HETF was undertaken by the hospitals that initiated the treatment. As the number of patients receiving such treatment increased, commercial organizations have established an organizational infrastructure for delivering feed and ancillary equipment through a national and international network. Some companies employ doctors, nurses, and other staff so that they can provide most of the care, although this practice varies from country to country. In many countries, there is joint care between commercial companies and the national health care systems.

HETF is initiated by many centers or hospitals, and some patients are followed up as outpatients. However, it is impractical to follow up many severely disabled patients in hospital, because they are house bound. Patients receiving HPN are often managed by centres with expertise in nutritional problems (e.g., in France, Denmark, and the United Kingdom). It has been suggested that all patients on HPN should be managed at such centers, but travelling to distant centers may require considerable time, effort, and expense. It is possible for patients to be managed more locally, especially if they are uncomplicated. It remains to be demonstrated if locally managed patients have better satisfaction and similar outcomes as those managed by larger centers. Of course, it is possible to have a system that combines local care and more distant specialist care when required.

Funding arrangements also vary. In several countries, home nutritional support is either totally or partially funded by the national health service, but payment may also be provided by private insurance and individual patients. The overall pattern of funding differs considerably among countries. Sometimes, confusion exists about the funding arrangements even in the same country, and this may limit and delay the use of HETF or HPN.

Patient organizations have developed in some countries, such as Patients on Intravenous and Nasogastric Nutrition Therapy (PINNT) in the United Kingdom. This organization provides support and information to people on home feeding, and it contributes to all levels of the operation of the British Association for Parenteral and Enteral Nutrition (BAPEN), through which it influences policy and decision making. Furthermore, since the feeding equipment for use at home was found to be impractical because it was originally designed for hospital use, PINNT has redesigned the equipment specifically for home use.

Standards of Care

Several surveys have identified inadequacies in training, support, and follow-up of patients receiving HETF and HPN. Specific problems include lack of written instructions about how to manage simple problems that may arise during feeding, lack of telephone contacts for use in emergency, lack of confidence, and inadequacy of equipment for home use. Such surveys have also highlighted the importance of a multidisciplinary approach and the need to undertake home visits to assess the status of severely disabled patients who cannot easily attend a hospital. Pressure on hospital beds has meant that some patients are discharged home before they have been adequately trained, and the care of such patients is sometimes passed on to other health care workers who have little experience of home nutritional support. Since HPN is relatively uncommon in the population, general practitioners may have never encountered patients on this form of therapy and are therefore poorly equipted to manage them. Patients' needs may change during the course of their treatment; therefore, there is a need to establish an organisational infrastructure for continuity of care for HETF and HPN over time and from one health care setting to another. Many hospitals do not have a nutrition team or policies that embrace the needs of people receiving artificial nutrition at home.

A series of guidelines for the management of artificial nutrition in the community have been developed by BAPEN (**Tables 3** and **4**). The guidelines cover aspects of training prior to discharge from hospital (although training can take place at home) and the support required from trained specialist staff once the patient is at home. A national and local organizational structure for delivering the support would aid the process.

Monitoring

The basic elements of monitoring are similar for both HETF and HPN. They include an assessment

Table 4 Standards of practice for home parenteral nutrition (HPN)

Structure	*Process*	*Outcome*
There will be a training program for health care professionals involved in the care of patients receiving HPN.	Discharge planning will be performed only by professionals who have the necessary experience or who have undertaken a course of training in the topic.	The patient has confidence in the hospital team planning his/her discharge.
There will be a model of care for patients needing home intravenous nutrition.	All members of the multidisciplinary team will be involved in writing the 'mission statement' on which the model is based.	The patient will know the beliefs, aims and objectives of the HPN Care Team.
There will be a relaxed, quiet area suitable for private discussion.	There will be a caring and compassionate atmosphere with adequate time for discussion.	The patient will feel able to express his/her fears and expectations.
The discharge planning documentation will include sections on domestic, family and social circumstances.	The nutrition team will evaluate with the patient and family how the HPN will alter his/her way of life.	The patient will believe that the feeding system can be integrated into an acceptable way of life.
There will be written patient/carer learning goals for HPN.	A designated nurse will be responsible for teaching the patient according to his/her capacity for learning.	The patient/carer will be able to demonstrate the necessary skills and achieve all the individual learning goals.
There will be an instruction manual for HPN.	Information and procedures will be regularly updated in order to reflect developments and innovations in venous access, nutrient solutions and delivery systems.	The patient will perform therapy based on current practice.
A relative, friend or appropriate health care professional will be available to deliver therapy if the patient is unable to do so (e.g., parent or guardian of a child).	The health care professional will help the patient to identify the most appropriate carer. The district nurse will be given the opportunity to visit the patient in hospital and observe therapy before the patient is discharged.	The patient has confidence that safe care will be available at home.
Venous access will be achieved by a central venous catheter suitable for long-term use.	The patient, nurse and doctor will choose the most appropriate catheter and access site.	The patient will use a central venous catheter that is acceptable and accessible.
There will be written procedures for the management of central venous catheters.	The nurse will adapt the procedures according to the patient's physical skills and domestic circumstances.	The patient's daily life will not be restricted by prolonged inappropriate procedures.
There will be a policy for sharing care with the patient's general practitioner (GP).	The GP will be contacted and a shared care protocol agreed.	The patient will know the responsibility of each health care professional.
Written information describing HPN will be available for the GP.	The hospital teams will provide the GP with the information before the patient is discharged, together with the discharge date, and on-call telephone numbers.	The patient will have confidence in his/her GP's knowledge of HPN.
There will be a written prescription for the nutrition solutions (and other prescribable items).	The patient's GP will be contacted and advised on how to prescribe the feed.	The patient will have the feeding solution available at home on the day of discharge.
There will be a list of the required equipment, e.g., refrigerator, infusion pump, syringes, sterile gloves, telephone.	Before discharge, the patient's home health authority will be provided with the list and asked to arrange supply by making local arrangements or establishing a contract with a commercial supplier.	The patient will have all the necessary supplies at home on the day of discharge.
There will be an on-call system for providing expert advice to the patient by telephone day and night.	The nurse will explain the system to the patient and identify the professions involved.	The patient/carer will know the names and telephone numbers to contact in case of emergency by day or night.
Information will be available describing how the nutrient solutions and supplies will be provided following discharge.	The nurse will explain the chosen supply system and discuss storage depending on the patient's home circumstances.	The patient will know how to obtain supplies, store them and dispose of unwanted material.
There will be a post-discharge monitoring protocol, established by the nutrition team.	Monitoring will be supervized by the nutrition team.	The patient will know the date of the first outpatient visit and what monitoring will be performed.

of the activity of the underlying disease, the nutritional and metabolic state of the patient, and complications associated with nutritional support (**Table 5**). The clinical history alerts the attending health professional to the general well-being, as well as the likelihood of specific problems, such as dehydration, electrolyte imbalance (e.g., diarrhoea), local infection (e.g., local redness and swelling near the

Table 5 Some complications associated with parenteral nutrition and enteral tube feeding

	Parenteral	*Enteral tube feeding*
Mechanical	Catheter malposition. Insertion trauma (e.g., pneumothorax, brachial plexus injury, cardiac arrhythmia) Catheter blockage, kinking or occlusion Catheter embolus Air embolus Clot embolus (from catheter tip) Lack of access site	Tube malposition (e.g., into lung) Insertion trauma: drainage to stomach and bowel: peritonitis and peristomal leakage and inflammation Tube blockage, e.g., kinking or occlusion
Feed/flow	Nutrient overload (e.g., hyperglycemia, infusional hyperlipidemia)	Diarrhea or constipation Bloated adbomen/cramps Regurgitation/aspiration of feed
Infections	Catheter-related sepsis Infected feed/administration set	Infected feed administration set Infection around gastrostomy
Metabolic	Fluid and electrolyte disturbances Hyperglycemia Deficiency syndromes, e.g., trace elements and vitamins Nutrient overload (see above) and toxicity (e.g., some trace elements)	Fluid and electrolyte disturbances Deficiency syndromes (rate with standard feeds given to typical patients) Hyper/hypoglycemia
Organ tissue dysfunction	e.g., Abnormal liver function, intestinal atrophy, metabolic bone disease	Mainly disease related, abnormal liver function Aspiration pneumonia
Psychological	Anxiety, depression, disturbance in self-image, social isolation	Anxiety, depression, disturbance in self-image, social isolation
Financial	Economic issues vary from centre to centre and country to country	Economic issues vary from center to center and country to country

catheter exit site or peristomal area), blocked tubes and catheters, and so on. Catheter-related sepsis is an important complication of PN, and aspiration pneumonia is an important complication of ETF. The patient/caregiver should have written instructions about basic procedures, which aim to reduce complication rates, and how to deal with simple problems and to recognize those that they cannot readily deal with. Specialist advice should be available 24 h a day. The frequency of complications depends at least partly on the support provided by health professionals.

Dietary intake should be monitored, especially in patients whose clinical status is changing. Appropriate dietary advice may facilitate return to normal oral feeding in some patients. In those with a swallowing difficulty, it may be necessary to assess whether swallowing has improved, with input from speech and language therapists, so that unnecessary HETF is not continued when full oral feeding becomes possible. Studies in the United Kingdom suggest that 15% of patients receiving HETF can revert to full oral feeding after 1 year. Blood tests should be carried out at intervals to check for metabolic stability and specific nutrient deficiencies (e.g., vitamins, minerals, and trace elements) and toxicities. The frequency with which tests are carried out depends on the patient (e.g., whether the patient is receiving HETF or HPN), the duration of feeding, the extent of oral intake, and disease activity.

Outcome

The most important predictor of outcome in patients receiving home artificial nutritional support (enteral or parenteral) is the underlying disease. Therefore, mortality statistics strongly depend on the initial indications. Nevertheless, a few conclusions can be made. First, the complications associated with artificial nutritional support vary but are reported to be responsible for less than 3–5% of deaths. Second, the outcome is dependent not only on the type of disease but also on the stage of the disease (e.g., patients with advanced HIV who start HPN are only expected to survive a few months, whereas patients with less advanced disease are expected to survive longer). Third, the outcome of patients receiving HPN and HETF for a variety of conditions is available from the British Artificial Nutrition Survey (**Table 6**). For patients on HPN, overall mortality at 1 year is 11%, with 16% returning to oral feeding and the majority continuing with HPN. Patients with Crohn's disease often have a good prognosis (with 4% mortality and 38% returning to oral feeding within 1 year). For patients on HETF, typically an older patient group, mortality

Table 6 Twelve-month outcomes for patients receiving home parenteral nutrition (HPN) and home enteral tube feeding (HETF)

	Continuing		*Discontinuing*		
	Continues (%)	*In hospital (%)*	*Transferred to oral (%)*	*Withdrawn/refused (%)*	*Died (%)*
HETF					
All adults ($n=26\,501$)	45.7	0.6	16.2	1.1	36.3
– CVA ($n=9326$)	49.2	0.5	11.5	0.8	38.0
– Oesophageal cancer ($n=2050$)	26	0.7	22.6	2.0	48.8
All children ($n=5419$)	72.9	0.6	18.7	0.9	6.8
– Cerebral palsy ($n=903$)	87.5	0.3	5.2	0.9	6.3
– Congenital handicap ($n=561$)	86.8	1.2	6.4	0.4	5.2
HPN					
All adults ($n=765$)	71.2	0.4	15.7	1.7	11
All children ($n=68$)	77.9	1.5	10.3	0	10.3

Based on British Artificial Nutrition Survey (2004).

is higher overall (36% at 1 year) and outcome varies according to age and condition. The outcome data for two common conditions in adults and children receiving HETF are shown in **Table 6**.

Assessments of quality of life, using EuroQol, suggest that the majority of patients receiving HETF and HPN have some problems (moderate or extreme) with mobility, self care, usual activities, pain/discomfort, and anxiety/depression (five EuroQol dimensions). Mean quality-of-life scores (0, 'worst imaginable health state'; 100, 'best imaginable health state') in adults receiving HPN (53 ± 18) are higher than those for adults receiving HETF (42 ± 27), but both are considerably lower than the scores obtained from the general population, even when adjusting for age. For HETF patients, quality-of-life scores have been found to be similar for those living at home and those in nursing care.

Intestinal Transplantation

In some patients with irreversible intestinal failure, intestinal transplantation can be considered as an alternative to long-term PN. The first intestinal transplantation in humans was undertaken in the early 1960s. Limitations in technical expertise and immunosuppressive therapy meant that none of the original patients survived beyond 76 days. From 1985 to 1990, a series of 20 patients were given cyclosporine but only 2 patients were able to resume normal nutrition and most of the grafts failed. The development of new immunosuppressive agents, particularly tacrolimus, resulted in renewed interest in intestinal transplantation. Furthermore, since 1990, there has been greater standardization of patient selection, operative procedures, and postoperative care mainly in centers specializing in intestinal transplantation. The total international experience is still limited, involving less than 1000 transplants by 2004 (some of the transplants were isolated intestinal grafts, others were intestinal–liver transplants, and the remaining few were multivisceral transplants that included the intestine). Better graft and patient survival rates have been reported in the more experienced centers. In a series of 165 intestinal transplants at the University of Pittsburgh, patient survival was reported to be more than 75% at 1 year, 54% at 5 years, and 42% at 10 years. More than 90% of patients resumed an unrestricted oral diet.

It appears that intestinal transplantation has become a realistic life-saving option for some people who cannot be maintained on HPN. However, it is not yet the treatment of choice in patients who can be successfully maintained on HPN without noteworthy complications. Nor is it the treatment of choice in patients who are likely to deteriorate rapidly from other causes, such as aggressive multisystem disease, or likely to improve so that they can resume oral nutrition (e.g., patients with healing intestinal fistula or those with short bowel syndrome, in which benefits from intestinal adaptation may continue for up to 1–3 years). A better understanding of the immune response to the transplanted intestine and better immunosuppressive therapy, surgical techniques, and postoperative management are required. Appropriate selection and referral of patients to specialist centers are also important criteria that affect clinical outcomes.

Ethical Issues

The provision of nutritional support to people who are chronically sick, who have rapidly progressive disabling diseases, or who are terminally ill raises many ethical questions. Opinions about withholding or withdrawing artificial nutritional support vary

from country to country because of different clinical, religious, and social beliefs and differences in national economies, some of which cannot support large-scale expensive long-term treatments. Thus, there is little home artificial nutrition in countries with poor economies. In more developed economies, the types of patients being fed may also vary considerably. For example, parenteral and enteral nutrition in patients with cancer are used more frequently in Italy than in the United Kingdom, suggesting that clinical attitudes to this type of nutritional support vary. The sanctity of human life is a belief that is strongly held by many religions, but when these conflict with medical judgment, public policies normally override personal religious beliefs. A common ethical controversy concerns the need to provide food and fluid to prolong life in severely disabled patients, such as those with severe neurological problems (e.g., cerebrovascular accident) or those approaching the end of their lives. Although health professionals have a duty to prolong life, it seems inappropriate to prolong suffering. There has been controversy as to whether the provision of food and fluid by a feeding tube placed in the stomach or small intestine should be regarded as an essential part of care or medical treatment. The highest legal authorities in countries such as the United States and England have ruled that this is medical treatment. From an ethical perspective, there is no difference between withholding and withdrawing treatment, but in practice it is often more difficult to withdraw treatment once it has begun than to not initiate it. Joint discussions at the outset between mentally capable patients, family members, and health care workers can do much to prevent future ethical dilemmas.

Conclusions

Home nutritional support, including both oral and artificial (enteral and parenteral) methods of feeding, is an important modality of treatment that is being used for an increasing number of people with disease and disability who are managed in the community. The identification of individuals who are at increased risk of malnutrition and who may benefit from additional nutritional support is a vital first step, which can be undertaken using a validated screening tool (such as MUST; **Figure 1**). Oral nutritional support, including liquid multinutrient supplements, is of value in improving the nutritional intake and functional well-being of patients with malnutrition in the community. Without ETF, many patients with persistent swallowing difficulties would die; similarly, without PN, many patients with persistent intestinal failure would not survive. Although these forms of home therapy can be life-saving, they may restrict normal lifestyle and lead to life-threatening complications. These complications can be prevented or treated by establishing an adequate organizational infrastructure. This should include education and training of both health workers and patients/caregivers as well as a management structure that allows all patients to be followed up and, when necessary, admitting patients to the hospital for more intensive investigations and therapy. Ethical difficulties about withholding or withdrawing artificial nutritional support are likely to continue and to vary with time and from country to country. Intestinal transplantation is becoming a potentially realistic option for a few patients with irreversible intestinal failure who cannot be adequately maintained on long-term PN, but it has not yet become part of routine clinical care in the same way as renal transplantation has become routine in patients with renal failure, who would otherwise receive a lifelong treatment with dialysis.

See also: **Food Fortification**: Developed Countries; Developing Countries. **Malnutrition**: Secondary, Diagnosis and Management. **Nutritional Support**: Adults, Enteral; Adults, Parenteral; Infants and Children, Parenteral. **Supplementation**: Dietary Supplements; Developing Countries; Developed Countries.

Further Reading

British Medical Association (1999) *Withholding and Withdrawing Life-Prolonging Medical Treatment. Guidance for Decision Making*. London: British Medical Association.

Elia M (2003) *Screening for Malnutrition: A Multidisciplinary Responsibility. Development and Use of the 'Malnutrition Universal Screening Tool' ('MUST') for Adults*. Redditch, UK: BAPEN.

Elia M (Chairman), Russell CA, Stratton RJ, and British Artificial Nutrition Survey (BANS) Committee (2001) *Trends in Artificial Nutrition Support in the UK during 1996–2000*. Maidenhead, UK: BAPEN.

Elia M, Stratton RJ, Holden C *et al.* (2001) Home enteral tube feeding following cerebrovascular accident. *Clinical Nutrition* **20**: 27–30.

Glencorse C, Meadows N, Holden C, and British Artificial Nutrition Survey (BANS) Committee (2003) *Trends in Artificial Nutrition Support in the UK between 1996 and 2002*. Redditch, UK: BAPEN.

Langnas AN (2004) Advances in small-intestine transplantation. *Transplantation* **77**: S75–S78.

Lennard-Jones JEB (1998) *Ethical and Legal Aspects of Clinical Hydration and Nutritional Support*. Maidenhead, UK: BAPEN.

Moreno JM, Shaffer J, Staun J *et al.* (2001) Survey on legislation and funding of home artificial nutrition in different European countries. *Clinical Nutrition* **20**: 117–123.

Stratton RJ and Elia M (1999) A critical, systematic analysis of the use of oral nutritional supplements in the community. *Clinical Nutrition* **18**(supplement 2): 29–84.

Stratton RJ, Green CJ, and Elia M (2003) *Disease-Related Malnutrition: An Evidence Based Approach to Treatment*. Oxford: CABI Publishing.

Adults, Enteral

K N Jeejeebhoy, University of Toronto, Toronto, ON, Canada

Nutrients are normally taken by eating a diet composed of a variety of natural foods. In instances in which a normal oral diet cannot be taken, it becomes necessary to nourish the individual by either the enteral or parenteral (intravenous) route. In this article, the use of the enteral route is considered.

Definition of Enteral Nutrition

Enteral nutrition (EN) is the process of nourishing an individual by the administration of a liquid diet of defined composition, usually through nasogastric (NG), nasointestinal (NI), gastrostomy, or jejunostomy tubes (tube feeding). However, palatable enteral products may be taken as supplemental or complete enteral feeding by mouth.

Indications for Enteral Nutrition

Enteral nutrition is the preferred way of feeding patients who cannot eat, absorb, or use a normal diet in the presence of a usable gastrointestinal tract. The following are indications for EN:

1. Critical care patients, including those with trauma and burns, and also after major surgery.
2. Anorexia in patients with malignant disease, sepsis, liver and renal failure, and inflammatory bowel disease (IBD).
3. Upper gastrointestinal obstruction or ulceration of the pharynx, esophagus, stomach, and duodenum may prevent the ingestion of normal food. Examples of these conditions are cancer, central nervous system disorders, and stenosis following ulceration.
4. Pancreatic disease: In patients with pancreatitis it may be possible to feed a low-fat enteral formula through a NI route beyond the duodenum without causing increased disease activity or pain.
5. Short bowel and severe malabsorption: In controlled trials enteral diets are not better absorbed than normal food. Therefore, the presence of a short bowel per se is not an indication for enteral feeding. On the other hand, some patients with severe malabsorption may benefit from the use of elemental diets.
6. Inflammatory bowel disease: In IBD, enteral feeding is useful under the following situations:
   i. Profound anorexia preventing the ingestion of a normal diet.
   ii. Abdominal discomfort due to partial bowel obstruction or intestinal inflammation.
   iii. Growth retardation resulting from insufficient nutrient intake.
   iv. In Crohn's disease some controlled trials have suggested that enteral feeding induces a remission comparable to that seen with steroids.
7. Dementia: Patients unable to feed themselves because of profound mental changes.

Selection of Patients and Timing of Nutritional Support

The process of nutritional support should be an integrated continuum from normal diet to EN based on a plan. This plan should be developed at the moment of each patient's entry and implemented with modifications until discharge. It is undesirable to have to make an urgent decision on a weekend after realizing that the patient has been starving for the previous 2 weeks. I favor an approach in which the nutritionist examines the needs of each patient at entry and follows the algorithm given in **Figure 1**.

Target Nutrient Intake Possibly Achievable

The nutritionist discusses with the patient alternatives to dietary intake and the use of oral nutritional supplements that may include enteral diets taken by mouth. These supplements include liquid formula diets as well as specific supplements (e.g., potassium, magnesium, calcium, zinc, and vitamins). If during a trial period there is progressive improvement in intake or the patient meets the target, this process is continued. If the patient cannot meet the target or is clearly unable to progress toward it, then formal EN is started.

Target Nutrient Intake Achievement Failed or Impossible

These patients are prime candidates for NG or NI feeding if the gastrointestinal tract is normal, as in the case of critical care, anorexia (usually secondary

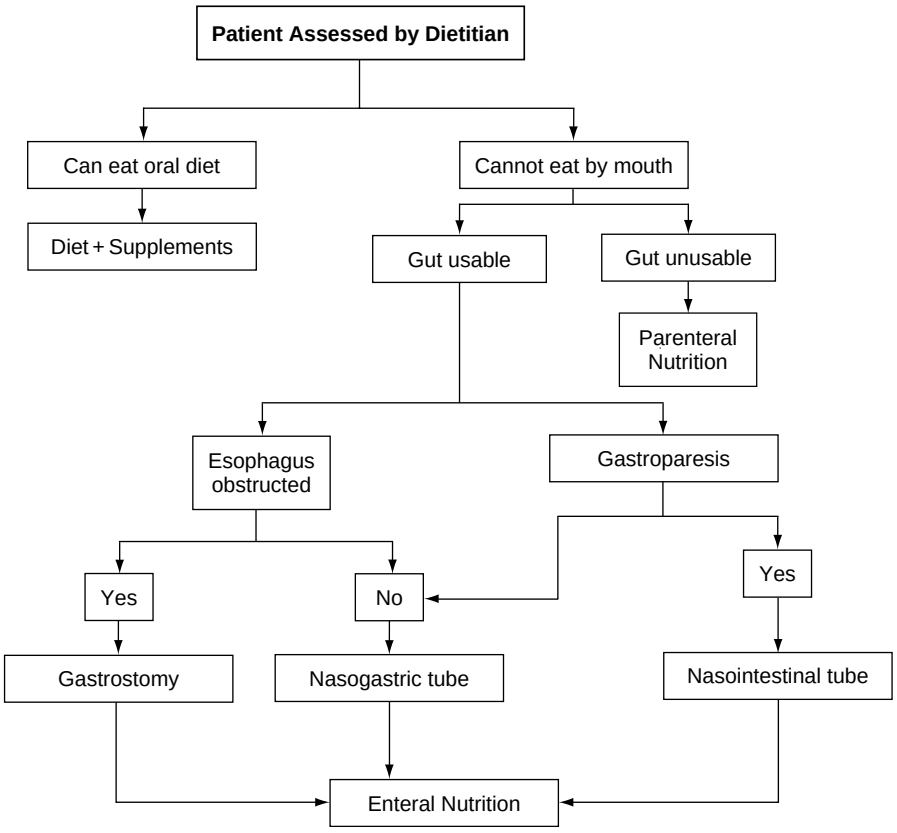


Figure 1 Algorithm for nutritional support.

to disease and malnutrition), neurological impairment preventing oral feeding, substantially increased requirements with relative anorexia (e.g., in burn cases), or chronic obstructive lung disease with severe dyspnea. However, for diseases of the pharynx, esophagus, or stomach or in cases of surgery of the esophagus, stomach, or pancreas, patients usually require intubation of the stomach or intestine by percutaneous gastrostomy or operative jejunostomy to allow feeding beyond the site of obstruction. If there is an abnormality of the intestinal tract, such as short bowel with more than 60 cm of available small intestine, IBD, or chronic partial bowel obstruction, diets must be delivered carefully with the aid of a pump to avoid surges of delivered fluid diets and consequent distension of the bowel. Despite careful selection, a proportion of patients expectantly fed via the nasogastric or nasoenteral route will show intolerance, complications, or inability to meet target nutrient intake without clinically unacceptable side effects. For example, in trials of patients with Crohn's disease, approximately 20% of patients could not tolerate nasogastric feeding. When EN fails or cannot be used for reasons given previously, then parenteral nutrition (PN) must be used.

Home Enteral Nutrition

In some patients, the oral route may be unusable for long periods of time or even permanently. This includes patients with indications given by Nos. 2, 3, and 5 in the list near the beginning of the article. In these patients, a percutaneous endoscopic gastrostomy is placed, and the patients are trained to feed themselves during the night and disconnect themselves during the day to go to school or work. This technique is invaluable to get patients with "gut failure" out of the hospital and rehabilitated.

Enteral Diets

The enteral diets are all complete and will meet the RNIs when fed to meet total energy requirements but may be deficient in meeting micronutrient requirements if given as supplements or in hypocaloric (not meeting total requirements) amounts. The types of diets are as follows:

1. Polymeric: Composed of whole proteins and oligosaccharides with fat partly as long-chain triglycerides and partly as medium-chain triglycerides. They are low in osmolality and palatable and can be taken by mouth.

Table 1 Specialized formulations

Formulation	*Use*
Branched-chain amino acids	Treat hepatic encephalopathy.
Glutamine	Reduce intestinal permeability, improve immunity, and promote mucosal regeneration.
ω-3 fatty acids	Reduce inflammatory response.
Arginine	Improve immunity.
High-fat diets	Reduce insulin requirements in diabetics. Reduce CO_2 output in patients with respiratory disease.

2. Peptide-based diets: The same as above but the protein is given in the form of peptides, which in theory are absorbed better than amino acids or proteins.
3. Elemental diets: They contain amino acids, are very low in fat, and the carbohydrate may be in the form of glucose. These diets are hyperosmolar and unpalatable. They are best given through a tube as a continuous infusion.
4. Special diets: They may be enriched in branched-chain amino acids, glutamine, ω-3 fatty acids, and arginine. Others have a high-fat content (**Table 1**).

Techniques of Administration

To ensure full calorie intake and to avoid gastrointestinal discomfort, enteral diets are best infused through a tube at a constant rate using a pump. The objective is to infuse at a rate that is equal to the rate of absorption so that intestinal distension does not occur. The diet should be stored in a sterile container so as to prevent bacterial growth while it is being infused. The routes of administration are as follows:

1. Nasogastric tube made of silicone rubber or polyurethane, 9–12 Fr in size: It is passed into the stomach and positioned in the antrum.
2. Nasointestinal tube: These tubes are similar to nasogastric tubes but longer. They are advanced under fluoroscopy into the duodenum. They may also be advanced by endoscopic guidance into the duodenum.
3. Percutaneous gastrostomy and buttons: Under sedation and local anesthesia, a gastrostomy can be placed using an endoscope. It can also be advanced into the duodenum. In long-term feeding the gastrostomy is replaced with a 'button' that is flush with the skin and can be intubated for feeding.
4. Percutaneous jejunostomy: This is placed at operation and can be used to feed into the intestine.

Evidence of the Benefits of Enteral Nutrition

EN has not been compared to standard care (SC) in the same systematic way as PN. Systematic reviews of EN compared to PN have consistently shown increased infectious complications with PN. However, all showed significantly elevated blood glucose in the PN group. It is likely that hyperglycemia was more frequent with PN because patients randomized to PN received more energy than those on EN, despite the intent to make both groups isocaloric. Data from a large controlled trial in intensive care unit (ICU) patients showed that keeping blood glucose below 7 mmol/l irrespective of the route of feeding significantly reduced mortality and multisystem organ failure arising from sepsis. This study indicated that hyperglycemia in the PN arm of the study would have significantly increased the risk of sepsis. None of these studies prove that EN is better than standard therapy; they show that it is less likely to cause infection than PN given without regard to the rigid control of blood glucose. This conclusion is supported by a large study (562 patients) comparing EN and PN that mirrors the conventional practice of NS. Using modest energy intake and avoiding hyperglycemia, the study showed that nutritional intake below 80% of the target was observed in 75% of randomized EN patients and 25% of randomized PN patients ($p < 0.001$). There was no significant differences in the incidence of septic morbidity between patients receiving PN and those given EN. The inability of EN to deliver target energy intake was also seen in several other trials, in which outcome was also no different between PN and EN.

Early Enteral Nutrition, Parenteral Nutrition, and Bacterial Translocation

In animal models, burns and trauma have been associated with the appearance of organisms in the mesenteric lymph nodes (MLNs). This process has been called bacterial translocation. Enteral feeding has been associated with reduced translocation in guinea pigs. In other animal studies, early enteral feeding reduced nitrogen loss and the level of catabolic hormones. However, human studies in patients who have been traumatized have not show any benefit of early (<24 h) enteral feeding. In addition, a prospective sampling of portal blood in trauma patients failed to confirm that translocation occurs in traumatized humans. A meta-analysis of early vs late enteral feeding showed no difference in outcome. In obese patients, a quasi-randomized trial

showed that early feeding with a higher energy intake increased sepsis.

Special Formulations

Immunonutrition

Enteral formulations enriched in arginine, omega-3 fatty acids, and glutamine nucleotides are considered to enhance the immune response, and treatments with these formulations are collectively referred to as immunonutrition. The formulations under consideration vary in composition. They are distinguished by high (12–15 g/l) or low (4–6 g/l) arginine, the presence or absence of glutamine and nucleotides, and the concentration of omega-3 fatty acids. The proceedings of the summit on immune-enhancing enteral therapy concluded that immunonutrition should be given to malnourished patients undergoing elective gastrointestinal surgery and trauma patients with an injury severity score of $\geq$18 or those with an abdominal trauma index of $\geq$20. Despite lack of evidence, it was recommended for patients undergoing head and neck surgery and aortic reconstruction, those with severe head injury and burns, and for ventilator-dependent nonseptic patients. The summit did not recommend it for patients with splanchnic hypoperfusion, bowel obstruction distal to the access site, and after major upper gastrointestinal hemorrhage.

In contrast to the conclusions of the summit, systematic reviews of the evidence have given mixed results. The reviews suggested that although immunonutrition did reduce septic complications, the reduction did not result in reduced mortality. In a meta-analysis of 22 randomized controlled trials performed in 2419 critically ill or surgical patients, it was concluded that the amount of arginine in the formulations influenced the results. Taken as a whole, in critically ill patients there were no treatment effects on mortality or rates of infectious complications. In fact, there was a suggestion that in critical illness these formulations may increase mortality. To support this possibility, a trial suspended the use of immunonutrition in seriously ill patients after an interim analysis showed increased mortality with immunonutrition in these patients. However, in elective surgical patients immunonutrition reduced complications and length of stay. Since many trauma and septic patients may be critically ill, the recommendations made in the two publications referred to previously are at variance. Other meta-analyses have not separated critically ill and nonseptic patients and have concluded that immunonutrition reduced septic complications and length of stay but criticized the component studies as being variable and overall not altering mortality.

Enteral Glutamine Supplementation

Glutamine is released into the circulation from muscle continuously in healthy people and especially in those with catabolic illness. The glutamine in the circulation is an important nutrient for immunocytes such as lymphocytes and for the mucosa of the intestine. In septic and malnourished patients, muscle glutamine is depleted, and it is hypothesized that in these patients the availability of glutamine for lymphocytes and the gut is reduced, resulting in increased risk of sepsis. Although enteral mixtures designed to improve immunity have given variable results, glutamine supplementation has not been shown to be harmful and has reduced complications in patients with bone marrow transplantation, after surgery, and in those with critical illness and burns.

Enteral Branched-Chain Amino Acid Formulations

Patients with hepatic encephalopathy have low levels of enteral branched-chain amino acids (BCAAs) and increased levels of phenylalanine and tryptophane. It has been postulated that since BCAAs compete with phenylalanine and tryptophane for transport through the blood–brain barrier, reduced levels of BCAAs promote the accumulation of these amino acids in the brain, where they are metabolized to false neurotransmitters that then cause encephalopathy. To reverse this state, BCAA-enriched formulas were developed, and in a meta-analysis of randomized controlled trials of these formulations in hepatic encephalopathy, they were shown to reduce the duration of encephalopathy in comatose patients.

Enteral Formulation Enriched in Protein

High-protein formulations are available to feed patients with adequate amounts of protein at a lower energy intake. These formulations are especially useful in patients who are critically ill and become hyperglycemic on standard formulations. Using high-protein formulations has allowed the administration of up to 2 g/kg protein in patients with protein losses, such as those with fistulas, burns, and abscesses, and in calorie-intolerant patients. One controlled trial has shown that a high protein intake reduces mortality in burn patients, whereas high energy intake in sepsis appears to increase mortality.

High-Fat Formulations

The average fat content of enteral formulations is approximately 30% of total energy. The main source of energy in these formulations is carbohydrate. The high carbohydrate has two relevant metabolic effects. First, it increases the need for insulin secretion. Second, when fed in amounts that exceed energy requirements, it increases CO_2 production. Diabetics potentially would therefore have reduced insulin requirements and a lower risk of hyperglycemia if fed high-fat formulations. Similarly, malnourished patients with chronic obstructive pulmonary disease (COPD) would have a lower risk of CO_2 retention if given a high-fat diet with a high calorie intake to promote weight gain. Therefore, high-fat formulations have been developed to feed patients with diabetes and COPD. However, there is little evidence that they are significantly better than standard formulations.

Enteral Omega-3 Fat Supplementation

Omega-3 fats are composed of polyunsaturated fatty acids, in which the first double bond 3 carbon atoms are located away from the methyl end of the fatty acid chain. The fatty acids found in fish oil, called eicosapentanoic and docosahexaenoic, are precursors of prostaglandins and thromboxanes that antagonize the prothrombotic effects of similar compounds derived from linoleic acid. In humans, when infused, they reduce the production of proinflammatory cytokines from stimulated mononuclear cells. They potentially have anti-inflammatory effects and have been shown in controlled trials to benefit patients with ARDS.

Enteral Nutrition with Probiotics

The administration of a probiotic, *Lacobacillus plantarum*, indicates that probiotics with EN may have a role in reducing septic complications in patients with pancreatitis and those after liver transplantation.

Optimizing Enteral Nutrition and Reducing Risk of Aspiration Pneumonia

Enteral feeding is associated with several factors that may result in reflux of gastric contents and aspiration: the supine position of the patient, the presence of a nasogastric tube, gastric contents, and delayed emptying of the stomach. Intuitively, placing the tip of the feeding tube into the intestine rather than the stomach should reduce aspiration, and the Canadian clinical practice guidelines recommend postpyloric feeding. On the other hand, a meta-analysis comparing gastric and postpyloric feeding did not show any significant difference in the incidence of pneumonia between patients fed into the stomach and those fed beyond the pylorus. Although enteral feeding is widely practiced as the route of choice, in a study of 103 patients admitted to an ICU who were observed prospectively for the development of nosocomial pneumonia, there was evidence that feeding contributed to pneumonia. In that study, a multivariate analysis concluded that continuous enteral feeding, but not the nasogastric tube, was an independent risk factor for nosocomial pneumonia and patients who developed pneumonia had a significantly higher mortality of 43.5% compared to 18.8% for those who did not develop pneumonia. Clearly, more studies need to be done to determine the best approach to prevent pneumonia.

The use of prokinetics is another way of promoting gastric emptying. In a placebo controlled randomized trial of 305 patients receiving enteral feeding, giving metoclopramide did not reduce the incidence of pneumonia. Erythromycin, a motilin receptor agonist, is another powerful prokinetic agent. In a randomized controlled trial the benefit of erythromycin was questionable. There was no difference in the rate of pneumonia between the placebo and erythromycin-treated patients. The previous studies unfortunately involved small numbers of patients, and there is a need for larger trials of small bowel feeding and prokinetics to establish their role in promoting enteral feeding and reducing the risk of aspiration.

Perioperative Enteral Nutrition

Infusing a diet enriched with arginine, omega-3 fatty acid, and RNA preoperatively and postoperatively has resulted in a significant reduction in total but not major complications. In patients undergoing abdominal surgery, progressive postoperative oral supplementation without formal EN reduces complications and raises the question as to whether EN, total parenteral nutrition (TPN), or immunonutrition are even necessary for the majority of patients. Complications can be reduced by introducing early sip feeding of liquid diets without formal EN.

A small but provocative study from India raises the same questions about the routine use of EN. Sixty postoperative patients were randomized to either a standard ward diet or a diet with a home-made liquid supplement (10 patients per group). They were also stratified to mild, moderate,

and severe malnutrition groups of 20 patients each. There was no mortality in the study and patients were discharged after approximately 10 days. The supplemented groups received significantly more energy and protein. However, only in the severely malnourished patients was there a difference in the incidence of complications—7/10 in the control and 4/10 in the supplemented group. In the mild malnutrition group, there were 1/10 complications in each arm and 2/10 in the moderate malnutrition group. This study suggests that a very modest oral intake of supplements reduced complications but did so only in the severely malnourished group. It is likely that the aggressive nutritional support, as is practiced currently, may not be necessary and may even be detrimental in some situations. Larger randomized trials of oral supplements should be considered.

Enteral Nutrition and Head Injury

A systematic review of controlled trials of nutritional support in cases of head injury showed that the relative risk (RR) for death with early nutritional support was 0.67 (95% confidence interval (CI), 0.41–1.07), and the RR for death or complications at the end of follow-up was 0.75 (95% CI, 0.50–1.11). The findings suggested that early nutrition showed a trend toward reduced mortality and complications.

Nutritional Support of Bone Marrow Transplant Patients

A review found that although EN is the current standard for nutritional support, it has not found favor for patients undergoing bone marrow transplant because these patients have severe mucositis, often vomit the tube, and do not tolerate nasogastric tubes because of discomfort and ulceration. Veno-occlusive disease with encephalopathy may occur in bone marrow patients, which is another indication for TPN with branched-chain amino acids rather than EN. A controlled trial of EN versus TPN in bone marrow transplant patients showed that outcome was no different but body composition and magnesium levels were better maintained on TPN. In contrast, TPN patients had more fluid overload and hyperglycemia. It should be noted that 'enteral nutrition' in this trial was not tube feeding but a combination of snacks, diet counseling, and tube feeding. The authors concluded that TPN should be reserved for patients with severe mucositis. Review of the Cochrane database concluded that the relative effectiveness of EN versus TPN could not be evaluated. In addition, patients with gastrointestinal failure should consider TPN with the addition of glutamine if EN is not possible.

Alcoholic Hepatitis and Enteral Nutrition

Seventy-one patients with severe alcoholic hepatitis were randomized to prednisone 40 mg/day or EN giving 2000 kcal/day for 28 days and then followed for 1 year or until death. The EN was a branched-chain-enriched diet and patients on steroid therapy were encouraged by dietitians to eat 2000 kcal/day with 1 g/kg/day of protein. No patients from the steroid arm dropped out, whereas 8/35 patients from the EN arm did not receive EN for the entire period but were included in the analysis (intent to treat analysis). It is of interest that all patients in the steroid arm ate 80% of the prescribed diet. Using intent to treat analysis, there were no differences in mortality or complications in the hospital between groups. After discharge, even when confounding variables were adjusted, the EN group had a significantly better survival. Since both groups seemed to receive the same energy intake, the reason for better long-term survival with EN needs further study. Was it because of the use of branched-chain amino acids or because steroid therapy had an undesirable catabolic effect?

Pancreatitis and Enteral Nutrition

Oral feeding is known to increase abdominal pain in patients with pancreatitis. Therefore, TPN has been used in these patients to 'rest' the pancreas. One study aimed to define the indications and evaluate the cost-effectiveness of nutritional support in a series of patients with pancreatitis. The patients were given nothing orally and only intravenous fluids for 48 h. Those who improved were fed orally (O group). The remainder were randomized to receive nutrients either infused into the jejunum (EN group) or by vein (TPN group). A total of 156 patients were included, of whom 75% improved (O) and were discharged within 4 days. In the randomized patients, 56% of the EN group received inadequate energy intake but were fed for a significantly shorter period (mean, 6.7 versus 10.8 days) and had less metabolic ($p < 0.003$) and septic complications ($p < 0.01$). More than 50% of TPN patients were hyperglycemic, in contrast to only approximately 15% of the EN group. Despite fewer complications in the EN group, the mortality was similar in the

two groups. The authors concluded that hypocaloric enteral feeding is better than TPN. This study is similar to many others, showing that EN providing less than estimated energy intake is associated with reduced hyperglycemia and sepsis. The conventional interpretation is that the EN route reduces sepsis. The trial by van den Berghe *et al.* showed that irrespective of the route of nutritional support, control of hyperglycemia reduced mortality in the ICU. Their findings support the alternative explanation that the EN route protects the patient because it results in hypocaloric feeding, which prevents hyperglycemia. The study shows that EN can be given as a cheaper source of nutrition, but since EN was needed only for 6.7 days with less than adequate energy intake, it raises the question as to whether any nutritional support was required. Another important question is whether TPN should be hypocaloric rather than meet target energy intake in patients who are unable to take oral nutrition or EN for periods exceeding 7–10 days. In order to settle these issues, larger multicenter randomized trials are required.

See also: **Burns Patients**. **Colon**: Disorders; Nutritional Management of Disorders. **Diabetes Mellitus**: Dietary Management. **Eating Disorders**: Anorexia Nervosa; Bulimia Nervosa. **Microbiota of the Intestine**: Probiotics. **Nutritional Support**: Adults, Parenteral; Infants and Children, Parenteral. **Supplementation**: Dietary Supplements.

Further Reading

Alexander JW, MacMillan BG, Stinnett JD *et al.* (1980) Beneficial effects of aggressive protein feeding in severely burned children. *Annals of Surgery* **192**: 505–517.

Artigas AT, Dronda SB, Valles EC *et al.* (2001) Risk factors for nosocomial pneumonia in critically ill trauma patients. *Critical Care Medicine* **29**: 304–309.

Bertolini G, Iapichino G, Radrizzani D *et al.* (2003) Early enteral immunonutrition in patients with severe sepsis. Results of an interim analysis of a randomized multicentre clinical trial. *Intensive Care Medicine* **29**: 834–840.

De Jonghe B, Appere-de-Vechi C, Fournier M *et al.* (2001) A prospective survey of nutritional support practice in intensive care unit patients: What is prescribe? What is delivered? *Critical Care Medicine* **29**: 8–12.

Eyer SD, Micon LT, Konstantinides FN *et al.* (1993) Early enteral feeding does not attenuate metabolic response after blunt trauma. *Journal of Trauma* **34**: 639–643.

Gadek JE, DeMichele SJ, Karlstad MD *et al.* (1999) Effect of enteral feeding with eicosapentaenoic acid, gamma-linolenic acid, and antioxidants in patients with acute respiratory distress syndrome. Enteral Nutrition in ARDS Study Group. *Critical Care Medicine* **27**: 1409–1420.

Han-Geurts IJM, Jeekel J, Tilanus HW, and Brouwer KJ (2001) Randomized clinical trial of patient-controlled versus fixed regimen feeding after elective surgery. *British Journal of Surgery* **88**: 1578–1582.

Heyland DK, Dhaliwal R, Drover JW, Gramlich L, and Dodek P (2003) Canadian clinical practice guidelines for nutrition support in mechanically ventilated, critically ill adult patients. *Journal of Parenteral and Enteral Nutrition* **27**: 355–373.

Heyland DK, Novak F, Drover JW *et al.* (2001) Should immunonutrition become routine in critically ill patients? *Journal of the American Medical Association* **286**: 944–953.

Ibrahim EH, Mehringer L, Prentice D *et al.* (2002) Early versus late enteral feeding of mechanically ventilated patients: Results of a clinical trial. *Journal of Parenteral and Enteral Nutrition* **26**: 174–181.

Jeejeebhoy KN (2001) TPN potion or poison. *American Journal of Clinical Nutrition* **74**: 160.

Kudsk KA and Moore FA (chairpersons) (2001) Proceedings from the summit on immune-enhancing enteral therapy. *Journal of Parenteral and Enteral Nutrition* **25**: S1–S63.

Marik PE and Zaloga GP (2003) Gastric versus post-pyloric feeding: A systematic review. *Critical Care* **7**: R46–R61.

Montejo JC, Zarazaga A, Lopez-Martinez J and the Spanish Society of Intensive Care Medicine and Coronary Units (2003) Immunonutrition in the intensive care unit. A systematic review and consensus statement. *Clinical Nutrition* **22**: 221–233.

Moore FA, Moore EE, Poggetti R *et al.* (1991) Gut bacterial translocation via the portal vein: A clinical perspective with major torso trauma. *Journal of Trauma* **31**: 629–636.

Murray SM and Pindoria S (2002) Nutrition support for bone marrow transplant patients. *Cochrane Database of Systems Review* **2**: CD002920.

Naylor CD, O'Rourke K, Detsky AS, and Baker JP (1989) Parenteral nutrition with branched-chain amino acids in hepatic encephalopathy. A meta-analysis. *Gastroenterology* **97**: 1033–1042.

Nelson JK and Fleming CR (1990) Home enteral nutrition for adults. In: Rombeau JL and Caldwell MD (eds.) *Clinical Nutrition Enteral and Tube Feeding*, 2nd edn, pp. 450–462. Toronto: WB Saunders.

Saluja SS, Kaur N, and Shrivastava UK (2002) Enteral nutrition in surgical patients. *Surgery Today* **32**: 672–678.

Sanders DS, Carter MJ, D'Silva J *et al.* (2000) Survival analysis in percutaneous endoscopic gastrostomy feeding: A worse outcome in patients with dementia. *American Journal of Gastroenterology* **95**: 1472–1475.

Souheil A-Ai, Kimberly C, and O'Keefe SJD (2002) Hypocaloric jejunal feeding is better than total parenteral nutrition in acute pancreatitis: Results of a randomized comparative study. *American Journal of Gastroenterology* **97**: 2255–2262.

van den Berghe G, Wouters P, Weekers F *et al.* (2001) Intensive insulin therapy in the critically ill patients. *New England Journal of Medicine* **345**: 1359.

Woodcock NP, Zeigler D, Palmer MD *et al.* (2001) Enteral versus parenteral nutrition: A pragmatic study. *Nutrition* **17**: 1–12.

Yanagawa T, Bunn F, Roberts I, Wentz R, and Pierro A (2002) Nutritional support for head-injured patients. *Cochrane Database of Systems Review* **3**: CD001530.

Zaloga GP and Roberts P (1994) Permissive underfeeding. *New Horizons* **2**: 257–263.

Adults, Parenteral

J Binkley, S Daniell and G L Jensen, Vanderbilt Center for Human Nutrition, Nashville, TN, USA

Parenteral nutrition (PN) is a compounded formulation of amino acids, dextrose, and lipid emulsions, along with electrolytes, multivitamins, and trace elements. The development of this form of nutrition intervention gave new hope to patients who suffered from intestinal compromise or failure.

PN had its modern beginnings in the mid-1960s as Dr. Stanley Dudrick and colleagues researched infusion of hypertonic glucose and protein solutions into the superior vena cava of beagle puppies. Normal growth and development of the beagles were maintained for 36 months using this approach. The first research that established the use of intravenous feedings in humans was reported by Dudrick and colleagues in 1969. From its early beginnings, PN was known as 'hyperalimentation,' with the belief that it was desirable to feed in excess of standard requirements. Years of clinical practice has shown that PN should be provided in more limited amounts in order to prevent some of the complications associated with its overzealous use.

This article focuses on components of PN for adult nutrition support, indications and contraindications for its use, implementation and monitoring for safety and efficacy, complications associated with use of PN, and consideration of its use for home patients. Prudent patient selection and careful monitoring will help to ensure the safe and effective administration of PN. A multidisciplinary approach to PN management is suggested to help optimize the use of this therapy.

Indications for PN

PN should be considered for patients when oral intake or enteral feedings are not possible or are contraindicated for a prolonged period of time. Enteral feedings are the preferred route of administration for specialized nutrition support for many reasons. Enteral feedings are more physiologic and facilitate maintenance of gastrointestinal integrity and function. In comparison with PN, enteral feedings are considerably less expensive and are associated with fewer serious adverse effects. Enterohepatic circulation and barrier function of the gastrointestinal mucosa can be preserved with even small quantities of enteral stimulation. Enteral feedings are associated with a decreased incidence of bacterial translocation and associated sepsis in animal models. PN is indicated when the gastrointestinal tract is not functional, when the safe placement of an enteral feeding access device is not possible, or when the enteral route cannot adequately meet the nutritional needs of a patient. **Table 1** lists common indications for PN. When enteral feedings cannot be established within 7–10 days, PN should be considered. **Table 2** lists contraindications to the use of enteral feedings.

Bowel Rest

PN is often used when continued use of the gastrointestinal tract may not be advisable. PN may be selected for inflammatory bowel disease patients with severe acute exacerbations or for perioperative care. For patients with Crohn's disease, PN may aid the management of complications such as intestinal obstruction, fistula formation, short bowel syndrome, and severe diarrhea. Otherwise, enteral nutrition support is frequently used for nutrition support in inflammatory bowel disease with comparable efficacy.

PN can be used for bowel rest in severe acute pancreatitis, when its duration is anticipated to be more than 7–10 days. Various scoring systems are used to classify the severity of pancreatitis and together with sound clinical judgment can help to

Table 1 Common diagnoses with indications for PN

Perioperative support in severe malnutrition
Inflammatory bowel disease and related complications
Short bowel syndrome
Severe acute pancreatitis
Mechanical intestinal obstruction or pseudo-obstruction
High-output entercutaneous fistula
Prolonged postoperative ileus
Severe malabsorption
Bone marrow transplant/peripheral stem cell transplant
Severe hyperemesis gravidarum

Table 2 Contraindications to enteral nutrition

Diffuse peritonitis
Intestinal obstruction that prohibits use of the bowel
Intractable vomiting
Paralytic ileus
Intractable diarrhea
Gastrointestinal ischemia

Adapted from ASPEN Board of Directors and the Clinical Guidelines Task Force (2002) Guidelines for the use of parenteral and enteral nutrition in adult and pediatric patients. *Journal of Parenteral and Enteral Nutrition* **26**(1 supplement): 18SA.

assess the need for bowel rest with PN. Studies using nasojejunal or jejunostomy feeding tubes with elemental low-fat enteral formulas in patients with mild to moderate pancreatitis have demonstrated effectiveness. Patients who exhibit feeding intolerance with enteral nutrition should be considered for PN therapy.

Bowel rest may also be indicated for selected enterocutaneous fistulas, which can occur as a result of complicated Crohn's disease, gastrointestinal or abdominal abscesses, abdominal surgery or trauma, ischemia, or tumors or their accompanying treatment regimens such as chemoradiation. Bowel rest can help to promote potential closure of fistulas and can improve nutritional status of these patients. Depending on the output of the fistulas, many of these patients are at risk for malnutrition as well as dehydration and electrolyte abnormalities.

Perioperative Support in Severe Malnutrition

Increased morbidity and mortality risks are associated with malnourished surgical patients. The Veteran's Affairs Parenteral Nutrition Cooperative Trial evaluated the benefits of preoperative PN in patients with varying degrees of malnutrition. Significant benefit was demonstrated only among those patients who were severely malnourished (albumin <3.0 g/dl). Interestingly, an increased rate of infectious complications was observed in mildly and moderately malnourished patients receiving PN compared to the control group. If enteral access is available and feeds are tolerated, preoperative enteral nutrition support has been found to be equally effective when utilized for 7–14 days or longer in malnourished surgical patients. Early studies that suggested increased adverse outcomes with parenteral compared to enteral support were often based on management techniques that are no longer consistent with standard of practice, which included the overfeeding of macronutrients, rapid infusion rates, and a much less aggressive approach to the prevention of hyperglycemia. Therefore, many of the studies demonstrating more favorable outcomes with enteral support need to be repeated using current standards of care.

Postoperatively, in patients who have undergone intraabdominal procedures with extensive bowel manipulations, paralytic ileus is commonly observed. Use of narcotics or other pain medications, paralytic agents, and electrolyte abnormalities may also contribute to slow gastrointestinal motility and ileus. If a patient's gastrointestinal tract is not functional for more than 7–10 days, PN should be considered.

Gastrointestinal Inability to Absorb Adequate Nutrients

PN is indicated if the patient is unable to absorb adequate nutrients via the gastrointestinal tract. Short bowel syndrome can result from extensive bowel resection or dysfunction. Initially, fluid and electrolyte management is often the most critical aspect of nutrition care. Losses can be extensive. The resulting degree of malabsorption, and therefore the specific nutrition management requirements, depends on the remaining intestinal length or function, the presence or absence of large bowel continuity, and the presence or absence of the ileum. Enteral nutrition support is preferred over parenteral support; however, the adaptive phase for the remaining gastrointestinal tract often necessitates a period of PN requirement, often months to a few years. Sometimes patients with extensive bowel resection (less than 100 cm of remaining bowel) may require lifelong PN or small bowel transplantation.

Other Possible Indications for Use

The importance of adequate nutrition in pregnancy is widely recognized. Hyperemesis gravidarum (HG) is severe nausea and emesis that can persist throughout the gestation period, preventing the patient from receiving adequate calories and protein for fetal growth. HG usually occurs before the 20th week of gestation and is characterized by weight loss, changes in fluid and electrolyte status, and disturbances in acid–base balance. After attempting to achieve weight gain and resolution of symptoms with antiemetics, intravenous hydration, and diet modifications without success, specialized nutrition support with enteral feedings or PN may be appropriate for refractory cases. Enteral feedings are preferred, but PN is indicated if enteral feedings are not tolerated.

Another potential need for PN occurs in patients treated with high-dose chemotherapy with or without radiation, followed by hematopoetic stem cell transplantation. This treatment regimen may be associated with significant gastrointestinal side effects, including nausea, emesis, diarrhea, mucositis, esophagitis, xerostomia, and odynophagia. Nutritional compromise is possible due to the persistence of these symptoms for protracted time periods. Specialized nutrition support with enteral nutrition or PN may be indicated if patients are unable to maintain adequate oral intake or if they demonstrate persistent gastrointestinal side effects. PN has been suggested to improve posttransplantation survival, reduce disease relapse, and shorten

hospital stay, although enteral feedings have also been successfully administered to such patients.

Nutrition Components of PN

Amino Acids

Amino acids yield 4 kcal/g when oxidized for energy. Nitrogen content varies somewhat, depending on the individual amino acid formulation and mixture of amino acids. Mixtures of buffered essential and nonessential amino acids are available as stock concentrations ranging from 3 to 20%. Specialty amino acid products are also available for specific disease states or pediatric populations. For example, formulations containing a higher concentration of branched-chain amino acids may be considered for patients with hepatic insufficiency with accompanying encephalopathy.

Dextrose

The carbohydrate energy source for PN is hydrated dextrose, which yields 3.4 kcal/g. Dextrose is available commercially in concentrations ranging from 2.5 to 70%. Dextrose solutions are acidic, and their osmolarity depends on concentration. Higher concentrations of dextrose must be administered directly into central veins instead of peripheral veins in order to prevent thrombophlebitis.

Lipid Emulsions

PN regimens include lipid emulsions as the source for fat calories and essential fatty acids. In the United States, commercially available intravenous lipid products contain largely n-6 long-chain fatty acids derived from vegetable oils. Long-chain lipid emulsions provide the most concentrated source of calories in PN (9 kcal/g). Other lipid-containing products are currently available in Europe, including lipid emulsions containing medium-chain fatty acids and structured triglycerides. The latter are custom synthesized triglycerides that contain both long-chain and medium-chain fatty acids on the same glycerol moiety. These lipid substrates may offer certain metabolic and immune tolerance advantages.

Controversy surrounds the use of intravenous lipid emulsions (IVLEs) due to early reports of adverse immune function and pulmonary effects in critically ill patients. These early reports of alterations in immune function were associated with excessive infusion rates or doses of IVLEs in comparison with today's standard dosing regimens. Concerns have also been raised that n-6 fatty acids may fuel inflammatory eicosanoid pathways. Lipids are contraindicated in patients with significant hypertriglyceridemia. When lipids are restricted, modest doses of lipids (30–40 g twice weekly) should be provided to prevent essential fatty acid deficiency (EFAD). Linoleic acid (18:3n6) and α-linolenic acid (18:3n3) are required to prevent EFAD. To prevent EFAD, 2–4% of calories should be provided as linoleic acid and 0.5% as α-linolenic acid.

Electrolytes

Specific patient electrolyte requirements can be determined and added, as feasible, to PN solutions. Usually, maintenance doses of electrolytes are determined and provided in daily amounts of electrolytes added to the PN solution. Additional repletion and replacement of losses can occur in electrolyte doses given outside the PN admixture. Sodium and potassium can be added as chloride or acetate, depending on acid–base needs. Physicochemical incompatibilities exist for large quantities of electrolytes added to the same PN formulation. In particular, calcium and phosphorus concentrations must be carefully scrutinized to prevent precipitation. Many factors may influence the solubility of these electrolytes in the PN solution, including the concentration of electrolytes, the pH of the final formula, temperature, and the presence of other components. A nutrition support pharmacist can be a resource to address compatibility concerns.

Multivitamins

Multivitamin preparations, including both water-soluble and fat-soluble vitamins, are available for inclusion in the PN admixture. These products have been formulated to meet the guidelines established by the American Medical Association Nutrition Advisory Group and the Food and Drug Administration. **Table 3** lists the composition of

Table 3 Contents of parenteral multivitamin preparations

Vitamin component	*Current FDA requirements*
Vitamin A	3300 IU
Vitamin D (ergocalciferol or cholecalciferol)	200 IU
Vitamin E (α-tocopherol)	10 IU
Vitamin K (phylloquinone)	150 μg
Vitamin C	200 mg
Folic acid	600 μg
Niacin	40 mg
Vitamin B_2	3.6 mg
Vitamin B_1	6 mg
Vitamin B_6	6 mg
Vitamin B_{12}	5 μg
Pantothenic acid	15 mg
Biotin	60 μg

standard adult multivitamin products. Some individual vitamin preparations, such as vitamin K, are available for injection as well.

Trace Elements

Intravenous trace element preparations are commercially available as single-entity products as well as a variety of combination products. Commonly used trace elements include zinc, copper, manganese, chromium, and selenium (**Table 4**). Other elements may be included as single additives.

Titration of Volume

Final fluid volumes of PN can be titrated to a desired amount using sterile water for injection. Concentrated substrates may be used to minimize volume of PN to provide the desired components. Typical PN volumes range from 800 to 2500 ml daily.

Compounding and Technical Requirements for PN

Access Devices and PN Concentrations

PN is administered into the venous system either through peripheral venous lines or through centrally placed access devices. Lower concentrations of dextrose and amino acids may be administered through peripheral veins for a short duration of therapy. Such formulas usually do not provide the patient's full nutrition needs, may require large volumes of fluid, and can only be used for short durations due to the difficulty of maintaining peripheral intravenous access. Osmolarity of peripheral formulas is best maintained at approximately 600 mOsm/l or less. This requirement means that peripheral PN formulas should contain no more than 5–10% dextrose and 3.5–5% amino acids. Potential complications of peripheral PN include phlebitis, infiltration, or fluid-overload issues. When higher concentrations of dextrose and amino acids are used, such as those generally needed to provide adequate daily nutrient requirements via PN, the hyperosmolar formula must be administered directly into the superior or inferior vena cava to facilitate rapid dilution. Commonly used central venous catheters that may be used to administer PN include subclavian vein catheters, peripherally inserted central catheters, subcutaneously tunneled percutaneous catheters, or implanted subcutaneous infusion ports. Catheter type will be determined by expected duration of need, specific patient condition, patient care setting, as well as physician or patient preference.

Table 4 Contents of a common parenteral trace element preparation

Component	*Dose*
Zinc	5 mg
Copper	1 mg
Manganese	0.5 mg
Chromium	10 μg
Selenium	60 μg

Standard versus Individualized Preparations

Institutions or patient care providers may choose to provide PN solutions as standardized formulas or as customized admixtures, specially tailored to the individual's needs. Commercially available premixed PN solutions typically contain 5–25% dextrose and 2.75–5% amino acids, and they may vary by electrolyte content. Individualized formulations are selected to ensure the highest quality in patient safety and product efficacy; however, premixed solutions are often used in settings in which the demand for PN is low.

Two-in-One versus Total Nutrient Admixture

PN admixtures can be compounded by one of two methods: as a dextrose–amino acid solution with lipids infused separately or as a three-in-one formulation, also known as total nutrient admixture, in which all three macronutrients are combined in the same infusion bag. There are benefits and limitations with both methods, and the choice of administration depends on the care setting and institution or practitioner preference.

Cyclic PN

PN therapy may be infused continuously 24-h daily, or the same volume may be infused over a shorter period of time, such as a cycle of 12-h PN infusion and 12-h free of infusion. Infusion pumps can be programmed to adjust infusion rates according to the desired volumes and administration times. Continuous infusion is generally selected when PN is first initiated for an acutely ill patient. Benefits for a cyclic total PN regimen are particularly notable for long-term patients. A cycled PN regimen allows more mobility for the patient, thus enabling the patient to achieve a more active lifestyle. Limitations to a cycled PN regimen include fluid intolerance or glucose intolerance. When initiating a PN cycle, blood glucose concentrations should be checked to ensure that hyperglycemia or hypoglycemia is not an issue for the patient. Because of the potential for 'rebound hypoglycemia' upon abrupt cessation of infusion,

the rate of administration is often tapered down at the end of the cycle to allow for downregulation of pancreatic release of insulin. Many infusion pumps have programmable taper functions.

Quality Control

Safety of PN solutions is a paramount objective and includes ensuring accuracy in compounding and avoiding both particulate matter and microbial contamination. PN solutions must be prepared using a strict aseptic technique in a class 100 environment using a laminar flow hood. All PN additives should always be added in the sterile environment to prevent risk of contamination. Many incompatibility issues exist when considering mixtures of PN solutions with other medications. Practitioners should assume that medications are incompatible unless data otherwise prove compatibility exists. Some medications have been demonstrated to be compatible as a component of PN solutions, such as heparin, regular insulin, H_2-receptor antagonists, and corticosteroids.

Calculating Nutritional Needs

Accurate height and weight measurements are important to make appropriate estimations of energy and protein needs. Several formulas exist to calculate ideal body weight (IBW). The following are commonly used:

$$\text{Men} = 50\,\text{kg} + 2.3\,\text{kg} \times \text{each inch over 5 feet}$$

$$\text{Women} = 45.5\,\text{kg} + 2.3\,\text{kg} \times \text{each inch over 5 feet}$$

Adjusted body weight can be calculated utilizing the following formula:

$$\text{Adjusted body weight} = \text{IBW} + 25\%\ \text{of (actual body weight} - \text{IBW)}$$

Calculating Energy Needs

Calorie requirements are estimated to meet a patient's energy requirements, which are dependent on the patient's size, clinical condition, concurrent organ failure, and activity level. Determination of the appropriate energy prescription for a patient is crucial for meeting metabolic demands and helping to prevent erosion of lean body mass. Overzealous feeding is associated with significant risks, including difficulties with glucose control and other metabolic complications such as excessive carbon dioxide production. Common methods to estimate calorie requirements include simple weight-based algorithms (e.g., 25–30 kcal/kg body weight/day) (**Table 5**). The Harris–Benedict equations are also frequently used to estimate basal energy expenditure (BEE) using the following formulas:

Table 5 Determination of energy needs

Condition	*Need (kcal/kg)*
Overnourished/obese	20 (upper end IBW)
Maintenance	25
Undernourished	30
Stressed/critically ill	25

IBW, ideal body weight.
Adapted from the National Advisory Group on Standards and Practice Guidelines for Parenteral Nutrition (1998) Safe practices for parenteral nutrition formulations. *Journal of Parenteral and Enteral Nutrition* **22**: 49–66.

$$\text{Males: BEE (kcal)} = 66.5 + [13.8 \times \text{weight (kg)}] + [5 \times \text{height (cm)}] - 6.8 \times \text{age (years)}]$$

$$\text{Females: BEE (kcal)} = 655.1[9.6 \times \text{weight (kg)}] + [1.8 \times \text{height (cm)}] - [4.7 \times \text{age (years)}]$$

The BEE is then adjusted for the perceived degree of stress. Recent trends of providing fewer calories to seriously ill patients in order to prevent complications such as hyperglycemia, hypercapnia, and hepatic steatosis have been described as 'permissive underfeeding.' Trials have been conducted in obese hospitalized patients using hypocaloric regimens. These studies demonstrated that most patients achieved positive nitrogen balance and improved clinically with high-protein, hypocaloric PN formulations without experiencing significant adverse effects.

When an accurate assessment of energy needs is desired, indirect calorimetry may be considered. Such patients may include those who are otherwise difficult to assess or those who will require protracted nutrition support. A metabolic cart is used to measure oxygen consumption (VO_2) and carbon dioxide production (VCO_2). The modified Weir equation is used to estimate resting energy expenditure. The respiratory quotient (RQ) is determined as $RQ = VCO_2/VO_2$. The RQ gives an indication of net substrate oxidation, with $RQ > 1.0$ consistent with carbohydrate oxidation associated with overfeeding, and $RQ < 0.68$ consistent with lipid oxidation or starvation ketosis. Gas leaks or elevated FiO_2 requirements (>0.60%) are common limitations to this approach.

Protein Requirements

Determining appropriate protein goals is important in order to help maintain lean body mass and to promote positive nitrogen balance. Initial requirements are adjusted by a subjective assessment of the patient's degree of catabolism and an evaluation of renal and hepatic function (**Table 6**). Subsequently, the protein prescription should be adjusted based on clinical response and continued reevaluation. The likelihood of achieving nitrogen balance or improvement in visceral protein status will depend on the degree of ongoing inflammatory response. Reductions in protein doses may be warranted in patients with significant hepatic failure with encephalopathy. Reductions may also be indicated for renal insufficiency, depending on the severity of renal failure and whether dialysis is initiated. For the morbidly obese patient, ideal body weight should be used to estimate protein needs.

Dextrose Prescription

Dextrose in PN solutions generally provides 40–60% of total energy requirements. Hyperglycemia is a common complication of PN due to diabetes, medications, or stress response, so the dextrose load is often initiated below goal until tolerance is demonstrated. The maximum glucose utilization rate is 5–7 mg/kg/minute. Doses that exceed this may result in glucose intolerance or hepatic steatosis. Studies have demonstrated that aggressive blood glucose management is associated with fewer septic complications in critically ill patients.

Fat Requirements

Fat calories usually comprise 20–30% of total PN energy. Doses may represent 0.5–1.0 g/kg body weight/day in the PN regimen. Patients with preexisting lipid disorders should be evaluated for potential hypertriglyceridemia. Baseline triglyceride levels will aid the practitioner in safe determination of lipid dosing. Rapid piggyback lipid infusion with PN should be avoided in critically ill subjects.

Table 6 Estimation of protein needs

Condition	*Need (g/kg/day)*
Mild stress	1.0
Moderate stress	1.2–1.5
Severe stress	1.5–2
Acute renal failure, no dialysis	0.6
Hemodialysis/CVVHD	1.1–1.5
Peritoneal dialysis	1.2–1.5
Liver failure without encephalopathy	1.2–1.5
Liver failure with encephalopathy	0.4–0.6

Adapted from the National Advisory Group on Standards and Practice Guidelines for Parenteral Nutrition (1998) Safe practices for parenteral nutrition formulations. *Journal of Parenteral and Enteral Nutrition* **22**: 49–66.

Fluid Requirements

A patient's fluid intake and output should be considered when determining fluid requirements. Urinary losses, along with other losses such as diarrhea, nasogastric suction, emesis, fistula, or other drainage losses, can significantly increase the patient's need for additional fluid. In general, adult patients will need approximately 30 ml/kg body weight daily to meet volume requirements. Monitoring should include adequate urine output, skin turgor, and adequate mucous membrane hydration. Fluid overload in conditions such as compromised cardiac, hepatic, or renal function may dictate use of volume-concentrated formulations.

PN Implementation

Monitoring and Management

Appropriate monitoring of PN therapy is critical to ensure optimal nutrition therapy is achieved and to prevent complications of PN. Evolution of clinical course and patient condition may warrant changes in the frequency of tests and reevaluation of therapy. Additionally, as PN is transitioned to enteral feedings, tolerance should be monitored and PN should be weaned and discontinued.

PN tolerance should be carefully evaluated upon initiation of therapy. The managing practitioner should consider the patient's clinical condition and concurrent organ function, laboratory measurements, nutrition, and fluid status parameters. Laboratory measurements should include a complete metabolic profile and liver function tests at baseline and subsequent measurements after initiation of PN (**Table 7**). Electrolytes should be monitored daily until the patient is stable. Monitoring for acute care patients usually requires more frequent laboratory evaluations and more frequent changes to the PN formula than in long-term patients.

Evaluation of visceral proteins, such as albumin or prealbumin, has historically been an important part of nutrition assessment. These markers, however, may not adequately reflect an accurate picture of nutritional status or response to therapy because of other conditions, such as nephropathy, enteropathy, liver disease, or volume overload. Additionally, these visceral protein levels are often

Table 7 Suggested laboratory monitoring

Parameter	*Baseline*	*Initiation*	*Critically ill patients*	*Stable patients*
CBC with differential	Yes		Weekly	Weekly
PT, PTT	Yes		Weekly	Weekly
Electrolytes (Na, K, Cl, CO_2, Mg, Ca, PO_4, BUN, Cr)	Yes	Daily $\times 3$	Daily	1 or 2 times per week
Serum triglycerides	Yes	Day 1	Weekly	Weekly
Transferrin or prealbumin	Yes		Weekly	Weekly
Serum glucose	Yes	Daily $\times 3$	Daily	1 or 2 times per week
Capillary glucose		As needed	TID until consistently <200 mg/dl	As needed
Weight	Yes	Daily	Daily	2 or 3 times per week
Intake and output	Yes	Daily	Daily	Daily unless fluid status assessed by physical exam
ALT, AST, ALP, total bilirubin	Yes	Day 1	Weekly	Monthly
Nitrogen balance	As needed		As needed	As needed

Reprinted from Mirtallo JM. Introduction to parenteral nutrition. In: Gottschlich MM, ed. *The Science and Practice of Nutrition Support: A Case-Based Core Curriculum*. Dubuque, IA: KendallHunt Publishing Company; 2001: 221, with permission from the American Society for Parenteral and Enteral Nutrition (A.S.P.E.N.). A.S.P.E.N. does not endorse the use of this material in any form other than its entirety.

reduced by stress or injury due to a cytokine-mediated inflammatory response.

Because of significant protein binding, serum calcium levels should be evaluated in light of the patient's visceral protein status. The unbound calcium or 'ionized' calcium is physiologically active. Both bound and unbound calcium are included by standard serum calcium measurements. Laboratory testing for ionized calcium provides a more accurate depiction of calcium status in comparison to standard calcium samples.

Blood glucose levels should be carefully monitored throughout the course of PN infusion in order to detect and prevent hyperglycemia or hypoglycemia. Capillary blood glucose monitoring devices provide a convenient means of determining blood sugars. Blood capillary glucose levels should be obtained more frequently during the initial days of PN therapy and subsequently as needed for 'spot checks' or to verify glucose levels obtained by serum blood sampling. Insulin management may warrant a separate intravenous insulin infusion, subcutaneous coverage with sliding-scale insulin, or the addition of insulin as a component of PN. Because insulin needs are often acutely elevated in infection or stress, sliding-scale subcutaneous insulin or a separate insulin infusion may be used in combination with the addition of insulin to the PN.

Baseline triglyceride levels may be obtained prior to PN administration and periodically during the duration of therapy. Lipid doses should be reduced or held temporarily in adult patients with triglyceride levels >400 mg/dl. EFAD may develop after prolonged administration of PN without lipids; therefore, at least weekly or biweekly doses should be considered for all patients.

Complications of PN

Short-term complications of PN therapy may be divided into three classes: mechanical, infectious, and metabolic. Longer term complications can include overfeeding, hepatobiliary complications, and metabolic bone disease.

Central catheter placement can be associated with serious mechanical complications, including pneumothorax, arrhythmias, catheter-related thrombosis, and catheter occlusion. Radiologic confirmation of line placement is necessary before initiating PN therapy. Catheter occlusion is the most common mechanical complication and may require thrombolytic treatment or line replacement. Catheter flushing protocols should be carefully followed to reduce risk of occlusion.

Infection due to catheter-related sepsis is another serious complication of PN that is associated with appreciable morbidity and cost. Prudent and meticulous catheter care and sterile technique should be emphasized. Catheter infections comprise a significant percentage of all nosocomial infections. Fever and unexplained hyperglycemia may be potential warning signs of catheter-related sepsis. It is important to note that there are frequently no external signs of catheter infection visible at the insertion site. Aseptic technique in manipulating the central line and related administration lines can help prevent introduction of infectious sources such as endogenous skin flora or contamination of the catheter

hub. Appropriate methods and a sterile environment in compounding PN will reduce chances of contamination of the PN admixture during preparation. Treatment of catheter-related sepsis often includes access device removal and administration of appropriate antibiotic or antifungal therapy. With selected pathogens in patients with limited access options, salvage antimicrobial therapy may be considered to prevent the necessity of line removal.

The incidence of metabolic complications in PN patients is estimated to be 5–10%. Metabolic complications may include intolerance to fluid or macronutrients, or imbalances in electrolyte or vitamin and trace element homeostasis or function. Hyperglycemia is the most common complication associated with PN therapy.

Refeeding syndrome is a potential phenomenon of metabolic complications that may be observed when severely malnourished patients are re-fed in an overzealous manner. The rapid provision of macronutrients is associated with serum depletion and intracellular shifts of phosphorus, potassium, and magnesium as well as fluid retention and vitamin derangements. These abnormalities may result in the development of clinical sequellae such as arrhythmias, heart failure, respiratory failure, and death. Prevention of this phenomenon is achieved by identifying the patient at risk, repletion of electrolytes prior to initiation of nutrition support, and the slow advancement of PN with careful daily monitoring of electrolytes, including phosphorus and magnesium levels, as well as weights and fluid intake and output.

Electrolyte adjustments warrant close clinical evaluation. For example, hyponatremia can represent sodium deficiency or water excess. Electrolyte loss or shifts may occur from renal or gastrointestinal losses, hormonal imbalances, medication use, or acid–base disturbances. Accumulation of electrolytes may occur with fluid or acid–base shifts, renal insufficiency, or overzealous exogenous replacement. Generally, a consistent PN formula is recommended, with additional acute electrolyte replacements provided separately from the PN. Lower concentrations or even elimination of selected electrolytes from PN are often indicated in patients with renal failure.

Hepatobiliary complications, including steatosis and cholestasis, are associated with PN patients due to the lack of enteral stimulation and limited gastrointestinal motility. Cholestasis is universal in patients receiving PN for more than 6 weeks without enteral feedings. Transition to enteral or oral feedings will help prevent the potential development of gallstones associated with cholestasis. In adults, hepatic steatosis, or fatty liver, is generally associated with normal or mildly increased bilirubin levels and mild elevations in alkaline phosphatase and hepatic transaminases. In rare cases, hepatic steatosis may progress to steatohepatitis. In order to reduce the potential for hepatobiliary complications, the practitioner should attempt enteral feedings as soon as possible. Overfeeding with excessive lipid and dextrose loads should be avoided. The PN infusion can also be cycled to provide a rest period for the liver's macronutrient processing. Administration of a cholecystokinin–octapeptide may also help to reduce cholestasis.

Osteoporosis or osteomalacia may develop in long-term PN patients. Metabolic bone disease may develop due to underlying disease, inadequate intakes or malabsorption of calcium and vitamin D, corticosteroid therapy or other medications, and hypercalciuria. For selected long-term PN patients, treatment with bisphosphonates, calcium, and vitamin D should be considered to prevent the development of complications.

Home Parenteral Nutrition

Many patients are able to receive PN in the home setting. If the patient's medical condition is stable and careful patient selection has been employed, home parenteral nutrition (HPN) can be considered. The patient and caregiver must be able to be taught to use an infusion device, administer PN safely, and search for signs of infection, fluid issues, or other complications. Appropriate education should begin in the hospital and continue in the home setting. Psychosocial and socioeconomic issues, such as family support, private or government payer status, and patient emotional status, are important to consider when assessing the appropriateness of HPN.

Nocturnal administration of PN over 12 h can aid with patient mobility and quality of life and may also help to minimize hepatobiliary complications. Cycled PN infusion should generally begin in the inpatient hospital environment so tolerance of cycling can be safely evaluated. Since greater infusion rates are required for cycling, close monitoring during the transition phase is of particular importance for those with glucose or volume tolerance concerns.

Conclusions

PN offers a viable way to provide essential nutrients to individuals who are unable to use their gastrointestinal tracts effectively for a prolonged time period. Although this medical intervention is

associated with significant risks, for many patients it can be a life-saving or life-prolonging therapy. Practitioners who manage patients who receive PN warrant specialized nutrition support training. PN management is most effective through a multidisciplinary approach, utilizing the expertise of physicians, pharmacists, dietitians, nurses, case managers, and social workers.

See also: **Energy**: Requirements. **Fatty Acids**: Metabolism. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **Nutritional Support**: In the Home Setting; Infants and Children, Parenteral. **Protein**: Requirements and Role in Diet. **Supplementation**: Dietary Supplements.

Further Reading

American Gastroenterological Association (2001) AGA technical review on parenteral nutrition. *Gastroenterology* **121**: 970–1001.

American Society for Parenteral and Enteral Nutrition Board of Directors and the Clinical Guidelines Task Force (2002) Guidelines for the use of parenteral and enteral nutrition in adult and pediatric patients. *Journal of Parenteral and Enteral Nutrition* **26**(1 supplement): 1SA–138SA.

Flancbaum L, Choban PS, Sambucco S *et al.* (1999) Comparison of indirect calorimetry, the Fick method, and prediction equations in estimating the energy requirements of critically ill patients. *American Journal of Clinical Nutrition* **69**: 461–466.

Heyland DK, MacDonald S, Keefe L *et al.* (1998) Total parenteral nutrition in the critically ill patient: A meta-analysis. *Journal of the American Medical Association* **280**: 2013–2019.

Horattas MC, Trupiano J, Hopkins S *et al.* (2001) Changing concepts in long-term central venous access: Catheter selection and cost savings. *American Journal of Infection Control* **29**(1): 32–40.

Klein CJ, Stanek GS, and Wiles CE (1998) Overfeeding macronutrients to critically ill adults: Metabolic complications. *Journal of the American Dietetic Association*, vol. **98**: 795–805.

Klein S, Kinney J, Jeejeebhoy K *et al.* (1997) Nutrition support in clinical practice: Review of the published data and recommendations for future research directions. *Journal of Parenteral and Enteral Nutrition* **21**: 133–157.

McCowen K, Friel C, Sternberg J *et al.* (2000) Hypocaloric total parenteral nutrition: Effectiveness in prevention of hyperglycemia and infectious complications—A randomized clinical trial. *Critical Care Medicine* **28**: 3606–3611.

Mirtallo JM (2001) Introduction to parenteral nutrition. In: Gottschlich MM (ed.) *The Science and Practice of Nutrition Support: A Case-Based Core Curriculum*. Dubuque, IA: Kendall-Hunt.

Quigley EM, Marsch MN, Shaffer JL, and Markin RS (1993) Hepatobiliary complications of total parenteral nutrition. *Gastroenterology* **104**: 286–301.

Ranson JHC, Rifkin KM, Roses DF *et al.* (1974) Prognostic signs and the role of operative management in acute pancreatitis. *Surgical Gynecology & Obstetrics* **139**: 69–81.

Rombeau JL and Caldwell MD (eds.) (2001) *Parenteral Nutrition*, 3rd edn. Philadelphia: WB Saunders.

Shikora SA, Martindale RG, and Schwitzberg SD (eds.) (2002) *Nutritional Considerations in the Intensive Care Unit: Science, Rationale and Practice*. Dubuque, IA: Kendall-Hunt.

Solomon SM and Kirby DF (1990) The refeeding syndrome: A review. *Journal of Parenteral and Enteral Nutrition* **14**: 90–97.

Souba W (1997) Nutritional support. *New England Journal of Medicine* **336**: 41–48.

Task Force for the Revision of Safe Practices for Parenteral Nutrition (2004) Special report: Safe Practices for Parenteral Nutrition. *Journal of Parenteral and Enteral Nutrition* **28**: S39–S70.

Veterans Affairs Total Parenteral Nutrition Cooperative Study Group (1991) Perioperative total parenteral nutrition in surgical patients. *New England Journal of Medicine* **325**: 525–532.

Infants and Children, Parenteral

S Collier and C Lo, Children's Hospital, Boston, Harvard Medical School, and Harvard School of Public Health, Boston, MA, USA

Parenteral nutrition (PN) is a technique that allows provision of complete nutrient requirements intravenously, containing adequate amounts of energy, carbohydrate, protein, fat, minerals, and vitamins, while bypassing the gastrointestinal tract. It has allowed survival of many thousands of patients who cannot or will not eat or absorb enough to maintain their weight or nutritional balance because of disease or surgery. From its beginnings only approximately 30–40 years ago, it has expanded rapidly to become available for many patients, especially in the United States and then in Europe, but its expense and complications preclude wide availability in many countries. Before the availability of PN, as many as 30–50% of hospitalized patients had unrecognized malnutrition from chronic diseases and would remain for weeks without adequate nutrition to maintain weight or lean body mass, making them susceptible to infections and poor wound healing. Surprisingly, it has been difficult to demonstrate substantial reductions in morbidity or mortality with PN except in moderately or severely malnourished patients or those with long-term intestinal failure. It is difficult to estimate the exact impact, but a US registry, the Oley Foundation, enumerated 10 035 Medicare beneficiaries on home PN in 1992, giving a rough estimate of 40 000 patients on home PN in the United States. Approximately 15–20% of these patients were children.

One of the first attempts at PN was carried out by Sir Christopher Wren in 1656. He infused ale, opium, and beer intravenously into animals. Complete intravenous nutrition that we are most familiar with for patient support has been available for approximately 40 years. The research carried out by Dr. Stanley Dudrick and others allowed the support of the first pediatric patient on intravenous nutrition. The provision of intravenous nutrition was challenged by the development of several factors prior to its completed use in patient support, including catheter access, sterility of solutions, and the optimal form of each macro- and micronutrient.

Indications for Parenteral Nutrition

The often repeated adage continues to be true, "If the gut works, use it." However, there are many circumstances in which PN is necessary and life sustaining. The indications for use have not changed dramatically throughout the years since the development of PN. Congenital malformation of the intestine, specifically small bowel atresia, was the diagnosis the first time PN was used in the infant and young child. Congenital malformations of the gastrointestinal tract continue to be one of the leading reasons for its use. Other indications include severe malabsorption, intestinal dysmotility, other congenital defects, and patients with hematology–oncology diseases (**Table 1**).

Table 1 Conditions commonly requiring parenteral nutrition

Condition	*Examples/comments*
Surgical gastrointestinal disorders	Gastroschisis, omphalocele, tracheoesophageal fistula, intestinal atresias, meconium ileus, peritonitis, malrotation and volvulus, diaphragmatic hernia, prolonged postoperative ileus, Hirschsprung's disease, intestinal dysmotility
Short bowel syndrome	
Prematurity	
Congential heart disease	
Pancreatitis	
Gastrointestinal fistulas	
Bone marrow transplantation	
Acute intestinal disease	Antibiotic colitis, necrotizing enterocolitis, inflammatory bowel disease, chronic or secretory diarrhea
Hypermetabolic states	Burns, multiple trauma
Chronic idiopathic intestinal pseudo-obstruction	

Adapted from Hendricks KM, Duggan C, and Walker WA (eds.) (2000) *Manual of Pediatric Nutrition*, 3rd edn., London: BC Decker.

Dextrose

The primary source of energy during intravenous therapy is usually provided by dextrose (D-glucose). This is especially true in infants and children when higher energy requirements often necessitate glucose infusion rates of up to 15 mg/kg/min or more. Not until 1945 did Zimmerman report the first attempt at infusing intravenous solutions through a catheter placed in the superior vena cava. Experiments performed by Dudrick in beagle puppies advanced the glucose infusion solutions closer to what is utilized currently with hypertonic dextrose solutions. In current practice, hypertonic solutions are infused through a catheter with its tip centrally located in the superior vena cava or inferior vena cava. It continues to be the major energy component of intravenous support.

Initial doses of glucose should be approximately 5–7 mg carbohydrate/kg/min with incremental increases by 2–5 mg/kg/min. Frequent monitoring of blood glucose and urine for glucosuria is important to assess tolerance to increasing glucose infusion rates. It is important to avoid excessive carbohydrate intake to minimize complications from potential hyperglycemia with subsequent osmotic diuresis and over the long-term hepatic steatosis from increased fat synthesis that can occur with overfeeding. Hyperglycemia may ensue even without excess carbohydrate infusion in certain clinical situations, such as sepsis and renal failure, and also with the use of medications such as steroids. Glucose infusion rates should be decreased if hyperglycemia ensues; however, it may still be necessary to add insulin to control blood glucose to provide adequate support.

Protein

Another vital macronutrient that needed to be provided was protein. Initial experiments in the 1930s were done with plasma as the protein source, and investigators achieved positive nitrogen balance. In the early 1900s, research began on the development of protein hydrolysates and crystalline amino acids. Vitrum, a Swedish company, produced the first commercially available casein hydrolysate solution. It was developed by Arvid Wretlind, who hydrolyzed casein enzymatically and then dialyzed the mixture to remove large polypeptides. The crystalline amino acid solutions that we are more familiar with were first developed by Bansi in 1964. Wretlind went on to modify it further and eventually replaced the hydrolysates in the 1970s.

The development of amino acid solutions specifically for infants occurred in the early 1980s. These solutions provided conditionally required amino acids for the immature organ systems of premature infants and newborns. They were formulated based on the postprandial plasma amino acid levels of breast-fed infants. Special amino acid solutions for renal or liver failure are also available that have increased amounts of branched-chain amino acids. Studies on the solutions for liver failure have shown that they are probably beneficial in adult patients with encephalopathy. Glutamine is a much researched amino acid that could not initially be added to PN solutions due to shelf instability in liquid form. When added as a dipeptide, it has been found to be more stable. Not all studies have shown clear benefit to its addition in patients for gut adaptation or prevention of bacterial translocation.

The recommendation for initiation of protein is 1 g/kg/day and that for advancement is 1 g/kg/day to goal (**Table 2**). Blood urea nitrogen is monitored for tolerance to amino acid infusion.

Lipid Emulsions

Glucose was the only nonprotein source of energy until intravenous lipids were developed during the 1920s to the 1960s. The first emulsion available for clinical use was Lipomul, a cottonseed oil-based formulation. Because there were many adverse effects from its use, it was withdrawn from clinical use in the mid-1960s. After extensive testing, Wretlind developed Intralipid, a soybean-based emulsion, in 1961. It was well tolerated and is the most familiar intravenous fat emulsion currently available. It is available in 10, 20, and 30% solutions. The advantage of the 20 and 30% solutions over the 10% solution is the lower ratio of phospholipids to triglyceride, which minimizes the increase in plasma lipoprotein X levels. Lipid emulsions provide essential fatty acids in addition to a concentrated energy source, which is particularly advantageous for patients requiring fluid restriction. Trials are under way on the use of emulsions that contain a blend of long-chain fat with medium-chain fats and those with fish oil blends. Also, structured fat emulsions are being studied for clinical use. These specialized emulsions may have advantages in patients with liver disease and those with sepsis.

Lipid emulsions are usually initiated at 1 g/kg/day and advanced to 2 or 3 g/kg/day or 30–50% of total energy. Serum triglyceride levels are monitored for tolerance. Hypertriglyceridemia may occur in situations of stress, sepsis, and renal and liver insufficiency/failure. In addition, a number of medications can cause hypertriglyceridemia. In these situations, a reduction in fat infusion is warranted, usually by infusing over 18–20 h instead of 24 h.

A minimum of 3–5% of total energy requirement is necessary to meet essential fatty acid requirements. In infants with indirect hyperbilirubinemia, it may be prudent to lower intravenous fat infusion to avoid potential risk of kernicterus since free fatty acids may displace bilirubin from albumin binding sites.

Table 2 Pediatric parenteral nutritional requirements

	<2000 g	*0–4 years*	*5–18 years*
Energy (kcal/kg/day)	120	90–108	40–70
Protein (g/kg/day)	3–3.5	2.0–3.0	1–1.5
Fat (g/kg/day)	<3	<3	<2
Sodium (mEq/kg/day)	2–3	2–4	2–4
Potassium (mEq/kg/day)	2–3	2–4	2–4
Chloride (mEq/kg/day)	2–3	2–4	2–4
Calcium			
mEq/kg/day	3–4.5	2–3	0.5–2.5
mg/kg/day	60–90	40–60	10–50
Magnesium (mEq/kg/day)	0.35–0.6	0.25–0.5	0.25–0.5
Phosphate (mM/kg/day)	1.5–2.5	1–2	1–2
Zinc (μg/kg/day)	400	300	100
Selenium (μg/kg/day)	1–3	1–3	1–2
Trace elements (ml/l)	2	2	2
Multi vitamins (ml/day)	5	5	5–10

Micronutrients

To provide complete nutritional support, micronutrients, electrolytes, and minerals also need to be in the parenteral solution. The addition of adequate amounts of calcium and phosphate in one solution may be particularly problematic since precipitation may occur. Solubility guidelines are available that account for the brand and percentage of amino acids, which impact the pH of the solution. Compounding guidelines for the order of addition of calcium and phosphorus, amounts of other additives, and the temperature of the solution are other factors to optimize the solubility. Filters in the delivery system also help to minimize the risk of occlusion of the catheter if a solution should precipitate, especially with trimix solutions, in which lipids are mixed with the glucose/amino acid solution. Studies have evaluated the stability of the variety of nutrient components in trimix solutions.

Before the availability of vitamins and minerals, plasma levels of micronutrients decreased rapidly

while infusing only macronutrients. The only commercial preparation initially available was a trace element solution that had iron and iodide. The first commercial preparation of multivitamins for intravenous use, introduced in the 1960s, lacked folic acid, vitamins B_{12} and K, and biotin. It also had very high concentrations of vitamins A and D and thiamin. Because of the variability in practice, there was increased risk of toxicity to vitamins A and D and deficiencies of other vitamins. Recommendations were made for intravenous pediatric and adult intravenous preparations in 1975. By 1978, there was a commercial multivitamin preparation that met these recommendations. Current preparations contain all vitamins for which there are Dietary Reference Intake values, with the exception of choline. A recent Food and Drug Administration mandate requires the addition of vitamin K to all preparations. Differences between the pediatric and adult forms of MVI include amounts of B vitamins and vitamin D (**Table 3**). There is currently no multivitamin preparation specifically for the premature infant. Dosing recommendations of Pediatric MVI for this group are based on weight (one-third vial for <500 g, two-thirds vial for 500–1000 g, and full vial for more than 1000 g).

A few trace element deficiencies have been noted in patients receiving long-term PN support. The first case of chromium depletion was reported in 1977, that of selenium deficiency in 1979, and that of molybdenum in 1981. There are now many trace element solutions available with a variety of combinations of minerals and that are appropriate to meet the needs of premature infants through the adult population. They are also available as single elements to tailor a solution as necessary. Contamination of trace elements can occur in parenteral solutions. Aluminum is one element that has been under scrutiny by the Food and Drug Administration mandate to minimize the amount patients receive for safety issues. It was initially found in high concentrations in the casein hydrolysates and continues to be found in high concentrations in a variety of intravenous preparations. Over time, aluminum can deposit in the bone, interfering with bone calcium uptake, and deposition in the brain may impair neurological development.

Table 3 Comparison of parenteral multivitamin preparations for pediatric and adult populations

Vitamin	*MVI Pediatric (Mayne)*	*Infuvite (Baxter) and MVI Adult with vitamin K (Mayne)*
A (IU)	2300	3300
D (IU)	400	200
E (IU)	7	10
K (μg)	200	150
Ascorbic acid (mg)	80	200
Thiamine (mg)	1.2	6
Riboflavin (mg)	1.4	3.6
Niacin (mg)	17	40
Pantothenate (mg)	5	15
Pyridoxine (mg)	1	6
B_{12} (μg)	1	5
Biotin (μg)	20	60
Folate (μg)	140	600

Metabolic Complications

Liver Disease

Although PN may be life sustaining, long-term use may be detrimental to the liver. The severity of injury ranges from reversible transaminase elevations to severe cholestasis and cirrhosis, especially in infants with short bowel syndrome. It is not clear whether this is due mainly to a nutrient deficiency, toxicity, or some physiological process missing because of the lack of enteral feeding. Prevention and treatment strategies continue to include minimizing or preventing episodes of sepsis, providing enteral feedings, moderating energy intake to provide for adequate growth but not to overfeed, cycling parenteral nutrition infusion, reduction of copper and manganese, use of an amino acid solution developed for infants, treatment/prophylaxis for bacterial overgrowth, and the use of ursodeoxycholic acid. Another drug that has been studied but is not available for clinical use is cholecystokinin, which promotes gallbladder contraction. A recent and controversial recommendation is the adjustment of the dose of intravenous lipid emulsion to ≤ 1 g/kg/day. Intravenous lipid emulsions are a rich source of linoleic acid, an omega-6 polyunsaturated fatty acid, and may enhance production of the proinflammatory cytokines. Increased leukotriene B4 synthesis by the hepatic macrophages will draw additional polymorphonuclear leukocytes that intensify the inflammatory response to endotoxin by release of reactive oxygen species.

Bone Disease

The development of osteopenia is another complication that is common with long-term PN support. The reasons are multifactorial and include relative immobility, inability to provide adequate calcium and phosphorus with solubility limitations, and hypercalciuria. It has also been suggested that the dose of vitamin D in the multivitamin preparation

may contribute to bone disease. Excessive vitamin D may suppress parathyroid hormone secretion and directly cause bone resorption. Although aluminum is still present in some intravenous solutions, including calcium gluconate, vitamins, and trace elements, the amounts are much less than those seen with the casein hydrolysates and are not believed to be a significant contributor to the development of metabolic bone disease. Prevention and treatment strategies include maximizing calcium and phosphorus in PN solutions, especially for growing children; providing enteral supplementation of these minerals as feasible; and providing weight-bearing physical therapy if possible.

Micronutrient Deficiency and Excess

If a patient is entirely PN dependent, certain micronutrients need to be provided. Some PN solutions require the addition of carnitine and selenium (if not provided in multi-trace element solutions) and iron dextran (if the patient is not receiving transfusions). All serum levels should be monitored on a monthly basis or every 6–12 months if in the long-term phase of support. There may be other micronutrients not yet identified that may be deficient in the purified PN solution, which is another reason to begin enteral feedings as soon as feasible. Monitoring for excess losses is also important. For example, with increased stool/ostomy losses, the patient may require increased zinc in the PN solution (**Table 4**).

Excess micronutrients can be caused by contamination, such as the case with aluminum, or clearance. Copper and manganese can accumulate and become directly hepatotoxic since both elements depend on the biliary pathway for excretion. Therefore, in the presence of cholestasis, there will be increased intrahepatic accumulation. Manganese has also been reported to deposit in brain tissue, so copper and manganese levels should be monitored routinely.

Table 4 Suggested monitoring schedule for inpatients receiving parenteral nutrition

Parameter	*Daily*	*Weekly*[a]	*Periodically*[a]
Weight	x		
Fluid balance	x		
Vital signs	x		
Urine sugar	x		
Catheter site/function	x		
Laboratory (serum)			
Sodium		x	
Potassium		x	
Chloride		x	
Bicarbonate		x	
Glucose		x	
Urea nitrogen		x	
Creatinine		x	
Triglycerides		x	
Calcium		x	
Magnesium		x	
Phosphorus		x	
Albumin and/or prealbumin		x	
Transaminases		x	
Bilirubin		x	
Selenium			x
Copper			x
Zinc			x
Iron			x

[a]Or more often as necessitated by clinical course.
Adapted from Hendricks KM, Duggan C, and Walker WA (eds.) (2000) *Manual of Pediatric Nutrition*, 3rd edn., London: BC Decker.

Catheter Complications

Complications with central venous catheters most frequently include obstructions, infections, and occasional leakage and perforation. Although PN can be temporarily provided through peripheral intravenous catheters, the high osmolarity of intravenous glucose–electrolyte solutions often causes phlebitis and loss of access. Therefore, long-term access requires placement of a central venous catheter placed via the internal or external jugular vein or a subclavian vein. There is also increased placement of peripherally inserted catheters by a team of specially trained staff and/or an interventional radiologist. Tip position in the superior vena cava or right atrial junction should be verified radiographically to reduce complications from venous thrombosis or rare perforations. Central placement allows rapid dilution of hypertonic solutions in a large-diameter vein to minimize obstruction or thrombosis. Catheters for central venous access have been made of polyvinyl chloride, polyurethane, and silastic, often with a Teflon cuff to anchor the catheter subcutaneously. However, formation of a fibrin sheath is still common, often with a biofilm that may harbor infectious organisms and prevent penetration of antibiotics. Central catheter obstructions can often be visualized by ultrasound or inserting radio-opaque dye in the catheter. A thrombus can often be lysed with installation of a small bolus of tissue plasminogen activator. Long-term anticoagulation with coumadin, low-dose coumadin, or low-molecular-weight heparin has been advocated by some to avoid repeated catheter obstruction, venous thrombosis, superior vena cava obstruction, and potential pulmonary emboli.

Obstructions caused by precipitation of calcium phosphate salts or medications may be susceptible to installation of a small amount of dilute acid, and those due to fatty material may be dissolved with dilute ethanol. For long-term home parenteral use, some patients prefer the use of implantable ports, which can be accessed through the skin daily with a special needle. Recently, peripherally inserted central catheters have been used for periods up to 1 month or longer without requiring a surgical procedure.

Infections

Patients who require PN are often predisposed to infectious complications. The catheter hub is often the entry site, with skin flora such as *Staphylococcus epidermidis*, *Staphylococcus aureus*, or *Candida* being the most common organisms, along with gram-negative enteric bacteria possibly from bacterial translocation. Antibiotic treatment through the central line is often successful without replacement of the catheter using antibiotic combinations such as vancomycin and gentamicin or with an antibiotic lock.

Summary

Advancements in the technology, production, and manufacturing of intravenous solutions have progressed during the past 40 years. In addition to improvement in the solutions available for dextrose, amino acids, and lipid emulsions, there has been progress with the delivery systems, catheters, and sterile techniques for line and skin care to reduce overall complications.

Ongoing research and product development in areas associated with long-term PN support are vital for future patient management to be able to continue to provide optimal support with minimal risk for those patients for whom PN is life sustaining.

See also: **Aluminum**. **Amino Acids**: Chemistry and Classification. **Bone**. **Children**: Nutritional Requirements; Nutritional Problems. **Chromium**. **Copper**. **Infants**: Nutritional Requirements. **Lipids**: Chemistry and Classification. **Liver Disorders**. **Manganese**. **Protein**: Quality and Sources. **Selenium**. **Supplementation**: Developing Countries.

Further Reading

American Gastroenterological Association Clinical Practice Committee (2001) AGA technical review on parenteral nutrition. *Gastroenterology* **121**: 970–1001.

American Society of Parenteral and Enteral Nutrition (ASPEN) (1993) Guidelines for the use of parenteral and enteral nutrition in adult and pediatric patients. *Journal of Parenteral and Enteral Nutrition* **17**: 1–52.

Buchman A (2002) Total parenteral nutrition associated liver disease. *Journal of Parenteral and Enteral Nutrition* **26**: S43–S48.

Dudrick SJ (2003) Early developments and clinical applications of total parenteral nutrition. *Journal of Parenteral and Enteral Nutrition* **27**: 291–299.

Forchielli ML, Gura K, Anessi-Pessina E *et al.* (2000) Success rates and cost-effectiveness of antibiotic combinations for initial treatment of central venous line infections during total parenteral nutrition. *Journal of Parenteral and Enteral Nutrition* **24**: 119–125.

Kaufman SS (2002) Prevention of parenteral associated liver disease in children. *Pediatric Transplantation* **6**: 37–42.

Kinney JM (2000) *Clinical Nutrition Parenteral Nutrition*, 3rd edn., pp. 1–20. Philadelphia: WB Saunders.

Klein S, Kinney S, Jeejeebhoy K *et al.* (1997) Nutrition support in clinical practice: Review of published data and recommendations for future research directions. *Journal of Parenteral and Enteral Nutrition* **21**: 133–156.

Kleinman RD, Barness LA, and Finberg L (2003) History of parenteral nutrition and fluid therapy. *Pediatric Research* **54**: 762–772.

Oley Foundation (1994) *North American Home Parenteral and Enteral Nutrition Patient Registry, Annual Report with Outcomes Profiles 1985–1992*. Albany, NY: Oley Foundation.

Rombeau JL and Caldwell MD (1993) *Parenteral Nutrition*, 2nd edn. Philadelphia: WB Saunders.

Seidner DL (2002) Parenteral nutrition associated metabolic bone disease. *Journal of Parenteral and Enteral Nutrition* **26**: S37–S42.

Shils ME (2000) Recalling a 63 year nutrition odyssey. *Nutrition* **16**: 582–628.

Veterans Affairs Total Parenteral Nutrition Cooperative Study Group (1991) Perioperative total parenteral nutrition in surgical patients. New England Journal of Medicine **325**: 525–532.

Vinnars E and Wilmore D (2003) History of parenteral nutrition. *Journal of Parenteral and Enteral Nutrition* **27**: 225–232.

NUTRITIONAL SURVEILLANCE

Contents

Developed Countries

N R Sahyoun, University of Maryland, College Park, MD, USA

There has been increasing recognition in the past three decades that dietary intake patterns are associated with the development of chronic diseases and that improving nutritional intake may be a means of improving the well-being of the population and of reducing the cost of health care. Hence, nutritional surveillance has become an important topic on the health political agenda in many areas of the world and may become an integral part of surveillance systems. In addition, as new technology has enabled faster data collection and analysis, surveillance systems have evolved and become more sophisticated in the past decade.

This article defines nutrition surveillance and its usefulness. It also describes the types of surveillance activities and systems in place in the industrialized countries of Europe, the United States, Australia, New Zealand, and Canada. Emerging issues in nutrition and health are also discussed. Note that although the structures of surveillance systems in industrialized and developing countries have many similarities, the type of activities, target populations, and outcomes may be different. Thus, there is a need to describe the nutritional surveillance systems in place in industrialized countries separately.

Nutrition Surveillance and Its Usefulness

Nutrition surveillance is a system established to continuously monitor the dietary intake and nutritional status of a population or selected population groups using a variety of data collection methods whose ultimate goal is to lead to policy formulation and action planning. The term 'nutrition monitoring' is often used in addition to or interchangeably with 'nutrition surveillance' and is defined as surveillance that is carried out on selected individuals. In this article, the term nutrition surveillance is used to include all data collection methods that are described.

The information obtained through nutrition surveillance is used for three broad purposes: policy development, nutrition research, and monitoring. As **Figure 1** illustrates, there are strong interrelationships between these three purposes. Specifically, the information generated by nutrition surveillance activities is used to describe the nutritional status of the population and identify population groups at high nutrition risk. Programs are then targeted to those in need. The efficacy of the programs is assessed and nutrition policy developed. Trends in health status and food intake are monitored and food supply needs are estimated. Also, linkages between food consumption, nutritional status, and health status are examined. For example, normative data collected from surveys in the United States have been used to develop new growth charts, released in 2000, to monitor nutritional status and health of children. Similarly, the World Health Organization, using international data, is also in the process of developing new international growth charts. Monitoring trends in child

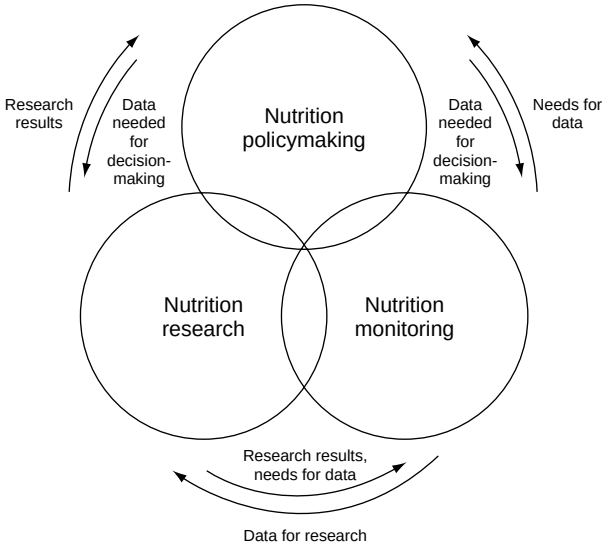


Figure 1 Relationships among nutrition policymaking, nutrition research, and nutrition monitoring. (From the US Department of Health and Human Services/US Department of Agriculture (1993) Ten-year comprehensive plan for the National Nutrition Monitoring and Related Research Program. *Federal Register* **58**: 32752–32806.)

growth helps to identify populations in need, evaluate nutritional and health interventions, and raise political awareness of nutritional problems.

Food fortification is another example of an interaction between monitoring, research, and policy, and it highlights the importance of nutrition surveillance systems. Folic acid supplementation, in addition to normal dietary folate intake, was recognized to significantly reduce the incidence of neural tube defects, one of the most common birth defects. This led to the mandatory fortification in 1998 of enriched grain products with folic acid in Canada and the United States. Survey data were used to determine the amount of folic acid that needed to be added to the food supply to provide beneficial effects without the harmful effect of potentially excessive intake. Surveys that examined the impact of folic acid supplementation showed a 19% reduction in neural tube defects. Continuous monitoring is essential to ensure that the added intake of folic acid does not have longer term negative impacts on the different population groups.

Nutrition Surveillance Systems

A nutrition surveillance system ideally collects information on all components in the relationship between food and health. This includes collecting data on food production, food supply and availability for consumption (national and household), food consumption patterns, dietary composition of foods, nutrient intake, nutrient utilization, and nutritional status. It also includes variables that may influence these processes, such as food culture, food security, lifestyle, knowledge, attitude and behavior toward food, and sociodemographic factors. **Figure 2** depicts the relationship of food to health outcome and illustrates the various levels of influence. A nutrition surveillance system would ideally obtain nutrition information along that continuum from food supply to health. At the core of a nutrition surveillance system is the collection of dietary intake patterns because they provide a basis for nutritional risk assessment. These dietary data include information obtained from the national food supply and from food consumption by households and by individuals. Each type of data collected corresponds to a different stage in the food distribution chain and is obtained by different methods. Some of these methods are described next.

Dietary Data Collection Methods

Food supply data Food supply data provide information on the type and amount of food available for human consumption to the country as a whole. The most common method of measuring this available food is through the use of food balance sheets. It is a method of indirectly estimating the amounts of food consumed by a country's population at a certain time. It provides data on food disappearance rather than on actual food consumption. It is calculated by using beginning and ending inventories, and the

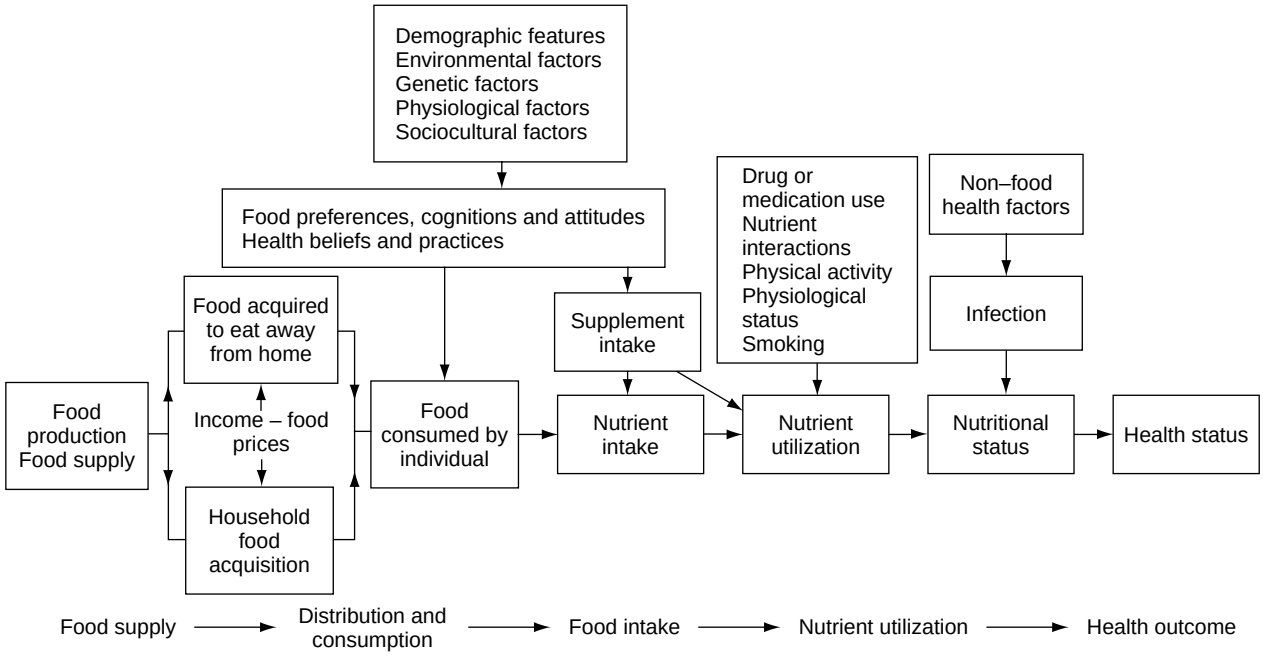


Figure 2 A conceptual model of the relationships of food to health. (Life Sciences Research Office, Federation of American Societies for Experimental Biology. Nutrition Monitoring in the United States – An Updated Report on Nutrition Monitoring. Washington, D.C.: U.S. Government Printing Office, 1989.)

difference is the amount of food consumed. The beginning inventory includes data on food production, imports and exports, and adjustments for non-human food consumption and an estimate of waste. Results are then converted to per capita basis food availability. To obtain the mean per capita annual consumption of food, total disappearance of food is divided by the country's population. Quantities of each food commodity are then multiplied by the appropriate nutrient values, and the results are expressed either in kilograms per year or in grams per day of individual food commodities and nutrient availability per person. No deduction is made for household food waste or the loss of nutrients in food preparation. Food is not distributed equally among a country's population, so this method only indicates the amount of food that leaves the food distribution system and is available for human consumption to the country as a whole. The best and most useful use of this data is to compare available food supply within and between countries and to monitor trends and forecast food consumption patterns over time. For example, **Table 1** illustrates the use of food balance sheets data and shows the worldwide and regional increase in the average supply of dietary fat from 1967–1969 to 1997–1999. Caution is needed, however, in comparing data between countries because food balance sheets, although compiled in a similar manner, may differ in food groupings, level of processing of commodities, and nutrient conversion of factors.

The Food Agriculture Organization (FAO) has published international food balance sheets yearly since 1949 and also covering the period 1934–1948. FAO food balance sheets are compiled from data supplied by approximately 200 countries. FAO uses the United Nations Population Division mid-year estimates of population size for its food balance sheet data and to calculate per capita values. International food balance sheets have also been compiled and published by the Organization for Economic Cooperation and Development on 23 countries (18 European countries, Australia, Canada, Japan, New Zealand, and the United States), whereas the Commission of the European Communities (EURO-Stat) publishes data for its 12 member countries. In addition, individual countries publish their own data. For example, Canadian data on per capita food availability have been prepared annually by Statistics Canada since 1979. In the United States, annual estimates of commodity foods consumed by the civilian population are calculated and have been reported since 1909.

Food consumption by households Methods have been devised to obtain information on the availability of food and beverages for consumption by a household, family group, or institution. The basic concept is to collect the types and amounts of food that enter a household and that are available for consumption. These methods vary by the level of respondent burden and extent of recall expected, and there are four main ones: the food account, list-recall, inventory, and food record methods. Generally, the information is collected for a period of 7 days. In the food account method, the head of household records daily all types and quantities of food that enter the household within a 7-day period. In the list-recall method, the head of household is interviewed and must recall the foods used by the household on an 'as purchased' basis. The inventory and food record methods require daily recording of food acquired and of changes in the food inventory and also detailed weighing and measuring of food, placing a heavy burden on the respondent.

These household food consumption methods do not provide actual food intake by individuals within the household. Instead, individual food consumption and nutrient intake are calculated by dividing household food consumption by the number of members in the household regardless of age or sex. This information is then reported in terms of household income level, family size, and other general characteristics of interest. Several countries have used household methods for their national

Table 1 Supply of dietary fat by region

Region	*Supply of dietary fat (g per capita per day)*			
	1967–1969	*1977–1979*	*1987–1989*	*1997–1999*
World	53	57	67	73
North America	117	125	138	143
European Community	117	128	143	148
Oceania	102	102	113	113

Adapted from Food and Agriculture Organization of the United Nations (2003) *Diet, Nutrition and the Prevention of Chronic Diseases. Report of a Joint WHO/FAO Expert Consultation*, WHO Technical Report Series No. 916. Geneva: World Health Organization.

food consumption surveys. In Canada, household food consumption surveys were conducted at approximately 4-year intervals and then yearly since 1997. The United Kingdom originally used the household inventory method but in 1950 changed methods because it had too high of a respondent burden. In the United States, nationwide household food surveys started in the 1930s and used the household food record method, but in the 1980s household food consumption surveys collected food intake of individual household members in addition to household food expenditure. In Europe, an ongoing project of the European Union is to harmonize dietary exposure from household surveys to improve comparisons of these data between countries.

Food consumption by individuals Collecting information on dietary intake of individuals provides a level of detail that allows the exploration of relationships between dietary intake, nutritional status, and health outcomes. It also provides information on intake distribution and patterns by age, gender, and other well-defined criteria, thereby identifying population groups at risk. Individual dietary intake data is collected in a variety of ways. The three most common methods are food record, 24-h recall, and food frequency questionnaires (FFQs). Food records are used to measure dietary intake over a single time period, usually 3–7 days. Respondents are asked to measure and weigh their food intake. Although this method provides details on the amount and kind of food consumed, it places a heavy burden on respondents, requiring motivated, trained, and literate individuals. The 24-h dietary recall also requires a trained interviewer who asks and probes respondents on the kind and amount of all food and drink consumed during the previous day. This method is used to monitor group mean intakes in the population. It is currently recommended that a second recall be collected to better estimate the population distribution of usual intake of nutrients and to correct for reporting error. Methods of collecting this data have become progressively more sophisticated as computerized systems have been developed to include standardized probes and multiple passes of intake over the day to prompt recall. These innovations have led to improved estimates of nutrient intake. FFQs are most often self-administered instruments in which respondents are presented with a list of food items and asked to report usual frequency of consumption over a specific time period (usually 1 year). FFQs are designed to obtain data regarding usual intake, are less costly to administer and code than recalls or records, and vary in the number of foods included in the food list. FFQs are semiquantitative and lack the detail of records or recalls.

Several variations of these methods of individual data collection have been devised. However, each of these dietary assessments provides advantages and limitations, and their applicability depends on the setting and the purpose of the data collected. Collecting individual dietary intake patterns is more time-consuming than collecting information on the food supply or household intake; however, most industrialized countries now collect these data or are making plans to do so. **Table 2** presents some of the surveys conducted by various industrialized countries and the method of dietary data collection selected by each country. The methods are quite varied, which limits comparability across surveys.

Examples of Nutrition Surveillance Activities

The 1990s saw a proliferation of nutrition surveys as local governments and international bodies such as the European Commission, the World Health Organization (WHO) and the International Union of Nutritional Scientists called for the establishment of a nutrition surveillance system. Most countries in Europe have a health interview survey, and some countries have a health examination survey as well. The dietary information gathered from these surveys varied considerably, from including one question on diet to including a FFQ. A few countries even measured nutritional biomarkers, whereas others did not collect any dietary intake data at all. Again, these methodological differences limit comparability of the results.

Most of the European Union member states have or are establishing a nutritional policy as an independent field of multidisciplinary research. The European Food Consumption Survey Method (EFCOSUM) was created as a project of the European Union Programme on Health Monitoring. One of its functions is to define a method for monitoring food consumption in nationally representative samples of all age–sex categories in Europe in a comparable way. It made recommendations about the best and most cost-effective data collection methods to use and the minimum variables needed to assess the nutritional status of populations. In addition, another objective of EFCOSUM is to indicate how to make existing food consumption data comparable and available to the health monitoring system. Also, WHO is stimulating regional and international networking and is strengthening community-based activities to prevent major

Table 2 Nationwide food consumption surveys with individual-based dietary intake data

Country	*Year*	*Survey*	*Population (ages in years)*	*Sample size*	*Dietary method*	*Other information*[a]
Australia	1983	National Dietary Survey of Adults	25–64	6295	24 h Rcl	A, BC, CE, MH
	1985	National Dietary Survey of Schoolchildren	10–15	5224	1d FR	A, BC, BP
	1995	National Nutrition Survey	2+	13 858	24 h Rcl (2nd 24 h Rcl from subsample) FFQ	A, BP
Canada	1970–1972	Nutrition Canada	0–65+	12 795	24 h Rcl, FFQ	A, BC, MH
	2004	National Population Health Survey	12+			
Europe						
Austria	1991–1994	Austrian Study on Nutritional Status (ASNS)	6–18	2173	7d FR	
	1993–1997	ASNS	19–65	2065	24 h Rcl, DH	A
	1998–2002	ASNS	19–60 55+	2580 645	24 h Rcl	A
Belgium	1980–1985	Belgium Interuniversity Research on Nutrition and Health	25–74	10 971	1d FR (DH in subsample)	A, BC, MH
	2003–2005	National Food Consumption Survey	All	3200 planned	24 h Rcl, FFQ	Fieldwork in 2004
Denmark	1985	Dietary Habits in Denmark	15–80	2442	DH	A
	1995	Danskernes Kostvaner	1–80	3098	7d FR	
	2000–2002	National Continuous Dietary Survey	4–75	1500 (2000) 1500 (2001) 1000 (2002)	7d FR	
Finland	1992	Dietary survey of Finnish adults (FINDIET 1992)	25–64	1861	3d FR	
	1997	Dietary survey of Finnish adults (FINDIET 1997)	25–64 65–74	2862 290	24 h Rcl	
France	1993–1994	Etudes Nationale des Consommations Alimentaires	2–85	1500	7d FR	A
	1998–1999	Individual National Food Consumption Survey	3–14 15+	1018 1985	7d FR	
Germany	1985–1989	National Nutrition Survey in Former West Germany	4–65+	24 632	7d FR KN, ATT, BH	A, BC
	1991–1992	National Nutrition Survey in East Germany	18–79	1897	DH	
	1998	German Nutrition Survey	18–79	4030	DH (4-week recall and FFQ)	
Ireland	1990	Irish National Nutrition Survey	10–65+	1214	DH	
	1997–1999	North–South Food Consumption Survey	18–64	1379	7d FR, ATT	A (self-reported)
	2003–2005	National Children's Food Survey	5–12	600		
Italy	1994–1996	INN-CA 1994–96	0–94	2734	7d FR	

Continued

Table 2 Continued

Country	*Year*	*Survey*	*Population (ages in years)*	*Sample size*	*Dietary method*	*Other information*[a]
The Netherlands	1987–1988	The Dutch National Food Consumption Survey (DNFCS-1)	1–85	5898	2d FR	A (self-reported)
	1992	The Dutch National Food Consumption Survey (DNFCS-2)	1–92	6218	2d FR	A (self-reported)
	1997–1998	The Dutch National Food Consumption Survey (DNFCS-3)	1–97	6250	2d FR	A (self-reported)
Norway	1993	National Dietary Survey	13 18	1705 1564	FFQ ATT, BH	A (self-reported)
	1993–1994	National Dietary Survey among Adults NORKOST	16–79	3144	FFQ, ATT, BH	A (self-reported)
	1997	National Dietary Survey among Adults NORKOST	16–79	2672	FFQ, ATT	A (self-reported)
	1999	National Dietary Survey	6 and 12 months, 2 years	2400 2010	FFQ	
Portugal	1980	Portuguese Food Consumption Survey	1–65+	13 080	1d FR, 24 h Rcl, FFQ	A, BC,CE, MH
Sweden	1989	Household Food Survey, HULK	1–74	2036	7d FR	A (self-reported)
	1997–1998	Riksmaten	18–74	1215	7d FR	A (self-reported)
United Kingdom	1986–1987	The Dietary and Nutritional Survey of British Adults	16–64	2197	7d FR	A, BC, CE, BP
	1992–1993	National Diet and Nutritional Survey (NDNS)	1.5–4.5	1675	4d FR	
	1994–1995	NDNS	65+	1687	4d FR	A, BC, CE, BP
	1997	NDNS	4–18	1701	7d FR	
	2000–2001	NDNS	19–64	2000	7d FR, BH	A, BC, CE, BP
	2003–2005	Low Income Diet and Nutrition Survey		2000	7d FR	Fieldwork in 2003
New Zealand	1977	National Diet Survey	20–74	1938	24 h Rcl	A
	1989	Life in New Zealand Survey	15+	1702	24 h Rcl, FFQ, ATT, BH	A, BC, CE
	1997	National Nutrition Survey	15+	4636	24 h Rcl, FFQ, KN, ATT, BH	A, BC, CE
	2002	Children's Nutrition Survey	5–14	3200	24 h Rcl, FFQ, BH	A, BC
United States	1970–1974	National Health and Nutrition Examination Survey (NHANES I)	1–74	20 749	24 h Rcl, FFQ	A, BC, CE, MH
	1976–1980	NHANES II	1–74	20 322	24 h Rcl, FFQ	A, BC, CE, MH
	1988–1994	NHANES III	2 month+	33 994	24 h Rcl, FFQ	A, BC, CE, MH
	1999+	National Health and Nutrition Survey	2 month+	9965 (1999–2000) 5500 (2001) 5000 (planned/year)	24 h Rcl, FFQ	A, BC, CE, MH Continuous data collection
	1992	NHANES I Epidemiologic Follow-Up Study	25–74	9281		Follow-up interviews of NHANES I participants in 1982, 1986, 1987

Continued

Table 2 Continued

Country	Year	Survey	Population (ages in years)	Sample size	Dietary method	Other information[a]
	1992	NHANES II Mortality Follow-Up Study				Mortality Follow-up of NHANES II participants
	1977–1978	Nationwide Food Consumption Survey		30 467	24 h Rcl, 2d FR	
	1987–1988	(NFSC)		25 100		
	1985–1986 (annual)	Continuing Survey of Food Intakes by Individuals (CSFII)	19–50 F 1–5 years 19–50 M	6400 3200 1100	24 h Rcl, 2d FR	
	1989–1991 (annual)	CSFII	All	15 192	24 h Rcl, 2d FR (subsample) KN, ATT, BH (subsample)	
	1994–1996 (annual)	CSFII	All	16 103	24 h Rcl, 2d FR (subsample) KN, ATT, BH (subsample)	
	1998		0–9	5559		

[a]Information other than sociodemographic. 24 h Rcl, 24-h dietary recall; 1d FR, 1-day food record; FFQ, food frequency questionnaire; DH, dietary history; KN, dietary knowledge; ATT, dietary attitude; BH, dietary behavior; A, anthropometry; BC, biochemical tests; BP, blood pressure; CE, clinical exam; MH, medical history; F, Female; M, Male.

noncommunicable diseases (NCDs) or chronic conditions. An international assessment of national capacity for the prevention and control of these diseases conducted in 2001 indicated that many countries lack policies to deal with the prevention of chronic diseases and lack legislation for food and nutrition.

The United States has the most extensive and comprehensive nutrition surveillance system in the world. Food consumption surveys were initiated in the 1930s and the surveillance system has expanded since then to include many cross-sectional and longitudinal surveys and surveillance systems. Due to the large number of surveys conducted in the United States, only a partial list is presented in **Table 2**. In 1990, the National Nutrition Monitoring and Related Research Program was established by the US Congress to strengthen food and nutrition data collection efforts via a 10-year plan. One of the outcomes of this effort was the establishment in 1999 of a continuous nutrition survey, the number one source of nutrition status information on the US population, which aims to collect a representative sample of 5000 individuals yearly. Similarly, in 1992, Australia launched a national food and nutrition policy for ongoing monitoring and surveillance of the food and nutrition system, which was established in 1998 and aimed to consolidate and strengthen data collection efforts. In Canada, the Canadian Community Health Survey began in 2000 to provide timely cross-sectional estimates of health determinants, health status, and the health system in a 2-year collection cycle. A nutrition component of the survey was implemented in 2004. The National Nutrition Survey of Japan has been conducted annually since 1946. Recently, Japan modernized the handling and processing of the data, which will allow for more rapid and greater data gathering capabilities and more accurate and timely reporting.

The assessment of nutritional status not only includes collecting dietary intake but also anthropometric measures, biochemical tests, and clinical examination. The measures of nutritional status collected in the different surveys vary considerably, as shown in **Table 2**. The simplest and most common anthropometric measures of nutrition status are height and weight, which are used to calculate body mass index, a widely accepted measure of overweight and obesity. Waist circumference and hip-to-waist ratio are also frequently measured to obtain an estimate of body fat distribution. Alternatively, a number of surveys have collected self-reported height and weight measures. Also, biological samples such as blood, urine, saliva, and hair have been collected, particularly in US surveys, as biomarkers of dietary intake to validate dietary data collection instruments, to relate to environmental exposure, and to study diet–health relationships. The selection of

biomarkers for inclusion in a survey is subject to budgetary constraints and survey logistics as well as health and methodological priorities determined by each country. For example, Germany has analyzed blood for ferritin, minerals, and certain vitamins such as folate and vitamin B_{12}, whereas the United Kingdom has analyzed blood for folate and vitamins D and C. In an effort to prioritize and standardize data collection, EFCOSUM has recommended at a minimum the analysis of biomarkers for folate, vitamin D, iron, iodine, and sodium. Finally, the most common clinical measure incorporated in surveys is that for blood pressure. Additionally, self-reported or physician-reported medical conditions are frequently collected from survey participants. In the United States, a wide array of clinical exams are conducted as part of the National Health and Nutrition Examination Survey, including dental health and vision.

Emerging Nutrition and Health Issues

Several emerging health issues that are related to dietary intake and lifestyle choices have made it essential to track eating habits and nutritional status over time. For example, obesity is an escalating epidemic through the world among both children and adults and a major concern because of its health consequences. Obesity has been linked to an array of health disorders, such as type 2 diabetes, cardiovascular disease, and disability. In 1995, WHO estimated that there were approximately 200 million obese adults worldwide and 18 million children younger than 5 years old classified as overweight. As of 2000, the number of obese adults had increased to more than 300 million. In the United States, 64.5% of adults were classified as overweight in 1999–2000, up from 46.0% in 1976–1980. Also, the percentage of overweight children in the United States (aged 5–14 years) has doubled in the same period from 15 to 32%. In England, the prevalence of obesity has doubled since 1980. Australia, New Zealand, Canada, and European countries have all reported an increase in the proportion of obese adults and children. WHO has begun to formulate a Global Strategy on Diet, Physical Activity, and Health under a 2002 mandate from the World Health Assembly. The overall goal of the strategy is to improve public health through healthy eating and physical activity.

Another issue of emerging international importance is that both the number and the proportion of people 60 years of age or older are increasing in almost all areas of the world, and these worldwide trends are expected to continue. In 2002, there were an estimated 605 million older people in the world. **Table 3** shows the countries with the highest percentage of the population older than 60 years of age. As a consequence of this demographic change, NCDs have been estimated to account for approximately 60% of global deaths and 45% of the global burden of disease. Attention to this demographic change has resulted in changes in health policies in order to help the population achieve healthy and active aging. WHO has developed a policy framework that focuses on preventing and reducing the burden of disabilities and reducing the risk factors associated with NCDs. These policies include healthy eating and physical activity.

Table 3 Countries with more than 10 million inhabitants in 2002 with the highest percentage of people older than age 60 years and projections for 2025

2002		*2025*	
Country	*%*	*Country*	*%*
Italy	24.5	Japan	35.1
Japan	24.3	Italy	34.0
Germany	24.0	Germany	33.2
Greece	23.9	Greece	31.6
Belgium	22.3	Spain	31.4
Spain	22.1	Belgium	31.2
Portugal	21.1	United Kingdom	29.4
United Kingdom	20.8	Netherlands	29.4
Ukraine	20.7	France	28.7
France	20.5	Canada	27.9

Data from the United Nations (www.who.int/hpr/ageing/ActiveAgeingPolicyFrame.pdf).

Research has consistently demonstrated that sufficient daily intake of fruit and vegetables could help prevent major NCDs, such as cardiovascular diseases, type 2 diabetes, obesity, and certain cancers. According to the 2002 World Health Report, up to 2.7 million lives could potentially be saved each year if fruit and vegetable consumption were increased. However, surveys conducted in Europe and the United States have indicated that the consumption of fruit and vegetables is lower than recommended for health. According to Kraisid Tontisirin, the director of FAO's Food and Nutrition Division, "FAO faces the challenge to increase worldwide awareness of the health benefits of increased fruits and vegetable consumption. To effectively promote more consumption of fruit and vegetables, prevailing diets need to be more systematically assessed for their nutrition and health implications."

These emerging issues linking nutrition and health outcomes reinforce the importance of developing and maintaining a nutrition surveillance system. However, direct and indirect methods of dietary data collection that can easily be applied in the field need to be developed further.

See also: **Dietary Intake Measurement**: Methodology; Validation. **Dietary Surveys**. **Folic Acid**. **Food Fortification**: Developed Countries; Developing Countries. **Nutritional Surveillance**: Developing Countries. **Pregnancy**: Prevention of Neural Tube Defects. **World Health Organization**.

Further Reading

Australian Centre for International and Tropical Health and Nutrition (2000) *Plan for the Development and Management of a National Food and Nutrition Monitoring and Surveillance System*, Australian Centre for International and Tropical Health and Nutrition, The University of Queensland, Herston, Australia. Available at ftp://www.sph.uq.edu.au/pdf.ftp/P1_0802_Long.pdf.

Briefel RR (2001) Nutrition monitoring in the United States. In: Bowman BA and Russell RM (eds.) *Present Knowledge in Nutrition*, pp. 617–635. Washington, DC: ILSI Press.

Canadian Community Health Survey (accessed 2004) www.statcan.ca/english/sdds/3226.htm.

EFCOSUM Group (2002) European Food Consumption Survey Method. *European Journal of Clinical Nutrition* **56**(supplement 2): S1–S94.

Food and Agriculture Organization of the United Nations (2003) *Diet, Nutrition and the Prevention of Chronic Diseases. Report of a Joint WHO/FAO Expert Consultation,* WHO Technical Report Series 916. Geneva: World Health Organization.

Margetts BM and Nelson M (1998) *Design Concepts in Nutritional Epidemiology*. New York: Oxford University Press.

New Zealand Ministry of Health (2003) *Food and Nutrition Monitoring in New Zealand*. Wellington, New Zealand: Ministry of Health.

U.S. Department of Health and Human Services/U.S. Department of Agriculture (1993) Relationship among nutrition policymaking, nutrition research, and nutrition monitoring. Ten-year comprehensive plan for the National Nutrition Monitoring and Related Research Program. *Federal Register* **58**: 32752–32806.

Woteki CE, Briefel RR, Klein CJ *et al.* (2004) Nutrition monitoring: Summary of a statement from an American Society for Nutritional Sciences Working Group. *Journal of Nutrition* **132**: 3782–3783. Data supplement available at www.nutrition. org/cgi/data/132/12/3782/DC1/1.

Developing Countries

L M Neufeld and L Tolentino, National Institute of Public Health, Cuernavaca, Mexico

Nutritional surveillance was defined in the previous article as a system established to continuously monitor the dietary intake and nutritional status of a population or selected population groups using a variety of data collection methods, with the ultimate goal of having a direct impact on actions to improve the situation. The challenges to meet this goal in developing countries are many, and they differ from those of more industrialized countries for a number of reasons. First, in most developing countries the prevalence of problems related to nutritional deficiency is higher than in industrialized countries and the prevalence varies greatly within and between regions and countries. Second, continuous national monitoring (e.g., through the health care system) is not well established in many countries and resources in many countries are scarce. Finally, during the past decade there has been a dramatic increase in the prevalence of overweight and obesity and their related morbidities. Together, the problems related to under- and overnutrition present unique challenges to national and international policymakers and heighten the need for nutritional surveillance systems that are able to provide useful information to policymakers.

To date, the major shortfall of nutritional surveillance has been the link between the data collected and its use in policy and programs, particularly in developing countries, where the need for nutrition interventions is great. Health and nutrition policies and programs should use information from nutritional surveillance systems to identify needs of specific populations within regions and countries and to help design appropriate interventions that address the relevant causes of these problems. This implies an open communication between those involved in data collection and those who would ultimately use the information. Unfortunately, the number of concrete examples in which this link has resulted in nutritional surveillance information being directly used to influence policy is still limited.

The responsibility for making nutritional surveillance action-oriented lies with all parties involved—donors, agencies or researchers involved in data collection and analysis, and policymakers. In many developing countries, the lack of existing information systems and limited local resources implies that external funds, often from donor agencies, will be required for surveillance activities. At all stages of planning, those responsible for data collection should interact directly with policymakers to ensure that the information is collected, analyzed, and presented in a way that is meaningful to them. Once data on the nutrition situation become available, policymakers should seek technical assistance from experts in the field to assist with the design of interventions with high potential for impact. Evaluation of policy and programs is essential to complete the cycle and permit new assessments and analyses based on these outcomes. This again implies the need for external funds in many cases. Researchers or national or

international agencies may need to become advocates to promote dialogue with policymakers and to convince donors of the importance of this process.

The Nature of Nutritional Surveillance Data in Developing Countries

Information collected as part of a nutritional surveillance system should include not only documentation of the nutritional problems but also an analysis of their direct and indirect causes. A conceptual framework, such as that of UNICEF, for understanding the causes of malnutrition should be used to determine the types of information that are needed. The exact information needed therefore may be context specific, depending on what preexisting information is available. The following are examples: (1) If recent national data show an adequate national food supply, but a high prevalence of malnutrition in children younger than 5 years of age still exists in the country, information related to household food security, individual food consumption, as well as other causes of childhood malnutrition such as infections may be needed; and (2) if the prevalence of obesity has increased recently in a country, information on dietary intake and physical activity patterns will be needed to understand the causes of this increase and to design appropriate interventions.

Details of different methods to collect data at the national, household, and individual level are described in the previous article and will not be reviewed here. Rather, this article focuses on some of the specific strengths, limitations, and applications of each type of data as they apply to nutritional surveillance in developing countries and describes some additional methods that have been adapted for use in developing countries.

National Food Supply Data

The Food and Agriculture Organization (FAO) compiles and monitors food supply data for many developing countries. The estimates are typically based on food balance sheets supplied from each country's national records. The information is usually converted to per capita food availability and is presented for developing countries as a whole, by region, subregion, and for more than 100 individual countries. Information is also available for many countries (e.g., in Latin America—Mexico, Brazil, Argentina, and Chile) from national statistical institutes as well as regional statistical organizations (e.g., the Council for Statistics in Latin America). Much of this information can be accessed free of charge through local Internet sites.

Food supply data are essential to make comparisons between and across regions and to monitor trends. For example, according to the FAO food supply data, approximately 10% of the worlds' population now lives in countries where the food supply is low (<2200 kcal/person/day). This is down from 57% in the mid-1960s. Nonetheless, according to 2001 FAO data, there are still 30 countries with low food supply. Ideally, this type of information should be used to promote agricultural policies that will enhance food supply.

Despite these important uses, these data should be interpreted with caution, particularly in the developing country context. Although many countries show national increases in food supply, this does not address the issue as to how food is distributed within the country. Increases in access among the most vulnerable groups may not parallel increases in national production. Thus, although food supply data are useful for trend analysis, they should not be used to assess changes in food consumption or food security.

Trends toward a decline in the food supply can also be identified using food balance information. This may be a reflection of an unstable political environment or some severe natural disease that influenced food production. Ideally, this information should be used to influence agricultural policy to stimulate higher levels of production. However, war or other political strife may impede this process. The data should not be used to predict food shortages or famine because it is not useful to identify vulnerable groups within a population and because vulnerable groups may already be experiencing shortages by the time that this information is available and processed.

The quality of data used to generate food balance sheets can vary greatly between countries. In general, the methods are thought to underestimate total per capita energy availability in developing countries. In some countries, particularly those where small-holder agriculture is still common, this may be related to underestimates of true production due to a less centralized economy.

Household Food Consumption Data

The documentation of household food security in developing countries continues to be of great interest because of its relationship to specific health and nutrition indicators and as a means of monitoring the impact of political and environmental change on these outcomes. There have also been a number of efforts to document the impact of poverty alleviation programs on food security. Food security is

often measured by quantifying household food consumption, which provides an estimate of the food available to be consumed on a per capita basis.

Traditional methods to assess household food consumption include those that collect data over a period of time, often 7 days, by asking the respondent to keep a record of food entering the home (food account method) or by quantifying the food consumed at each meal (household food record method). Other methods may include an inventory of food available in the home over a period of time or the list-recall method, whereby the respondent is asked to recall all food purchased, quantity, and purchase price over a given time. These methods have many limitations in a developing country. For example, respondents may have limited literacy and numeracy skills. In this case, field-workers would be responsible for data collection, resulting in increased survey time and costs. Many poor households have little or no food stores in the home and inventory methods may not provide an adequate estimate of household consumption. Furthermore, these methods often rely on telephone or costly house-to-house surveys, the resources for which may not be available. Thus, these types of household consumption methods have been excluded from many large-scale surveillance systems in developing countries and efforts have been made to develop more appropriate methods.

In the past decade or so, there has been considerable interest in dietary diversity as an indicator of household consumption. (The dietary diversity score is also used to assess intake of individuals. The principle is the same, but the respondent is asked to list all foods or food groups consumed by the individual.) This method provides qualitative information on all foods or food groups, including meals and snacks, that were consumed over a given period of time (often 1, 3, or 7 days) by all members of the household. Each food or food group is assigned a value based on its nutrient density, bioavailability, and typical portion size. Portion size is included because although some foods (e.g., nuts) may have high nutrient density, they are typically consumed in small quantities. Points are then summed and the adequacy of dietary diversity is assessed based on this score. Reasonable correlations have been found between dietary diversity, household socioeconomic status, and household consumption as assessed by more traditional methods. The major advantage of this type of instrument is that it is simple and less time-consuming than other household consumption methods, with important implications for its use in large surveys. Although the use of this type of instrument in nutritional surveillance systems is still limited, its potential as a simple method to assess and monitor household food security appears promising.

Individual Nutritional Status and Dietary Intake Data

Information on the dietary intake and nutritional status of individuals in a population is essential for monitoring trends in these indicators over time and in response to political and environmental changes, as a means of identifying groups for intervention, and to assess the impact of interventions on nutritional status of the population. Although dietary intake and simple anthropometric measurements, such as weight and height, have often been the focus of health and nutrition surveys, it is essential that other indicators of nutritional status such as micronutrient deficiencies also be documented because they continue to be important public health problems in most developing countries. Furthermore, as discussed previously, information on factors that are direct (e.g., the prevalence of infections) and indirect (e.g., maternal education and family socioeconomic status) causes of nutritional problems increases the usefulness of nutritional surveillance information for policymakers.

Many nutritional surveillance systems have dealt with this daunting list of indicators by focusing efforts on specific high-risk groups—a logical decision in light of limited resources. Thus, more information is available for children younger than 5 years of age and women of reproductive age than for olden children, adolescents, adult men, and older adults. With the increasing prevalence of overweight and obesity, particularly in school-age children and adults, this strategy may need to be modified. Although not evident in all developing countries, this paradox is particularly striking in some middle-income Latin American countries, such as Mexico and Chile, but is also documented in India and many other countries.

The coexistence of malnutrition and "over-" nutrition represents an important challenge to all those involved in nutritional surveillance. For funders, the population groups being monitored may need to be expanded, with important cost implications; malnutrition in children and pregnant and lactating women has not disappeared in developing countries with the increase in overweight and obesity. For those involved in data collection, these additional nutritional problems imply the development and validation of new instruments to measure causes of overweight and obesity (e.g., physical activity). For policymakers, the burden lies in the need for policies and programs that respond to two extremes of nutrition problems, often occurring in the same communities and even households. For example, programs designed to improve dietary intake in household members at risk for nutritional deficiencies

(e.g., children younger than 2 years of age) should not cause an increase in energy intake among those members of the household at risk for overweight and obesity (e.g., school-aged children). Thus, program evaluations must be designed to detect both desirable and unexpected or undesirable outcomes.

The choice of which indicators are most appropriate for monitoring the nutritional status and dietary intake of the population depends on the country context and the specific objective of the surveillance system. For example, in countries where food shortages are common, indicators that are particularly sensitive to change, such as the prevalence and severity of malnutrition in children younger than 5 years of age, should be used. If the objective is to determine the impact of improving the nutritional status of a population in which stunting and anemia are the principal problems, then the prevalence of these should obviously be monitored.

Information on the intake and nutritional status of individuals is available from a variety of sources in developing countries. We present a description of the types of information available for children younger than 5 years of age (**Table 1**) and adults (**Table 2**) from a variety of information sources. Much of this information is obtained from large-scale multination health and nutrition surveys and from databases maintained by international organization, such as FAO and the World Health Organization (WHO). A number of countries conduct periodic nationally representative health and nutrition surveys, and information may also be available from smaller scale health and nutrition surveys and from routine growth monitoring and promotion programs. The following sections provide a brief discussion of each of these types of information in the developing country context.

Multination Health and Nutrition Surveys

During the past few decades, the Demographic and Health Surveys (DHS) have been conducted in many countries in all regions of the world. The DHS surveys are nationally representative surveys that include household and individual health and nutrition indicators. The surveys are large, typically 5000 to 30 000 households, and are conducted periodically, often at 5-year intervals. The data included in the survey vary slightly by country (**Tables 1** and **2**) but typically include as a minimum anthropometric measurements and hemoglobin concentration (prevalence of anemia) of children and women of reproductive age and breast-feeding and complementary feeding practices. One of the major strengths of the DHS surveys is that they use standard questionnaires that allow for comparisons across survey years and between countries. Information from DHS surveys is readily available on the Internet.

The WHO Global Database on Child Growth and Malnutrition provides a compilation of information from nationally representative and smaller scale surveys conducted in a number of countries. In order to be included in the database, a number of criteria must be met for data collection, analysis, and presentation. This facilitates the comparison of information that has been collected in different countries and regions. Nutrition Country Profiles are also compiled by FAO and include national-, household-, and individual-level data. The national-level data are obtained from the United Nations global data banks and are supplemented for many countries by data from local institutions and independent experts. Considering this broad range of sources, many differences in methodology of data collection, analysis, and presentation may exist and should be taken into consideration when comparing data from different countries.

National Health and Nutrition Surveys and Small-Scale Surveys

Both the WHO and FAO databases may include information obtained from nationally representative health and nutrition surveys conducted by individual countries and from small-scale health and nutrition surveys. The former has the major advantage that data may be representative of the population in the country. The latter does not usually provide representative data but has the strength that the survey may be targeted to specific high-risk groups, thus providing data for those to whom policymakers may need to target interventions.

Information on nutritional status of individuals, particularly children, may also be collected at the local community level through national growth monitoring and promotion activities conducted as part of government or nongovernmental agency development activities. Many such activities stress a high level of local involvement in data collection and can be very useful to provide feedback for decisions on resource allocation that need to be made at a local level. Data can then be aggregated to higher administrative levels and can be used for regional and national resource allocation. Although this type of surveillance may not have the same level of data quality control as the larger, more heavily supervised surveys, they have the advantage of being readily available and may promote a higher level of community involvement.

Table 1 Surveys in developing countries with individual nutritional status data for children younger than 5 years of age

Region/country	*Survey year*	*Age (years)*	*Sample size*[a]	*Data included*	*Source*
Africa					
Benin	2001	0–4.99	5305	Anthropometry, dietary intake, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), complementary feeding practices, fertility and birth interval, vaccine coverage, morbidity, mortality	DHS, FAO
Burkina Faso	1998–99	0–4.99	3792	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), complementary feeding practices, morbidity, mortality, sanitation	DHS
Congo	1997	0–4.99	NA	Anthropometry	FAO
Eritrea	2002	0–4.99	5241	Anthropometry, use of nutritional supplements, use of iodized salt, complementary feeding practices, parental education, vaccine coverage, morbidity, mortality, sanitation	DHS
Ethiopia	2000	0–4.99	9814	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), use of iodized salt, complementary feeding practices, fertility and birth interval, parental education, vaccine coverage, morbidity, mortality	DHS
Guinea	1999	0–4.99	2939	Anthropometry, complementary feeding practices, morbidity, mortality	DHS
Malawi	2000	0–4.99	9318	Anthropometry, use of nutritional supplements, use of iodized salt, complementary feeding practices, fertility and birth interval, parental education, vaccine coverage, morbidity, mortality, sanitation	DHS
Mali	2001	0–4.99	9408	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), use of iodized salt, complementary feeding practices, fertility and birth interval, parental education, vaccine coverage, morbidity, mortality, sanitation	DHS
Mauritania	2000–01	0–4.99	3554	Anthropometry, use and nutritional supplements, use of iodized salt, complementary feeding practices, morbidity	DHS
Mozambique	1997	0–2.99	4206	Anthropometry, complementary feeding practices	DHS
Namibia	2000	0–4.99	4123	Anthropometry, use and nutritional supplements, complementary feeding practices	DHS
Niger	2000	0–4.99	4616	Anthropometry	WHO

Continued

Table 1 Continued

Region/country	*Survey year*	*Age (years)*	*Sample size*[a]	*Data included*	*Source*
Rwanda	2000	0–4.99	6490	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), use of iodized salt, complementary feeding practices, fertility and birth interval, parental education, vaccine coverage, sanitation	DHS
Tanzania	1999	0–4.99	2582	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), use of iodized salt, complementary feeding practices, fertility and birth interval, parental education, vaccine coverage, morbidity, mortality, sanitation	DHS
Togo	1998	0–2.99	3260	Anthropometry	WHO
Uganda	2000–01	0–4.99	5604	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), use of iodized salt, complementary feeding practices, fertility and birth interval, parental education, vaccine coverage, morbidity, mortality, sanitation	DHS
Zambia	2001–02	0–4.99	5216	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), use of iodized salt, complementary feeding practices, fertility and birth interval, parental education, vaccine coverage, morbidity, mortality, sanitation	DHS
Zimbawe	1999	0–4.99	3559	Anthropometry	WHO
Asia and Southwest Pacific					
Bangladesh	2001	0–4.99	71 931	Anthropometry	WHO
Bangladesh	1999–2000	0–4.99	5421	Anthropometry, micronutrient deficiencies (Fe, I, vitamin A)	DHS
Bhutan	1999	0.5–4.99	2981	Anthropometry, micronutrient deficiencies (Fe, I, vitamin A)	FAO, WHO, NS
Cambodia	2000	0–4.99	3372	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), complementary feeding practices	DHS
China	2000	0–4.99	16 491	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe)	FAO, WHO
Fiji	1993	0–4.99	618	Anthropometry, micronutrient deficiencies (Fe, I, vitamin A), complementary feeding practices	FAO, WHO, NS
India	1998–99	0–2.99	24 396	Anthropometry, micronutrient deficiencies (Fe, I, vitamin A)	FAO, WHO, NS
Lao People's Democratic Republic	2000	0–4.99	1347	Anthropometry, use of nutritional supplements, micronutrient deficiencies (I, vitamin A), complementary feeding practices	WHO, NS
Nepal	2001	0–4.99	6409	Anthropometry, dietary intake, use of nutritional supplements, complementary feeding practices	WHO, DHS, NS

Continued

Table 1 Continued

Region/country	*Survey year*	*Age (years)*	*Sample size*[a]	*Data included*	*Source*
Pakistan	1995	0–4.99	7368	Anthropometry, micronutrient deficiencies (I, vitamin A)	FAO, WHO
Papua New Guinea	1982–83	0–4.99	27 464	Anthropometry, complementary feeding practices, micronutrient deficiencies (I, vitamin A)	WHO, NS (rural)
Philippines	1998	0–4.99	24 308	Anthropometry	FAO, NS
Sri Lanka	1995	0.25–4.99	2782	Anthropometry, micronutrient deficiencies (Fe, I, vitamin A)	FAO, NS
Vanuatu	1996	0–4.99	1194	Anthropometry, dietary intake, micronutrient deficiencies (Fe, I, vitamin A)	FAO, NS
Vietnam	2000	0–4.99	94 469	Anthropometry	WHO
Vietnam	2001	0–2.99	1321	Complementary feeding practices	DHS
Near East					
Egypt	2003	0–4.99	5761	Anthropometry, use of nutritional supplements, dietary intake, micronutrient deficiencies (Fe, I, vitamin A), use of iodized salt, complementary feeding practices, vaccine coverage, morbidity	DHS
Iran	1998	0–4.99	2536	Anthropometry	FAO, WHO, NS
Jordan	2002	0–4.99	5484	Anthropometry, use of nutritional supplements, dietary intake, micronutrient deficiencies (Fe, I, vitamin A), complementary feeding practices	DHS
Morocco	1997	0–4.99	3555	Anthropometry	WHO
Turkey	1998	0–4.99	2677	Anthropometry, dietary intake, complementary feeding practices	DHS
Latin America and the Caribbean					
Antigua y Barbuda	1981	0–5.99	463	Anthropometry	WHO
Argentina	1995–96	0–5.99	16 981	Anthropometry, complementary feeding practices	WHO
Barbados	1981	0–4.99	597	Anthropometry, complementary feeding practices	NS
Bolivia	1998	0–4.99	5773	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), complementary feeding practices, fertility and birth interval, parental education, vaccine coverage, morbidity, mortality, sanitation	DHS
Brazil	1996	0–4.99	3815	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), complementary feeding practices, fertility and birth interval, parental education, vaccine coverage, morbidity, mortality, sanitation	DHS
Chile	2002	0–4.99	51 572	Anthropometry, complementary feeding practices	WHO
Colombia	2000	0–4.99	4060	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), use of iodized salt, complementary feeding practices, fertility and birth interval, parental education	DHS
Costa Rica	1996	1–6.99	1008	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A)	NS

Continued

Table 1 Continued

Region/country	*Survey year*	*Age (years)*	*Sample size*[a]	*Data included*	*Source*
Dominica	1984	0–4.99	245	Anthropometry, complementary feeding practices	WHO
Dominican Republic	2002	0–4.99	2086	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), complementary feeding practices, fertility and birth interval, parental education	DHS
Ecuador	1998	0–4.99	2998	Anthropometry, complementary feeding practices	PAHO
El Salvador	2002–03	0.25–4.99	—	Anthropometry, complementary feeding practices	WHO
Guatemala	2002	0.25–4.99	6308	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), complementary feeding practices, fertility and birth interval, parental education	DHS
Guyana	1997	0–4.99	289	Anthropometry, complementary feeding practices	PAHO
Haiti	2000	0–4.99	6176	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), complementary feeding practices, fertility and birth interval, parental education	DHS
Honduras	2001	0.25–4.99	5613	Anthropometry	WHO
Jamaica	1999	0–4.99	574	Anthropometry, complementary feeding practices	PAHO
Mexico	1999	0–4.99	8011	Anthropometry, micronutrient deficiencies (Fe, Zn, I, vitamin A, vitamin C, folic acid), complementary feeding practices	NS
Nicaragua	2001	0–4.99	171	Anthropometry	World Bank
Panama	1997	0–4.99	2049	Anthropometry, complementary feeding practices	PAHO
Paraguay	1990	0–4.99	3389	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), complementary feeding practices, fertility and birth interval, parental education	DHS
Peru	2000	0–4.99	10 477	Anthropometry, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A), complementary feeding practices, fertility and birth interval	DHS
Trinidad & Tobago	2000	0–4.99	781	Anthropometry, complementary feeding practices	WHO
Uruguay	1992–93	0–4.99	11 521	Anthropometry, complementary feeding practices	SISVEN
Venezuela	2000	0–4.99	321 257	Anthropometry, complementary feeding practices	SISVAN

[a]Total sample size reported. Actual sample sizes differ by variable. All samples include both sexes.
Fe, anemia and/or iron deficiency; I, iodine deficiency; DHS, Demographic and Health Surveys; FAO, Food and Agriculture Organization of the United Nations; WHO, World Health Organization; NS, nationally representative health and/or nutrition survey not included in FAO or WHO database; PAHO, Pan American Health Organization; SISVEN, Sistema de Vigilancia Epidemiológica Nutricional [Nutritional Epidemiology Monitoring System]; SISVAN, Sistema de Vigilancia Alimentaria y Nutricional [Food and Nutrition Monitoring System].

Table 2 Surveys in developing countries with *per capita* consumption data and individual nutritional status data for adults

Region/country	*Survey year*	*Age (years)*	*Sample size*[a]	*Data included*	*Source*
Africa					
Benin	2001	15–49	2579	Anthropometry, dietary intake, micronutrient deficiencies (Fe), per capita consumption	FAO
Burkina Faso	1999	15–49	3416	Anthropometry, dietary intake, micronutrient deficiencies (Fe, I), per capita consumption	FAO, NS
Ethiopia	2000	15–49	13 447	Anthropometry	DHS
Guinea	1990	—	779	Anthropometry, dietary intake, micronutrient deficiencies (I), per capita consumption	FAO
Mali	2001	—	10 049	Anthropometry, micronutrient deficiencies (Fe, I)	DHS
Mauritania	1990	>18	2112	Anthropometry, dietary intake, micronutrient deficiencies (I), per capita consumption	FAO, NS
Namibia	1992	15–49	2249	Anthropometry, dietary intake, micronutrient deficiencies (Fe, I, vitamin A), per capita consumption	FAO
Niger	1995	18–60	NA	Anthropometry, dietary intake, micronutrient deficiencies (I), per capita consumption	FAO
Togo	1997	>19	375	Anthropometry, dietary intake, micronutrient deficiencies (I), per capita consumption	FAO
Zimbawe	1999–2000	15–49	5590	Anthropometry	DHS
Asia and Southwest Pacific					
Bangladesh	1996–97	15–49	3921	Anthropometry	FAO, DHS
Cambodia	1998	15–49	1109	Anthropometry	FAO
China	1996	>20	28 706	Anthropometry, dietary intake, micronutrient deficiencies (Fe, I) per capita consumption	FAO, NS
Fiji	1993	18–65	2573	Anthropometry, dietary intake, micronutrient deficiencies (Fe, I), per capita consumption	FAO, NS
India	1996	All	NA	Anthropometry, dietary intake, micronutrient deficiencies (Fe, vitamin A), per capita consumption	FAO
Lao People's Democratic Republic	2001	>15	5942	Anthropometry, dietary intake, use of nutritional supplements, micronutrient deficiencies (Fe), complementary feeding practices	FAO
Nepal	2001	15–49	7774	Anthropometry, dietary intake, use of nutritional supplements, micronutrient deficiencies (vitamin A), complementary feeding practices	DHS
Papua New Guinea	1997	21–50	1041	Anthropometry, dietary intake, per capita consumption	FAO
Philippines	1998	20–39	3123	Anthropometry	FAO, NS
Sri Lanka	1997	—	2624	Anthropometry, dietary intake, per capita consumption, micronutrient deficiencies (Fe)	FAO
Vanuatu	2000	20–60	800	Anthropometry	FAO
Viet Nam	1997	15–49	4212	Anthropometry, micronutrient deficiencies (vitamin A)	FAO, NS

Continued

Table 2 Continued

Region/country	*Survey year*	*Age (years)*	*Sample size*[a]	*Data included*	*Source*
Near East					
Egypt	2003	15–49	8078	Anthropometry, use of nutritional supplements, use of iodized salt	DHS
Iran	1995	20–74	NA	Anthropometry, dietary intake, per capita consumption, micronutrient deficiencies (Fe)	FAO
Jordan	2002	15–49	7682	Anthropometry, dietary intake, use of nutritional supplements, micronutrient deficiencies (Fe, I, vitamin A)	DHS
Morocco	1992	20–35	2751	Anthropometry, dietary intake, micronutrient deficiencies (I), per capita consumption	FAO
Turkey	1998	15–49	2183	Anthropometry, micronutrient deficiencies (I)	FAO
Latin America and the Caribbean					
Argentina	1990	19–64	504	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Bahamas	1988–89	15–64	1771	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Brazil	1996	15–49	2951	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Chile	1996	25–64	2127	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Colombia	2000	15–49	3070	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Costa Rica	1996	20–59	NA	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Cuba	1995	20–59	9815	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Dominican Republic	1997	15–49	2492	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Guatemala	1999	15–49	2585	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Jamaica	1999	25–74	2075	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Mexico	2000	20–99	45 200	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO, NS
Nicaragua	1999	15–49	4793	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Panama	1995	21–60	2448	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Peru	1997	15–49	9600	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO

Continued

Table 2 Continued

Region/country	Survey year	Age (years)	Sample size[a]	Data included	Source
Trinidad & Tobago	1999	>20	803	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Uruguay	1991	20–50	1079	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO
Venezuela	1997	20–50	14 084	Anthropometry, dietary intake, micronutrient deficiencies, per capita consumption	FAO

[a]Total sample size reported. Actual sample sizes differ by variable. All samples include both sexes with the exception of Costa Rica and Peru, for which data were found for women only.
Fe, anemia and/or iron deficiency; I, iodine deficiency; NA, not available; DHS, Demographic and Health Surveys; FAO, Food and Agriculture Organization of the United Nations; NS, nationally representative health and/or nutrition survey not included in FAO database.

See also: **Dietary Intake Measurement**: Methodology; Validation. **Dietary Surveys**. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **Nutritional Assessment**: Anthropometry; Biochemical Indices; Clinical Examination. **Nutritional Surveillance**: Developed Countries. **Obesity**: Definition, Etiology and Assessment. **United Nations Children's Fund**. **World Health Organization**.

Further Reading

Comisión Económica para América Latina y el Caribe [Council for Statistics in Latin America] (CEPAL) (2004) www.cepal.org. Accessed 23 September 2004.

Demographic and Health Surveys (2004) http://measuredhs.com. Accessed 27 September 2004.

Food and Agriculture Organization (FAO) (2004) *World Agriculture: Towards 2015/2030—An FAO Perspective*. http://www.fao.org/docrep/005/y4252e/y4252e04.htm. Accessed 21 September 2004.

Gibson RS (1990) *Principals of Nutritional Assessment*. New York: Oxford University Press.

Jonsson U (1995) *Towards an improved strategy for nutritional surveillance. Food and Nutrition Bulletin* **16**(2). [www.unu.edu/unupress. Accessed 20 September 2004].

Latham M (2004) *Human Nutrition in the Developing World*. http://www.fao.org/DOCREP/W0073e07.htm. Accessed 20 September 2004.

Rose D, Meershoek S, Ismael C *et al.* (2002) Evaluation of a rapid field tool for assessing household diet quality in Mozambique. *Food and Nutrition Bulletin* **23**: 181–191.

Ruel MT (2003) Operationalizing dietary diversity: A review of measurement issues and research priorities. *Journal of Nutrition* **133**: 3911S–3926S.

Sistema de Vigilancia Alimentaria y Nutricional [Food and Nutrition Monitoring System] (SISVAN) (2004) http://www.sisov.mpd.gov.ve/articulos/23/. Accessed 25 September 2004.

Sistema de Vigilancia Epidemiológica Nutricional [Nutritional Epidemiology Monitoring System] (SISVEN) (2004) http://165.158.1.110/english/sha/ururstp.htm. Accessed 25 September 2004.

WHO/FAO (2003) *Joint WHO/FAO Expert Consultation on Diet, Nutrition and the Prevention of Chronic Diseases*. Geneva: World Health Organization.

World Health Organization, Department of Nutrition for Health and Development (2004) *WHO Global Database on Child Growth and Malnutrition*. http://www.who.int/nutgrowthdb/. Accessed 22 September 2004.

NUTS AND SEEDS

J Gray, Guildford, UK

In botanical terms, the word 'nut' is used to describe a wide range of seeds, mostly from trees, with a tough, often lignified, seed coat or shell. True nuts include the chestnut, brazil nut, and hazelnut. In practice, these are usually classified together with certain other so-called nuts, for example the almond, cashew, and peanut, and other seeds which are all used in similar ways in the diet. Nuts and seeds come from a diverse range of different plants, so their nutritional composition is quite varied, but like most plant seeds they contain a food reserve designed to meet the needs of the developing plant embryo. In many nuts and seeds this is fat, but in others it is starch or other polysaccharides.

Therefore, these foods are concentrated sources of dietary energy, as well as sources of protein, unsaturated fatty acids, various micronutrients, and fiber (nonstarch polysaccharides, NSP).

Nuts and seeds have a wide range of uses. In the typical Western omniverous diet they tend to be used either as snack items or added as minor ingredient to savory and sweet dishes, but they have wider applications in vegetarian diets as important sources of protein and other nutrients. Certain nuts and seeds are also made into spreads, for example peanut butter and tahini (sesame seed spread).

Types

The major types of nuts and seeds grown for human consumption are shown in **Table 1.**

Nuts

Almond The almond (*Prunus amygdalis* var. *dulcis*), sometimes called the sweet almond, is one of the oldest nut crops. It is believed to have originated in Southeast Asia but is now grown more widely, including in southern Europe, Africa, southern Australia, and California. It is closely related to peaches and plums, but in the almond, in contrast to these other fruits, the 'flesh' or mesocarp becomes hard and dry as it matures, and splits open to leave the thin shell or endocarp which contains the edible almond seed or 'nut.' The nuts are eaten fresh, often in the ground form in prepared dishes, as well as roasted and salted.

Another species, *Prunus amara* or the bitter almond, is inedible but is cultivated for its oil, which is also present in the sweet almond and in the kernels of apricots and peaches. This oil contains benzaldehyde, the essential oil, and hydrocyanic acid, from which the benzaldehyde is separated to be used in flavorings and perfumes.

Brazil nut The triangular-shaped Brazil nut (*Bertholletia excelsa*) grows in large forests in the Amazon river basin in South America. The nuts are actually hard-shelled seeds which are produced in groups of between 12 and 30 within a large, hard, thick-walled woody fruit or pod. The sweet-tasting nut meat is consumed in the fresh state and Brazil nut oil may be extracted for use as a lubricant.

Table 1 Major types of nuts and seeds grown for human consumption

Almond	Pecan
Brazil	Pine nuts
Cashew	Pistachio
Chestnut	Walnut
Coconut	Pumpkin seeds
Hazelnut	Sesame seeds
Macadamia	Sunflower seeds
Peanut	

Cashew nut The cashew (*Anacardium occidentale*) originated in Brazil but is now cultivated extensively in all tropical areas, notably in India and East Africa. The cashew fruit, which contains the seed or 'nut,' hangs at the end of what is referred to as the cashew 'apple'—the edible swollen fruit stem or pedicel. The fruit itself is kidney-shaped, about the size of a large bean, and has a two-layered shell. The outer layer of this shell contains a caustic oil that must be burned off before the nut is touched. The nuts are then roasted again or boiled to remove other toxic substances and the second shell is removed. The nuts may also be used as a source of oil.

Chestnut The sweet or Spanish chestnut (*Castanea sativa*) is a native tree of southern Europe, believed to have been introduced into Britain by the Romans. The fruit consists of two to four compartmentalized seeds or burrs, covered with numerous needle-sharp branched spines and containing the seeds or 'nuts,' which are covered with a tough outer coat. The flesh of the nut is hard and inedible and is cooked, often by roasting or boiling, before being eaten. The cooking process changes the texture so that the chestnut becomes much softer than other nuts and more like a vegetable, largely as a result of its high carbohydrate content (see below).

Coconut The coconut (*Cocos nucifera*) grows on the coconut palm, which is common in tropical areas throughout the world. The native origin of the palm is uncertain, as the nuts were easily dispersed between both islands and continents by ocean currents and by early explorers. The fruits are borne on the tree in clusters of about 15 to 20 and are enclosed in a thick outer husk and covered in a mass of fibers (the mesocarp and exocarp), which is normally removed when the coconut is harvested. The familiar hard shell of the coconut is the endocarp, or inner layer, of the mature ovary of the fruit, and within the shell is the actual seed, covered with a thin brown seed coat. The white coconut 'meat,' which can be eaten either fresh or desiccated, is actually part of the endosperm (storage tissue) of the seed. Coconut 'milk,' which is found in the unripe nut and is drunk or used in cooking, is the liquid form of the endosperm, which solidifies as the fruit ripens. The coconut meat may be dried

to produce copra, which is pressed to remove the coconut oil used widely as a food oil and in soap and cosmetic manufacture.

Hazelnut (cobnut; filbert) The most widely grown hazelnut (*Corylus avellana*) is a native of Europe, although about 10 different species of *Corylus* grow throughout Europe, North America, and Asia. There is evidence that these nuts were cultivated in Ancient Greece and collected by Mesolithic peoples. The shell of the hazelnut is the matured ovary wall of the flower and the edible nut meat within this is the matured embryo.

Macadamia nut The macadamia nut (*Macadamia integrifolia*), smooth-shelled; *M. tetraphylla*, rough-shelled) is native to eastern tropical Australia but was subsequently introduced to Hawaii, which is now the leading producer of these nuts, and also to parts of Africa and South America. It is the smooth-shelled variety that has been developed commercially. The edible kernel of the nut is the seed, consisting mostly of the cotyledons of the embryo. It is enclosed in a hard, thick, brown shell, which is itself encased in a fibrous husk that splits open when the husk dries. This occurs after the fruit falls, or when it is removed from the tree at maturity. After harvesting, the nuts are dried (to a moisture level of 1.5%), roasted (traditionally in coconut oil, or dry-roasted), and salted.

Peanut The peanut (*Arachis hypogaea*), sometimes referred to as the ground nut or monkey nut, originated in South America. Although referred to as a nut, it is in fact part of the legume family. The plant was introduced to Africa by early European explorers and to North America by the slave trade; it was also introduced to India and China. The name 'ground nut' derives from the fact that the flower withers after pollination to leave a stalk-like part of the plant, which pushes under the soil and carries the fertilized ovules in its tip. Underground, the tip continues to develop into the characteristic pod of the peanut, containing the seeds, or 'nuts.' The shape and size of the pod, and the number and color of the seeds, are variable, depending on the peanut cultivar. On a worldwide basis, two-thirds of the peanut crop is crushed for oil (arachis oil) and peanut products are used widely in both food processing, with peanut butter as an important product, and for animal feed. The peanut itself may be eaten fresh or roasted and salted.

Pecan The pecan (*Carya illinoinensis*) is a member of the walnut family, and the tree is classified botanically as a hickory. The tree is a native of North America, grown in the southern central states. After harvesting, the nuts are air-dried to remove 10–20% of their moisture. The nut is similar to the walnut, but with a more mild and sweet flavor. The pecan nut kernel is eaten fresh and in the US it is used widely in confectionery and baked goods.

Pine nuts Pine nuts or kernels are small edible seeds which are extracted from the cones of various species of pine. The most commonly eaten variety is that from the European stone pine (*Pinus pinea*), which is native to northern Mediterranean regions. The small, oil-rich seeds are encased in a hard shell. The seeds are sometimes referred to as pignolia nuts, whereas the seeds of the pinyon pines (*Pinus edulis* and *Pinus monophylla*), which grow in the south-western US and in northern Mexico, are known as pinon nuts.

Pistachio nut The pistachio nut is the seed of the pistachio tree (*Pistacia vera*). It is a native of central Asia, Pakistan, and India, where it was cultivated 3000 years ago, and it has also been cultivated for many years in Mediterranean regions and more recently in California. The pistachio fruit is similar to a peach; the outer 'husk' (the exocarp and mesocarp of the fruit) encloses a hard but thin off-white shell (the endocarp). This splits open just before the nut matures to reveal the edible embryo, which consists mainly of two green cotyledons covered in a thin seed coat. The green nut kernels are highly prized and are eaten roasted and salted as well as in various Middle Eastern dishes.

Walnut The walnut (*Juglans* spp.) is the common name given to about 20 species of trees in this family. The most important species is *Juglans regia*—the English or Persian walnut—which is believed to have originated in Ancient Persia, later taken to Greece, and eventually distributed throughout the Roman empire. There are records of its growth in England in the sixteenth century. It was taken to America and called the English walnut to distinguish it from the native American black walnut (*Juglans nigrans*) and the butternut (*Juglans cinerea*), both of which have much thicker, less brittle shells. The walnut fruit has an outer leathery husk and an inner furrowed stone, which is the shell of the nut, within which is the edible seed.

Seeds

Pumpkin seeds The large flat seeds of the members of the pumpkin family (*Cucurbita maxima; C. moschata*

and related species) can be dried and eaten raw, used in both sweet and savory cooked dishes, or roasted.

Sesame seeds The sesame plant (*Sesamum indicum*), which is a native of Africa, grows in tropical and subtropical regions and is now common in Asia. The seeds are small and off-white in color. They may be eaten whole or used in confectionery and baked goods and as a source of oil used in cooking. The seeds are also ground to a paste called tahini.

Sunflower seeds The sunflower (*Helianthus annus*) is a member of the Compositae or daisy family. It is believed to have originated in North America, where it was cultivated by the native Indians, and was introduced to Europe in the sixteenth century. The flat seeds may be dehusked and eaten raw or cooked, but the plant is generally cultivated for the oil they contain, which is a rich source of polyunsaturated fatty acids (see below), and is widely used for cooking and in margarine manufacture. The residual oil-cake is used for animal feed.

Macronutrient Content

Green nuts, as harvested, may contain 50% or more water, but these nuts must be cured or semidried for storage, so the moisture content of most nuts, as eaten, is low (1–6%). The exceptions are fresh coconut and chestnuts, with a moisture content of 45 and 52%, respectively. The water, macronutrient, and energy content of the nuts and seeds discussed in this article are shown in **Table 2**.

Fat

The total fat content of most nuts and seeds is high because, as the seed ripens, the fat store increases and its starch content declines. However, the amount of fat is quite variable, ranging from about 78% in the macadamia nut and 70% in the pecan to around 50–55% in nuts such as the almond, cashew, hazelnut, and pistachio, and as low as 3% in chestnuts. The fat content of the edible seeds is between 45 and 60%.

The different fatty acid fractions contained in these nuts and seeds are also quite variable, as shown in **Table 3**. The vast majority of nuts and seeds are rich in monounsaturated and polyunsaturated fatty acids. However, in some nuts, such as the peanut, hazelnut, and macadamia nut, monounsaturated fatty acids predominate, whereas in the walnut and in sunflower seeds polyunsaturated fatty acids predominate. The exception is the coconut, in which saturated fatty acids constitute the major fat fraction.

Carbohydrate

With the exception of the starch-rich chestnut (almost 37% carbohydrate), the carbohydrate content of most nuts is relatively low at around 3–7%. However, peanuts, cashews, pumpkin, and sunflower seeds contain more carbohydrate (13–19%). In most nuts and seeds this carbohydrate is a variable mixture of starch and sucrose, although in some there are small quantities of glucose and fructose as well, and in sunflower seeds there are some oligosaccharides.

Table 2 Water, macronutrient, and energy content of selected nuts and seeds (per 100 g, kernel only)

	Water (g)	*Protein (g)*	*Fat (g)*	*Carbohydrate (g)*	*Energy*	
					(kJ)	*(kcal)*
Almond	4.2	21.1	55.8	6.9	2534	612
Brazil	2.8	14.1	68.2	3.1	2813	682
Cashew	4.4	17.7	48.2	18.1	2374	573
Chestnut	51.7	2.0	2.7	36.6	719	170
Coconut	45.0	3.2	36.0	3.7	1446	351
Hazelnut	4.6	14.1	63.5	6.0	2685	650
Macadamia (salted)	1.3	7.9	77.6	4.8	3082	748
Peanut	6.3	25.6	46.1	12.5	2341	564
Pecan	3.7	9.2	70.1	5.8	2843	689
Pine nuts	2.7	14.0	68.6	4.0	2840	688
Pistachio (roasted, salted)	2.1	17.9	55.4	8.2	2485	601
Walnut	2.8	14.7	68.5	3.3	2837	688
Pumpkin seeds	5.6	24.4	45.6	15.2	2360	569
Sesame seeds	4.6	18.2	58.0	0.9	2470	598
Sunflower seeds	4.4	19.8	47.5	18.6	2410	581

Data from Holland *et al.* (1992).

Table 3 Total fat and fatty acid composition of selected nuts and seeds (g per 100 g, kernel only)

	Total fat	*Saturated fatty acids*	*Monounsaturated fatty acids*	*Polyunsaturated fatty acids (total)*	cis *n-6 Polyunsaturated fatty acids*	cis *n-3 Polyunsaturated fatty acids*
Almond	55.8	4.7	34.4	14.2	13.3	0.1
Brazil	68.2	16.4	25.8	23.0	22.9	0.1
Cashew	48.2	9.5	27.8	8.8	—[a]	—[a]
Chestnut	2.7	0.5	1.0	1.1	1.0	0.1
Coconut	36.0	31.0	2.0	0.8	0.5	0
Hazelnut	63.5	4.7	50.0	5.9	5.4	0.1
Macadamia (salted)	77.6	11.2	60.8	1.6	—[a]	—[a]
Peanut	46.1	8.2	21.1	14.3	—[a]	—[a]
Pecan	70.1	5.7	42.5	18.7	16.0	0.7
Pine nuts	68.6	4.6	19.9	41.1	—[a]	—[a]
Pistachio (roasted, salted)	55.4	7.4	27.6	17.9	—[a]	—[a]
Walnut	68.5	5.6	12.4	47.5	—[a]	—[a]
Pumpkin seeds	45.6	7.0	11.2	18.3	—[a]	—[a]
Sesame seeds	58.0	8.3	21.7	25.5	23.6	0.4
Sunflower seeds	47.5	4.5	9.8	31.0	—[a]	—[a]

[a]No data available.
Data from Holland *et al.* (1992) and The Ministry of Agriculture, Fisheries and Food.

Protein

The protein content of nuts is quite variable, but most nuts are considered to be a good source of protein. It is low (2–3%) in the chestnut and coconut, between 8 and 15% for most other nuts, but high (18–26%) in the cashew, pistachio, almond, and peanut, so that the amount of protein in many nuts is about the same as in meat, fish, or cheese. Pumpkin, sesame, and sunflower seeds are also rich in protein.

However, the proportions of indispensable amino acids in any one particular type of nut or seed, and in fact all plant foods, differ from those needed in the human diet, with one or sometimes more 'limiting amino acids.' In most nuts and seeds, with the exception of pistachio nuts and pumpkin seeds, it is lysine that is the limiting amino acid. Thus, although the total amount of protein in nuts and seeds may be high, these foods must be complemented by other sources of plant protein, such as legumes and/or animal sources of protein (meat, fish, eggs, milk, cheese), to ensure that the overall protein quality of the diet is adequate.

Micronutrient Content

The vitamin and mineral contents of the nuts and seeds discussed in this article are shown in **Tables 4** and **5**, respectively.

In general, nuts and seeds are a good source of the B vitamins, including folic acid, and of the tocopherols (vitamin E), although some, such as almonds, hazelnuts, and sunflower seeds, contain much more vitamin E than others. Nuts and seeds do not contain vitamin C, and many nuts have little or no vitamin A activity.

Nuts and seeds contain quite large amounts of many minerals. In particular, many nuts and seeds, especially sesame seeds, are good sources of calcium. They are also generally rich in potassium, magnesium and phosphorus, iron, and in trace elements such as copper, zinc, manganese, and others such as chromium. Brazil nuts are particularly rich in selenium.

Fiber Content

Compositional values for the total amount of fiber (nonstarch polysaccharides, NSP), and the different fiber fractions where available, are shown in **Table 6** for the nuts and seeds discussed in this article. It can be seen that nuts and seeds contain significant amounts of fiber, similar to the amounts found in vegetables and fruit. Although nuts and seeds do contain some soluble fiber, most of the fiber in these foods is of the insoluble type, much of which is cellulose. Of the insoluble noncellulosic polysaccharides, arabinose predominates in most nuts, although the coconut contains large quantities of mannose. Most nuts and seeds are likely to contain quite large amounts of lignin, particularly those with a tough seed coat such as sesame seeds, although actual values are not available.

Table 4 Vitamin content of selected nuts and seeds (per 100 g, kernel only)

	Carotene (μg)	*Vitamin E (mg)*	*Thiamin (mg)*	*Riboflavin (mg)*	*Niacin (mg)*	*Vitamin B_6 (mg)*	*Folate (μg)*
Almond	0	23.96	0.21	0.75	3.1	0.15	48
Brazil	0	7.18	0.67	0.03	0.3	0.31	21
Cashew	6	0.85	0.69	0.14	1.2	0.49	67
Chestnut	0	1.20	0.14	0.02	0.5	0.34	N[a]
Coconut	0	0.73	0.04	0.01	0.5	0.05	26
Hazelnut	0	24.98	0.43	0.16	1.1	0.59	72
Macadamia (salted)	0	1.49	0.28	0.06	1.6	0.28	N
Peanut	0	10.09	1.14	0.10	13.8	0.59	110
Pecan	50	4.34	0.71	0.15	1.4	0.19	39
Pine nuts	10	13.65	0.73	0.19	3.8	N	N
Pistachio (roasted, salted)	130	4.16	0.70	0.23	1.7	N	58
Walnut	0	3.85	0.40	0.14	1.2	0.67	66
Pumpkin seeds	230[b]	N	0.23	0.32	1.7	N	N
Sesame seeds	6	2.53	0.93	0.17	5.0	0.75	97
Sunflower seeds	15	37.77	1.60	0.19	4.1	N	N

[a]Nutrient present in significant quantities but no reliable information available on the amount.
[b]Estimated value.
Data from Holland *et al.* (1992).

Table 5 Mineral and trace element content of selected nuts and seeds (per 100 g, kernel only)

	Sodium (mg)	*Potassium (mg)*	*Calcium (mg)*	*Magnesium (mg)*	*Phosphorus (mg)*	*Iron (mg)*	*Copper (mg)*	*Zinc (mg)*	*Manganese (mg)*	*Selenium (μg)*
Almond	14	780	240	270	550	3.0	1.00	3.2	1.7	4
Brazil	3	660	170	410	590	2.5	1.76	4.2	1.2	1530[a]
Cashew	15	710	32	270	560	6.2	2.11	5.9	1.7	29
Chestnut	11	500	46	33	74	0.9	0.23	0.5	0.5	Tr
Coconut	17	370	13	41	94	2.1	0.32	0.5	1.0	1[b]
Hazelnut	6	730	140	160	300	3.2	1.23	2.1	4.9	Tr
Macadamia (salted)	280	300	47	100	200	1.6	0.43	1.1	5.5	7
Peanut	2	670	60	210	430	2.5	1.02	3.5	2.1	3
Pecan	1	520	61	130	310	2.2	1.07	5.3	4.6	12
Pine nuts	1	780	11	270	650	5.6	1.32	6.5	7.9	N[c]
Pistachio (roasted, salted)	530	1040	110	130	420	3.0	0.83	2.2	0.9	6[b]
Walnut	7	450	94	160	380	2.9	1.34	2.7	3.4	19
Pumpkin seeds	18	820	39	270	850	10.0	1.57	6.6	N	6[b]
Sesame seed	20	570	670	370	720	10.4	1.46	5.3	1.5	N
Sunflower seeds	3	710	110	390	640	6.4	2.27	5.1	2.2	49[b]

[a]Range, 230–5300 μg per 100 g.
[b]Estimated value.
[c]Nutrient present in significant quantities but no reliable information available on the amount.
Data from Holland *et al.* (1992).

Toxins and Contaminants

Phytic Acid

Phytic acid (*myo*-inositol hexaphosphoric acid) is present in all seeds, where it is believed to act as a store of phosphate and trace elements for the developing plant embryo. The phytate content of the commonly eaten nuts and seeds is variable. In general, the oil seeds, such as sesame and sunflower, and a number of the tree nuts, have higher phytate levels than the leguminous peanut, although the oils expressed from the seeds do not contain phytate. The phytate content of the coconut and chestnut is particularly low.

Because of its molecular structure, phytic acid is a highly effective chelator, which forms insoluble complexes with mineral cations. Its presence in plant foods has led to concerns that it may reduce the bioavailability of various dietary minerals and trace elements, including calcium, magnesium, iron,

Table 6 Total dietary fiber, as measured by the Englyst method, and fiber fractions in selected nuts and seeds (g per 100 g, kernels only)

		Fiber fractions			
			Noncellulosic polysaccharide		
	Total fiber	*Cellulose*	*Soluble*	*Insoluble*	*Lignin*
Almond	7.4[a]	1.9[a]	1.1[a]	4.4[a]	N[b]
Brazil	4.3	1.6	1.3	1.4	N
Cashew	3.2	0.6	1.6	1.0	N
Chestnut	4.1	1.1	1.3	1.7	N
Coconut	7.3	0.8	1.0	5.5	N
Hazelnut	6.5	2.2	2.5	1.8	N
Macadamia (salted)	5.3	1.4	1.9	2.0	N
Peanut	6.2	2.0	1.9	2.3	N
Pecan	4.7	1.2	1.5	2.0	N
Pine nuts	1.9	N	N	N	N
Pistachio (roasted, salted)	6.1	1.3	2.7	2.1	N
Walnut	3.5	1.1	1.5	0.9	N
Pumpkin seeds	5.3	1.1	1.7	2.5	N
Sesame seeds	7.9	N	N	N	N
Sunflower seeds	6.0	1.4	1.8	2.8	N

[a]Estimated value.
[b]Nutrient present in significant quantities but no reliable information available on the amount.
Data from Holland *et al.* (1992).

zinc, and copper. Although nuts are rich in iron, there is evidence that the addition of nuts to a meal can have a substantial inhibitory effect on iron absorption, presumably because of their phytate and polyphenol content. However, it appears that this can be overcome by the addition of a source of vitamin C to the meal, thereby underlining the need to mix different groups of foods within a meal, particularly when plant foods are the main source of nutrition.

The significance of dietary phytate intake to overall mineral nutriture is still uncertain. It is likely that in a mixed diet of animal and plant foods, dietary phytate may be of less significance than among people consuming diets where plant foods are the sole source of nutrition (vegans). Available data suggest that the trace element status of most adult vegetarians is adequate, but because of increased requirements for growth, vegetarian children may be more vulnerable to the reduced bioavailability of minerals and trace elements, notably zinc, which could be a consequence of the ingestion of large amounts of phytate-containing plant foods.

Intolerances/Allergies to Nuts

Intolerances to nuts, or more specifically, allergies to nut proteins, occur in a relatively small minority of people. However, there is evidence that such adverse reactions have become more common, and the severity of the reaction that occurs in these sensitive individuals means that they must be taken very seriously. Peanuts are the most commonly cited cause of these severe reactions, estimated to affect between 0.1 and 0.2% of the population, but allergic reactions to tree nuts, incuding Brazil nuts, almonds, hazelnuts, and cashews, and also to sesame seeds, have been reported.

Contaminants

Nuts and seeds may be subject to mould growth during storage if the conditions are inappropriate. Certain moulds produce secondary metabolites which are toxic to humans and animals, known as the mycotoxins. Of these mycotoxins, the aflatoxins, notably aflatoxin B1, are produced by three closely related species of mould: *Aspergillus flavus*, *A. parasiticus*, and *A. nomius*. These moulds may contaminate various food commodities in tropical and subtropical regions, including tree nuts, but one of the most important crops to be affected is the peanut. Aflatoxins are acutely toxic to the liver and may also be involved in the etiology of human liver cancer in certain parts of the world. Ochratoxins, which are produced by other *Aspergillus* species, have also been found to contaminate nuts.

Some species of mould are able to proliferate within growing crops even before they are harvested, forming an endophytic relationship with the plant. This relationship has been found to exist between *Aspergillus parasiticus* and peanuts. It appears that

when the plant is growing normally, no aflatoxin is produced by the mould, but when the plant is stressed, as occurs in drought conditions, then the mycotoxin may be produced. The concentrations of aflatoxins produced in this way are lower than would ensue from poor postharvest storage, but the economic consequences still may be considerable.

There are regulatory limits for the aflatoxin levels in foods. In the UK, the sale of nuts for direct consumption is prohibited if the aflatoxin content exceeds $4\,\mu g\,kg^{-1}$ or $10\,\mu g\,kg^{-1}$ for nuts which are to be subjected to further processing before being sold. A proportion of nuts imported into the UK, especially peanuts, are contaminated with aflatoxin. In 1994, 3% of samples examined under a European surveillance program were found to exceed the UK limit. Nonetheless, such findings should be kept in perspective: The numbers are low and their significance in public health terms, relative to other diet-related risks, is small.

Role in the Diet

Nuts and seeds can make a useful contribution to the dietary intake of macronutrients, notably protein and unsaturated fatty acids, micronutrients, dietary fiber, and energy. Although these commodities play a relatively minor role in the average Western diet, they are more important in the diets of Western vegetarians, especially vegans. Even on a worldwide basis, the nutritional contribution of nuts and seeds is relatively small: Plant foods are estimated to supply around 65% of edible protein, but only 8% of protein and 4% of total dietary energy is estimated to derive from pulses, oil crops, and nuts (Young and Pellett, 1994).

In the UK, average weekly household consumption of nuts and their products, as recorded by the National Food Survey, is about 14 g per capita, with only 11% of households purchasing these commodities; there are no separate data for nuts eaten as out of home snacks. Data from the Dietary and Nutritional Survey of British Adults indicate that average weekly intake of people consuming unsalted nuts and nut mixes is 63 g per week, but again only 12% of the adults surveyed were consuming these commodities. Therefore, even for nutrients which are present in relatively large amounts in nuts, such as vitamin E, magnesium, and copper, these foods only provide about 1% of the average daily intake in the UK.

See also: **Dietary Fiber**: Physiological Effects and Effects on Absorption. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Folic Acid**. **Food Allergies**: Etiology. **Food Safety**: Mycotoxins. **Protein**: Quality and Sources. **Vegetarian Diets**.

Further Reading

Englyst HN, Bingham SA, Runswick SA, Collinson E, and Cummings JH (1988) Dietary fibre (non-starch polysaccharides) in fruit, vegetables and nuts. *Journal of Human Nutrition and Dietetics* **1**: 247–286.

Gibson RS (1994) Content and bioavailability of trace elements in vegetarian diets. *American Journal of Clinical Nutrition* **59**(supplement): 1223S–1232S.

Gregory J, Foster K, Tyler H, and Wiseman M (1990) *The Dietary and Nutritional Survey of British Adults*. London: HMSO.

Harland BF and Oberleas D (1987) Phytate in foods. *World Review of Nutrition and Dietetics* **52**: 235–259.

Hartmann HT, Flocker WJ, and Kofranek AM (1981) *Plant Science. Growth, Development, and Utilization of Cultivated Plants*. Englewood Cliffs, NJ: Prentice Hall.

Holland B, Unwin ID, and Buss DH (eds.) (1992) *Fruit and Nuts. The First Supplement to McCance & Widdowson's The Composition of Foods*, 5th edn. London: The Royal Society of Chemistry.

Macfarlane BJ, Bezwoda WR, Bothwell TH *et al.* (1988) Inhibitory effect of nuts on iron absorption. *American Journal of Clinical Nutrition* **47**: 270–274.

Ministry of Agriculture, Fisheries and Food (1994) *The Dietary and Nutritional Survey of British Adults—Further Analysis*. London: HMSO.

Morris ER (1993) Phytic acid. In: Macrae R, Robinson RK, and Sadler MJ (eds.) *Encyclopaedia of Food Science, Food Technology and Nutrition*, pp. 3587–3591. London: Academic Press.

Moss MO (1996) Mycotoxins. *Mycological Research* **100**: 524–526.

Ryden P and Selvendran RR (1993) Phytic acid. In: Macrae R, Robinson RK, and Sadler MJ (eds.) *Encyclopaedia of Food Science, Food Technology and Nutrition*, pp. 3582–3587. London: Academic Press.

Young VR and Pellett PL (1994) Plant proteins in relation to human protein and amino acid nutrition. *American Journal of Clinical Nutrition* **59**(supplement): 1203S–1212S.

O

OBESITY

Contents

Definition, Etiology and Assessment

A Pietrobelli, Verona University Medical School, Verona, Italy

Obesity is a situation of excess body fat accumulation, and a clinical diagnosis of obesity should be based on an accurate direct or indirect measure of total body fat. The most widely used measurement to define obesity in adults is body mass index (BMI; weight in kg/height in m^2). It is a predictor of body fat from a population perspective, but it has limitations on an individual level and is only a proxy measurement of body fat. BMI shows significant variations during childhood; thus, age- and gender-specific reference standards must be used, and in adolescents the pubertal status should also be evaluated. An expert committee convened by the International Obesity Task Force (IOTF) in 1999 determined that although BMI is not an ideal measure of adiposity, it has been validated against other measures of body fat and may therefore be used to define overweight and obesity in children and adolescents. Because it is not clear at which BMI level adverse health risk factors increase in children, the group recommended cutoffs based on age-specific values that project to the adult cutoffs of 25 kg/m^2 for overweight and 30 kg/m^2 for obesity. Using data from six different reference population (Great Britain, Brazil, The Netherlands, Hong Kong, Singapore, and the United States), Cole and colleagues derived centile curves that passed through the points of 25 and 30 kg/m^2 at age 18 years. **Table 1** is useful for epidemiological research because children and adolescents can be categorized as non-overweight, overweight, and obese using a single standard tool.

There are differences in body composition across adult ethic groups, with one study of whites and Asians showing a difference of 2 or 3 BMI units in adults with the same body composition. It has been found that African American, Mexican American, and Mohawk Indian children carry more central fat than white children. Several studies have compared the US NHANES criteria for defining overweight or obesity using age- and gender-specific 85th and 95th percentile cutoffs with those of the Centers for Disease Control and Prevention (CDC) using similar percentile cutoffs and the IOTF alternative set of cutoffs based on centiles passing through BMI 25 and 30 at age 18 years. Using the NHANES III data, the different methods (i.e., NHANES/WHO, CDC, and IOTF) give approximately similar results but with some discrepancies, especially among younger children.

Definition

Obesity is an increase in body fat. This increase in fat can be evenly distributed over the body, or it can be concentrated in specific regions. Differences in body fat distribution are gender specific. Women tend to deposit fat more on their buttocks (gynoid distribution), and men tend to deposit fat on their waist (android distribution).

Table 1 International cutoff points for body mass index (BMI) for overweight and obesity by sex between 2 and 18 years

Age (years)	*BMI 25 kg/m²*		*BMI 30 kg/m²*	
	Males	*Females*	*Males*	*Females*
2	18.41	18.02	20.09	19.81
2.5	18.13	17.76	19.80	19.55
3	17.89	17.56	19.57	19.36
3.5	17.69	17.40	19.39	19.23
4	17.55	17.28	19.29	19.15
4.5	17.47	17.19	19.26	19.12
5	17.42	17.15	19.30	19.17
5.5	17.45	17.20	19.47	19.34
6	17.55	17.34	19.79	19.65
6.5	17.71	17.53	20.23	20.08
7	17.92	17.75	20.63	20.51
7.5	18.16	18.03	21.09	21.01
8	18.44	18.35	21.60	21.57
8.5	18.76	18.69	22.17	22.18
9	19.10	19.07	22.77	22.81
9.5	19.46	19.45	23.39	23.46
10	19.84	19.86	24.00	24.11
10.5	20.20	20.29	24.57	24.77
11	20.55	20.74	25.10	25.42
11.5	20.89	21.20	25.58	26.05
12	21.22	21.68	26.02	26.67
12.5	21.56	22.14	26.43	27.24
13	21.91	22.58	26.84	27.76
13.5	22.27	22.98	27.25	28.20
14	22.62	23.34	27.63	28.57
14.5	22.96	23.66	27.98	28.87
15	23.29	23.94	28.30	29.11
15.5	23.60	24.17	28.60	29.29
16	23.90	24.37	28.88	29.43
16.5	24.19	24.54	29.14	29.56
17	24.46	24.70	29.41	29.69
17.5	24.73	24.85	29.70	29.84
18	25	25	30	30

Modified from Cole TJ, Bellizzi MC, Flegal KM and Dietz WH (2000) Establishing a standard definition for child overweight and obesity worldwide: International survey. *British Medical Journal* **320**(7244): 1240–1243.

Measures of Body Fatness

Studies of adipose tissue distribution, its causes, and its effects on morbidity and mortality are fundamental in the field of obesity. An ideal measure of body fat should be "precise with small measurement error; accessible, in terms of simplicity, cost, and easy to perform; acceptable to the subject; well documented with published reference values." There is no consensus as to which methods best define and describe adipose tissue and its distribution. Several studies have noted that the increased risk of obesity is related to mesenteric and portal depots of adipose tissue. However, subcutaneous adipose tissue, particularly around the hips and buttocks, appears not to increase health risk.

The different methods used to estimate total fat and adipose tissue are discussed next and presented in order of decreasing accuracy.

Cadaver analysis The main use of cadaver studies is to validate other methods that can be used to study patients *in vivo*.

Imaging techniques Total adipose tissue and its distribution can be quantified using imaging techniques such as computed tomography (CT) and magnetic resonance imaging (MRI). Both methods produce high-resolution cross-sectional images from signals resulting from exposure of the subject to an X-ray source (CT) or electromagnetic field (MRI). Total body fat volume, total fat mass, and percentage fat mass can be estimated. In addition to providing total adipose tissue, imaging techniques are able to separate adipose tissue into subcutaneous, visceral, and intraorgan components. An accuracy of better than 1% error for body fat measurement is possible with these techniques.

Dual-energy X-ray absorptiometry This method is based on the principle that transmitted X-rays at two energy levels are differentially attenuated by bone mass and soft tissue mass, and the soft tissue mass is subdivided into fat mass and lean mass. Reproducibility of dual-energy X-ray absorptiometry (DXA) is approximately 0.8% for bone, 1.7% for fat, and 2.0% for body weight. One concern regarding DXA is whether changes in soft tissue hydration influence body fat estimates. A few studies have shown small but systematic and predictable errors in DXA soft tissue composition analysis with body fluid balance changes. Using DXA, it is possible to obtain abdominal fat estimates. Unfortunately, these cannot be separated into subcutaneous and visceral components.

Bioimpedance analysis This method measures the resistance in the body to an imperceptible electrical current. The measurement is based on the relationship among the volume of the conductor (body), the conductor's length (height), and its electrical impedance. Bioimpedance analysis assumes fat mass is anhydrous and that conductivity reflects fat-free mass. Conceptually, a human devoid of adipose tissue could have minimum impedance, and impedance would increase to a maximum when all lean tissue is replaced by fat/adipose tissue. This approach estimates total body water, which can be transformed using appropriate formulas and in turn can estimate fat-free mass and hence fat mass.

Anthropometric measurements Anthropometric measurement can be used to estimate total body fat, regional fat, and fat distribution. Anthropometric measures of relative adiposity or fatness are BMI, skinfold thickness, waist, hip, and other girth measurements. BMI is widely used as an index of relative adiposity among children, adolescents, and adults. The World Health Organization classifies a person with a BMI of 25 kg/m^2 or higher as overweight, whereas a person with a BMI of 30 kg/m^2 or higher is classified as obese. This measurement has low observer error, low measurement error, and good reliability and validity. However, BMI may not be a sensitive measure of fatness in subjects who are short, tall, or who have highly developed muscle. There may also be racial differences in the relationship between proportion of body fat and BMI.

The amount of subcutaneous fat can be estimated by measuring thickness directly using a skinfold caliper at different sites on the body. The sites most often used are the upper arm (biceps and triceps), under the scapula (subscapular), and above the iliac crest (suprailiac). Increasing the number of measurement sites reduces errors and corrects for possible differences in fat distribution among individuals within the same age and gender group.

Anthropometric methods are also applicable as 'surrogate' measurements of visceral adipose tissue. Circumferences are more reliable than skinfolds, and in recent years the most widely used anthropometric technique has been the waist circumference. Waist circumference is measured at the minimum circumference between the iliac crest and the rib cage using an anthropometric tape. It is an indirect measure of visceral adiposity, which is strongly correlated with risk for cardiovascular disease in adults and an adverse lipid profile and hyperinsulinemia in children.

Etiology

Obesity is a multifactorial disease. The relative contributions of genetics and the environment to the etiology of obesity have been evaluated in several studies. Approximately 30–40% of the variance in BMI can be attributed to genetics and 60–70% to the environment. Clearly, the interaction between genes and the environment is fundamental. In a given population, some subjects are genetically predisposed to become obese, but the genotype may be expressed only with adverse environmental situations (i.e., high-fat, energy-dense diets and sedentary lifestyle).

Development of obesity occurs when caloric intake is higher than energy expended. Three metabolic factors have been reported to be predictive of weight gain: a low sedentary energy expenditure, a high respiratory quotient, and a low level of physical activity. Resting metabolic rate is highly correlated with fat-free mass. Sedentary lifestyle has an impact on weight gain. Several other factors are also associated with overweight. Gender, age, race, and socioeconomic status could influence weight gain, with overweight and obesity being more likely among women, older subjects, minority races, those with a low socio-economic status, and those with low levels of education.

It is well-known that obesity runs in families. In fact, high birth weight, maternal diabetes, and obesity in family members are factors that may influence the degree of adiposity. For a subject, if one parent is obese, the odds ratio is approximately 3 for obesity in adulthood, and if both parents are obese, the odds ratio increases to 10. There are critical periods of development for excessive weight gain. The duration of breastfeeding was found to be inversely associated with risk of being obese later in life, possibly mediated by physiologic factors present in human milk. Adolescence is another critical period for development of obesity. The risk of obesity persisting into adulthood is higher among obese adolescents than among younger children.

Lifestyle changes during the past several decades have affected childhood patterns of physical activity as well as diet. Leisure activity (e.g., television and computer games) is increasingly sedentary, and there is generally a decreased amount of routine physical activity. Taken together, these factors play a potential role in the development of the overweight epidemic.

Assessment for Therapy

One of the goals of assessment of overweight/obesity is to decide whom to treat. Three main issues must be evaluated: whether treatment is indicated, whether treatment is safe for the patient, and whether the patient is ready and motivated to lose weight. In addition, routine assessment of eating and activity patterns in adults as well as in children must be considered. Recognition of excessive weight gain relative to linear growth is essential throughout childhood.

Proper identification and classification of obesity through body composition assessment are important steps to initiate before beginning weight-loss treatment. Dietary management, physical activity,

surgery, pharmacotherapy, and psychological and familial support must be considered together as part of obesity assessment. Before beginning a weight-loss program, patients should be evaluated for number and severity of cardiovascular risk factors. These conditions may require that treatment be initiated along with weight-loss strategies.

See also: **Adolescents**: Nutritional Requirements. **Body Composition**. **Children**: Nutritional Requirements. **Dietary Intake Measurement**: Methodology; Validation. **Obesity**: Fat Distribution; Childhood Obesity; Complications; Prevention; Treatment.

Further Reading

American Academy of Pediatrics (2003) Policy statement. Prevention of pediatric overweight and obesity. *Pediatrics* **112**: 424–430.

Anonymous (2004) Who pays in the obesity war. *Lancet* **363**: 339.

Aronne LJ (2002) Obesity as a disease: Etiology, treatment, and management consideration for the obese patient. *Obesity Research* **10**(S2): 95–130.

Dietz WH (2004) Overweight in childhood and adolescence. *New England Journal of Medicine* **350**: 855–857.

Eckel RH (2003) Obesity: A disease or a physiologic adaptation for survival. In: Eckel RH (ed.) *Obesity Mechanisms and Clinical Management*, pp. 3–30. Philadelphia: Lippincott Williams & Wilkins.

Friedrich MJ (2002) Epidemic of obesity expands its spread to developing countries. *Journal of American Medical Association* **287**: 1382–1386.

Hedley AA, Ogden CL, Johnson CL *et al.* (2004) Prevalence of overweight and obesity among U.S. children, adolescents, and adults, 1999–2002. *Journal of the American Medical Association* **291**: 2847–2850.

Heshka S and Allison DB (2001) Is obesity a disease? *International Journal of Obesity* **25**: 1401–1404.

Lobstein T, Baur L, and Uauy R (2004) Obesity in children and young people. A crisis in public health. *Obesity Reviews* **5**(S1): 1–104.

Ogden CL, Flegal KM, Carroll MD, and Johnson CL (2002) Prevalence and trends in overweight among U.S. children and adolescents, 1999–2000. *Journal of the American Medical Association* **288**: 1728–1732.

Peskin GW (2003) Obesity in America. *Archives of Surgery* **138**(4): 354–355.

Pietrobelli A and Heymsfield SB (2002) Establishing body composition in obesity. *Journal of Endocrinological Investigation* **25**: 884–892.

Pietrobelli A, Heymsfield SB, Wang ZM, and Gallagher D (2001) Multi-component body composition models: Recent advances and future directions. *European Journal of Clinical Nutrition* **55**(2): 69–75.

Pietrobelli A and Steinbeck KS (2004) Paediatric obesity. What do we know and are we doing the right thing? *International Journal of Obesity* **28**: 2–3.

Roux L and Donaldson C (2004) Economics and obesity: Causing the problem or evaluating solutions? *Obesity Research* **12**: 173–179.

Fat Distribution

J Stevens and K P Truesdale, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

In 1956, the French physician, Jean Vague, noted that an upper body, or masculine, fat distribution was associated with adverse health consequences. It has now been clearly demonstrated that obesity-related chronic diseases are associated with the location, as well as the amount, of adipose tissue on the body. Although the relative importance of total adiposity versus type of adiposity continues to be debated, the notion that an 'apple-shaped' (or android) body is associated with greater obesity-related health risks than a 'pear-shaped' (or gynoid) body is well accepted (**Figure 1**). Imaging techniques such as computed tomography (CT) allow measurement of visceral adipose tissue and layers of subcutaneous fat. Anthropometric studies do not provide precise measures of fat depots but nevertheless have provided clues to the causes and consequences of differences in fat distribution. Guidelines are being developed for the use of anthropometric assessments of fat distribution in clinical and public health settings.

Measurement of Fat Distribution

Fat patterning, the distribution of fat, is measured using either imaging or anthropometric

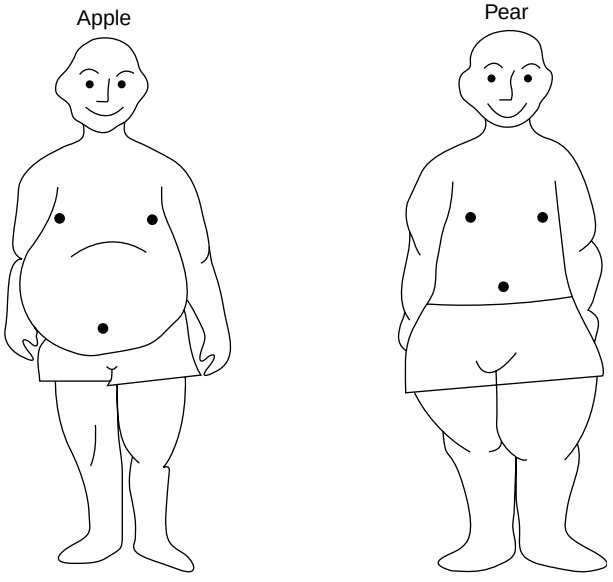


Figure 1 The apple (android) and pear (gynoid) body shapes.

techniques. Measurement has focused on assessment and differentiation of subcutaneous and intraabdominal (visceral) depots; however, recently measurement of fat residing in muscle has become of interest. Imaging techniques have the advantage of providing separate measurements of fat in these three different depots, but they remain too expensive for use in most clinical and community settings. Anthropometric measurements cannot provide a direct assessment of the amount of fat in different depots, but they can provide variables that correlate with assessments from imagining techniques and are quick, inexpensive, and noninvasive.

Imaging Techniques

Computed tomography and magnetic resonance imaging (MRI) are considered the most precise methods for measuring body fat distribution. MRI has the advantage of not exposing subjects to radiation. Dual energy X-ray is primarily used to measure bone mineral content and total body fat. This technique can measure total abdominal fat, but it cannot differentiate between visceral and subcutaneous fat.

Figure 2 shows two different cross-sectional images of the abdomen obtained by MRI. These images are constructed from 256×256 pixels, which vary from white to black with different shades of gray. Each pixel represents $2.4\,mm^2$. The fat regions are depicted as the lighter portions of the images. The subcutaneous fat area delineates the perimeter of the abdomen, whereas the visceral area is contained within the subcutaneous area. **Figure 2A** represents a cross section of an abdomen with a relatively small subcutaneous fat area in comparison with an enlarged visceral fat area. **Figure 2B** shows a subject with a small visceral fat-to-subcutaneous fat depot ratio.

CT scans have shown that approximately 12% of fat in normal weight subjects is among and inside muscles. Some researchers advocate considering this fat in a separate compartment, which, from a metabolic standpoint, is more closely related to visceral fat. Some researchers have suggested that subcutaneous fat be separated into deep and superficial layers separated by the 'fascia superficialis.'

Anthropometric Techniques

Anthropometric indices used to measure fat patterning include skinfold thicknesses, circumferences, sagittal diameter, and ratios such as waist-to-hip, waist-to-thigh, waist-to-height, and subscapular-to-triceps skinfolds. Skinfold thicknesses and skinfold ratios have not been found to be very well correlated with metabolic measurements or with visceral fat and are not recommended for use as indicators of fat patterning. Numerous equations using combinations of anthropometric measurements to predict the amount of visceral fat have not offered substantial improvement over the simpler measurements, and an accurate equation has yet to be developed.

Waist circumference (WC) alone and waist-to-hip ratio (WHR) are the most popular anthropometric methods used to measure fat distribution in both clinical and community settings. Both measures are correlated with visceral fat, with a correlation coefficient (r) generally ranging from 0.5 to 0.8. It is problematic that there is no uniform method of defining the location at which the waist and hip measurement should be assessed (**Table 1**). Waist

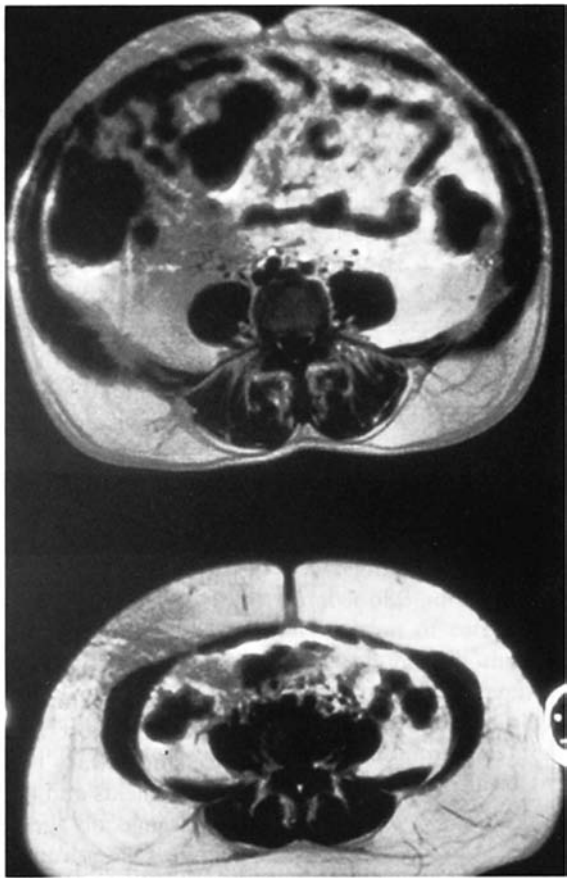


Figure 2 Cross-sectional images of the abdomen obtained by MRI. (A) Small subcutaneous fat area and enlarged visceral fat area. (B) Small visceral fat area in comparison with subcutaneous fat depot.

Table 1 Anatomical locations used to measure waist and hip circumferences

Waist circumference
One-third between the xyphoid process and umbilicus
Narrowest part of torso
Midway between xyphoid process and umbilicus
Midway between lower rib and iliac crest
One inch (~2.5 cm) above umbilicus
Level of umbilicus
Level of iliac crest
Immediately below the lowest rib
Immediately above the iliac crest
Hip circumference
Largest horizontal circumference around the buttocks
Level of iliac crest
Maximal circumference between superior border of iliac crest and thigh region 4 cm below superior iliac crest

circumferences measured at four sites (immediately below the lowest rib, at the narrowest point, midpoint between the lowest rib and the iliac crest, and immediately above the iliac crest) have been compared and found to differ from each other. Other work has shown that the highest correlations with risk factors were obtained when WHR was calculated as the waist measured at the point midway between the lower rib margin and iliac crest (approximately 1 inch (~2.5 cm) above the umbilicus) or when the waist was measured at the umbilicus and hips measured at the widest point of the buttocks. Although two different waist measurements have been demonstrated to perform equally well, the bony landmark measurement (the point midway between the lower rib margin and iliac crest) may be preferred since the umbilicus may shift position when an individual gains or loses weight. The World Health Organization (WHO) has recommended measuring the waist at the midpoint between the lowest rib and the iliac crest, whereas immediately above the iliac crest is the site recommended by the National Institutes of Health (NIH).

Sagittal diameter, the height of the abdomen measured with the subject in the supine position, can be measured anthropometrically or by imaging. **Figure 3** shows a technique for the anthropometric measurement of sagittal diameter using a caliper. Measurement is usually taken at the largest supine anteroposterior diameter between the xyphoid process and umbilicus. Some studies have found sagittal diameter to be a better indicator of visceral fat than WHR. Correlations between sagittal diameter and amount of visceral fat range from $r = 0.51$ to $r = 0.87$, with higher correlations occurring when sagittal diameter is measured using imaging techniques. In general, correlations tend to be higher in men than women.

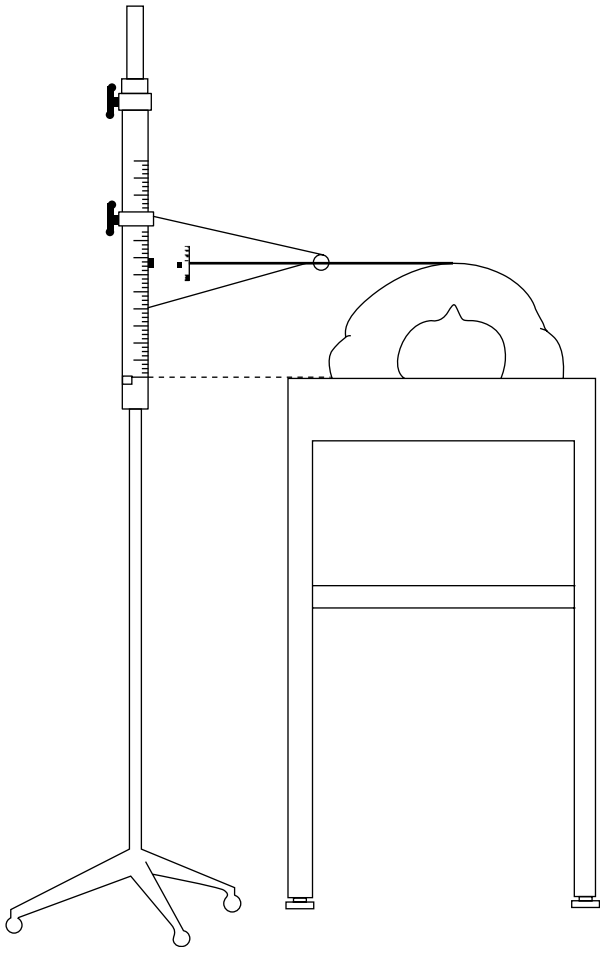


Figure 3 Sagittal diameter measured anthropometrically using calipers.

Metabolic Characteristics of Visceral and Subcutaneous Fat

The main function of adipose tissue is to store and break down fat based on energy excess or need, respectively. The uptake of fat is regulated by the enzyme lipoprotein lipase (LPL). This enzyme hydrolyzes triacylglycerols into free fatty acids, which can then be transported into the adipocyte and reesterified for storage. Greater LPL activity is associated with greater accumulation of fat. In premenopausal women, its activity is higher in the gluteal–femoral adipose areas than in the abdominal areas. The opposite is true in men, in whom LPL activity is the same or higher in the abdominal adipose areas than in the gluteal–femoral regions.

The breakdown of fat (lipolysis) is regulated by the enzyme hormone-sensitive lipase (HSL). This

enzyme releases free fatty acids, which are then released into the bloodstream and taken up by tissues, with the exception of the brain and red blood cells, for energy use or storage. The rate of basal lipolysis is higher in gluteal–femoral fat tissue than in abdominal tissue in both men and women. This may be due to greater cell size in that region. In the abdominal area, basal lipolysis is higher in subcutaneous fat than in visceral fat. However, when stimulated hormonally, rates of lipolysis may differ between men and women. Lipolytic rates have been shown to be higher in the visceral compared to the subcutaneous region in men, whereas the opposite trend is seen in women.

Regulators of Lipolysis and Fat Storage

The processes of lipolysis and fat storage are regulated by hormonal factors, which either enhance or suppress the activities of HSL and LPL. Through the action of glucocorticoid receptors, glucocorticoids enhance LPL activity and promote abdominal deposition of fat. The density of glucocorticoid receptors is greater in the visceral abdominal depot than in the subcutaneous abdominal depot. Therefore, an increase in glucocorticoid secretion is associated with increases in abdominal fat deposition compared to other fat depots.

Insulin favors fat storage by increasing LPL and decreasing HSL activity. Insulin has stronger antilypolytic effects in adipose located in the abdominal region compared to the femoral regions in both men and women. Paradoxically, insulin binding is stronger in the gluteal–femoral region than the abdominal region. Therefore, it has been hypothesized that insulin regulates lipolysis at the postreceptor level.

Catecholamines regulate lipolysis through α_2- and β-adrenoreceptors. The β-adrenoreceptors increase lipolysis, whereas the α_2-adrenoreceptor inhibits it. Although both the α_2- and the β-adrenoreceptors coexist in adipose tissue, they are regionally specific such that there may be an excess of one type of receptor relative to the other in various adipose regions. The lipolytic effect of catecholamines is 10–20 times greater in the abdominal region than in the gluteal–femoral region, as marked by a twofold increase in the number of β-adrenoreceptors in both sexes. The lipolytic action of catecholamines is more pronounced intraabdominally than in the abdominal subcutaneous tissue. Sex differences are displayed with the α-adrenoreceptor. Although the number of receptors is similar in both sexes, the sensitivity of the receptors is reduced by a factor of 10–15 in the abdominal compared with the gluteal–femoral region.

Sex hormones, such as estrogen, testosterone, and progesterone, also affect the balance of fat accumulation/mobilization, although their effects vary in men and women and the mechanisms are not clearly understood. Studies show that estrogen decreases LPL expression and activity in adipose tissue. It has been shown that testosterone stimulates lipolysis by increasing the number of β-adrenoreceptors. Estrogen and progesterone, on the other hand, stimulate fat storage and inhibit lipolysis, preferentially in the gluteal–femoral area compared to the abdominal area.

An increased androgenic profile is associated with upper body fat accumulation in women, but studies on men are conflicting. Significant inverse associations between fat distribution and testosterone have been found in population studies on men. Reduced visceral fat has also been observed when testosterone treatment was administered to men. These findings challenge the hypothesis that an androgenic hormone profile contributes to a more 'male type' of fat pattern and the associated metabolic sequelae.

The controversy over the effect of sex hormones on fat distribution is complicated by the metabolism of sex hormones. Sex hormone-binding globulin (SHBG) binds circulating testosterone and estrogen. Decreased SHBG concentration may be associated with an android shape. Therefore, studies need to distinguish between total circulating and unbound sex hormones and SHBG.

Sequelae of Altered Metabolism in Visceral Fat

Intraabdominal adipose tissue has metabolic characteristics that are different from those of adipose tissue from other sites. These differences seem to be most pronounced in the regions that are drained by the portal circulation. These 'portal adipose tissues' have a sensitive system for the mobilization of free fatty acids due to a preponderance of β-adrenergic receptors and little α-adrenergic inhibition.

The hypothesis has been advanced that the heightened responsiveness of intraabdominal fat to lipolytic agents results in increased lipolysis with venous drainage of the released free fatty acids directly to the liver. These fatty acids may contribute to increases in triacylglycerol synthesis and hyperinsulinemia secondary to decrements in insulin degradation. Hyperinsulinemia could produce insulin resistance and eventually type 2 diabetes in susceptible individuals. However, the hypothesis that increased release of free fatty acids from intraabdominal adipose tissue leads to insulin resistance through effects on the liver lacks supporting evidence *in vivo*.

The proposed mechanism of action of fat patterning on metabolic syndrome is linked to

hyperinsulinemia. Hyperinsulinemia may lead to increased blood pressure through increased sympathetic stimulation of the vessels, heart, and kidneys. In addition, insulin resistance combined with a relative increase in androgenic activity may lead to an unfavorable lipid profile. In addition to the effects of free fatty acids on insulin and glucose, an increased visceral depot decreases the activity of LPL. This causes an increase in very low-density lipoprotein (VLDL) secretion and a decrease in its catabolism. The production of high-density lipoprotein (HDL) therefore decreases, the transfer of lipids (i.e., VLDL to LDL and HDL) increases, and an enrichment of triacylglycerols results.

In obesity and type 2 diabetes, there is an increased content of lipids within and around muscle fibers. Researchers have suggested that the accumulation of triaclyglycerols within the skeletal muscle may play an important role in insulin resistance. In obese individuals with elevated amounts of visceral adipose tissue, there is a strong correlation between visceral adipose tissue and insulin resistance independent of subcutaneous (abdominal and nonabdominal) adipose tissue and cardiovascular fitness. It has been suggested that the discrepancies in the literature regarding the independent effect of visceral or subcutaneous adipose tissue on insulin resistance are due to the large variations of abdominal obesity within the study populations.

Leptin is a hormone that is produced in the adipose cells and can act on the hunger center in the hypothalamus to reduce hunger and appetite and thereby lower food intake. Plasma leptin levels are correlated with body fat. Researchers have discovered a leptin receptor gene that is responsible for obesity due to the mutation or absence of the gene. This condition is extremely rare in humans. In general, in obese humans the leptin levels are elevated (hyperleptinemia).

There is a progressive increase in plasma levels during puberty in girls due to the increase in body fat during this period and in response to the effect of estrogens. Circulating leptin levels tend to decrease in response to testosterone in boys, thus resulting in higher plasma leptin levels in women compared to men. Leptin levels are also affected by insulin and glucocorticoids.

Correlates and Possible Determinants of Fat Distribution

A large number of studies have examined correlations between fat distribution and genetic, behavioral, and physiological variables. Many factors, including heredity, overall fatness, gender, age, smoking, alcohol consumption, physical activity, and ethnicity, are associated with either an android or a gynoid shape. The underlying reasons for the observed associations between these variables and fat patterning remain to be elucidated. Correlates of fat distribution are important to understand since they may confound relationships between fat patterning and physiological outcomes or morbidity or mortality outcomes. There is evidence that body shape and amount of visceral fat are partially determined by genetics. After eliminating effects of age and overall fatness, studies have shown that heritable factors can account for as much as 20–50% of the variability in waist-to-hip ratio.

Fat distribution becomes more central or android as overall fatness increases. The correlation between overall fatness and fat patterning indices ranges from $r = 0.5$ to 0.9, depending on which measure of fat patterning is used. The more obese an individual, the more difficult it is to measure the waist and hip circumferences and the higher the measurement error.

Fat distribution has long been known to vary by gender, with men more android (apple shaped) than women, who are more gynoid (pear shaped). Men have a higher WHR and significantly more intra-abdominal adipose tissue than women. During weight gain in normal weight men, fat is preferentially deposited abdominally in the subcutaneous and visceral regions—proportionately more in the upper compared to the lower abdomen. In men, little fat is deposited in the gluteal–femoral regions until they become obese. Women have a higher percentage of body fat and higher proportion of fat in the gluteal–femoral regions than men. The gender differences are sufficiently large that recommended cutpoints for indices of fat distribution must be gender specific.

Aging is accompanied by changes in both weight and fat distribution. The largest increase in body weight occurs between young adulthood and middle adulthood. Independent of weight gain, abdominal fat increases with aging. This increase tends to be most pronounced between young adulthood and middle age in men and between middle age and old age in women (related to menopausal status).

Although cigarette smokers tend to be leaner than nonsmokers, they have more central adiposity (as indicated by larger waist circumference and WHRs) compared to nonsmokers, after the effects of age and body mass index (BMI) are eliminated. Furthermore, WHR increases progressively with an increase in the number of cigarettes smoked daily. The WHR increases with increasing 24-h cotinine excretion, indicating that central fat accumulation is dependent on the dose of smoke inhaled. Some studies have found a more androgenic hormone profile

in cigarette smokers, although this finding has been inconsistent. Increased cortisol secretion, an endocrine response to stressors associated with upper body fat deposition, may explain some of the association between smoking, alcohol consumption, and fat distribution.

Although alcohol consumption has been postulated to be correlated with fat distribution, studies have been inconclusive. There is evidence that beer and spirits are associated with higher levels of WHR, whereas wine is not. Not only frequency but also intensity of alcohol consumption may be important. One study showed that frequency of alcohol consumption was inversely associated, but intensity was positively associated, with abdominal adiposity (measured by sagittal diameter), even after the effects of age, education, physical activity, smoking, and grams of alcohol had been controlled. After combining the effects of frequency and intensity, the high frequency (daily) but low intensity (<1 drink/day) group had the lowest sagittal diameter, whereas the low frequency (<weekly) but high intensity (>3 drinks/day) group had the largest sagittal diameter.

Physical activity is inversely correlated with fat distribution in both men and women. Negative associations exist between WHR and various sports and exercise indices after controlling for the effects of BMI, smoking, and education. There is evidence that activity may be associated with a preferential mobilization of abdominal fat. Endurance training has been shown to increase aerobic fitness and decrease body mass and fat mass. Resistance training results in an increase in fat-free mass and muscle strength.

Fat distribution varies by ethnicity. African Americans have less visceral fat than whites at the same BMI, whereas Asians have a larger percentage of visceral fat. At the same BMI, African Americans have greater bone density and muscle mass than whites. Asians have smaller body frames, less muscle mass, and a larger percentage of fat mass at the same BMI as African Americans and whites. A study comparing migrant and British-born South Asian women to a general population of women in Scotland found that after controlling for the effect of age, migrant South Asians had larger waist circumference and WHR. However, after also controlling for physical activity, cigarette smoking, alcohol consumption, and parity, only WHR remained different between the two groups.

Fat Distribution and Disease Risk

Numerous studies have examined associations between fat patterning and mortality and morbidity. Since fat distribution is correlated with age as well other risk factors for disease, such as smoking, alcohol consumption, physical activity, and menopause in women, it is important to control for the effects of these variables in order to obtain an estimate of the independent effect of central obesity on morbidity. The impact of some of these correlates of fat distribution may be subtle and unlikely to seriously distort relationships between fat patterning and disease. However, age, the ultimate risk factor for disease and death, is sufficiently highly correlated with fat distribution to result in substantial distortion. Similarly, cigarette smoking is related adequately strongly to fat patterning and to various diseases and outcomes to make analyses that do not adjust for smoking difficult to interpret.

The large correlation between fat patterning and overall adiposity also influences the interpretation of results, making it difficult to differentiate between the two effects. Some researchers compare the size of the correlation between fat distribution (usually measured as WHR) and total adiposity (usually measured as BMI) in an attempt to show the relative importance of each. Others examine effects within tertiles (or other categories) of BMI and WHR simultaneously or test for an independent effect of WHR or BMI in multiple regression models that include both variables. In the latter type of analysis, the associations of both WHR and BMI with an outcome can be greatly reduced or even disappear because of collinearity between the two measures.

Researchers have found positive correlations between fasting glucose, insulin, blood pressure, total cholesterol, LDL cholesterol, and triaclyglycerols using imaging techniques, sagittal diameter, waist circumference, and WHR in most, but not all, studies. Visceral fat and HDL cholesterol are inversely associated. The strength of the associations varies but tends to be largest for triaclyglycerols. Associations are reduced after controlling for BMI and age.

There is strong evidence to link waist circumference and WHR with the risk of developing type 2 diabetes, even after eliminating the effects of age, smoking, BMI, and other important correlates. An individual who is obese (>150% ideal body weight) and has an elevated WHR (>0.8) may have as much as a 10-fold increased risk for developing type 2 diabetes compared with an individual who is of normal weight (<120% ideal body weight) and has a low WHR (<0.72).

Elevated WHR has been positively associated with cardiovascular disease in some population studies, although not as consistently as diabetes. Scientists have recognized that several of the cardiovascular

disease risk factors, including abdominal obesity, cluster in individuals. This cluster of risk factors is referred to as metabolic syndrome. The other risk factors in metabolic syndrome are insulin resistance/glucose intolerance, dyslipidemia (high triaclyglycerols and low HDL cholesterol), and high blood pressure.

Applications

Waist circumference can be used to assess obesity-related health risks in public health and clinical settings. Because it consists of only one measurement instead of two, it introduces less measurement error than WHR. A large waist has been shown to reflect both generalized obesity and centralized body fat distribution, which suggests that waist circumference could replace both BMI and WHR as a simple indicator of the need for weight management. Also, waist circumference tends to be more highly correlated with visceral fat than WHR.

It has been shown that hip circumference alone is inversely associated with cardiovascular disease risk after controlling for age, BMI, smoking, and waist circumference. Therefore, some predictive information may be lost if hip circumference is not assessed. If an index of body shape, independent of total body fatness, is desired the WHR may be preferred over waist alone because it is less highly correlated with total adiposity. Waist-to-hip ratio is a widely accepted form of fat patterning assessment. It is a good predictor of disease and metabolic disorders, with an increasing WHR indicating increased risk. Cutpoints used to define elevated WHR range from 0.90–1.00 in men to 0.80–0.90 in women.

Guidelines for the use of waist circumference in combination with BMI have been issued by NIH and WHO. NIH guidelines use BMI cutoffs for an initial assessment of overweight and obesity and recommend waist circumference cutoffs as a supplementary indicator of health risk. Increased relative risk for the development of obesity-associated risk factors in most adults is predicted for adults within the BMI range of 25–35 when the waist is $\geq$102 cm (40 in.) in men and $\geq$88 cm (35 in.) in women.

Conclusions

The relationship of body fat distribution to metabolic abnormalities and disease has now been well recognized. Individuals with a more android than gynoid body shape tend to have a more adverse metabolic profile and an increased risk for type 2 diabetes and cardiovascular disease.

Although much has been learned about fat distribution, there are several issues that need further exploration. The question of whether body type can be changed through behavior needs to be more fully addressed. Standardized anthropometric measurements need to be established, and the implications of differences in fat distribution among ethnic groups need to be elucidated. A better understanding of the risks associated with fat residing in muscle is needed, and this information must be integrated with what is known about visceral and subcutaneous fat. Finally, more research is needed to identify mechanisms of action. An increased abdominal depot may not necessarily be the cause of metabolic disturbances but an effect of underlying genetic and endocrine abnormalities.

See also: **Adipose Tissue**. **Body Composition**. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology. **Exercise**: Beneficial Effects. **Hyperlipidemia**: Overview; Nutritional Management. **Lipids**: Chemistry and Classification. **Obesity**: Definition, Etiology and Assessment.

Further Reading

Aronne LJ and Segal KR (2002) Adiposity and fat distribution outcome measures: Assessment and clinical implications. *Obesity Research* **10**(supplement 1): 14S–21S.

Bosello O and Zamboni M (2000) Visceral obesity and metabolic syndrome. *Obesity Review* **1**: 47–56.

Deurenberg P, Deurenberg-Yap M, and Guricci S (2002) Asians are different from Caucasians and from each other in their body mass index/body fat per cent relationship. *Obesity Review* **3**: 141–146.

Jakicic JM, Donnelly JE, Jawad AP *et al.* (1993) Association between blood lipids and different measures of body fat distribution: Effects of BMI and age. *International Journal of Obesity* **17**: 131–137.

Kelley DE, Goodpaster BH, and Storlien L (2002) Muscle triglyceride and insulin resistance. *Annual Review of Nutrition* **22**: 325–346.

Lean MEJ, Han TS, Bush H *et al.* (2001) Ethnic differences in anthropometric and lifestyle measures related to coronary heart disease risk between South Asian, Italian and general-population British women living in the west of Scotland. *Internation Journal of Obesity* **2001**: 1800–1805.

Lissner L, Björkelund C, Heitmann BL, Seidell JC, and Bengtsson C (2001) Larger hip circumference independently predicts health and longevity in a Swedish female cohort. *Obesity Research* **9**: 644–646.

Molarius A and Seidell JC (1998) Selection of anthropometric indicators for classification of abdominal fatness—A critical review. *International Journal of Obesity* **22**: 719–727.

National Institutes of Health, National Heart Lung and Blood Institute (1998) Clinical guidelines on the identification,

evaluation, and treatment of overweight and obesity in adults. The Evidence Reports. *Obesity Research* **6**(supplement 2): 53S. [Available at www.nhlbi.nih.gov/guidelines]

Seidell JC, Kahn HS, Williamson DF, Lissner L, and Valdez R (2001) Report from a Centers for Disease Control and Prevention workshop on use of adult anthropometry for public health and primary health care. *American Journal of Clinical Nutrition* **73**: 123–126.

Smith SR, Lovejoy JC, Greenway F *et al.* (2001) Contribution of total body fat, abdominal subcutaneous adipose tissue compartments, and visceral adipose tissue to the metabolic complications of obesity. *Metabolism* **50**: 426–435.

Stevens J, Couper D, Pankow J *et al.* (2001) Sensitivity and specificity of anthropometrics for the prediction of diabetes in a biracial cohort. *Obesity Research* **9**: 696–705.

Turcato E, Bosello O, Francesco V Di *et al.* (2000) Waist circumference and abdominal sagittal diameter as surrogates of body fat distribution in the elderly: Their relation with cardiovascular risk factors. *International Journal of Obesity* **24**: 1005–1010.

Wang J, Thornton JC, Bari S *et al.* (2003) Comparisons of waist circumferences measured at 4 sites. *American Journal of Clinical Nutrition* **77**: 379–384.

Childhood Obesity

E M E Poskitt, London School of Hygiene and Tropical Medicine, London, UK

This article discusses obesity in children and adolescents with regard to prevalence, epidemiology, clinical features, and management/prevention. Obesity not associated with a recognized underlying clinical condition is the focus of this article since this represents far the majority of children with obesity. However, obesity associated with congenital or acquired medical conditions is discussed briefly. Management and prevention are also discussed.

Obesity is increasing in prevalence among children in virtually all developed countries. In the United Kingdom, 8.5% (twice the rate of 10 years ago) of 6-year-old and 15% (three times the rate of 10 years ago) of 15-year-old children are obese. Childhood obesity is also increasing in prevalence among the affluent in less well-developed countries. Since it has been estimated that in Western countries one-third of obese adults were obese in childhood, and since both adult and adolescent obesity carry significant risk of health complications, obesity in childhood is currently seen as a concern for families, communities, and nations.

Body Composition in Childhood and Definition of Childhood Obesity

Obesity is an excess of body fat. However, the percentage of body weight that is fat varies normally throughout childhood (**Table 1**). The infant is born with modest amounts of fat. More than 50% of the energy in breast milk comes from fat, and young infants lay down fat very rapidly so that in the 4 or 5 months that it takes a normal infant to double birth weight, the weight of fat in the body has tripled. By 6 months of age, infants are increasing weight-bearing activity and fat deposition slows relative to lean tissue growth. From 1 year onward, there is a natural process of slimming with less fat than lean tissue deposited so that the child of 5 years often has a lower percentage body weight as fat than at any other time in life. This is followed by the 'adiposity rebound,' when fat deposition accelerates only to slow again with the onset of the pubertal growth spurt in males. In pubertal girls, very brief slimming early in the female growth spurt is followed by vigorous fat deposition particularly around the breasts and hips.

Assessment of Overweight and Obesity in Childhood

Precise methods of estimating body fat are complicated and expensive. There are no accepted age-related 'norms' for percentage body weight as fat in childhood. For these reasons, anthropometric indices involving weight and height are widely used to estimate relative fatness. Such methods are relatively simple, noninvasive, well tolerated, and can be used in clinical practice and large population studies. However, they provide only indirect measures of fatness.

Weight can be related to height and age in various ways. Until recently, there was no consensus

Table 1 Percentage of body weight as fat at different ages in childhood

Age (years)	*% body weight as fat*	
Birth	11[a]	
0.3	25[a]	
1.0	24[a]	
	Males	*Females*
5	12.5[b]	15.3[b]
10	17.6[b]	16.0[b]
15	11.4[b]	23.3[b]

[a]Fomon SJ (1974) *Infant Nutrition. 2nd edition*, Philadelphia: WB Saunders, p. 69.

[b]Widdowson EM (1974) *Changes in body proportions and compositions during growth.* In Davis JA, Dobbing J. *Scientific Foundations of Paediatrics.* London: Wm Heinemann. pp. 152–63.

definition of overweight/obesity from weights in relation to height and age. In adults, body mass index (BMI; weight in kg/(height in m)^2) is used as a proxy for fatness. A BMI >25 kg/m^2 (overweight) is associated with a significant increase in the risk of mortality and with an increased prevalence of complications of obesity. In childhood, BMI varies with age in a nonlinear fashion. Since at different ages children tend to retain their growth positions in relation to those of their peers, the International Obesity Task Force has defined childhood overweight and obesity as those points on the BMI centile, or standard deviation, for age distribution charts that, if followed to the age 18, would meet the adult cutoff points for overweight and obesity (BMI, 25 and 30 kg/m^2). This definition involves no direct assessment of body fat or lean body mass for age. It needs evaluating against other evidence of excessive body fat and the prevalence of complications of obesity, particularly since it presumes a constant prevalence of obesity in childhood at every age, which clinically seems unlikely. Nevertheless, the method does allow the opportunity to compare relative fatness between different studies and to demonstrate changes in population distribution of BMI for age over time.

Waist circumference is widely used in adult assessment of obesity because high waist circumference is associated with increased abdominal fat and increased risk for the morbid complications of obesity in adult life. Consensus regarding cutoff points for normal waist circumference measurements in childhood has not been reached, but high (compared with age-related populations) waist circumferences do seem to predispose to developing obesity comorbidities.

Risk Factors for Childhood Obesity

There is no clear evidence that obese individuals eat more or exercise less than their nonobese peers. Methods of measuring energy intakes and outputs are not precise when used over time and in community settings. The range of normal requirements and normal basal metabolic rates is large and obscures the energy imbalances of individuals. However, for the individual, obesity occurs when energy intake (food) exceeds the energy expenditure (basal metabolism, physical activity, growth, counteracting infection, maintaining body temperature, and thermodynamic action of food).

Familial Obesity

Most studies from developed countries show that approximately 80% of obese children have at least one parent, and 40% have both parents, overweight or obese. Twin and adoption studies indicate that genetic factors play a role in this family predisposition to obesity, although lifestyles almost certainly also influence familial similarities in habitus. In most cases of familial obesity, there is no recognised genetic explanation or apparent Mendelian inheritance, suggesting a genetic susceptibility expressed in an obesogenic environment.

Socioeconomic and Environmental Deprivation

Although it is the affluent who tend to become obese in countries undergoing industrialization, it is children from socioeconomically deprived environments in Europe and from families in which child care and nurture are poor irrespective of income who show the greatest predisposition to obesity.

Early Feeding

There is no consistent evidence that breast feeding protects children from later obesity. Any associations between breast feeding and a low prevalence of obesity may simply indicate that both obesity and a low prevalence of breast feeding are common in socioeconomically deprived communities. Furthermore, breast feeding is not a passive process but one that involves maternal emotions and close mother–child contact. The process of feeding and recognizing readiness to feed may teach a mother subtle subconscious understanding of her child's needs. Thus, the process of breast feeding may have positive influences on mothers' attitudes to child nurture—attitudes that are less readily acquired through formula feeding. Likewise, studies of early weaning, although occasionally showing evidence of an association with later obesity, are certainly not consistent in finding relationships between weaning practices and later overweight. Early feeding studies can never be double-blind controlled, and differences may only reflect common aspects of nurture rather than specific effects of a particular infant feeding procedure.

Diet and Dietary Change

Studies from several countries suggest that the childhood obesity epidemic has developed despite secular trends toward lower energy intakes by children. These estimates may have failed to account for recent increases in food eaten outside the home in the United Kingdom and other countries. The eating habits of most families in industrialized countries have changed during the past 30 years in ways that seem likely to make it easy for individuals to overeat. Foods are readily available and children have money to buy them. Much advertizing of snack foods is aimed at children. Manufactured foods

have varied forms and packaging. Small differences in flavor or appearance may reduce the satiety effect usually associated with eating large amounts of the same food. Most snacks aimed at children and the well-advertised prepared-before-sale meals are energy dense and high in saturated fats, refined carbohydrates, and sugar. In addition, the portion sizes in restaurants and of confectionery items have increased. It is too easy to eat without being aware of energy intake.

Physical Activity

Trends to lower energy expenditure, as well as dietary change, must have significance for the development of obesity. Opportunities for vigorous physical activity in sport have declined in many schools and communities, but the increase in long periods of almost complete inactivity (such as when watching television) may be having greater effects on children's nutritional status than the loss of relatively brief periods of intense activity. Studies in the United States and Mexico indicate that in adolescent boys obesity increases in proportion to the hours spent watching television.

Characteristics of Obese Children

Children without Recognizable Pathology

Obese prepubertal children are relatively tall for age (many in the upper quartile and most in the upper half of the population distribution for height). Advanced growth may be associated with advanced maturity of bones (advanced bone age), early onset of puberty, and cessation of growth with only average stature in adult life. However, some children remain tall and obese into adult life, and others slim dramatically with the adolescent growth spurt. It is not clear whether obesity drives accelerated maturation or whether obesity is one manifestation of a predisposition to exuberant growth of both lean and fat tissue also expressed by early puberty.

Obesity Associated with Recognized Medical Condition

There are conditions in which obesity is part of a recognized genetic defect, clinical syndrome, or acquired pathological condition (**Table 2**). Together, these conditions account for only a very small

Table 2 Specific conditions associated with obesity in childhood

Conditions	*Inheritance*	*Clinical example*
Congenital conditions		
Congenital obesity	Single gene defect affecting leptin metabolism	Congenital leptin deficiency Leptin receptor defect Prohormone convertase-1 defect Melanocortin-4 receptor defect Peroxisome proliferators activated receptor POMC deficiency
Inherited syndromes associated with childhood obesity	Autosomal dominant Autosomal recessive	Biemond's syndrome Alstrom's syndrome Bardet–Biedl syndrome Biemond's syndrome (some) Carpenter's syndrome Cohen's syndrome
	X-linked recessive	Borjeson–Forssman–Lehmann syndrome
Inherited syndromes affecting mobility	X-linked recessive Polygenic inheritance	Duchenne muscular dystrophy Spina bifida
Inherited disorders of growth	Autosomal dominant	Achondroplasia
Chromosomal abnormalities	Deletion or uniparental disomy for q11–q13 fragment of chromosome 15	Prader–Willi–Labhart syndrome
	Trisomy 21	Down's syndrome
	Abnormalities of sex chromosomes	Klinefelter's syndrome Turner's syndrome
Acquired conditions		
	Hormonal abnormalities	Hypothyroidism Growth, hormone deficiency Cushing's syndrome Polycystic ovarian syndrome
	Hypothalamic damage	Hydrocephalus Meningoencephalitis
	Drug treatment	Steroid treatment Sodium valproate

proportion of obese children. With the exception of very rare single gene defects in leptin metabolism, obesity is a secondary feature in these conditions and presentation is usually for some other aspect of the condition. Single gene defects affecting leptin are associated with progressive gross obesity from early life and may respond with dramatic fat loss with leptin treatment. Where obesity is only a part of a spectrum of abnormalities, common associated features are short stature, developmental delay, and craniofacial and other bony abnormalities.

Chromosomal abnormalities are more frequent causes of a predisposition to obesity. Prader–Willi syndrome, due to deletion or uniparental disomy of part of the long arm of chromosome 15, is associated with characteristic facies, small hands and feet with tapering fingers, hypogonadism, early hypotonia, difficulty feeding, and initially failure to thrive. From the second year of life many of these children show voracious appetite, progressive obesity, and negative behavior (stealing food and refusing to follow a diet). Many also commonly have psychodevelopmental problems with moderate mental retardation that exacerbates the difficulties maintaining normal weight for height and age. Gross obesity commonly leads to early death associated with hypoventilation (Pickwickian syndrome) and/or complications of type 2 diabetes mellitus.

Down's syndrome children are also prone to develop obesity in late childhood and adolescence. This is generally unrelated to recognized pathophysiological explanations for the obesity, although the syndrome is associated with an increased incidence of autoimmune thyroiditis and hypothyroidism (which exacerbates obesity).

Obesity may be an associated feature of other pathology in childhood. Endocrine problems, such as hypothyroidism and Cushing's syndrome, lead to obesity, but linear growth retardation does also, which often draws attention to the problem before obesity is severe. Hypothalamic damage (e.g., hydrocephalus and meningoencephalitis) and problems leading to immobility (e.g., spina bifida and Duchenne's muscular dystrophy) may also predispose to obesity. Nonpathological childhood obesity is usually associated with normal intelligence, relatively tall stature before puberty, and no overt abnormalities, so brief assessment of growth, general health, and intelligence usually distinguishes obese children for whom investigation for possible underlying pathology is required.

Complications of Childhood Obesity

Childhood obesity used to be considered relatively free of serious medical complications compared to adult obesity, although psychological consequences were recognised as common. Today, many obese children and adolescents show evidence of significant pathophysiological changes. Thus, the increasingly gross obesity of children and adolescents in North America, western Europe, and some other affluent societies has become a matter of major public health concern.

Cosmetic Problems

Orthopedic problems Flat feet and knock knee, perhaps related to the excess weight and need to internally rotate the knees to accommodate fat thighs when bringing the legs together, are common and can lead to ungainly gait. Slipped upper femoral epiphysis is a more serious problem, which is particularly common in overweight young adolescents and may also be associated with hormonal abnormalities such as hypothyroidism.

Skin problems Intertrigo, seborrheic eczema, and thrush are common in the thick heavy skinfolds of severely obese children. Pink or pale cutaneous striae, distinct from the purplish striae resulting from thinning of subcutaneous tissues in Cushing's syndrome, are common on the abdomen and upper limbs and may be a source of embarrassment. Hirsutes (abnormal facial and body hair) occurs particularly in adolescent girls with polycystic ovarian syndrome, which is associated with obesity and insulin resistance. Acanthosis nigricans, a velvety, pigmented, thickening of the skin usually at the back of the neck, is another important marker for insulin resistance, affecting up to 90% of children with type 2 diabetes mellitus.

Psychological problems Some overweight/obese children maintain high self-esteem and have little concern about their body image. These children may excel in sports in which their excess weight and tall stature are advantageous. However, many obese and overweight children have low self-esteem, dissatisfaction with their body image, and difficulty with peer relationships. Often, they underachieve at school. For some obese children, psychological problems antedate the obesity. Low self-esteem and difficulty with peer relationships have led to withdrawal, inactivity, and seeking solace in food. For other obese children, however, obesity is the prime cause of their psychological problems. Studies using silhouettes of figures with different body builds show that most children perceive obese silhouettes very negatively, prefering those portrayed by slimmer figures as friends.

Severe Complications

Adult obesity The extent to which childhood obesity progresses to adult obesity depends on the ages of children and adults at the time of study, the severity of obesity, the duration of obesity, and the family history of obesity. In one study fewer than 20% of males younger than 17 years of age remained obese as 35-year-old adults, whereas 20–39.9% of females younger than 17 years of age were still obese at 35 years of age. The probability of being obese at age 35 increased with increasing age and increasing BMI in childhood at the time of study. Where there is a strong family history of obesity in adult life, it seems likely that the obese child will follow the family pattern. Progression from child to adult obesity still only accounts for a minority of obese adults, although with the increasing prevalence of childhood obesity, this may change since it is highly unlikely that equal proportions of fat and thin children become obese adults.

Type 2 diabetes mellitus and the metabolic syndrome Although hyperinsulinemia has long been recognized from research studies in obese children, overt type 2 diabetes mellitus has been considered a rarity in childhood until recently. Studies in the United States show that among grossly obese children, type 2 diabetes mellitus is now disturbingly common, not only in adolescence but also in children younger than 10 years old. The problem is less common, but certainly present, in Europe also. Although hyperinsulinemia seems most prevalent in obese children from the Indian subcontinent, hyperinsulinemia and overt type 2 diabetes mellitus are also described in obese Caucasian children. Seventy-five percent of UK children with type 2 diabetes mellitus are overweight and 50% have a family history of type 2 diabetes mellitus. Girls are proportionally more likely (3:2) to develop type 2 diabetes than boys.

The metabolic syndrome (insulin resistance syndrome; syndrome X) is a clustering of problems associated with resistance to insulin and/or hyperinsulinemia that includes obesity, high central (i.e., intra- and peri-abdominal) distribution of fat, hypertension, and dyslipidemia. Females with polycystic ovarian syndrome also show clustering of these features. The criteria for diagnosis of the insulin resistance syndrome in childhood have not been defined, but some obese children show clustering of extreme values for the parameters of the metabolic syndrome. Hypertension, hyperinsulinemia, and dyslipidemia in obese children are indications for vigorous intervention to prevent morbidity and early mortality.

Pickwickian syndrome Very severe obesity may be associated with hypoventilation and/or upper respiratory obstruction with sleep apnoea. (The sleepy fat boy in Charles Dicken's *Pickwick Papers* is the origin of the syndrome's name.) Underventilation leads to increased circulating carbon dioxide levels, which may precipitate pulmonary hypertension and right-sided heart failure. Rising circulating carbon dioxide levels may result in the respiratory centre of the brain ceasing to respond to carbon dioxide buildup and instead responding to falling oxygen levels as stimulus to breathe. Thus, if affected individuals are given oxygen because of increasing cyanosis, the stimulus to breathe may be removed with potentially disastrous consequences.

Management

Goals

Ideally, the goal of fat reduction in obesity should be to restore normal body composition and retain it for the rest of life. However, evidence suggests that morbidity and mortality are reduced with even small reductions in excess fat. Thus, loss of some excess fat and the pursuit of healthy eating and activity may be beneficial even if normal fatness is not restored. Parents and children need realistic guidance on achievable goals and on the time required to achieve them. Fat reduction programs should be sustainable, able to maintain normal linear growth, and follow overall healthy lifestyle practices.

For young children, it may not be necessary to lose weight since the normal rates of weight and height gain mean that keeping weight stationary while linear growth occurs allows children to grow 'into their weight.' However, most children presenting for help with obesity are so overweight that it would require years of static weights for current weights to decrease to normal for their heights. Gradual weight reduction should aim for weight losses of approximately 500–1000 g/month. Dramatic weight losses suggest excessive energy deficit with perhaps reduced lean tissue deposition, shorter adult height, and potentially reduced peak bone mass. The fattest children are unlikely to ever achieve normal BMI for age and normal fatness, but they need to be encouraged that significant fat reduction will improve their self-image, ability to exercise, and reduce late complications of obesity.

Dietary Management

Treatment must alter energy balance so that energy intakes are less than energy expenditures in metabolism and activity. Diets should be adequate for protein and micronutrients. They should aim to *change the quality and energy density* of the food eaten more than *reduce the quantity* of food eaten, although reducing

the amount of snacking will probably be appropriate (**Table 3**). There is no consistent evidence that reduction of any particular energy source is more effective than any other in promoting fat loss, so 'balanced diets' conforming to the 'healthy diet' principles of WHO and many governments should be followed.

Physical Activity

More time is spent in relatively minor activity than in strenuous physical activity, so policies that increase energy expenditure in activity must include reductions in sedentary 'activities.' People, not only children, tend to eat more when they are inactive

Table 3 Management of childhood obesity: dietary measures

Purpose of action	*Policy*	*Action*
Organize eating	Control number of eating events	Restrict eating to recognized meal and snack periods with perhaps two snacks only for children and three snacks for adolescents
	Eat meals, as a family whenever possible, at table rather than in front of the television	Where possible, eat meals prepared at home and served on a plate rather than ready-to-eat, microwaved individual meals
Be aware of the nutrient content of meals	Meals prepared at home	Where possible, prepare meals at home so that the cook at least is aware of the nutritional makeup of the meal
	Precooked/ready to eat meals	Read the nutritional information given on the packet and observe not only the content/100 g but also the weight (and thus nutrient content) of the food bought and fed to each member of the family
Reduce the energy content of the food intake	Portion sizes	Portion sizes can be reduced—using smaller plates may make this less obvious; avoid second helpings
	Change the form of food used to low-energy density versions	Use 'low-calorie' margarines, spreads, mayonnaise, yoghurts, soups, baked beans, etc. Use semi-skimmed milk, sugar-free fruit squashes, etc. Grill and bake and boil without added fat rather than frying foods
	Avoid added fats and sugars	Do not add fats to vegetables when preparing them for table Avoid (or reduce) added sugar to stewed fruit dishes; sweeteners dissolved in boiled water can be used instead if necessary
Reduce energy content of drinks	Fruit juices, etc.	Eat whole fruit rather than fruit juices (which are usually many fruits compressed and often with added sugar) Use 'low-calorie' fruit squashes Preferably drink water Avoid added sugar
	Tea, coffee, etc.	Try to avoid sweeteners so as to accustom child to less sweet tastes
Increase satiety	Increase intake of foods that require chewing, that take time to eat, or that increase satiety	Increase vegetable, salad, and fruit intake Increase whole-meal cereal intake Encourage 'jacket' potatoes, boiled potatoes rather than chips, crisps, and mashed potatoes
	Take more time over meals	Eat as a family when possible to allow social interaction during eating, slower eating, and thus greater sense of satiety after the meal

Table 4 Management of childhood obesity: increasing energy expenditure

Purpose	*Type of action*
Reduce sedentariness	Reduce time spent watching television
	Develop interests/hobbies that give children things to occupy them at home and that may give them activities outside the home
	Encourage children to participate in family life by helping parents around homes, doing simple domestic tasks, running up- and downstairs to fetch for other members of family, etc.
Increase activity in everyday life	Walk or cycle rather than go by car whenever possible
	Use public transport rather than car so at least have to walk to bus stop
	Use stairs rather than elevators and escalators when practical
	Walk up escalators
	Do short walking errands for family as much as possible
	Send child out into garden for activity when he or she comes home from school before doing homework, etc.
Increase family activity	Make a habit of going for walks, taking part in physical activity in garden or parks, etc. in leisure time
	Plan activities during holidays and weekends
Encourage and support child to participate in physical activity at school	Obese children may be very successful at swimming (but may be too self-conscious to wear bathing suit)
	Dancing and aerobics may be more acceptable than contact sports, especially for girls
Increase energy expenditure as heat	Reduce home heating a few degrees to increase need for energy to keep warm in cold weather
	Encourage family to become accustomed to relatively cool environments

and relaxing rather than when they are occupied and active. Keeping children from being bored or from spending their leisure time watching television, when food can be consumed almost unnoticed, should reduce eating opportunities. Overweight children should be encouraged to take up hobbies in order to keep their minds off eating. **Table 4** outlines how their physical activity can be increased without necessarily subjecting them to the often perceived misery of sports and gym (although these should also be encouraged). Embarrassment and fear of ridicule as well as the high energy expenditure required for activity on the sports field are exacerbated by mechanical difficulties associated with gross weight.

Television

It is important to reduce time spent watching television for most of these children. Energy utilization is very low when viewing, and much advertising is aimed at encouraging children to eat foods that are energy dense, high in fat, and of low satiety. Viewing as a family should be encouraged, with the television in the living room rather than in children's bedrooms, so parents are involved in their children's viewing and can advise on the significance and nature of advertizements. Viewing time should be limited, but wise negotiation rather than didactic action will probably be necessary to avoid intrafamily conflict. Indeed, parents should be involved in children's slimming regimens, particularly because so many parents are overweight. Many children who watch television express preference for other activities but indicate that they are not given the opportunities to participate in other activities. Children cannot be expected to implement slimming behavior if an obesogenic family lifestyle continues unchanged around them.

Very severe childhood obesity (particularly if accompanied by a life-threatening complication such as Pickwickian syndrome) may require more dramatic interference than described previously. Very low-energy diets have been used quite successfully for short-term weight reduction. However, such diets are intrusive, carry some risk for nutrition and growth, and unacceptable to many obese. No drugs are currently approved for treatment of obesity in childhood. Drug treatment has not been associated with notable successes in the past.

Prevention of Obesity in Childhood

The prevention of obesity involves creating lifestyle changes at the family, school, community, and national level. Initiatives need to be affordable and sustainable so that those most at risk of obesity are reached and feel ownership of community programmes. **Table 5** suggests changes needed to reduce the obesogenic factors in the current Westernized environment. If the obesity epidemic is to be halted, governments and international industries have to work with communities to bring about effective change.

Table 5 Possible national and community measures to reduce epidemic of childhood obesity in Western societies

Purpose	*Action*
Reduce snacking on energy-dense foods	Act to reduce all advertising of energy-dense foods to children
	Possibly ban advertising to children on television
	Remove sweetened drinks and confectionery dispensing machines in schools
	Review foods on sale at school
Increase children's and parents' knowledge of nutrient content of foods	Programmes to educate parents and children on interpreting nutrition labels on foods
	Consider indicating energy content of foods in terms of minutes/hours of activity necessary to balance energy intake from food
	Practical nutrition teaching in schools
Encourage intake of whole foods, fruits and vegetables, and home-prepared foods so there is more awareness of content of foods eaten	Consider subsidising fresh fruits/vegetables and whole-meal cereals and making them more accessible in deprived communities
	Teach families how to cook rather than purchase ready-to-eat meals
Reduce energy intakes generally	Review nutritional content of school dinners
	Develop policies to encourage and make consumption of whole foods, fruits, and cereals attractive and fashionable to children
Increase energy expenditure in activity	Increase play areas, safe parks, and playing fields in communities
	Consider opening school playing fields off hours and on holidays
	Develop safe integrated community transport systems so children can use public transport
	Develop bike paths
Increase energy expenditure in heat	Reduce environmental temperature of public places by a few degrees; encourage people to wear more clothes if they find this uncomfortable

See also: **Adolescents**: Nutritional Problems. **Appetite**: Psychobiological and Behavioral Aspects. **Breast Feeding**. **Children**: Nutritional Requirements; Nutritional Problems. **Diabetes Mellitus**: Etiology and Epidemiology. **Exercise**: Beneficial Effects. **Food Choice, Influencing Factors**. **Nutritional Assessment**: Anthropometry; Clinical Examination. **Obesity**: Definition, Etiology and Assessment; Fat Distribution; Complications; Prevention; Treatment. **Socio-economic Status**. **Weight Management**: Approaches.

Further Reading

Burniat W, Cole TJ, Lissau I, and Poskitt EME (2002) In *Child and adolescent obesity: causes and consequences; prevention and management*. Cambridge: Cambridge University Press.

Cole TJ, Bellizzi MC, Flegal KM, and Dietz WH (2000) Establishing a standard definition for child overweight and obesity worldwide: international survey. *British Medical Journal* **320**: 1–6.

Dietz WH and Gortmaker SL (1985) Do we fatten our children at the television set: Obesity and television viewing in young children and adolescents. *Pediatrics* **75**: 807–812.

Farooqi IS and O'Rahilly S (2000) Recent advances in the genetics of severe childhood obesity. *Archives of Disease in Childhood* **83**: 31–34.

Lissau I and Sorensen TIA (1994) Parental neglect during childhood and increased risk of obesity in young adulthood. *Lancet* **343**: 324–327.

Lobstein T, Baur L, and Uauy R (eds.) (2003) *Childhood obesity. The new crisis in public health*. Report to WHO. London: IASO International Obesity Task Force.

Power C, Lake JK, and Cole TJ (1997) Measurement and long term health risks of child and adolescent fatness. *International Journal of Obesity* **21**: 507–526.

Reilly C, Methven E, McDowell ZC, Hacking B, Alexander D, Stewart L, and Kelnar CJH (2003) Health consequences of obesity. *Archives of Disease in Childhood* **88**: 748–752.

Complications

A Ahmed and R L Atkinson, Obetech Obesity Research Center, Richmond, VA, USA

Obesity is a serious chronic disease associated with complications and comorbidities that involve most systems of the body (**Table 1**). The common factor in all obese people is the presence of excess adipose tissue stores and an increased percentage of body fat. Even in the absence of complications and comorbidities, obesity increases the risk of early mortality. It has been estimated that there are 300 000 obesity-related deaths in the United States each year. In addition to medical complications, obesity is associated with psychological and social problems that may overshadow the medical problems in the quality of life for many obese people.

Table 1 Complications of obesity

Metabolic complications
Metabolic syndrome
Type 2 diabetes
Insulin resistance, hyperinsulinemia
Dyslipidemia
Gout
Abnormalities of hormones and other circulating factors
– Growth hormone
– Hypothalamic–pituitary–adrenal axis
– Cytokines
– Renin–angiotensin system
– Leptin
– Ghrelin
Diseases of organ systems
Cardiac and vascular diseases
– Coronary heart disease
– Hypertension
– Congestive heart failure
– Cerebrovascular disease
– Thromboembolic disease
Respiratory system abnormalities
– Obesity–hypoventilation syndrome
– Sleep apnea
Digestive system abnormalities
– Gall bladder disease
– Hepatic disease
Reproductive system abnormalities
– Hormonal complications: males
– Hormonal complications: females
– Obstetric complications
Nervous system
– Pseudotumor cerebri
– Adiposis dolorosa
– Alzheimer's disease
Immune system dysfunction
Skin disease
Eye disease
Cancer
Breast
Uterus
Gallbladder
Colon
Prostate
Others
Mechanical complications of obesity
Arthritis
Increased intraabdominal pressure
Surgical complications:
Perioperative risks: anesthesia, wound complications, infections
Incisional hernias
Psychosocial complications
Psychological complications
Social complications
Economic impact

Role of Distribution of Body Fat in the Complications of Obesity

The distribution of excess adipose tissue contributes to the complications of obesity. Obese individuals may be classified as those whose excess fat is deposited in the upper body versus those with increased lower body obesity. Upper body obesity may be localized to the subcutaneous space versus the intraabdominal space (visceral fat). Waist circumference and the ratio of waist to hip circumferences correlate with the morbidity and mortality of obesity. Individuals with increased visceral fat, as measured by the cross-sectional area on computed tomography or magnetic resonance imaging, are at greater risk for systemic complications of obesity compared to people with fat localized to abdominal subcutaneous depots or to the lower body. The mechanisms of these differences are not clear, but research has shown that visceral fat has a higher triglyceride turnover rate and releases greater amounts of fatty acids into the circulation than do other adipose tissue depots. Since blood vessels from the visceral fat drain into the portal vein, some investigators postulate that exposure of the liver to high levels of free fatty acids produces insulin resistance, which is known to be correlated with many of the complications of obesity described here. There are significant racial differences in deposition of visceral fat. Asians and Hispanics tend to selectively deposit fat in the abdominal cavity with excess energy intake, whereas blacks have less visceral fat than other groups.

Metabolic and Organ System Complications of Obesity

Obesity is a syndrome that resembles premature aging. Multiple metabolic, hormonal, and organ system dysfunctions occur in aging. Similar changes occur in obesity, but at an earlier age. This section reviews generalized metabolic changes that occur with obesity and discusses individual organ systems.

Metabolic Syndrome

The term 'metabolic syndrome' has been given to a cluster of abnormalities that classically include insulin resistance, glucose intolerance, hypertension, and dyslipidemia. Several other abnormalities, such as sleep apnea, gout, and pseudotumor cerebri, have been associated with insulin resistance and the metabolic syndrome.

Type 2 Diabetes

A strong association of obesity with the prevalence of type 2 diabetes mellitus (DM) is well documented. The risk of developing type 2 DM increases with the

degree and duration of obesity—as much as 50-fold with severe obesity. The US National Diabetes Commission reported that the risk of diabetes doubles for every 20% of excess body weight. The risk of type 2 DM is greater with visceral obesity. Type 2 DM is frequently associated with other complications, such as hypertension and dyslipidemia, resulting in additive risks for atherosclerosis and cardiovascular disease. Poor glycemic control in type 2 DM may lead to severe microvascular complications, including nephropathy, retinopathy, and neuropathy. Weight loss is a very effective treatment for type 2 DM and can prevent the onset of type 2 DM in susceptible individuals. Type 2 DM, once very rare in children, has increased greatly in prevalence with the obesity epidemic.

Insulin Resistance and Hyperinsulinemia

'Insulin resistance' refers to the phenomenon of insensitivity of the cells of the body to insulin's actions. Different tissues may have different insulin sensitivities. For example, adipose tissue may be more sensitive to insulin than muscle tissue, thus favoring the deposition of fatty acids in adipose tissue and diminished fatty acid oxidation in muscle. Insulin resistance is usually associated with hyperinsulinemia. Hyperinsulinemia is an independent marker that predicts the development of atherosclerosis. A causal relationship between hypertension and hyperinsulinemia has not been well established. Hypertension associated with hyperinsulinemia could be due to increased renal sodium retention, increased intracellular free calcium, increased sympathetic nervous system activity, or increased intraabdominal pressure due to increased visceral fat deposition.

The mechanisms of insulin resistance with increasing obesity are not clear, but increased production of cytokines such as tumor necrosis factor-α (TNF-α) and interleukin-6 (IL-6) is thought to play a role. Basal insulin levels increase with the degree of overweight, perhaps due to increased insulin secretion and/or reduced clearance by the liver. A reduced receptor number and/or post–insulin receptor defects may play a role in insulin resistance. Both basal hyperinsulinemia and insulin resistance decrease with weight reduction.

Dyslipidemia

Obesity, particularly visceral obesity, is associated with increased serum levels of cholesterol, triglycerides, low-density lipoproteins (LDL), very low-density lipoproteins (VLDL), apolipoprotein B, and reduced levels of high-density lipoprotein (HDL) cholesterol. Every 10% increase in relative body weight is associated with a 12 mg/dl increase in serum cholesterol concentration. The correlation of serum cholesterol with body mass index (BMI = kg/m^2) is greater for men than women. Increased serum triglycerides with weight gain may be due to increased intake of fats, hyperinsulinemia, and impaired removal of triglycerides into tissues because of low levels of lipoprotein lipase activity. Insulin resistance promotes lipolysis and increased circulating free fatty acids, which enhance the formation of VLDL in the liver. Dyslipidemia contributes to increased atherosclerosis in obesity. Weight reduction usually reduces serum cholesterol and triglycerides, increases HDL cholesterol, and may reduce atherosclerosis.

Gout

Serum uric acid and the prevalence of gout correlate positively with BMI. High serum uric acid levels correlate with insulin resistance and an increased risk of atherosclerotic cardiovascular disease in obesity. Serum uric acid levels may temporarily increase with acute weight loss, but they usually decrease with large amounts of weight loss. The lower uric acid levels are maintained with continued weight loss.

Abnormalities of Hormones and Other Circulating Factors

Growth hormone Obesity is typically accompanied by a decrease in growth hormone (GH) levels and an increase in growth hormone binding protein levels. An inverse relation exists between GH levels and percentage fat mass. GH levels fall with increasing age. GH is released by the anterior pituitary and affects lipid, carbohydrate, and protein metabolism. GH also controls the rate of skeletal and visceral growth. GH is lipolytic in adipose tissue. Animal studies show enhanced catecholamine-induced lipolysis and increased β-adrenoreceptors in adipocytes of GH treated animals. The rises in GH after meals, with sleep, and in response to secretogogues such as arginine or levodopa are blunted in obese people. GH stimulates secretion of insulin-like growth factor-1 (IGF-1). However, IGF-1 is increased in obesity, suggesting a difference in sensitivity to GH. The defects in GH and IGF-1 are reversed by weight reduction.

The hypothalamic–pituitary–adrenal axis The hypothalamic–pituitary–adrenal (HPA) axis may be abnormal in obesity. Patients with Cushing's syndrome display a number of clinical features that resemble those seen in patients with the metabolic

syndrome, including abdominal obesity, insulin resistance, impaired glucose homeostasis, hypertension, and lipid abnormalities. These similarities led to the hypothesis that a dysregulation of the HPA axis in the form of functional hypercortisolism could potentially be a cause for abdominal obesity and its different metabolic consequences. High levels of emotional or physical stress are thought to increase cortisol secretion or turnover and thereby increase visceral obesity.

Another potential mechanism involves the peripheral metabolism of cortisol. The enzyme 11-β-hydroxysteroid dehydrogenase, which converts steroid precursors to cortisol, is expressed in adipose tissue. With increasing obesity, more cortisol is derived from cortisone in adipose tissue due to the increased activity of this hormone. Urine studies in obesity also show an increase in the ratio of tetrahydrocortisol to tetrahydrocortisone, indicating a relative increase in the pathways leading to cortisol formation.

Cytokines Adipose tissue secretes a number of cytokines, such as TNF-α and interleukins, which may play a role in fat metabolism and insulin resistance. TNF-α has been shown to alter basal and glucose-stimulated insulin secretion and to produce insulin resistance in isolated cell lines. Adipocytes also produce IL-6, -10, and -11, which stimulate C-reactive protein, a systemic marker of inflammation. All of these ILs are increased in obesity. IL-6 and its subsequent inflammation have been postulated to play an etiologic role in the increased risk of thromboembolism observed in obese patients. Adipose tissue is also capable of producing plasminogen activator inhibitor-1, which may play a role in the increased risk of thromboembolism. Plasma IL-8 is increased in normoglycemic obese subjects and is related to fat mass and to TNF-a levels. Circulating IL-8 is also acutely upregulated by hyperinsulinemia. An increase in circulating IL-8 may be one of the factors linking obesity with greater cardiovascular risk.

Renin–angiotensin system Several components of the renin–angiotensin system are expressed by the adipose tissue. Angiotensinogen levels are increased and have been linked to hypertension and increased cardiovascular risk in obesity.

Leptin Leptin, the product of the *ob* gene, is made predominantly in adipose tissue. Leptin receptors are present in the hypothalamus. Leptin was postulated to act as a signal from adipose tissue to the brain to regulate fat stores. However, serum leptin levels correlate positively with body fat stores and are higher in obese people. Females have higher serum leptin levels than males, but this association does not appear to be due to estrogen levels. Leptin is found in greater concentrations in abdominal subcutaneous fat compared to visceral fat. The mechanisms for these differences are not known, but it is possible that this may play some role in the differential metabolic responses of subcutaneous and visceral fat.

Ghrelin Ghrelin is a potent growth hormone secretagogue that is produced mainly by the stomach. Administration of ghrelin increases food intake, and ghrelin levels increase with dieting and weight loss. However, serum ghrelin has a negative correlation with percentage body fat, so levels in obese people are lower than in nonobese people.

Diseases of Organ Systems

Atherosclerotic and Arteriosclerotic Vascular Diseases

Diseases of the vascular system provide the greatest contribution to the increased mortality associated with obesity. In both sexes, the excess mortality due to vascular disease increases linearly with BMI greater than 25 kg/m^2. The vascular complications of obesity can be categorized into five major groups: coronary heart disease, hypertension, congestive heart failure, cerebrovascular disease, and thromboembolic disease.

Coronary heart disease Longitudinal studies show a positive correlation of BMI with coronary heart disease (CHD), and obesity is an independent predictor of CHD. However, in the presence of other risk factors, such as hypertension, high serum cholesterol and triglycerides levels, low serum HDL cholesterol levels, and insulin resistance, all of which are increased by obesity, the risk of atherosclerotic CHD increases dramatically. Weight loss reduces all of these risk factors associated with cardiovascular disease, but because long-term reductions in body weight have been difficult to achieve, there are few long-term studies of changes in cardiovascular mortality due to weight loss. A very low-fat diet (10% of total calories as fat) has been shown to reduce the size of atherosclerotic plaques in coronary arteries. Such low-fat diets almost invariably produce weight loss.

Hypertension The prevalence of hypertension among overweight adults in the United States is 2.9 times higher than that of nonoverweight individuals.

Every 10-kg increase in body weight is associated with an increase of 3 and 2 mm Hg in systolic and diastolic blood pressures, respectively. Persistent hypertension can contribute to the development of left ventricular hypertrophy, coronary ischemia, and stroke.

The etiology of the association between hypertension and obesity is unclear. The following are some of the mechanisms offered to explain the association between obesity and hypertension:

Hyperinsulinemia due to insulin resistance leading to increased renal reabsorption of sodium
Sodium retention due to a decreased renal filtration rate, increased intraabdominal pressure, and/or increased plasma renin activity
Increased sympathetic nervous system activity

Except in long-standing cases, weight reduction is usually accompanied by a decrease in blood pressure. The reductions in blood pressure with weight loss are not dependent on decreases in salt intake. Many studies have shown that even modest weight losses, in the range of 5–10% of initial body weight, may produce reductions or even normalization of blood pressure in obese individuals.

Congestive heart failure Total blood volume increases with excess body weight. Higher oxygen consumption in obesity and increased blood flow to the splanchnic bed and adipose tissue increase cardiac output. Also, the transverse diameter of heart, thickness of the posterior wall, and thickness of the interventricular septum increase with body weight. Left ventricular mass is a stronger predictor of morbidity and mortality than blood pressure. A combination of these factors may result in the congestive heart failure seen in severely obese people. The heart rate, stroke volume, blood volume, cardiac output, and left ventricular work return to normal with weight reduction. One study that compared weight loss by dieting to treatment with antihypertensive drugs demonstrated a greater improvement in cardiac hypertrophy with weight loss, despite similar reductions in blood pressure.

Cerebrovascular disease Obesity-related atherosclerosis and arteriosclerosis increase the risk of cerebrovascular disease and strokes. Obesity is an independent risk factor for strokes, even in the absence of other comorbidities.

Thromboembolic disease The risks of venous stasis, deep vein thrombosis, and pulmonary embolism are increased in obesity, particularly in people with abdominal obesity. Lower extremity venous disease may result from increased intraabdominal pressure, impaired fibrinolysis, and the increase in inflammatory mediators described previously.

Respiratory System

Obesity is associated with reduced lung volume, altered respiratory patterns, and an overall reduction in the compliance of the respiratory system, including a diminished vital capacity and total lung capacity. More severe obesity is associated with the 'obesity–hypoventilation syndrome,' which is characterized by excessive daytime sleepiness and hypoventilation. The increased work required to move the chest wall, a decrease in arterial oxygenation in the lungs, and a diminished sensitivity of the respiratory center to the stimulatory effect of carbon dioxide are postulated to contribute to the obesity–hypoventilation syndrome.

The obesity–hypoventilation syndrome may be associated with, or exacerbated by, obstructive sleep apnea, a syndrome characterized by repeated collapse of the upper airway and cessation of breathing with sleep. Obstructive sleep apnea occurs when the tongue obstructs the glottis and prevents entry of air into the trachea. Up to 50% of massively obese people have sleep apnea. The risk of arrhythmias and sudden death increases during apneic episodes. Weight reduction usually reduces the severity of sleep apnea, and massive weight reduction, such as that after gastric bypass surgery, eliminates the disease in most patients.

Digestive System

Gallbladder disease The risk of gallbladder disease, particularly gallstone formation, is increased in obesity and occurs with greater frequency in women. The prevalence of gallbladder disease in obese individuals increases with age, body weight, and parity. The etiology of increased gallstones is unclear, but genetic factors play a role. Increased cholesterol production, which leads to increased excretion of cholesterol in bile, is known to occur in obesity and correlates with increases in body weight. Many obese people skip meals and the reduced number of meals may result in less frequent emptying of the gallbladder. The resulting bile stasis may contribute to gallstone formation. Although long-term weight loss and maintenance may reduce the occurrence of gallbladder disease, the risk of gallstone formation actually increases during the active weight loss phase. The etiology of this increase is thought to be the mobilization of cholesterol from adipose tissue during rapid weight loss. This increased load of cholesterol in the circulation produces supersaturation of the bile,

leading to gallbladder sludge in approximately 25% of patients and to symptomatic disease in approximately 1–3%. Treatment with ursodeoxycholic acid reduces or eliminates the risk of gallstone formation during weight loss.

Hepatic disease Abnormalities in hepatic function are commonly reported in obese people. Fatty liver, due to increased concentrations of fatty acids, diglycerides, and triglycerides in hepatocytes, is reported in obese people. The frequency of fatty liver has been reported to be as high as 94% in very obese subjects. A small number of very obese subjects will develop micronodular cirrhosis. Abnormal liver enzymes on laboratory screening are very common in obese people and do not require further evaluation unless they are markedly elevated. Weight loss results in disappearance of the excess fat and normalization of the liver function tests.

Reproductive System

Hormonal Complications: Males Obese men have elevated levels of plasma estrone and estradiol that correlate with the degree of obesity. Plasma total testosterone and free testosterone (the biologically active moiety) are reduced in obese men, and the reductions correlate negatively with the degree of obesity. The reduced levels of free and total testosterone are not generally accompanied by hypogonadism or a decrease in libido, potency, or sperm count in obese men. Free and total plasma testosterone levels normalize upon significant weight reduction. Also, estrogen levels are normalized if individuals attain normal weight but not if the weight loss is modest and significant obesity persists.

Hormonal Complications: Females Obese women have normal levels of total plasma estradiol but reduced levels of sex hormone binding globulins (SHBG). Thus, free estradiol (the biological active moiety) is significantly elevated. The high levels of free estradiol are postulated to increase the risks of endometrial and breast cancer and to reduce fertility. Estrone, derived in adipose tissue from androgen precursors, is also increased in obesity. Obesity in women is associated with the polycystic ovary syndrome (PCOS), characterized by hyperestrogenism, hyperandrogenism, polycystic ovaries, oligomenorrhea or amenorrhea, hirsutism, and infertility. Women with PCOS also have insulin resistance and are at high risk for developing impaired glucose tolerance and diabetes mellitus. Weight loss usually normalizes SHBG and estradiol levels for individuals with simple obesity, but weight loss may not restore fertility to patients with severe PCOS.

Obstetric complications Obesity increases the risk of complications during pregnancy and child birth. Increased body weight, hypertension, and fluid retention during pregnancy can lead to toxemia of pregnancy. Heavier women have a longer duration of labor and a greater frequency of abnormal labor and caesarian sections.

Nervous system

Pseudotumor cerebri This syndrome is characterized by increased intracranial pressure, headaches, blurred vision or loss of vision, and papilledema. It is most common in massively obese individuals and may be seen in association with sleep apnea or with the obesity–hypoventilation syndrome. It may be associated with retinal hemorrhage or loss of vision from severe papilledema. Some investigators believe that increased intraabdominal pressure with massive obesity is an etiologic factor for pseudotumor cerebri. Major weight loss, particularly after obesity surgery, results in dramatic improvement.

Adiposis dolorosa This is a syndrome of unknown etiology characterized by pain in subcutaneous adipose tissue. Adiposis dolorosa occurs predominantly in postmenopausal women (female: male ratio of about 30:1) and has been described over all areas of the body. The painful areas of fat may occur as subcutaneous lumps on physical examination, but more commonly there are no differences from normal adipose tissue. The disease usually begins gradually with mild pain and tenderness of the area involved, but it may progress to severe pain, particularly with movement or exercise. Intravenous infusions of lidocaine are reported to relieve pain short term or even permanently. The mechanism involved in the relief of pain from lidocaine is unknown.

Alzheimer's disease Obesity has been linked to an increased prevalence of Alzheimer's disease. The etiology of this increase is unknown.

Immune System

Animal studies have shown an increased rate of infection and mortality in obese dogs compared to lean animals experimentally infected with canine distemper virus. Cell-mediated immune response is impaired in obese individuals. Maturation of monocytes into macrophages after *in vitro* incubation is significantly

less for obese compared to lean subjects. Impaired cell-mediated immune response in children was demonstrated to be due to subclinical deficiencies of zinc and copper. The impairment in the immune response was reversed after 4 weeks of zinc and copper supplements. As described previously, there are changes in numerous cytokines with obesity. The role of these changes in immune function is not clear.

Skin

Obese people may have several disorders of the skin. The most common is stasis changes of the skin of the lower legs in massively obese people. The etiology of this finding is venous stasis, edema, and breakdown of the skin. Fragilitas cutis inguinalis is a condition of fragile skin in the inguinal area of obese people. This condition is diagnosed by stretching the skin of the inguinal area. A linear tear appears at right angles to an applied force that is insufficient to tear the skin of a normal person. This condition is unrelated to the sex and age of the person.

Acanthosis nigricans, seen occasionally in obesity, is characterized by darkening of the skin in the creases of the neck, axillary regions, and over the knuckles. An association between acanthosis nigricans and insulin resistance is reported in people who have circulating antibodies to the insulin receptors. Since acanthosis nigricans also may be associated with highly malignant cancers such as intraabdominal adenocarcinomas, physicians should be alert to this possibility and not attribute the condition simply to the presence of obesity.

Eye Disease

Obesity is associated with an increased prevalence of cataracts. People with abdominal obesity are at greater risk than those with lower body obesity, insulin resistance may be involved in the pathogenesis of cataract formation, and diabetes is a well-known risk factor.

Cancer

Obesity increases the risk of cancers of the breast, colon, prostate, endometrium, cervix, ovary, kidney, and gallbladder. Studies have also found a somewhat increased risk for cancers of the liver, pancreas, rectum, brain, esophagus, and non-Hodgkin's lymphoma.

Although there are many theories about how obesity increases cancer risk, the exact mechanisms are not known. The mechanisms may be different for different types of cancer. Also, because obesity develops through a complex interaction of heredity and lifestyle factors, researchers may not be able to determine whether the obesity or other factors led to the development of cancer.

Mechanical Complications of Obesity

Arthritis

Obesity is frequently complicated by degenerative arthritis (DJD). Increased body weight leads to trauma of the weight-bearing joints and speeds the development of osteoarthritis in obesity. Knee and hip joints are particularly affected. However, obese patients have increased DJD of the hands, perhaps due to cytokines produced by adipose tissue, which may damage the cartilage in joints. Flattening of the arc of the planter surface of the feet (flat feet) occurs more frequently in obese people, presumably due to the stress of carrying excess body weight. Flat feet may lead to unsteady gait and aches and pains after walking. Increased fat deposition, particularly in the abdominal region, can change the natural curvature of the spine, causing lordosis and resulting in backache in obese people.

Intraabdominal Pressure

In severely obese people, the excess visceral fat is thought to increase intraabdominal pressure. Animal research shows that experimentally induced acute increases in intraabdominal pressure to the levels seen in the abdomens of very obese people cause increases in pleural pressure, intracranial pressure, and central venous pressure. The investigators postulated that in humans, increased intraabdominal pressure may contribute to hypertension, insulin resistance and type 2 DM, obesity–hypoventilation syndrome, pseudotumor cerebri, incisional hernia, and urinary incontinence. Massive weight loss following obesity surgery normalizes the increased intraabdominal pressure and reduces or eliminates all the symptoms listed previously.

Surgical Complications

Obese patients are at an increased risk of surgical and perisurgical complications, including an increased risk of complications and death from anesthesia, longer operating times, delayed wound healing, increased postoperative wound infections and pneumonia, and a higher frequency of incisional hernias after surgeries involving the abdominal wall. Many surgeons recommend weight reduction before elective surgery, but there are few data to document that acute weight reduction improves the outcome of surgery.

Psychosocial Complications

Psychological Complications

Obesity is associated with negative emotions, low self-esteem, decreased marital satisfaction, and body image disparagement. All of these conditions and beliefs show improvement with weight reduction.

Dieting efforts correlate positively with the prevalence of eating disorders, particularly binge eating. A correlation of eating disorders with abuse of drugs and alcohol has been shown. In strictly dieting female college freshmen who were not alcohol abusers at baseline, the frequency of alcohol abuse was reported to increase after 1 year compared to nondieters.

Social Complications

Obesity carries a social stigma that dramatically affects the quality of life for obese individuals, particularly for women. Factors contributing to the social bias against obese people are beliefs that obesity is due merely to overeating and therefore obese people must lack will power. Many members of the general public, and even health professionals, ignore the evidence for the genetic contribution to obesity, believe that obese people are responsible for their own plight, and believe that they do not deserve sympathy for their disability. Despite similar intelligence (as judged by IQ values and the Scholastic Aptitude Test scores), a significantly lower number of obese females were admitted to certain colleges compared to nonobese females. The choice of mates is adversely affected by obesity. Obese individuals tend to marry mates with less education and from a lower socioeconomic class. It is more difficult for an obese person to find a job or to be promoted once hired, so lower earnings and a lower socioeconomic status are correlated with obesity. Obese employees are viewed as less competent, less productive, inactive, disorganized, and less successful by employers.

The bias against obesity has been shown to begin in early childhood. Obese children are considered lazy, stupid, slow, and self-indulgent by both children and adults. Because of these societal attitudes, many obese children and adolescents have lower self-esteem than do their nonobese counterparts.

Economic Impact

In the United States, the direct cost of obesity has been estimated at more than $100 billion per year. The indirect costs of early retirement and increased risk for disability requiring financial support are also considerable. Because obese people have more health problems, health care costs for the obese are higher than for nonobese individuals.

See also: **Arthritis**. **Cholesterol**: Sources, Absorption, Function and Metabolism; Factors Determining Blood Levels. **Coronary Heart Disease**: Lipid Theory; Prevention. **Cytokines**. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Gout**. **Hypertension**: Dietary Factors. **Lipoproteins**. **Obesity**: Definition, Etiology and Assessment; Fat Distribution; Childhood Obesity; Prevention; Treatment.

Further Reading

Atkinson RL (1982) Intravenous lidocaine for the treatment of intractable pain of adiposis dolorosa. *International Journal of Obesity* **6**: 351–357.

Bjorntorp P (1993) Visceral obesity: A 'civilization syndrome.' *Obesity Research* **1**: 206–222.

Flegal KM, Carroll MD, Ogden CL, and Johnson CL (2002) Prevalence and trends in obesity among U.S. adults, 1999–2000. *Journal of the American Medical Association* **288**(14): 1723–1727.

Grundy SM and Barnett JP (1990) Metabolic and health complications of obesity. *Disease of Month* **36**: 641–731.

Klein S and Romijn JA (2003) Obesity. In: Larsen PR, Kronenberg HM, Melmed S, and Polonsky KS (eds.) *Williams Textbook of Endocrinology*, 10th edn, pp. 1619–1641. New York: Saunders.

Kottke TE, Lambert A, and Hoffman RS (2003) Economic and psychological implications of obesity epidemic. *Mayo Clinic Proceedings* **78**: 92–94.

Ousman Y and Burman KD (2002) *Endocrine Function in Obesity*. Available at http://endotext.com/obesity/obesity12/obesityframe12.htm.

Sugerman HJ, Felton WL 3rd, Salvant JB Jr, Sismanis A, and Kellum JM (1995) Effects of surgically induced weight loss on idiopathic intracranial hypertension in morbid obesity. *Neurology* **45**: 1655–1659.

Tataranni PA and Bogardus C (2003) Obesity and diabetes mellitus. In: Porte D Jr, Sherwin RS, and Baron A (eds.) *Handbook of Diabetes Mellitus*, 6th edn, pp. 401–413. New York: McGraw-Hill.

Zumoff B and Strain GW (1994) A perspective on the hormonal abnormalities of obesity: Are they cause or effect? *Obesity Research* **2**: 56–67.

Prevention

T P Gill, University of Sydney, Sydney, NSW, Australia

There can be little doubt that obesity has become a major public health and economic problem of global significance. According to World Health Organization (WHO) estimates, approximately 1 billion people

throughout the world were overweight in 2002 and more than 300 million of these were obese. Prevalence rates continue to rise rapidly in all areas of the world, including low-income countries, and obesity-associated illness are now so common that they are replacing the more traditional public health concerns, such as undernutrition and infectious disease, as the most significant contributors to global ill health.

The health impact of obesity is considerable, and obesity impacts on both quality and length of life. Overweight and obesity are associated with a wide range of chronic conditions, such as diabetes, hypertension, cardiovascular disease (CVD), and certain cancers, as well as non-life-threatening but painful conditions, such as arthritis, back pain, and breathlessness. Obesity also places enormous financial burdens on governments and individuals and accounts for a significant proportion of total health care expenditure in developed countries. Analyses suggest that obesity is fast approaching cigarette smoking as the major preventable cause of mortality.

In recent years, our understanding of the epidemiology and causation of obesity has improved dramatically and there is an acceptance that urgent action is required to address the problem. However, there are very few examples of successful, large-scale obesity prevention initiatives from any area of the world. Despite these limitations, sufficient understanding has been gained from smaller scale obesity prevention initiatives together with experiences from the management of other epidemics of noncommunicable diseases to allow effective planning and implementation of obesity prevention programs to proceed.

Principles of Obesity Prevention

Rational for Obesity Prevention

There are a number of reasons why prevention is likely to be the only effective way of tackling the problem of overweight and obesity. First, obesity develops over time, and once it has done so, it is very difficult to treat. A number of analyses have identified the limited success of obesity treatments (with the possible exception of surgical interventions) to achieve long-term weight loss. Second, the health consequences associated with obesity result from the cumulative metabolic and physical stress of excess weight over a long period of time and may not be fully reversible by weight loss. Third, the proportion of the population that is either overweight or obese in many countries is now so large that there are no longer sufficient health care resources to offer treatment to all. It can be argued, therefore, that the prevention of weight gain (or the reversal of small gains) and the maintenance of a healthy weight would be easier, less expensive, and potentially more effective than to treat obesity after it has fully developed.

Objectives of Obesity Prevention

There remains a great deal of confusion regarding the appropriate objectives of an obesity prevention program. It is often assumed that to be effective, any intervention to address the problem of excess weight in the community should result in a reduction in the prevalence of overweight and obesity. However, such an objective is unrealistic and may be counterproductive. Most communities are experiencing significant increases in the average weight of the population as a result of a sizeable energy surplus resulting from reduced energy expenditure combined with an increased energy intake. This is leading to rapidly escalating rates of overweight and obesity. To reverse this trend will require not only the removal of this energy surplus but also the creation of a negative energy balance that will need to be maintained by the whole population for a significant period of time. Few (if any) interventions are capable of reducing energy intake, or increasing energy expenditure sufficiently, or are sustained long enough and with sufficient reach to achieve this effect. More appropriate objectives would relate to a reduction in the level of weight gain or the maintenance of weight stability in adults and the achievement of appropriate growth and development in children. The achievement of these objectives would result in a slowing in the rate of increase, followed by stabilization and then an eventual decline in the level of overweight and obesity in the community.

However, even the goal of weight stability within a population may be difficult to achieve in the short term because it would require the maintenance or reestablishment of energy balance in time of significant energy surplus. Therefore, it may be necessary to identify more sensitive short- and medium-term outcomes to evaluate obesity prevention programs. Such process outcomes may relate to the achievement of appropriate changes in energy intakes or outputs, food or physical activity behaviors, or changes to the environment that are significant enough to positively impact upon the achievement of energy balance.

Importance of Weight Gain Prevention in Adults

There are a number of important reasons why it is preferable to focus on weight gain prevention as the key individual and population objective of obesity prevention initiatives in adults (**Box 1**). The association between elevated body mass index (BMI) and increased risk of ill health is clear and consistent.

Box 1 Why focus on weight gain prevention?

- Weight gain in adulthood carries an independent risk of ill health.
- Risk for chronic disease begins to increase from low BMI levels and significant weight gain can occur within normal limits.
- Extended periods of weight gain are difficult to reverse.
- Weight gain in adulthood is mostly fat gain.
- The relationship between absolute BMI and health risk varies with age and ethnicity but no such variations occur in the relationship between weight gain and ill health.
- A focus on weight gain prevention avoids exacerbation of inappropriate dieting behaviors.
- Weight maintenance can serve as a first stage goal for weight treatment programs.
- The message is equally relevant to all sections of the adult population.
- It avoids further stigmatization of people with an existing weight problem.
- It avoids reference to poorly understood terms such as 'healthy weight.'

Figure 1 Levels of obesity prevention intervention. (Adapted from Gill TP (1997) Key issues in the prevention of obesity. *British Medical Bulletin* **53**(2): 359–388.)

However, research has demonstrated that weight gain per se is also associated with increased health risk, and that this risk is independent of absolute BMI (provided a person is not underweight). A number of studies have shown strong relationships between weight gain and increasing levels of diabetes, hypertension, gall bladder disease, and coronary heart disease. Therefore, a large weight gain in a lean individual may carry equivalent risk to maintaining a stable but slightly elevated BMI in an overweight individual. The combination of an elevated BMI and ongoing weight gain, however, leads to greatly magnified levels of risk.

Who Should Obesity Prevention Strategies Target?

Deciding where to invest limited time and resources in obesity prevention is a difficult task but finite health resources make this a necessity. WHO has identified three distinct but equally valid and complementary levels of obesity prevention (**Figure 1**). The specific 'targeted' approach directed at very high-risk individuals with existing weight problems is represented by the core of the figure, the 'selective' approach directed at individuals and groups with above average risk is represented by the middle layer, and the broader universal or populationwide prevention approach is represented by the outer layer. This replaces the more traditional classification of disease prevention (primary, secondary, and tertiary), which can be confusing when applied to a complex multifactorial condition such as obesity.

Universal prevention is the domain of public health, whereas selective and targeted prevention are predominantly dealt with in community and health care service settings. Community settings include schools, colleges, worksites, community centers, and shopping outlets.

Whole Community

Overweight and obesity are public health problems of relevance to the whole community and require strategies that focus on populationwide change rather than attempting to address individuals or small groups in isolation of the community in which they live. An effective population strategy needs to both improve population knowledge about obesity and its management and reduce the exposure of the community to obesity-promoting factors in the environment. Action at a population level requires coordination at a central level and the investment of resources to be maintained over a long period of time to achieve population change.

Family Focus

There are numerous reasons why children should be a major focus of any obesity prevention strategy. There is strong evidence that a high proportion of overweight or obese children will become obese adults. Childhood obesity also has immediate effects on health, and weight-related conditions are becoming more prevalent and their effect more pronounced as the rates of childhood obesity increase. However, children grow rapidly and increase the level of lean body mass as they age, and so reducing or keeping fat mass constant allows the normalization of weight over time. Thus, childhood

(particularly younger children) is a period during which prevention efforts have a higher chance of success.

However, children also have little direct control over the environment in which they live. Parents and other caregivers mostly control decisions regarding the food available and the opportunities for activity. In addition, the behaviors of parents and other siblings have a profound effect on the diet and physical activity behaviors of children. For this reason, it is preferable to focus childhood obesity prevention efforts on the family environment rather than directly on children.

High-Risk Groups

There are a number of groups that appear to be at higher risk of developing overweight and obesity (**Table 1**). These groups warrant special attention and include the following:

- Those with a family history of weight problems
- Socially disadvantaged and isolated communities
- Certain ethnic groups
- Smokers who have recently quit smoking
- Those who have recently lost weight

In addition, there are certain times in a person's life when the person is more prone to weight gain (**Table 1**). These age groups could be considered for selective prevention interventions. These times include the following:

Prenatal
Adiposity rebound (5–7 years)
Adolescence
Early adulthood
Pregnancy
Menopause

Table 1 Identifying at-risk groups for obesity

Critical ages and life stages	*Reason for increased risk*
Prenatal	There is evidence that in utero development has permanent effects on later growth and energy regulation.
Adiposity rebound (5–7 years)	Body mass index begins to increase rapidly after a period of reduced adiposity during preschool years. Food and activity patterns change as a result of exposure to other children and school. Early and rapid weight rebound often precedes the development of obesity.
Adolescence	Period of increased autonomy that is often associated with irregular meals, changed food habits, and periods of inactivity during leisure combined with physiological changes that promote increased fat deposition, particularly in females.
Early adulthood	Early adulthood usually correlates with a period of marked reduction in physical activity. In women this usually occurs between the ages of 15 and 19 years but in men it may be as late as the early 30s.
Pregnancy	Excessive weight gain during pregnancy often results in retention of weight after delivery, particularly with early cessation of breast feeding. This pattern is often repeated after each pregnancy.
Menopause	In Western societies weight generally increases with age but it is not certain why menopausal women are particularly prone to rapid weight gain. The loss of the menstrual cycle does affect food intake and reduce metabolic rate slightly.
High-risk groups	
Family history of weight problems	There is no longer any doubt that given the same environment some individuals are more prone to depositing fat. The basis of these differences in individual susceptibility to obesity is yet to be fully elucidated but is believed to involve a number of physiological processes associated with fat deposition and oxidation and involuntary energy expander.
Certain ethnic groups	In NSW, recent migrants from southern Mediterranean countries and the Middle East are more likely to be obese and their children are more likely to develop a more severe form of obesity that to immediate health consequences.
Socially or economically disadvantaged	In NSW, there is an inverse association between income and education level and obesity which is most pronounced among women and children. It is argued that cheaper foodstuffs are usually high in fat and energy dense and those with less financial resources spend more time in sedentary activities such as watching TV.
Recent successful weight reducers	Successful weight reduction is usually followed by the regain of one-half to one-third of the weight loss over the following year. It is believed that biological and behavioral processes act to drive body weight back to baseline levels.
Recent past smokers	Smokers are usually thinner than nonsmokers because smoking tends to depress appetite, increase the basal metabolic rate, and, after each cigarette, induce a surge in heart and metabolic rate. The effect on metabolism of smoking 24 cigarettes per day has been estimated at approximately 200 kcal per day.

Adapted from Gill TP (1997) Key issues in the prevention of obesity. *British Medical Bulletin* **53**(2): 359–388.

Those with an Existing Weight Problem

In developing weight gain prevention strategies, it is important not to neglect those with an existing weight problem who could benefit from more intensive efforts to help prevent further weight gain.

Key Elements of a Weight Gain Prevention Plan

Weight gain and obesity develop when the energy intake from food and drink exceeds energy expenditure from physical activity and other metabolic processes. It is often assumed that the prevention of weight gain should focus solely on attempting to alter these behaviors within individuals and communities. However, research has consistently shown that numerous and diverse factors, including environmental and social factors, influence the behaviors that lead to excessive weight gain. Addressing aspects of the obesogenic (obesity-promoting) environment, as well as individuals' eating and physical activity patterns, is considered to be critical to the success of any obesity prevention program.

The 2003 WHO report on diet, nutrition, and the prevention of chronic disease undertook a detailed review of the literature and identified a range of key factors that either increase or decrease the risk of weight gain and the development of obesity (**Table 2**). These factors were rated on the quality of evidence available to support their contributory role. This analysis serves as a very useful guide as to the focus of weight gain prevention initiatives.

Diet and Physical Activity Behaviors

The WHO analysis identified a number of key dietary and physical activity behaviors, amenable to change, that could conceivably influence energy balance sufficiently to contribute to the prevention of weight gain and obesity. Behaviors that reduced the risk of obesity included regular physical activity, high dietary fiber intake, and possibly breast-feeding and low glycemic index diets. Behaviors that increased the risk of obesity included a high intake of energy-dense foods, a high intake of sugar-sweetened drinks and juices, time spent in sedentary behaviors, and possibly large portion sizes, a high intake of fast foods, and a restrained eating pattern.

The area of dietary and physical activity antecedents to weight gain and obesity is still poorly understood and new research findings, which help clarify our understanding, are being presented on a regular basis. In addition, different behaviors are more prevalent or pronounced in different regions of the world. It is therefore difficult to give definitive recommendations on the most important and useful behaviors to target in obesity prevention strategies. However, strong evidence exists to support the inclusion of some key behaviors.

Reducing Energy Intake

Reducing the intake of high energy-dense foods (i.e., foods high in fat/sugar) There is a high level of agreement that the overconsumption of energy-dense foods is a major contributor to excess energy intake and weight gain and that restriction of energy-dense food items is a useful strategy for the prevention of weight gain. However, discussion continues as to whether fat or refined carbohydrate is the major contributor to energy density in the modern diet and thus should be the target of programs to control weight. The debate is being fuelled by dietary data from many developed countries that show that dietary fat intakes have leveled out or declined

Table 2 Summary of the strengths of evidence of factors that may promote or protect against weight gain and obesity

Evidence	*Decreases risk*	*Increases risk*
Convincing	Regular physical activity	High intake of energy-dense foods[a]
	High dietary fiber intake	Sedentary lifestyle
Probable	Home and school environments that support healthly food choices for children	Heavy marketing of energy-dense foods and fast-food outlets
	Promoting linear growth	Adverse social and economic conditions in developed countries (especially for women)
		Sugar-sweetened soft drinks and juices
Possible	Low glycemic index foods	Large portion sizes
	Breast-feeding	High proportion of food prepared outside of home
		Rigid restraint/periodic disinhibition eating patterns
Insufficient	Increased eating frequency	Alcohol

[a]Energy-dense foods are high in fat/sugar and energy-dilute foods are high in fiber and water, such as vegetables, fruits, legumes, and whole grain cereals.

Adapted from WHO (2003) *Joint WHO/FAO Expert Report on Diet, Nutrition and the prevention of Chronic Disease*, WHO Technical Report Series 916. Geneva: WHO.

slightly and intakes of carbohydrates have increased dramatically. However, research has shown that dietary fat (along with water and fiber) is a major contributor to the energy density of foods and that ad libitum low-fat diet plans are an effective dietary approach to weight gain prevention or moderate weight loss. There is also strong evidence that excess carbohydrate, particularly high glycemic index carbohydrate, contributes to weight gain and its restriction aids weight loss and improves cardiovascular risk factor profiles.

Increasing the intake of high-fiber, energy-dilute foods (especially vegetables and fruits) There is less evidence on the effectiveness of increasing the intake of energy-dilute foods such as vegetables and some fruits in the diet. Such a strategy would assist weight gain prevention only if the inclusion of such foods leads to a reduction in the intake of more energy-dense alternatives and thus creates a reduction in energy intake. Few studies have addressed this issue in a comprehensive manner, but the additional health benefits of these foods makes such a strategy low risk in nutritional terms.

Reducing the consumption of sugar-sweetened soft drinks and juices Evidence is accumulating from a variety of studies that energy consumed as sweetened drinks is less well compensated for than energy consumed as solid food. Longitudinal studies have also indicated that sweetened drinks (soft drinks or sodas) are associated with weight gain in both children and adults. Recent work has also demonstrated that the simple strategy of reducing the intake of sweetened drinks can be effective in preventing or limiting inappropriate weight gain.

Reducing the level of food prepared outside of the home The proportion of food purchased and consumed at food outlets outside of the home has increased dramatically in recent decades in both developed and developing nations. In the United States, approximately 40% of the household food budget is spent on food eaten away from home, and much of this is spent at fast-food outlets. A number of analyses have linked increased consumption of fast food with increased risk of obesity. Although few studies have evaluated the effect of reducing the consumption of fast food, it would seem to be a valuable strategy with few nutritional negatives.

Reducing portion sizes The portion size of packaged foods and snacks, as well as restaurant serving sizes, has increased rapidly in recent times and has been identified as an important factor in the consumption of excess energy. Evidence suggests that people will consume the portion of food they are provided rather than respond to satiety signals to stop eating and leave food. Also, as the serving size increases, the ability of consumers to estimate accurately how much they have consumed decreases. Reducing portion sizes is a simple but immediately effective mechanism for reducing energy intake.

Increasing Energy Expenditure

Regular physical activity Although it is difficult to obtain accurate assessments of physical activity, there is little doubt that energy expenditure from activity has decreased in the past 50 years in most countries throughout the world. In contrast to popular belief, participation rates in organized leisure-time physical activity have increased in recent times in many countries. This supports the contention that the greatest contributor to this reduction in energy expenditure is associated with substantial changes in occupational and incidental physical activity. Changes in employment patterns and work practices together with a reliance on motorized transport and the removal of almost all manual labor from our daily lives have led to a dramatic reduction in daily physical exertion.

Studies that have examined the association between physical activity and weight gain and the impact of increasing physical activity on weight gain prevention have been limited by the ability to accurately measure physical activity and to engage people in sufficient levels of physical activity to prevent weight gain. However, there is sufficient evidence to support an important role for increasing physical activity in any weight gain prevention strategy, although questions remain about how much exercise is necessary and what type of exercise is appropriate to promote. The issue of the amount of extra time that people should spend in moderate physical activity to prevent weight gain remains hotly debated, but it is clearly substantially more than the 30 minutes on 5 or more days each week recommended by experts to reduce cardiovascular disease risk. The type of exercise that should be the focus of weight gain prevention strategies is also under review. It has been suggested that the most effective ways to include regular physical activity in daily living are through increased incidental activity, increased participation in active recreation, and increased use of active transport.

Reduced time spent in sedentary behaviors (especially TV watching) Changes in societal structures and improvements in technology have allowed a reduction in time spent at work or on domestic chores, leaving a greater proportion of the day for leisure. At the same time, most of the entertainment options developed to fill this time, such as watching television, playing video games, and using computers, are sedentary activities that require very little energy expenditure. These forms of entertainment, which initially complemented other forms of leisure activity, are occupying more hours of the day and are displacing more active pursuits and games. As a consequence, a number of studies have identified clear links between time spent in this sedentary behavior and weight gain. However, it is important to make a distinction between a lack of physical activity and sedentary behavior because their mechanisms for impacting on body weight may be different and a person with a high level of physical activity can also have a high level of sedentary behavior. Although the precise pathway by which sedentariness influences weight gain is not known, it is believed to involved both a reduction in physical activity and an increase in dietary energy intake through inappropriate food intake that is often stimulated by and accompanies sedentary activities.

Some studies in children have shown that programs that seek to reduce time spent in sedentary behaviors are more effective in controlling weight than programs that aim to increase physical activity alone. In some cases, a simple program to reduce the amount of time spent watching television was sufficient to significantly limit inappropriate weight gain in children.

Creating Supportive Environments

The external physical, social, political, and economic environments in which people exist have a profound effect on their attitudes and behaviors. Each day, people interact with a wide range of services, systems, and pressures in settings such as schools, the workplace, home, restaurants, and fast-food outlets. In addition, laws, policies, economic imperatives, and the views of governments, industry, and society as a whole influence these settings. Each of the features of this complex system, which shapes the environment in which we live, has the capacity to inhibit or encourage appropriate dietary and physical activity patterns. The availability of open space, access to public transport, the design of suburbs, access to buildings, the perceived level of safety, provision of lighting, and many other factors influence our capacity and desire to be more physically active in our daily lives. Similarly, advertising pressures, access to appropriate food choices, school food policies, and nutrition information and labeling all potentially influence food selection. Today, there is also a large commercial drive to promote obesogenic behaviors (cars and food are the two most advertised products on television).

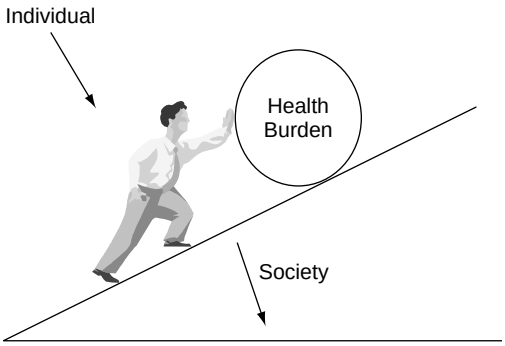


Figure 2 Influence of societal and environmental factors on development of obesity. (From House of Commons Health Committee (2004) *Obesity: Third Report of Sessions 2003–04. Volume 1. Report Together with Formal Minutes*. London: The Stationery Office Ltd.)

Trying to motivate people to make healthy choices when the external environment works against such behaviors is a recipe for failure. **Figure 2** illustrates the role that the social environment plays in assisting or inhibiting personal behavior choices made by individuals, which ultimately impact upon their health. Great success is likely to be achieved by creating a supportive environment and then promoting the healthy dietary and physical activity choices within such an environment.

Lessons from Past Prevention Efforts

Obesity Prevention Programs

A number of systematic reviews have assessed the current scientific literature on programs addressing the prevention of obesity in both children and adults and have identified only a limited number of evaluated programs. The reviews concluded that there was simply too small a body of research conducted in a limited number of settings to provide firm guidance on consistently effective interventions. However, reviews of childhood obesity prevention initiatives indicated that certain approaches appear to be associated with greater success. Intensive intervention in small groups was a successful management strategy in children, as was involving the entire

the family. Reducing levels of inactivity was successful at both treating and preventing weight gain. Some interventions that increased time spent in formal physical activity were successful in controlling weight gain, but generally multicomponent programs that addressed a range of strategies were deemed to hold the most promise.

There was general agreement that efforts should be heavily oriented toward preventing obesity in children because of the greater likelihood of success at a younger age. More effort needs to be directed at creating environmental and policy changes that will support the adoption of behaviors conducive to weight control rather than simply relying on education approaches.

Large-Scale Community Coronary Heart Disease and Diabetes Prevention Trials

Conducting large-scale, communitywide trials to address the prevention of obesity is a very expensive and difficult process; consequently, evidence of this nature is very limited. However, a number of large CVD and diabetes prevention trials have included weight as an intermediary outcome, which can also provide useful information about effective strategies to address obesity, and have demonstrated that it may be possible to prevent weight gain if not reduce weight at a population level.

The results of early large-scale community CVD prevention trials, such as the Stanford Three Community and Five Community studies as well as the Minnesota Heart Health Program, had limited impact on weight status and reinforced the difficulty of preventing weight gain in the community. However, later programs, such as the Pawtucket Heart Health Program, were able to make a modest impact on weight gain in the intervention community after 10 years. These programs demonstrate the time lag that can be expected between the implementation of a truly communitywide program and the extent of behavior change likely to be required to impact upon the weight status of the community. It has been suggested that unless weight is the primary outcome of the intervention, it is unlikely that sufficient focus will be placed on achieving the level of change required to impact on energy balance and community weight status.

Strong and consistent evidence of the success of large-scale weight gain prevention initiatives has been obtained from diabetes prevention trials that have addressed the progression to diabetes in people identified as glucose intolerant. Four large-scale trials have produced significant reductions in the rate of diabetes by focusing on exercise and diet, which resulted in small weight losses of approximately 3 or 4 kg on average. The largest trial conducted in the United States found that advice to reduce the energy and fat in the diet together with modest increases in physical activity, which was reinforced with regular follow-up from a 'coach,' led to an average weight loss of 5.6 kg and 58% reduction in the number of people progressing from impaired glucose tolerance to diabetes.

Lessons from Other Prevention Efforts

Although the number of successful large-scale obesity prevention programs is limited, there is a wealth of information from past public health programs that can be used to address other chronic diseases and risk factors. The International Obesity Task Force identified 10 key principles on which efforts to prevent obesity at a population level should be based. These are presented in **Box 2** and are drawn from experiences addressing cardiovascular disease, smoking, alcohol and drug problems, dental disease, road accidents, and other public health issues.

Although much has yet to be elucidated about the development of obesity and its effective management and prevention, there is a consensus that action to address the problem must not be delayed. Efforts to prevent weight gain need to be well

Box 2 IOTF principles for the development of population obesity prevention initiatives

1. Education alone is not sufficient to change weight-related behaviors. Environmental and societal intervention is also required to promote and support behavior change.
2. Action must be taken to integrate physical activity into daily life, not just to increase leisure time exercise.
3. Sustainability of programs is crucial to enable positive change in diet, activity, and obesity levels over time.
4. Political support, intersectoral collaboration, and community participation are essential for success.
5. Acting locally, even in national initiatives, allows programs to be tailored to meet real needs, expectations, and opportunities.
6. All parts of the community must be reached, not just the motivated healthy.
7. Programs must be adequately resourced.
8. Where appropriate, programs should be integrated into existing initiatives.
9. Programs should build on existing theory and evidence.
10. Programs should be properly monitored, evaluated, and documented. This is important for dissemination and transfer of experiences.

Source: Kumanyika S, Jeffery RW, Morabia A *et al.* (2002) Obesity prevention: The case for action. *International Journal of Obesity* **26**(3): 425–436.

designed, comprehensive, and appropriately evaluated so that the knowledge base improves with each new program.

See also: **Coronary Heart Disease**: Prevention. **Diabetes Mellitus**: Etiology and Epidemiology. **Energy**: Requirements. **Exercise**: Beneficial Effects. **Obesity**: Definition, Etiology and Assessment; Childhood Obesity. **Weight Management**: Approaches; Weight Maintenance. **World Health Organization**.

Further Reading

Campbell K, Waters E, O'Meara S *et al.* (2002) Interventions for preventing obesity in children. *Cochrane Database Systematic Review* **2**: CD001871.

Dietz W and Gortmaker S (2001) Preventing obesity in children and adolescents. *Annual Review of Public Health* **22**: 337–353.

Douketis J, Feightner J, Attia J *et al.* (1999) Periodic health examination, 1999 update: 1. Detection, prevention and treatment of obesity. *Canadian Medical Association Journal* **160**(4): 513–525.

Egger G and Swinburn B (1997) An 'ecological' approach to the obesity pandemic. *British Medical Journal* **315**(7106): 477–480.

French S, Story M, and Jeffery RW (2001) Environmental influences on eating and physical activity. *Annual Review of Public Health* **22**: 309–335.

Gill TP (1997) Key issues in the prevention of obesity. *British Medical Bulletin* **53**(2): 359–388.

House of Commons Health Committee (2004) *Obesity: Third Report of Sessions 2003–04. Volume 1. Report Together with Formal Minutes*. London: The Stationery Office Ltd.

James WPT and Gill TP (2004) Prevention of obesity. In: Bray G, Bouchard C, and James WPT (eds.) *Handbook of Obesity: Clinical Applications*, 2nd edn. New York: Marcel Dekker.

Kumanyika S, Jeffery RW, Morabia A *et al.* (2002) Obesity prevention: The case for action. *International Journal of Obesity* **26**(3): 425–436.

NHS Centre for Reviews and Dissemination (2002) The prevention and treatment of childhood obesity. *Effective Health Care Bulletin* **7**(6).

Saris W, Blair S, van Baak MA *et al.* (2003) How much physical activity is enough to prevent unhealthy weight gain? Outcome of the IASO 1st Stock Conference and consensus statement. *Obesity Reviews* **4**: 101–114.

Story M (1999) School-based approaches for preventing and treating obesity. *International Journal of Obesity and Related Metabolic Disorders* **23**(supplement2): S43–S51.

US Surgeon General (2001) *The Surgeon General's Call to Action to Prevent and Decrease Overweight and Obesity*. Washington, DC: Office of the Surgeon General.

World Health Organization (2000) *Obesity: Preventing and Managing the Global Epidemic. Report of a WHO Consultation*, WHO Technical Report Series 894. Geneva: WHO.

World Health Organization (2003) *Joint WHO/FAO Expert Report on Diet, Nutrition and the Prevention of Chronic Disease*, WHO Technical Report Series 916. Geneva: WHO.

Treatment

E C Uchegbu, Royal Hallamshire Hospital, Sheffield, UK

P G Kopelman, Queen Mary's, University of London, London, UK

Introduction

Increasing body weight is associated with increasing health risks (**Table 1**). Randomized controlled trials demonstrate that weight reduction reduces these health risks and confirm the value in treating overweight and obesity.

Obesity is a chronic disease of multiple etiologies characterized by an excess of adipose tissue. Recent research has begun to unravel the biochemical and genetic factors implicated in its etiologies. As a result of the factors that determine its severity, health risks, and response to therapy, treatment must be tailored to specific needs. The ability of a treatment to maintain long-term weight reduction is as important as its ability to cause the initial weight loss. In several studies inability in maintaining the lowered weight is the cause of the treatment failure.

Nevertheless, a successful program should also lead to an improvement in the quality of life, self-esteem, social functioning, anxiety, and depression.

Several professional, governmental, and other organizations have drawn up guidelines for obesity management. These strategies for providing care to the obese patient provide useful evidence-based guidance for clinical management.

Health Risks due to Overweight/Obesity

Increasing body fatness is accompanied by profound changes in physiological function. These changes are, to a certain extent, dependent on the regional distribution of adipose tissue. Generalized obesity results in alterations in total blood volume and cardiac function while the distribution of fat around the thoracic cage and abdomen restricts respiratory excursion and alters respiratory function. The intra-abdominal visceral deposition of adipose tissue, which characterizes upper body obesity, is a major contributor to the development of hypertension, elevated plasma insulin concentrations and insulin resistance, hyperglycemia, and hyperlipidemia. The alterations in metabolic and physiological function that follow an increase in adipose tissue mass are predictable when considered in the context of normal homeostasis.

Table 1 Obesity-associated diseases and conditions

Disorder	*Associated diseases and conditions*
Cardiovascular disorders	Coronary heart disease
	Hypertension
	Cerebrovascular disease
	Deep vein thrombosis
	Pulmonary embolism
Respiratory disorders	Obstructive sleep apneas
	Obesity hypoventilation syndrome
	Breathlessness
Gastrointestinal disorders	NASH (nonalcoholic steatohepatitis)
	Cirrhosis
	Gallstones
	Colorectal cancer
	Hiatus hernia/ gastroesophageal reflux
Renal disorders	Proteinuria
Reproductive disorders	Primary ovulatory infertility
	Development of gestational diabetes
	Increased risk of neural tube defects
Musculoskeletal disorders	Osteoarthritis
	Gout
	Nerve entrapment
Genitourinary	Endometrial cancer
	Prostate cancer
	Stress incontinence
Metabolic and endocrine disorders	Artherogenic lipid profile
	Insulin resistance
	Type 2 diabetes mellitus
	Polycystic ovary syndrome
	Postmenopausal breast cancer
	Hirsuitism
Skin disorders	Acanthosis nigricans
	Lymphoedema
	Sweat rashes

Ethnicity has an impact on body fat distribution and adipose tissue metabolism. Overweight currently is defined as a body mass index (BMI) $>25\,kg\,m^{-2}$ and obesity as a BMI $>30\,kg\,m^{-2}$. The evidence for this is drawn from large population studies that suggest people with a BMI of $19–25\,kg\,m^{-2}$ have the lowest mortality. However, there have been proposals to define race-specific standards according to ethnic background. Specifically, Asians have greater visceral fat and associated morbidity than do Caucasians. A BMI as low as $23\,kg\,m^{-2}$ may be associated with weight-related diabetes or insulin resistance in these groups. For any given weight category, the presence of certain complications moves the individual into a higher health risk category. Evaluation of such risks should be part of the intervention program.

Patient Selection

Obesity and overweight are chronic conditions. Short-term programs are likely to be ineffective, with rapid weight regain once treatment is stopped. Treatment programs must be for the longer term and include measures to prevent relapse. Preventing further weight gain in those at risk should also form part of obesity management and help ensure an appropriate use of resources. Those at risk will include moderately overweight subjects and those who have upper body obesity. Weight loss is indicated in adults with a BMI of more than $25\,kg\,m^{-2}$ and/or abdominal girth of more than or equal to 102 cm in males and more than or equal to 88 cm in females. Additional important treatment areas include weight gain in infancy, adolescence, and pregnancy. A family history of obesity or associated diseases, fat distribution, and risks for coronary heart disease are individually important factors that may influence treatment mode.

Treatment Aims and Realistic Weight Loss Goals

Treatment aims to improve health and well being and decrease the risks of ill health later in life, through reducing the amount and possibly distribution of body fat. The success or failure of any treatment program may be judged by an arbitrarily chosen weight or percentage weight loss. Hence, the evidence that modest degrees of weight loss produce significant health gain influences the success or failure of any treatment. With this background it is logical to redefine successful treatment in terms of a decrease in the severity of obesity rather than a return to normal weight. Even weight stabilization without weight loss represents a modestly successful outcome compared to the natural history of obesity, which is progressive weight gain. A weight loss of between 5 and 10% of the initial body weight is associated with clinically useful improvements in terms of blood pressure, plasma cholesterol, and a significant improvement in diabetic control (see **Table 2**). Weight loss should be approached incrementally with new weight goals negotiated with the patient if the original target is achieved. Goals for older patients (>65 years) will be different from those who are young; data suggest that a population becomes heavier with age whereas the risk from obesity does not increase proportionately.

Dietary Treatment of Obesity

The primary determinant of weight loss is energy deficit. Short-term weight loss has been achieved by

Table 2 Benefits of 10 kg weight loss

Condition	*Health benefit*
Mortality	Fall of more than 20% in total mortality
	Fall of more than 30% in diabetes-related death
	Fall of more than 40% in obesity-related cancer death
Blood pressure	Fall of 10 mm Hg systolic blood pressure
	Fall of 20 mm Hg diastolic blood pressure
Diabetes	Fall of 50% in fasting glucose
	Reduces risk of developing diabetes by 50%
Lipids	Fall of 10% in total cholesterol
	Fall of 15% in LDL cholesterol
	Fall of 30% in triglycerides
	Rise of 8% in HDL cholesterol

Adapted with acknowledgment from the Scottish Intercollegiate Guidelines Network (SIGN) Obesity in Scotland: integrating prevention with weight management. A national clinical guideline recommended for use in Scotland. Edinburgh (1996).

energy reduction in diets of varied macronutrient composition. Obesity is a chronic and relapsing disease; hence, it is the long-term efficacy of these dietary strategies in maintaining lowered weight (and minimizing the risk of diet-related chronic diseases) that is of fundamental importance.

Types of Dietary Treatment

There are several dietary strategies available both in a clinical and commercial setting. These diets vary greatly in the degree of caloric restriction, relative amounts of macronutrients (protein, carbohydrate, fat), medical supervision, scientific basis, and cost. These diets can be broadly divided into:

- low-calorie diets ($\geq$3400 kJ (800 kcal) day^{-1}, typically 3400–6300 kJ (800–1500 kcal) day^{-1})
- very low-calorie diets ($<$3400 kJ (800 kcal) day^{-1})

Traditionally, low-calorie diets that incorporate various methods for restricting food intake have been recommended for weight management.

Such treatment requires a period of supervision for at least 6 months. A review of 48 randomized control trials (RCTs) shows strong and consistent evidence that an average weight loss of 8% of the initial body weight can be obtained over 3–12 months with a low-calorie diet (LCD) and this weight loss causes a decrease in abdominal fat, the adipose tissue deposition that is associated with the highest disease risk. Very low-calorie diets (VLCD) have been shown to reduce weight at a greater rate in the first 2–3 months compared to low-calorie diets but have not been associated with superior maintenance of lost weight after a year. A review of weight loss trials of LCD and VLCD with available follow-up during 2–7 years showed that long-term weight loss in most trials is in the range of 2–6 kg.

Low-fat, high-carbohydrate diets Low-fat, high-carbohydrate diets have played a central role in the dietary management of overweight and obesity. Generally, these strategies aim to provide a macronutrient composition of 25–35% energy from fat, 45–60% from total carbohydrate, and 15–20% from protein, thereby moving individuals towards national dietary guidelines (COMA reports). A review of controlled clinical trials demonstrated that a 10% reduction of dietary fat leads to a $\sim$3–4-kg weight loss in normal overweight subjects and $\sim$5–6-kg weight loss in the obese. Evidence from a recent systematic review suggests that a low-fat diet is equally as effective in achieving long-term weight loss in overweight and obese subjects as alternative dietary strategies. Low-fat high-carbohydrate diets may have a role in weight maintenance. Combined with physical activity and behavioral strategies, the American Diabetes Prevention Program and the Finnish Diabetes Prevention Trial demonstrated maintenance of modest weight loss (3–4 kg) with a marked reduction in the risk of developing type 2 diabetes mellitus over a 4-year study period.

Low glycemic index diets The glycemic index (GI) is a dietary concept originally developed for the therapy of diabetes, which has recently become popular despite scant evidence of its effectiveness in weight management. The GI is a property that describes the effect of carbohydrate from a given food on postprandial blood glucose. It is measured by comparing the blood glucose response of the test food with that of a reference food (usually white bread). Low-GI foods are more slowly absorbed leading to an attenuated and prolonged insulin and metabolic response to foods; it is suggested that more moderate blood glucose and metabolic response may sustain satiety and energy balance to a greater extent than larger metabolic shifts would.

Epidemiological analyses link low-GI load diets to a more favorable lipids profile and reduced incidence of type 2 diabetes mellitus and cardiovascular disease. Evidence from interventional studies supports the benefits of low-GI diets in reducing the risks of coronary heart disease and diabetes but there are no long-term studies that have evaluated its weight-loss efficacy. Therefore, it is appropriate to promote the constituents of a low-GI diet (increased legumes, wholegrain cereals, and fruit consumption) as part of a well-balanced hypo-caloric diet for the long-term management of obesity and its metabolic complications.

High-protein, low-carbohydrate diets High-protein diets have recently been popularized as a means of rapid weight loss despite the lack of objective evidence in long-term efficacy and safety. Typically, these diets offer wide latitude in protein food choices,

and are restrictive in other food choices (mainly carbohydrate). Animal protein rather than plant protein is advocated leading to a higher intake of total fat – mainly saturated fat and cholesterol. Many of the popular high-protein diets promote protein intake of 28–64% of dietary energy, which exceeds established requirement of 10–15%, and severely limit carbohydrate dietary energy to 3–10%. A recent popular high-protein, low-carbohydrate diet, the Atkins diet, provides on average 27% energy from protein, 5% energy from carbohydrates, and 68% energy from fat. The diet results in the avoidance of important staple foods, such as bread, pasta, rice, potatoes, and cereals, as well as foods high in sugars. Consumption of fruits, vegetables, whole grains, and low-fat dairy products, foods associated with lowering blood pressure and protecting against cancer and heart disease, are all limited.

The initial weight loss in high-protein diets is high due to fluid and glycogen loss related to low carbohydrate intake, overall caloric restriction that is encouraged by structured eating plans, restricted range of foods allowed, and limited tolerance of high-protein foods. This often promotes a misconception about weight loss by suggesting that it is not related to total energy intake but is due to exclusion of certain foods.

A recent systematic review of the efficacy of low-carbohydrate, high-protein diets demonstrates that the amount of weight loss is principally associated with decreased caloric intake rather than reduced carbohydrate content. Researchers have yet to establish whether individuals can maintain long-term weight loss with a high-protein, low-carbohydrate diet because of the short duration of these studies, and long-term adverse effects are also unknown. Possible negative effects include increased risks of cardiovascular disease, renal disease, cancer, osteoporosis, and compromised vitamin and mineral status.

Energy prescribed diet This dietary strategy determines the daily energy requirement for weight loss by calculating energy expenditure, adjusting for physical activity, and subtracting an energy deficit to induce weight loss –usually 2100–2520 kJ (500–600 kcal) for 0.05 kg weight loss. As a result the prescribed diet will often be in excess of 3400–6300 kJ (800–1500 kcal). The popularity of this approach relates to the findings of improved compliance in those advised on a 2520 kJ (600 kcal) deficit diet compared to a traditional fixed energy intake of 5040 kJ (1200 kcal) day^{-1}.

Formulas and meal replacements Meal replacements are another category of calorie-controlled diets. These include nutritional fortified shakes, snack bars, and low-calorie frozen meals. An entire meal or snack is replaced with a portion controlled prepackaged meal or drink that provides approximately 840–1260 kJ (200–300 kcal), although formulations and nutrient content vary. Meal replacements are designed to be eaten with additions of conventional foods that supply dietary fiber, other nutrients, additional calories, and water. Most weight loss programs that use meal replacements recommend replacing two meals and one snack a day to lose weight and then replacing one meal per day to maintain weight loss. This strategy generally provides 5040–6729 kJ (1200–1600 kcal) day^{-1} and the regular meal should meet the recommendations of a healthy diet.

A recent meta-analysis that summarized the efficacy of this approach compared to conventional energy-restricted diets suggests that it is an effective weight-loss strategy both in the short and long term in a clinical trial setting. There is no information about the efficacy outside a clinical trial where meal replacement products need to be purchased, and are frequently discontinued at an early stage.

Very low-calorie diets Very low-calorie diets are formula foods; they are designed to provide larger and more rapid weight loss than the standard low-calorie diets. They are commonly given in liquid form to completely replace usual food and snack intake providing in the region of 1890–3400 kJ (450–800 kcal) day^{-1}. To reduce the potential risks from loss of lean body tissue, VLCDs are enriched in protein of high biologic value and also includes the full complement of recommended daily allowance for vitamins, minerals, electrolytes, and fatty acids. However, diets providing such low-energy intakes are often associated with a feeling of fatigue, constipation, nausea, and diarrhea. A most serious complication associated with VLCD is the development of symptomatic cholelithiasis associated with the rapid weight loss (1–2 kg $week^{-1}$).

Owing to the potential adverse effects of these diets, they are generally reserved for short-term treatment in individuals who are moderately to severely obese (BMI $>35\,kg\,m^{-2}$) and who have failed at more conservative approach to weight loss, in particular in those with medical conditions that may respond to weight loss such as obstructive sleep apnea, type 2 diabetes mellitus, or prior to surgical procedure.

Weight regain is common with the reintroduction of food. Studies show that in the long term, VLCDs are no more effective than more modest dietary restriction.

Commercial Slimming Organizations and Products Such organizations are profit-making ventures. However, they have been shown to be economical, practical, and an effective way of providing care for a large number of moderately obese people in the community. Weekly meetings serve to encourage and reinforce active participation by members, who learn through the exchange of ideas within the group. Weight losses achieved by commercial groups are comparable to those seen in general practice or hospital outpatient clinics. When behavioral techniques are added to the basic program of balanced diet, the results are further improved.

Over recent years there has been increasing use of weight loss-related web sites on the Internet, which are directed mainly at females. The content and structure of these web sites vary widely. They often lack professional contact and the expertise to deal with medical complications.

Behavior Treatment

Behavior therapy provides an important approach to losing and maintaining weight. The focus is on behaviors related to body weight, namely food intake and physical activity. It serves to identify the abnormal eating behaviors and life style developed over the years and helps to unlearn them and allow body weight to return to normal. The behavior techniques used include self-monitoring, stimulus control, and, recently, cognitive therapy, which involves identifying and changing negative thoughts.

The key difference between behavioral methods and other forms of treatment is that the individual must take responsibility for initiating and maintaining treatment rather than relying on external forces.

There are several elements of behavioral treatment (see **Table 3**). Evidence from RCTs confirms that behavioral strategies reinforce changes in diet and physical activity in obese subjects to produce weight loss of 10% over 4 months to 1 year. Longer term followup shows a return to baseline in the absence of continuing behavioral intervention.

Eating patterns and behaviors are, to a greater extent, acquired by learning, and for this reason there has been much interest in modifying the behavior within the family setting. Obese children are more likely than nonobese children to become obese adults. Behavior therapy seems to be effective in arresting this process in some children.

Table 3 Elements of behavioral treatment

Element	*Intervention strategy*
Self-monitoring	Observe, record, and provide feedback on: • food consumption (food diary) • physical activity (activity diary, pedometer) • weight record
Goal setting	Realistic weight-loss goals Separate short-term from long-term goals Focus on health benefits
Stimulus control	Identify and modify environmental barriers: • healthy eating, normalize eating pattern • increasing daily energy using activities
Problem solving	Handling emotional issues and social events: • examine situation • choose a solution and implement it • evaluate the outcome
Cognitive change	Changing inaccurate belief about weight loss • examine thought and feelings • challenge inaccurate ones • use positive self-affirmations

Exercise and Physical Activity

Exercise produces fat loss in obese and normal weight subjects, although losses rarely exceed 5% of body weight. For any given weight loss, fat-free mass (FFM) is better preserved in exercising than nonexercising subjects: this is likely to be important in the long term because FFM is the best predictor of resting metabolic rate, which is the largest contributor to daily energy expenditure for all but active athletes.

There are other beneficial effects of exercise that are independent of its effects on weight loss. Regular exercise reduces blood pressure, improves insulin sensitivity, both in association with or independent of weight loss. Favorable effects on the atherogenic lipid profiles have also been reported with exercise and physical training in obese subjects. Such benefits are substantial and should be emphasized to all patients; however, persuading an obese person to participate in regular physical activity and to maintain exercise as a part of daily routine is not easy.

One of the most consistent findings in studies of physical activity is enhanced weight maintenance for at least 2 years from the start of the intervention. It is not necessary to increase maximal oxygen uptake in the obese to derive benefit from exercise: metabolic evidence of fitness is achieved with less vigorous exercise.

Physical activity recommendations suggest 30 min of moderate activity on at least 5 days of the week. This level of activity is associated with improved fitness and protection from cardiovascular diseases. When using exercise solely as a strategy for weight reduction, longer duration of daily activity of a moderate intensity lasting 45–60 min is required.

Reduction in the time spent in sedentary behaviors (such as television watching) is an important

strategy for increasing physical activity and energy expenditure. Similarly, encouraging findings have been observed in children and adolescents advised to include more lifestyle activity (e.g., walking versus car use) compared to those with traditional programs of activity.

Drug Treatment of Obesity

Rationale

Diet restriction even when combined with behavioral therapy and increased exercise is often unsuccessful in achieving weight loss and maintenance in obese subjects. Obesity is not a single disorder but a heterogeneous group of conditions with multiple causes. Although genetic differences are of undoubted importance, the marked rise in the prevalence of obesity is best explained by behavioral and environmental changes that have resulted from technological advances. In such circumstances, it is appropriate to consider pharmacological treatment as an adjunct to the other treatment modalities.

In broad terms a pharmacological agent can cause weight loss by reducing energy intake or absorption/and by increasing energy expenditure. Current drug treatment of obesity is directed at reducing energy/food intake either by an action on the gastrointestinal system or via an action through the central nervous system control of appetite and feeding.

Selection of Patients

Pharmacological treatments of obesity have had a controversial history and are still regarded with skepticism and suspicion by some medical practitioners. This results from experiences with older agents that turned out to have serious side effects and were withdrawn as a result. Current agents approved for use have been shown to be safe and effective both in weight reduction and in the improvement of comorbidities of obesity. Nevertheless, it is important that doctors who prescribe such drugs are fully familiar with the mode of action and potential risks.

Several sets of guidelines have been developed for the use of drugs in the treatment of obesity. In the UK, The Royal College of Physicians' guidance on the use of anti-obesity drugs suggests that it may be appropriate to consider use of drugs after at least 3 months of supervised diet, exercise, and behavioral management. Exceptionally, this period may be shortened when the clinician judges that drug treatment is justified at an earlier stage due to over-riding medical circumstances. **Table 4** lists the criteria that should be applied to judge the suitability of a patient for drug treatment.

Table 4 Criteria for selecting obese patients suitable for obesity drug treatment

- Drug treatment may be appropriate where diet and exercise have not achieved acceptable weight loss relative to medical risk
- In such patients drug treatment may be appropriate for:
  - those whose BMI is more than 30
  - those with established comorbidities whose BMI is more than 27, if the drug license permits
- Weight-lowering drugs should be targeted at those at high risk from obesity, not obesity alone

The following groups will have priority for drug treatment

- Patients with established comorbidities such as type 2 diabetes, hypertension, and dyslipidemia
- Patients who are physically restricted by their weight either because of breathlessness or arthritis
- Patients considered to be at high risk – for example, those with a family history of overweight or obese parents who died prematurely from CHD or developed type 2 diabetes with complications

The criteria applied to the use of an anti-obesity drug are similar to those applied to the treatment of other relapsing disorders. It is important to avoid offering anti-obesity drug therapy to patients who are seeking a 'quick fix' for their weight problem. The initiation of drug treatment will depend on the clinician's judgement about the risks to an individual from continuing obesity. It may be appropriate after at least 3 months of supervised diet, exercise, and behavioral management, or at a subsequent review, if a patient's BMI is equal to or greater than $30\,kg\,m^{-2}$ and weight loss is less than 10% of the presenting weight. In certain clinical circumstances it may also be appropriate to consider anti-obesity drug treatment for those patients with established comorbidities whose BMI is $27\,kg\,m^{-2}$ or greater if this is permitted by the drug's licence (see **Figure 1**). An anti-obesity drug should not be prescribed for a patient whose BMI is less than that specified in the product licence for the drug – the licence indication does not presently take account of the morbidity from obesity seen in certain populations at a lower BMI.

The experience from the use of anti-obesity drugs during 12–24 month randomized controlled trials indicate that approximately 50% of the actively treated patients respond as judged by 5–10% reduction in body weight maintained over 12 months. The weight loss occurs in the 'responder' group within 12 weeks. This indicates a suitable time period when a response to drug treatment can be identified and a decision taken to continue the medication. Continuing assessment of drug therapy for efficacy and safety is essential. If the drug is efficacious in helping a patient to lose and/or maintain weight loss, and there are no serious side effects, it may be continued.

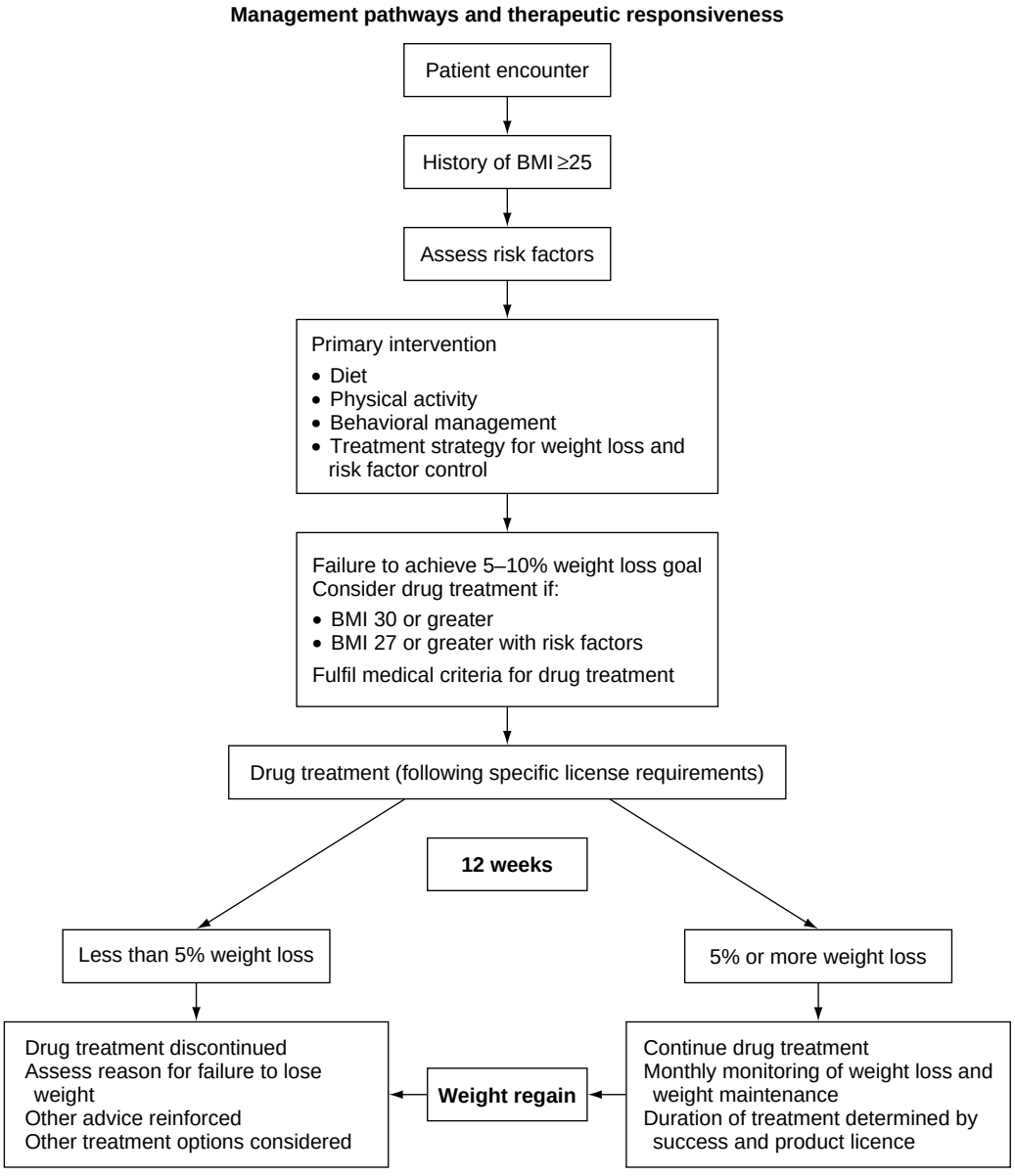


Figure 1 A management pathway for the appropriate prescription of an anti-obesity drug. (Adapted with permission from RCP Guidelines 2003.)

If not, it should be discontinued. Once a weight loss target has been achieved, there should be an opportunity for re-negotiation of a new target, if indicated, and/or long term monitoring with reinforcement.

Types of Drugs

The two categories of anti-obesity medication currently licensed for use in obese subjects are:

1. Those that act on the gastrointestinal system (pancreatic lipase inhibitors) as malabsorption agents to inhibit nutrient absorption.
2. Those that act on the central nervous system primarily to reduce hunger perception.

Drugs acting on the gastrointestinal system

Orlistat Orlistat is a gastric and pancreatic lipase inhibitor that reduces the absorption of dietary fat in a dose-dependent manner. At the therapeutic dose of 120 mg three times a day, it blocks the absorption of about 30% of dietary triacylglycerol resulting in an energy deficit of 850 kJ (200 kcal) day^{-1} for an individual on an average diet of 9240 kJ (2200 kcal) day^{-1} with 40% of calories from fat.

Adverse effects of orlistat are predominantly related to its gastrointestinal action of fat malabsorption and can be associated with a modest reduction in fat-soluble vitamins (A, D, E, and K). However, clinical deficiency has not been reported in clinical trials. Nevertheless, it is recommended that patients taking orlistat receive vitamin supplements. Patients may complain of loose or liquid stool, fecal urgency, anal leakage, and infrequently fecal incontinence due to undigested fat. These adverse effects become less common with longer duration of treatment suggesting that patients learn to avoid high-fat meals to avoid these side effects hence enforcing behavioral change. This may well contribute to the therapeutic effects of orlistat treatment. Orlistat is minimally absorbed (less than 1%) and systemic events are negligible.

Drugs acting on the central nervous system These drugs are commonly referred to as appetite suppressants, which is only one of their actions. Some of these agents have been proven to enhance satiety and slow gastric emptying; and an increase in energy expenditure has also been suggested. They act by increasing the neurotransmitter activity in the brain centers that regulate food intake.

Sibutramine Sibutramine enhances the sensation of satiety after a meal by its central action as a serotonin and/or epinephrine re-uptake inhibitor. Sibutramine is a beta-phenethylamine and is well absorbed following oral administration. It undergoes extensive first pass metabolism in the liver to produce two pharmacologically active metabolites that have long elimination half-lives of 14–16 h.

Side effects commonly reported in clinical trials include dry mouth, constipation, anxiety, rhinitis, and insomnia but these rarely led to withdrawal from the study. The noradrenergic actions of the drug may cause an increase in blood pressure and heart rate in some patients or prevent the expected fall in these parameters with weight loss. It should not be given in patients with uncontrolled hypertension. It should not be given concomitantly with monoamine oxidase inhibitors, nor other centrally acting anorexic drugs, or sympathomimetic agents including cold remedies such as pseudoephedrine.

Phentermine and diethylpropion Published evidence of the use of phentermine and diethylpropion indicates short-term induction of weight loss that is frequently followed by weight regain on cessation of the drug. There are no recently published randomized controlled trials of the drugs demonstrating efficacy beyond 26 weeks. Both drugs remain restricted to 3 month's use in the terms of their product license.

Rimonabant (SR 141716) Rimonabant is a selective central cannaboid (CB1) receptor antagonist. It is an appetite suppressant in advanced development for obesity treatment. The rationale behind this drug is to reduce appetite by blocking cannaboid receptors in the hypothalamus. The central cannaboid (CB1) receptors are believed to play a role in controlling food consumption and the phenomena of dependence/habituation.

Preliminary results from a 2-year international multicenter study confirm its effectiveness in weight reduction, reduction in waist circumference (a marker of the dangerous abdominal obesity), and improvements in lipids and glycemic profiles. The study also confirmed its good safety profile. The side effects reported were mainly mild and transient and most frequently involved nausea, diarrhea, and dizziness.

Rimonabant has potential as a treatment for smoking cessation because the central cannaboid system is also involved in the body's response to tobacco dependence.

Prescribing guidelines for anti-obesity drugs

Anti-obesity drugs should be prescribed in an appropriate clinical setting that includes systems for monitoring and follow-up of progress. The choice of anti-obesity drug is largely dependent on the experience of the prescriber in using one or another agent (see **Table 5**). For the two agents currently recommended for use there are no good clinical studies that have directly compared them or have explored which particular patient will benefit more from one than the other. A drug should not be considered ineffective because weight loss has stopped, provided the lowered weight is maintained.

The Elderly and Children

There is limited information about the use of anti-obesity drugs in patients over the age of 75 years. In such circumstances, the accepted practice is to aim for weight maintenance rather than weight loss. Neither drug is licensed for use in children.

Surgical Treatment for Obesity

Surgical treatment is an appropriate intervention for the management of morbid obesity. Criteria for selection of patients suitable for surgery are listed in **Table 6**.

Table 5 Comparison of actions and indications for use of sibutramine and orlistat

	Sibutramine	*Orlistat*
Mode of action	Promotes satiety Enhancing effect on thermogenesis	Dietary fat malabsorption
Indication	Adjunct to diet in obese patients with BMI $\geq$30 kg m^{-2} without comorbidities or BMI $\geq$27 kg m^{-2} with comorbidities	Adjunct to diet in obese patients with BMI $\geq$30 kg m^{-2} without comorbidities or BMI $\geq$28 kg m^{-2} with comorbidities
Suitable for	Those with uncontrollable appetite Frequent snackers Nocturnal eaters Those with need for immediate weight loss for medical reasons Those without contraindication to its use (specifically cardiac abnormalities or elevated blood pressure, i.e., >145/95 mm Hg) Patients with low HDL cholesterol	Those who have lost at least 2.5 kg through diet and lifestyle modification Patients requiring longer term behavioral changes whose dietary assessment suggests high-fat intake Patients with impaired glucose tolerance Those with elevated LDL cholesterol Chronic malabsorption Cholestasis
Specific contraindication	Tourette syndrome Cardiovascular disease Congestive cardiac failure Hypertension Hyperthyroidism, phaechromocytoma Pregnancy, breastfeeding	Pregnancy, breast feeding
Duration of treatment	Not more than 1 year	Maximum of 2 years

Types of Obesity Surgery

At least 30 surgical techniques have been developed for the treatment of obesity. Superficial cosmetic removal of adipose tissue (liposuction) will not be considered because it has no lasting benefit and it is not regarded as a treatment for obesity. Jaw wiring (intermandibular fixation) can restrict intake of food but it is no longer recommended for surgical treatment of obesity due to a lack of long-term efficacy.

The operative procedures currently used for the surgical treatment of obesity are outlined below.

Gastric restriction Gastric restriction can by achieved by gastroplasty or gastric banding. Gastroplasty techniques involve the fashioning of a proximal pouch of the stomach by vertical stapling and a constrictive band opening, thereby restricting the gastric volume to approximately 15–20 ml that empties into the remainder of the stomach.

Gastric banding involves the external 'pinching off' of the upper part of the stomach with a band usually made of Dacron. A modification of the gastric banding is an inflatable circumgastric band attached to a subcutaneous reservoir that allows access by a hypodermic syringe to inject or withdraw fluid thereby tightening or enlarging the bandwidth. This operation can be performed laparoscopically, significantly improving the perioperative safety of operating for the severely obese patients.

Gastric restriction operations require strict dietary compliance because an intake of high caloric liquids or soft foods are not inhibited by the narrow outlet and may explain a failure to lose weight. The advantage of these techniques is very low operative mortality (<1%) and relative lack of long-term nutritional deficiencies. The reported excess weight loss after 3–5 years is between 40 and 60% but there is a slow regain thereafter.

Table 6 Criteria for patient selection

- BMI $\geq$40 kg m^{-2}
- BMI $\geq$35 kg m^{-2} with serious comorbidity demonstrated to be responsive to weight loss
- Failure to achieve weight loss with conventional means
- Able to lose weight prior to surgery
- Have no evidence of psychiatric disease or maladaptive eating behaviors
- Absence of endocrine disorders that can cause morbid obesity
- Psychological stability:
  - Absence of alcohol and drug abuse
  - Understanding of how surgery achieves weight loss
  - preoperative psychological evaluation for selected patients

Gastric by-pass A 20–30-ml pouch is created by staples and connected to the jejunum transected 50 cm from the ligament of Treitz (Roux-en-Y gastric bypass). It results in weight loss by both restrictive and malabsorptive mechanisms. Published evidence confirms this procedure produces greater weight loss compared to gastric restrictive techniques but more frequent adverse effects including 'dumping' and nutritional deficiency may accompany it. Its operative mortality is approximately 1%.

Biliopancreatic diversion Biliopancreatic diversion includes a gastric resection and diversion of the biliopancreatic juice to the terminal ileum to reduce the absorption of nutrients. In this operation, an entero-entero anastomosis is performed between the proximal limb of the transected jejunum and ileum, 50–100 cm proximal to the ileocecal valve.

Biliopancreatic diversion achieves up to 78% excess weight loss at 5 years. Nutritional deficiencies are relatively common (between 5 and 40% of patients for the longer term). In addition, alterations in bowel movements are frequent with 3–5 motions, commonly offensive, occurring each day.

Efficacy of Surgical Treatment for Obesity

Surgery is usually successful in inducing substantial weight loss in the majority of obese patients. This is achieved primarily by a necessary reduction in calorie intake.

In a review of RCT comparing different treatment strategies of obesity, surgery resulted in greater weight loss (23–28 kg more weight loss at 2 years) with improvement in quality of life and comorbidities.

The Swedish Obese Subjects (SOS) study demonstrated long-term beneficial effects on cardiovascular risk factors. The development of type 2 diabetes mellitus is most favorably influenced with a 14-fold risk reduction in those obese patients undergoing surgical treatment.

A Multidisciplinary Approach to the Management of Overweight and Obesity

Published evidence confirms that patients do better whatever the treatment when seen more frequently and for a greater length of time. Moreover, strategies that involve expertise incorporating dietetic, behavioral, and exercise experts as well as physicians and surgeons are also more successful in sustaining weight loss. This underlines the importance of a multidisciplinary approach. Treatment programs should include a system for regular audit and the provision for change as a result of the findings. Any center that claims to specifically provide expertise in weight management should incorporate the essential elements outlined in Table 7.

Table 7 Essential elements of an appropriate setting for obesity management

- Trained staff directly involved in the running of the weight loss program. These staff (medical, nursing, and other healthcare professionals) should have attended courses on the management of obesity and must be given the opportunity to continue their education
- Printed program for weight management that includes clear advice on diet, behavioral modification technique, physical exercise, and strategies for long-term lifestyle changes. Such a program may include a family and/or group approach
- Suitable equipment, in particular accurate and regularly calibrated weighing scale and stadiometer
- Specific weight-loss goals for patients with energy deficit being achieved by moderating food intake and increasing physical expenditure
- Documentation of individual patients' health risks. This will include BMI, waist circumference, blood pressure, blood lipids, and cigarette smoking and comorbid conditions
- A clearly defined follow-up procedure that involves collaboration between the different settings of care, and provides regular monitoring and documentation of progress, along with details of criteria for judging the success of weight loss. This will allow a weight loss program to be properly supported, medical conditions to be monitored, and problems or issues to be addressed at the earliest opportunity. It is also advisable to have a checklist of possible adverse drug effects, e.g., anxiety, disturbances of sleep, breathlessness, depression, and diarrhea.

Strategies for Weight Loss Maintenance

Preventing regain of fat losses is the major challenge of weight maintenance. A program to enable the individual to maintain their lowered weight must follow any successful weight loss. Published evidence suggests that a combination of dietary and physical activity modifications and reinforcement of behavioral methods are the most effective in the long term. These modifications needs to be integrated and accepted as a way of life and the responsibility for following this must lie with the patient.

See also: **Coronary Heart Disease**: Prevention. **Diabetes Mellitus**: Dietary Management. **Energy**: Balance. **Exercise**: Beneficial Effects; Diet and Exercise. **Hunger**. **Hyperlipidemia**: Overview. **Obesity**: Definition, Etiology and Assessment; Fat Distribution; Prevention.

Further Reading

Astrup A, Ryan L, Grunwald GK *et al.* (2000) The role of dietary fat in body fatness: evidence from a preliminary meta-analysis of ad libitum low fat dietary intervention studies. *British Journal of Nutrition* **83**(supplement 1): S25–S32.

Bravata DM, Sanders I, Huang J *et al.* (2003) Efficacy and safety of low carbohydrate diets. A systematic review. *JAMA* **289**(14): 1837–1850.

Bray GA, Bouchard C, and James WPT (eds.) (1998) *Handbook of Obesity*. New York: Marcel Dekker.

British Nutrition Foundation (1999) *Obesity*. London: Blackwell Science.

Colquitt J, Clegg A, Sidhu M *et al.* (2003) Surgery for morbid obesity (Cochrane Review). In: *The Cochrane Library*, issue 3. Oxford: Update Software.

National Institutes of Health (NIH), National Heart, Lung, and Blood Institute (NLLBI) (1998). *Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity: the Evidence Report*. Washington: US Government Press.

Department of Health (1991) *Dietary Reference Values for Food Energy and Nutrients for the United Kingdom: Report on Health and Social Subjects*, vol. 41. London: HMSO.

Frost G, Masters K, King C *et al.* (1991) A new method of energy prescription to improve weight loss. *Journal of Human Nutrition and Dietetics* **4**: 369–373.

Haddock CK, Poston WSC, Dill PL *et al.* (2002) Pharmacotherapy for obesity: a quantitative analysis of four decades of published randomised clinical trials. *International Journal of Obesity* **26**: 262–273.

Harvey EL, Glenny AM, Kirk SF, and Summerbell CD (1999). A systematic review of interventions to improve health professionals' management of obesity. *International Journal of Obesity* **23**: 1212–1222.

James WPT, Astrup A, Finer N *et al.* for the STORM Study Group (2000). Effect of Sibutramine on weight management after weight loss: a randomised trial. *Lancet* **356**: 2119–2125.

Klem M, Wing R, McGuire H *et al.* (1997). A descriptive study of individuals successful at long-term maintenance of substantial weight loss. *American Journal of Clinical Nutrition* **66**(2): 239–246.

Kopelman PG (2000) Obesity as a medical problem. *Nature* **404**: 635–643.

Kopelman PG and Stock M (eds.) (1999) *Clinical Obesity*. London: Blackwell Science.

National Task Force on the Prevention and Treatment of Obesity (2002). Medical care for obese patients: advice for health care professionals. *American Family Physician* **65**: 81–88.

Robert SB (2000) High glycaemic index foods, hunger and obesity: Is there a connection. *Nutrition Review* **58**: 163–169.

Royal College of Physicians of London (2003) *Anti-Obesity Drugs. Guidance on Appropriate Prescribing and Management*. London: RCP.

Sjostrom CD, Peltonem M, Wedel IT *et al.* (2000) Differentiated long-term effects of intentional weight loss on diabetes and hypertension. *Hypertension* **36**(1): 20–25.

Sjostrom L, Rissanen A, Andersen T *et al.* (1998) Weight loss and prevention of weight regain in obese patients: a 2-year, European, randomised trial of Orlistat. *Lancet* **352**: 167–172.

Oils *see* **Fats and Oils**

OLDER PEOPLE

Contents

Physiological Changes

N Solomons, Center for Studies of Sensory Impairment, Aging and Metabolism (CeSSIAM), Guatemala City, Guatemala

The Aging of the Population and its People

The maximal human life span is about 120 years. Approaching this degree of longevity, however, was not a prominent feature in the evolutionary phases of our species, *Homo sapiens*. The imperative was to survive the various mortal hazards long enough to reproduce and provide initial care for the offspring. The twenty-first century has ushered in an unprecedented longevity. The life expectancy of infants born today in Western Europe or Japan is over 75 years. The most rapidly increasing population segment in the world today is the centenarian. By the year 2020, there will be over 1 billion people over 60 years of age, constituting 13.3% of the global population, and three-quarters of them will be living in developing countries.

Many people are living a long time, but not all of them are healthy and functional throughout their lifespan. Chronic disability and the cost of health

services and custodial care are a growing burden on the economies of developed and developing countries alike. In order to understand the pathological aspects of advancing age, the normative pattern of changes in physiological function in older persons is an essential benchmark.

The Nature of Senescence

Aging has been described as "a series of time-related processes that ultimately bring life to a close," that is, a process of physiological 'wearing out.' Physiology is the basis of human functionality, as well as of our susceptibility to disease. The late gerontologist, Nathan Shock, established the principle of a progressive decline in physiological reserves as a consequence of 'normal' aging, recognizing that the rate of decline differed markedly among the body's organ systems. In fact, one cannot really separate the concept of the physiology of older persons from the physiology of the aging process itself. Similarly, the high prevalence of chronic diseases in older persons challenges our ability to discriminate 'normative' senescence from pathophysiological changes.

The origin of physiological changes in older persons begins within the domain of cellular senescence. The extension to tissue and organ levels originates in what we interpret to be the physiological changes of human aging. Major advances in our cellular and molecular understanding of basic aging processes have been made in recent years.

Cellular Senescence

In most tissues, with the notable exception of neural tissue, healthy cells are replicating cells, which are capable of mobilizing at least 20 enzymes and proteins that must be preassembled to initiate DNA synthesis for cell division. An irreversible state of growth arrest known as replicative senescence is the fundamental basis of cellular aging. Such senescent cells remain viable and metabolically active, but their genomic function and protein expression are distinct from that of normal, proliferating cells. Iron accumulates in senescent cells, possibly contributing to the greater oxidative stress and cellular dysfunction seen in senescent cells. Senescent cells also express proinflammatory enzymes, an internal process that could possibly contribute to the aging process; intercellular adhesion molecules, which are part of the inflammatory response, are overexpressed in association with senescent cells and aging tissue.

Telomeres and Telomerase

Telomeres are small units composed of the tandem DNA repeats and associated proteins, which cap the end of linear chromosomes and are responsible for maintaining chromosome length. They provide stability to the chromosome and protect against DNA loss associated with cellular replication. The mechanism of replicative arrest of senescent cells has been related to changes in the function of telomerase, a nuclear enzyme that synthesizes and maintains the telomeres. Shortening and uncapping of these structures, related to the number of past cell divisions, renders the DNA strand incapable of replication.

Apoptosis

Another factor involved in aging at the cellular level is the orderly 'retirement' of cells. For every cell that divides in, another would somehow have to make space for the extra cell in order to maintain numerical stability in the organ. This is achieved by a process of programed cell death, known as 'apoptosis.' Cell senescence disrupts these apoptotic processes. Necrosis, by contrast, is cell death due to injury or noxious stimuli. Diseases of aging may favor the necrotic process.

Mitochondrial Senescence and Oxidative Stress

The intracellular mitochondria, organelles involved in energy metabolism, are central to the process of cell senescence. They are also involved in regulating thermogenesis, calcium buffering, and integrating apoptosis. With aging, mitochondria become less efficient, in part due to mutations in the cell nucleus, derepressing the expression of proteins that compete with mitochondrial function. This disrupts energy metabolism for the cell and makes the mitochondria more porous, releasing reactive oxygen species into the rest of the cell. The mitochondrial production of reactive oxygen species is inversely proportional to longevity in animals. The oxidative activity also damages the mitochondria themselves. Mitochondria have their own DNA strands, and these accumulate mutations with age. In tissues dependent on progenitor (stem) cells, mitochondrial DNA mutations can disrupt replication.

Free radicals and reactive oxidative species can produce mutations in nuclear material and oxidize proteins and lipids throughout the cells. Aging involves an accumulation of oxidative damage at the cellular level, if not an increase in its intensity as well. The thiol-containing antioxidant mechanisms, typified by glutathione but represented by a number of sulfur-containing species, represent an important buffer against intracellular free radicals, but decline

with age due to downregulation of their synthetic enzymes. Confirming the cellular trend to oxidative stress in aging cells, clinical biomarkers of oxidation and antioxidant mechanisms reveal that systemic oxidative stress increases with aging characterized by lower concentrations of vitamins E and C and carotenes as well as lower activities of Cu-Zn-superoxide dismutase, catalase, and glutathione peroxidase.

Physiological Changes Occurring in Tissues and Organ Systems with Human Aging

Physiology has classically been organized around organ systems. According to this convention, the important features of the age-associated changes are enumerated and synthesized, with implications for human nutrition.

Integumentary Tissues

The integumentary tissues (skin, hair, nails) cover and protect the body. Two of the more classical and reproducible manifestations of aging can be seen in this system. The depigmentation of hair to gray or white is an almost universal aging effect given sufficient survival. Wrinkling of the skin, due to alteration in connective tissue composition, is another consequence of aging; it should be assessed by the changes in skin texture only in the non-sun-exposed regions of the body. Beyond the cosmetic consequences of the aging integumentary tissues, wound healing is a health-relevant consideration. Healing of wounds is slower with increasing age, but the resulting scars have the same tensile strength. Reduced recruitment of vessels of the microvascular is a function of aging.

The skin is an endocrine organ. Vitamin D is produced from the conversion of 7-hydroxy-cholesterol to cholecalciferol in the dermis of the skin. The efficiency of vitamin D decreases with age, such that older persons need a longer exposure to solar radiation to produce a given quantity of the vitamin.

Pulmonary and Respiratory System

Compliance of the chest wall changes with age, which gets stiffer and less compliant. The muscular force of the diaphragm is reduced with advancing years. The combination of these two factors reduces the maximal amount of air that can be moved into and out of the lungs. This diminution in the so-called forced vital capacity (FVC) of the lungs occurs as one gets older. There is less compliance, less recoil, and greater dead space. The original lung capacity, however, is sufficient to allow for sufficient gas exchange throughout life in the absence of underlying pulmonary disease. Nonetheless, the longitudinal Framingham Heart Study found an association between decrease in lung capacity and all-cause mortality.

The hygiene of the respiratory airways is somewhat compromised by a decreasing function of the microcilia of the bronchial epithelial cells. Since this mechanism is used to clear microbial pathogens, it has a direct influence on host defenses. Finally, since the basis of the respiratory system is an exchange of gases (oxygen, carbon dioxide, trace gases) with the bloodstream, any cardiovascular changes involving the right-side chambers of the heart will influence the overall gas-exchange efficiency for the body.

Cardiovascular and Circulatory System

For this system, it is necessary to separate the aging effects on the cardiac muscle and its apparatus from the aging of the vessels of the circulatory system, which transports blood to and from the heart. A characteristic of aging is a diminished resting cardiac output, which can have the combined bases of lower force of the cardiac muscle and a lesser oxygen demand for metabolism with diminished active-cell mass. Aging of the myocardium reduces its capacity for cellular repair and replacement. With aging, elevations of noradrenaline (norepinephrine) associated with downregulation of beta-1 receptors mimics the process of the failing heart. The compliance of the arteries emanating from the heart decreases with age. Stiffening of these vessels produces a progressive rise in the systolic blood pressure.

It is the circulation through smaller blood vessels and the generation of new vessels (neovascularization) that is a major concern with advancing years. The process of angiogenesis, through which new blood vessels are formed, is impaired during aging. The integrity of endothelial cells lining the vessels, the cascade of coagulation factors, and growth factors and neurochemical mediators and their respective receptors are all altered by aging in the neovascularization processes.

Oral Cavity and Alimentary Tract

The digestive tract is subject to functional changes with aging. Beginning in the oral cavity, loosening and loss of teeth is a frequent companion of aging. Saliva secretion decreases leading to relative degrees of xerostomia or dry mouth.

Reduced parietal cell function develops in older persons, but prior *Helicobacter pylori* infections are now thought to be a major cause of hypochlorhydria in later life. An important nutritional consequence of reduced gastric acid secretion is a lesser biological

availability of iron. Since iron stores are generally replete in both men and women in later life, this has little practical nutritional impact. The reduced secretion of gastric intrinsic factor, however, contributes to vitamin B_{12} deficiency, which is an important nutritional problem of older persons.

The capacity of the liver for biliary secretion and the pancreas for digestive enzyme and bicarbonate secretion begins adult life with a >90% excess of the necessary minimum. Secretory function declines with increasing age, but rarely falls below the minimal reserve capacity. The metabolic and detoxifying capacity of the human liver also has a reserve capacity and is not usually compromised by normal aging.

Intestinal motility is reduced with aging as a result of functional changes in the visceral nerves. With decreased transit the residence time of the chyme on the absorptive surfaces is longer, compensating for any senescence in the mucosal uptake itself. The reduction in motility produces the most noticeable and notorious of the manifestations of intestinal health in older persons, namely reduced frequency of defecations.

Musculoskeletal System

Bone mineral content declines with age; this aging process is known as 'osteopenia.' (It should be distinguished from the related pathological process in which bone architecture is altered, producing 'osteoporosis.') From the peak in the third and fourth decades, a 30% average decline in bone mineral density occurs through the ninth decade. In women, there is well-characterized acceleration of the rate of bone mineral loss immediately following the menopause. Decreasing levels of anabolic hormones may be associated with musculoskeletal atrophy and decrease in function that is observed in older women. This change in skeletal mineralization with aging is not associated with any apparent change in vitamin D nutriture as reflected in circulating levels of the vitamin.

The joints of the body undergo changes with the senescence of replacement of the cartilaginous substance, complicated by the pathological effects of cumulative use over the life span.

Recently, increasing attention has been given to the loss of muscle strength and substance with increasing age. Sarcopenia loss of lean body mass skeletal muscle mass replacement by fat mass Decreased creatinine-to-height ratio in normative aging in healthy subjects diminished grip strength is a function of age. [Reduction in muscle mass (sarcopenia obesity) is an important determinant of physical function and metabolic rate.]

Renal and Urogenital System

That renal creatinine and inulin clearance decreases with aging has been demonstrated for decades. These functional changes in filtration are associated with changes in the glomerular structure in the kidney. Circulatory senescence decreases blood flow to the kidneys, which further reduces the efficiency of renal clearance. The reserve capacity of these organs is such, however, that age-associated glomerular decline *per se* does not compromise the net excretion of nitrogenous waste.

Urine flow at the outlet is another aging consideration. The male urogenital system undergoes a characteristic aging change in the hypertrophy of the prostate gland, associated with decreased secretion of prostatic fluid. The anatomical consequence is a constriction in the passage through which urine flows from the bladder.

Gonads and Reproductive System

It has been aptly stated by Harman that: "It is clear that aging results in alterations of endocrine physiology, which in turn appear to contribute to development of the senescent phenotype." Aging is associated with a decrease in pituitary hormone secretions. This decline explains, in part, the reduction in gonadal hormone production with aging. Primary aging of the testes and ovaries themselves accounts for the remainder of the changes. As the ovaries have a finite number of eggs, ovulation can only continue through the number of cycles that correspond to the original store of ova. Menopause ensues with the characteristic cessation of estrogenic hormone secretion. In both sexes, gonadal androgenic hormone production declines with consequent effects on libido.

Endocrine Systems and Metabolism

As stated above, the pituitary gland is the hub of endocrine regulation. Important among the decline stimulation within the axis is that growth hormone (GH) secretion declines with increasing age, a condition termed 'somatopause.' The changes in the growth hormone/insulin-like growth factor axis with aging produce changes in function, metabolism, and body composition analogous to the pathological growth hormone deficiency seen in younger adults. Another change with age is the efficiency with which physical activity stimulates the secretion of GH.

The availability of hormones is not the only variable in endocrine signaling. Cellular and intracellular receptor function is complementary. An attractive explanation for the disordering of hormonal axes is

oxidative damage to cell membranes, compromising the function of receptors.

Basal and resting metabolism and diet-induced thermogenesis are all reduced with increasing age. Changes in body composition, and the replacement of lean tissue with fat and the increasing visceral distribution of fat, as well as decreasing physical activity, influence these metabolic changes of aging. Basal metabolic rate (BMR) declines in aging more than can be attributed to body composition changes and intracellular mitochondrial senescence may explain part of this discrepancy. For practical purposes, the standard oxygen consumption value equivalent to one metabolic equivalent (MET), that is, $3.5\,\mathrm{ml\,min^{-1}\,kg^{-1}}$, is not appropriate for elderly people.

Hematopoietic and Immune System

The formation of new red and white blood cells and platelets is one of the most proliferation-dependent physiological processes of the body. The various classes of circulating white cells are the underpinning of the host defense system, together with tissue macrophages, hepatic proteins, and the alimentary tract's mucosa.

Hematological aging The blood-forming organ is the bone marrow. Aging is associated with fatty infiltration of the marrow spaces in the long bones, but enough marrow remains to support the turnover of erythrocytes and red blood cell lines. The circulating red blood cell mass does not normally change with advancing age, nor does the normative peripheral white cell count or platelet number. As noted, iron stores tend to be abundant in later life; nutritional problems influencing red blood cell production are based on alterations in gastric function (vitamin B_{12} malabsorption), which result in a macrocytic (megaloblastic) anemia.

Immunological aging Circulating phagocytic white blood cells counts do not reduce with aging but aging does influence the innate host defense system. Mucosal barrier functions are influenced by aging of the gut in its interaction with microflora. Although not reduced in number, aged macrophages and neutrophils have blunted intracellular signaling by specific receptors, decreased metabolic functions, and impaired bacterial killing. Production of superoxide anion, chemotaxis, and orderly apoptosis of neutrophils is also disrupted by the disordered signaling. The tumor cell-destroying capacity of natural killer (NK) cells in the elderly is diminished.

More profound changes occur in the adaptive immune functions, which rely on the memory (T cell) lymphocytic cell line. Life-long antigen exposure induces increases in the number of memory T cells, but with enhanced reactivity against self-antigens, priming the individual for autoimmune disease. In healthy adults, IgA concentration increases by $0.2\,\mathrm{g\,l^{-1}}$ per decade throughout life. The T lymphocytes, however, respond more poorly to ongoing antigen assault in later life. Thymic involution associated with neural and hormonal changes of aging is an impediment to T-cell maturation in older persons. The basis of intrinsic function deficits of memory cells, on the other hand, has been ascribed to defective signaling and includes hyporesponsiveness to mitogen-stimulated proliferation and decrease in genetic suppression, allowing increased stimulation of inflammatory cytokines; the balance between pro- and anti-inflammatory cytokines shifts with aging, favoring the inflammatory pole, especially with the greater expression of interleukin 6. This has a negative systemic effect on bone metabolism, as well as dysregulating overall immune function.

Aging of mitochondria in the immune cell lines produces increased intracellular reactive oxygen species burdens. Finally, there is diminished programed death (apoptosis) of immune cells and dysregulation of apoptosis-dependent functions.

Central and Peripheral Nervous System

The integration of all senses and origins of all systemic coordination is a function of the brain and central nervous system. This is the one system in which proliferation of the primary cells (neurons) is not an issue after early childhood, although the supportive, nerve-tending (glial) cells continue to depend on replication and apoptosis for normal function.

Central nervous system The neurons of the brain continue to divide only through to the second year of life. Thereafter, the goal is to preserve the number and health of the cerebral nerve cell mass. Myelination of axons of nerve cells must be maintained throughout life. This is the function of the supporting cells (oligodendrocytes), which over 40 years continue to differentiate into myelin-producing cells. Free radicals pose a threat to these axon-tending cells, whose metabolic demands for producing the brain's cholesterol and maintaining its array of myelin sheaths render them particularly vulnerable to stress.

Positron emission tomography (PET) imaging of the aging brain has revealed and mapped the plethora of changes in blood flow and neurotransmitter metabolism that occurs with advancing years.

Special senses The special senses related directly to the cranial nerves (vision, hearing, taste, and smell) experience age-related change. With respect to vision, the most typical of all biological aging changes is presbyopia, or the loss of accommodation function for the ocular lens with loss of capacity of the associated musculature. The consequence is loss of near-vision, which leads to the need for reading glasses or bifocal spectacles. A more important aging change related to the lens is the opacification that leads to cataract formation. The eye is designed to translate light energy into visual images, but the energy of light, particularly the ultraviolet β rays of solar energy, damages ocular tissue. Thus, there is as a strong environmental component to the disarranging of the laminar stacking of the fibrillar proteins of the lens, which imparts its clear, transparent basis; consumption of diets high in antioxidant vitamins has been associated with the delay in cataract formation.

Age-related hearing loss is a feature of biological aging. It affects the cochlear neural structures and leads to loss of acuity, especially for higher pitched tones. It is speculated that apoptosis of the most vital neural cells drives this hearing loss, based on mutations in the mitochondria due to life-long free-radical stress.

Taste and smell acuity decline with aging, both in sensitivity and in accuracy of recognition. Since these combined senses account for the recognition of flavors, their diminution with age could affect appetite and reduce the enjoyment of meals.

Cognitive function The intellectual, reasoning, and memory functions of the cerebral cortex decline with increasing age. This has been a universal observation in general elderly populations. The debate is whether this is a consequence of neurodegenerative diseases (pathological change) or a biological correlate of aging (senescence). Continued intellectual stimulation has been posited as an approach to retard cognitive decline, and a role for B-complex vitamins and antioxidants has been advanced.

Peripheral nervous system Vibratory perception in the peripheral extremities is the classical index of peripheral nervous decline with aging. Less well appreciated is the effect of aging on pain perception, in which there can be a numbing of sensation or, less commonly, an accentuation of perception. Pain perception from the visceral organs is often dulled, which can have adverse implications for the early detection of organic diseases. All of the peripheral nerve dysfunction can result from the compensatory sprouting of axonal limbs to compensate for the loss of motor neurons. This is well directed at first, but with further aging the synaptic connections are poorly directed and motor function suffers as a consequence.

Drug Metabolism

The metabolism of drugs and pharmacological agents is not the purview of any single organ system. Older persons tend to be prescribed increasing numbers of medications with advancing age. Important changes in drug metabolism occur with aging. Metabolism and disposition of drugs changes with age. This involves age-associated decrease in function of some, but not all, cytochrome P450 enzymes. Among the pharmacokinetic and pharmacodynamic changes that occur with advancing age are reductions in renal and hepatic clearance and an increased effective half-life of lipid-soluble drugs. The older population shows increased sensitivity to some psychotropic drugs and anticoagulants, with the frail elderly being more susceptible than healthy elders.

Synthesis and Conclusion

The number of older people is increasing in all regions and all societies of the world. Advancing age produces senescent changes in cellular function that are reflected in a declining capacity of all physiological systems. The increased prevalence of disease in older populations Aging is a major risk factor for disease but does not necessarily lead to age-related diseases.

All physiological systems are intrinsically interrelated in maintaining the health and function of the organism. Aging is associated with a loss of complexity in the dynamics of many physiological systems. It has been speculated that the basis for the syndrome of frailty in older persons may result from a reduced ability to adapt to internal and external stresses of daily life due to the loss of dynamic coordination among the interrelated physiological systems.

The alterations in physiological functions with aging have important implications for absorbing, retaining, and utilizing nutrients. The extent to which dietary patterns and nutrient intakes are accelerating or retarding the rates of functional decline is a matter of ongoing investigation in gerontological nutrition and physiology.

See also: **Aging**. **Brain and Nervous System**. **Cytokines**. **Older People**: Nutritional Requirements; Nutrition-Related Problems; Nutritional Management of Geriatric Patients. **Osteoporosis**. **Vitamin K**.

Further Reading

Ahluwalia N (2004) Aging, nutrition and immune function. *Journal of Nutrition, Health and Aging* **8**: 2–6.
Balin AK (ed.) (1994) *Practical Handbook of Human Biologic Age Determination*. Boca Raton: CRC Press.
Harman SM (2004) What do hormones have to do with aging? What does aging have to do with hormones? *Annals of the New York Academy of Science* **1019**: 299–308.
Hayflick L (2003) Living forever and dying in the attempt. Experimental Gerontology **38**: 1231–1241.
Hutchinson ML and Munro HN (1986) *Nutrition and Aging: Bristol-Meyer Nutrition Symposia*, vol. 5. Academic Press.
Leveille SG (2004) Musculoskeletal aging. *Current Opinion in Rheumatology* **16**: 114–118.
Lipsitz LA (2004) Physiological complexity, aging, and the path to frailty. *Science of Aging Knowledge Environment* **16**: 16.
Mishra SK and Misra V (2003) Muscle sarcopenia: an overview. Acta Myol **22**: 43–47.
Park HL, O'Connell JE, and Thomson RG (2003) A systematic review of cognitive decline in the general elderly population. *Internation Journal of Geriatric Psychiatry* **18**: 1121–1134.
Timiras PS (1994) *Physiological Basis of Aging and Geriatrics*, 2nd edn. Baton Raton: CRC Press.

Nutritional Requirements

N Solomons, Center for Studies of Sensory Impairment, Aging and Metabolism (CeSSIAM), Guatemala City, Guatemala

Human Aging and Nutrition

The World Health Organization defines the 'elderly' as persons of 60 years of age and older. The elderly constitute a rapidly expanding segment of populations in both developed and developing countries. This is the combined result of ever-longer survival and dramatic reductions in fertility rates. Regardless of age, people must respond to their feelings of hunger and thirst by consuming foods and beverages. This eating and drinking behavior also serves to provide the nutrients to nourish the body. The amount of a nutrient that must be ingested and absorbed to maintain an adequate and appropriate body composition varies with age across the life span, depending on basic underlying physiological and metabolic processes specific to the chronological stage of life. The degree to which we retain and conserve, or excrete or degrade, absorbed nutrients is influenced by chronological age and biological aging.

As a consequence of this new demographic reality, attention is being focused belatedly on gerontology and its nutritional biology; this, in turn, is reflected in very recent efforts to refine our knowledge of the amounts of various macro- and micronutrients that the aging body requires (nutrient requirements) and of the amounts that must be consumed in the diet to provide for sufficient uptake of these nutrients (nutrient recommendations).

Successful Aging, Normative Aging, and Frailty

From an epidemiologic and demographic, as well as an economic and humanitarian standpoint, the ideal contribution of life-long nutrition would be to a situation of 'compression of morbidity,' first enunciated by J. Fries. It strives to keep individuals free of chronic illness, functional, and independent until the final moments of their lives, and thus reduces the burden of disability and dependency suffered by individuals, their families, and the society that contributes to their maintenance to a minimum.

A disclaimer has traditionally been appended to the official pronouncements of recommended nutrient intakes; whether they are from national or international expert panels, the prescriptions are meant to apply to 'healthy' individuals. Nutrient needs in disease conditions are considered to be a clinical matter, and are related to the pathologies in question.

When it comes to older persons, the exigency of being 'healthy' becomes immediately problematic. Advanced age is associated with increased susceptibility to chronic and degenerative illnesses. Most persons over 60 years of age have two or three chronic illnesses diagnosed, and are receiving multiple medications. Maintaining a rigid definition of healthy for application of nutrient recommendations in later life would exclude almost everyone from coverage by nutrient-intake standards.

In fact, the older the cohort of individuals examined, the more heterogeneous are individuals of the same chronological age in their physical and cognitive functioning. Over the last two decades, general domains of classifications have come into usage to embrace the heterogeneity of aging populations: successful aging; usual aging; and frailty. Successful aging has been defined as multidimensional, "encompassing the avoidance of disease and disability, the maintenance of high physical and cognitive function, and sustained engagement in social and productive activities." It may involve aspects of resilience and wisdom, as well. Usual aging involves an accumulation of ailments and loss of function that is typical of older persons surviving to later life. Frailty is the far extreme of disability and dependency associated with major physical and cognitive decline in which disease and senescent processes become irreversibly established.

A prominent and optimistic school of thought suggests that exposures to behavioral and environmental

factors that modify risk of disease and dysfunction determine one's position in these alternative outcomes in the aging process. In this view, more optimal nutrient intake, food selection, and life-style choices could reduce the heterogeneity, retaining more individuals in the successfully aged category for most of their life span. Others consider that genetic constitution may be as important in determining the course of aging as any positive or negative influences during our lifetime.

Overview of Specific Factors of Aging Influencing Nutritional Requirements

The discussion of nutrient requirements and recommended dietary intakes of nutrients in older persons has proceeded on both the theoretical and empirical level. Since the peak years for human reproduction occur before advanced middle age, and well before older age begins, the forces of selective reproduction cannot exert themselves for Darwinian selection of traits favoring longevity in the evolution for any traits related to longevity *per se* or physiological sustained function. Hence, there is little evolutionary selection for nutrient requirements to achieve advanced age or for long-term survival. It is more for the preservation of comfort and function for those surviving to advanced age that optimization of nutritional intakes for the elderly would apply, that is for humanitarian and public health importance in the face of the physiological and anatomic changes of senescence.

As early as the 1970s, nutritional scientists advanced the proposition that requirements for different macro- and micronutrients changed with age. A large number of conjectures based on an emerging scientific understanding of senescent physiology have been advanced. It has been suggested that the decreased physical activity and physical conditioning associated with the body composition changes attendant to aging, sets the stage for alterations in requirements in both amounts and relative proportions of protein and the energy-yielding macronutrients. Decreased gastric secretory capacity has a negative influence on the absorption of calcium, iron, and vitamin B_{12}. Changing intestinal motility and digestive function evoked considerations of distinct increases and decreases of nutrients to compensate for the senescence of the intestinal tract, with particular interest in dietary fiber. Attention to compensatory intake for all of the nutrients involved in skeletal mineralization has come to the fore in relation to the recognized tendency to bone mineral loss with advancing age.

The immune and host defense system has been the focus of gerontological nutrition. Increased intakes of both vitamin E and zinc, well above the normally recommended level, have stimulated certain immune functions in studies involving older volunteers. Cognitive function declines with advancing age, and it has even been suggested that adjustment of nutrient intake can favorably affect the retention of memory and cognitive function in older persons. The adequate intake of B-complex vitamins, particularly those related to homocysteine metabolism (vitamin B_{12}, folic acid, vitamin B_6, riboflavin), are associated with mental function in older age. It has also been suggested that older individuals need more *n*-3 fatty acids for preserving cerebral cellular anatomy related to cognition.

Nutrient Intake Recommendations in Later Life

Comprehensive recommendations for macro- and micronutrients with differential attention to older persons have arisen from a collaboration between the US and Canada, and from expert panels serving the United Nations System. Each panel has set out its methodology and definitions and then presented tables of quantitative estimates. The recommendations for persons considered elderly in the respective systems are outlined below.

Definitions Surrounding Recommended Intakes of Nutrients

An important advance in establishing nutrient intake recommendations relates to the semantics. There has been a refining of the operational definitions of terms related to nutrient intakes. Recommended nutrient intakes (RNIs) are set by the agencies of the United Nations (UN) System and are considered to be the intakes of nutrients required to satisfy the requirements of nearly all healthy persons of a given age, sex, and physiological condition, and should be universal for all regions of the globe.

The Food and Nutrition Board of the Institute of Medicine in the US took a new approach in 1997 in which they applied the new dietary reference intakes (DRI) to micro- and macronutrient intakes. This work was undertaken jointly with Canada. It began with an assessment, where possible, of the estimated average requirement (EAR). This is defined as "the average daily nutrient intake level estimated to meet the requirement of half the healthy individuals in a particular life stage and gender group." The EAR is critical for an assessment of the risk of a nutrient deficiency problem at the

population level. The traditional criterion used for decades, the recommended dietary allowance (RDA), is preserved. It is defined in the DRI process as "the average daily nutrient intake level sufficient to meet the nutrient requirement of nearly all (97 to 98 percent) healthy individuals in a particular life stage and gender group." When an EAR cannot be established from which to derive a formal RDA, the DRI process has a 'fall-back' category known as adequate intake (AI); this is defined as "a recommended average daily nutrient intake level based on observed or experimentally determined approximations or estimates of nutrient intake by a group (or groups) of apparently healthy people that are assumed to be adequate." A new classification scheme involving a range of intakes was created specifically for energy, electrolytes, and liquids: the acceptable macronutrient distribution ranges (AMDRs).

For the first time, a specific and well-defined process to delimit levels of excess intake of nutrients and dietary substances was defined by the DRI process as the upper tolerable intake levels (UL). The UL is "the highest average daily nutrient intake level likely to pose no risk of adverse health effects to almost all individuals in the general population." It is considered that as intake increases above the UL, the potential risk of adverse effects increases. To date, the UN System's process has dealt much less explicitly with issues of excessive intake of nutrients and dietary substances.

Established Recommended Intakes for Older Persons

In earlier versions of the RDAs for the US population (up to the 10th edition in 1989), the nutrient recommendations for all healthy adults over 51 years of age were combined as a single value. For the UN System, the age threshold in the early editions was 50 years or older. Concerted efforts to refine our understanding of nutrient requirements for older adults have been made over the past two decades. This allowed the US-Canada DRI process to establish categories for men and women aged 70 years and older. For the WHO/FAO process, a specific estimation for individuals over 65 years has been provided in the 2002 micronutrient recommendations.

Given the magnitude of the theoretical considerations regarding senescence and aging physiology that have been raised by various authors, what is really surprising is the paucity of specific instances in which the recommended intakes of nutrients for men or women in the 'elderly category' are considered to be different from persons in the next youngest age category. Composite tables for men (**Table 1**) and women (**Table 2**) are given for all of the nutrients and dietary substances expressed in the US-Canada DRIs and in the UN system for RNIs.

Macronutrients In the DRI system, a universal, individual protein requirement was established as 0.80 g of good-quality protein per kilogram of body weight per day independent of age. No evidence for altered protein requirements with older age has been found. Moreover, it is recommended that the contribution of protein to total energy intake should not exceed 30%. The US Food and Nutrition Board also established an amino acid pattern in 2002. It specifies the density (mg per g protein) of seven indispensable (essential) amino acids (histidine, isoleucine, leucine, lysine, threonine, tryptophan, and valine) and for two amino acid combinations (methionine + cysteine, phenylalanine + tyrosine). This pattern is universal from age 1 year to the extremes of older age without modification.

It has long been recognized that energy recommendations cannot be made on a group basis, as each individual has his or her own daily energy requirement dependent on the amount of energy one is forced to expend with metabolic reactions, food processing, and physical exertion. In the DRI process, this is recognized in an effort to individualize the estimation of energy intake. Estimated energy requirement (EER) is based on the amount of energy needed to maintain energy balance in relation to one's total energy expenditure. The DRI process for the US and Canada has published general EER equations (multidimensional nomograms) by which a reasonable estimate of an individual energy requirement can be calculated. There are general equations for adult men and women (over 19 years of age), based on consideration of physical activity level, weight, and height. In addition, there is an age term in the general EER, which is attached to a negative (minus sign) term in the equation. This signifies that energy requirements decline as a function of advancing years.

Although dietary fiber is not considered to be an 'essential' nutrient, the DRIs give a recommended level for intake. Curiously, in light of the active discussion of the role of fiber for the elderly in colonic function, the recommendations for intake by men decline from 38 to 30 g per day and in women from 25 to 21 g per day after 50 years. These are continued throughout the 70 year period, as well. This is a consequence of the fiber recommendations being pegged to total average energy intake.

Table 1 Nutrient intake recommendations for older males

	UL[a]	*ERA*[a]	*RDA/AI*[a] *AMDR*[a]	*RNI*[b]
Macronutrients				
Water (l)	–	–	**2.1**[c]	–
Carbohydrate (g)	–	100	120	–
Protein (g)	–	46	56	–
Total fat (g)	–	–	20–35	–
n-6 PUFA (g)	–	–	14[c]	–
n-3 PUFA (g)	–	–	1.6[c]	–
Dietary fiber (g)	–	–	**30**	–
Vitamins				
Vitamin A (RAE)	3000	625	900	600 (μg RE)
Vitamin D (mg)	50	–	**15**[c]	**15**
Vitamin E (mg α-tocopherol)	1000	12	15	10 (mg α-TE)
Vitamin K (μg)	–	–	120[c]	65
Vitamin C (mg)	2000	75	90	45
Thiamin (mg)	–	1.0	1.2	1.2
Riboflavin (mg)	–	1.1	1.3	1.3
Niacin (mg)	35	12	16	16
Vitamin B_6 (mg)	100	1.4	1.7	1.7
Biotin (mg)	–	–	30[c]	–
Pantothenic acid (mg)	–	–	5[c]	5
Folic acid (μg)	1000	320	400	400
Vitamin B_{12} (μg)	–	2.0	2.4	2.4
Choline (mg)	3500	–	550[c]	–
Elements				
Sodium (g)	2.3	–	1.2[c]	–
Potassium (mg)	–	–	4.7[c]	–
Chloride (g)	3.6	–	1.8[c]	–
Calcium (mg)	2500	–	**1200**[c]	**1300**
Phosphorus (mg)	**3000**	580	700	–
Magnesium (mg)	(350)	350	420	**230**
Iron (mg)	45	6	8	14[d]
Zinc (mg)	40	9.4	11	7.0[e]
Iodine (μg)	1100	95	150	130
Copper (mg)	10	0.7	0.9	–
Fluoride (mg)	10	–	4[c]	–
Manganese (mg)	11	–	2.3[c]	–
Chromium (μg)	–	–	**30**[c]	–
Selenium (μg)	400	45	55	34
Molybdenum (μg)	2000	34	45	–

[a]In DRIs 70 years plus is considered as 'older'.
[b]In UN System (WHO/FAO/IAEA) 65 years plus is considered as 'older'.
[c]Recommendation in the form of adequate intake.
[d]Assumes a 10% bioavailability of iron from the diet.
[e]Based on the assumption of a moderate bioavailability of zinc.
The figures in **bold** denote recommendations specifically modified for ageing (see text).
UL, upper tolerable upper intake level; EAR, estimated average requirements; RDA, recommended dietary allowance; AI, adequate intake; AMDR, acceptable macronutrient distribution range; RNI, recommended nutrient intake; PUFA, polyunsaturated fatty acids; RAE, retinol activity equivalents; RE, retinol equivalents; α-TE, alpha-tocopherol.

Water It is recommended in the DRI as an adequate intake (AI) that males over the age of 70 require 2.6 l and females 2.1 l of water per day; this is a decline from the 51–70-year age group, where the daily water intake recommendations were 3.7 l and 2.6 l, respectively. It is further suggested that males and females over 70 years of age derive 81% of their daily water allowance from beverages and 19% as the metabolic water from foods. This is consistent throughout adulthood from age 19 years. Hence, there is no consideration of a higher requirement for water intake with older age. With respect to the electrolytes, no differences in AIs exist across the ages in adulthood.

Micronutrients A number of recommendations (RDAs or AIs) change with advancing age in the DRI system; this is indicated by the bold type in

Table 2 Nutrient intake recommendations for older females

	UL	*ERA*[a]	*RDA/AI*[a] *AMDR*[a]	*RNI*[b]
Macronutrients				
Water (l)	–	–	**2.6**[c]	–
Carbohydrate (g)	–	100	120	–
Protein (g)	–	38	46	–
Total fat (g)	–	–	20–35	–
n-6 PUFA (g)	–	–	11[c]	–
n-3 PUFA (g)	–	–	1.3[c]	–
Dietary fiber (g)	–	–	**21**	–
Vitamins				
Vitamin A (RAE)	3000	500	700	600 (μg RE)
Vitamin D (mg)	50	–	**15**[c]	**15**
Vitamin E (mg α-tocopherol)	1000	12	15	7.5 (mg α-TE)
Vitamin K (μg)	–	–	90[c]	55
Vitamin C (mg)	2000	60	75	45
Thiamin (mg)	–	0.9	1.1	1.1
Riboflavin (mg)	–	0.9	1.1	1.1
Niacin (mg)	35	11	14	14
Vitamin B_6 (mg)	100	1.3	1.5	1.5
Biotin (mg)	–	–	30[c]	–
Pantothenic acid (mg)	–	–	5[c]	5
Folic acid (μg)	1000	320	400	400
Vitamin B_{12} (μg)	–	2.0	2.4	2.4
Choline (mg)	3500	–	425[c]	–
Elements				
Sodium (g)	2.3	–	1.2[c]	–
Potassium (mg)	–	–	4.7[c]	–
Chloride (g)	3.6	–	1.8[c]	–
Calcium (mg)	2500	–	**1200**[c]	**1300**
Phosphorus (mg)	**3000**	580	700	–
Magnesium (mg)	(350)	265	320	**190**
Iron (mg)	45	5	**8**	11[d]
Zinc (mg)	40	6.8	8	4.9[e]
Iodine (μg)	1100	95	150	110
Copper (mg)	10	0.7	0.9	–
Fluoride (mg)	10	–	3[c]	–
Manganese (mg)	11	–	1.8[c]	–
Chromium (μg)	–	–	**20**[c]	–
Selenium (μg)	400	45	55	26
Molybdenum (μg)	2000	34	45	–

[a]In DRIs 70 years plus is considered as 'older'.
[b]In UN System (WHO/FAO/IAEA) 65 years plus is considered as 'older'.
[c]Recommendation in the form of adequate intake.
[d]Assumes a 10% bioavailability of iron from the diet.
[e]Based on the assumption of a moderate bioavailability of zinc.
The figures in **bold** denote recommendations specifically modified for ageing (see text).
UL, upper tolerable upper intake level; EAR, estimated average requirements; RDA, recommended dietary allowance; AI, adequate intake; AMDR, acceptable macronutrient distribution range; RNI, recommended nutrient intake; PUFA, polyunsaturated fatty acids; RAE, retinol activity equivalents; RE, retinol equivalents; α-TE, alpha-tocopherol.

Tables 1 and **2**. The change in recommendations occurs at either age 50 or 70 years. In women over 50 years, the RDA for dietary iron decreases from 18 mg to 8 mg day^{-1}; there is no change in requirement for the 70 year plus age group. This lower value is the recommendation for adult men of all ages. The fact that the menopause allows women to replete iron stores depleted by an adulthood of monthly menstrual blood loss accounts for this lower RDA in older women.

The senescence of the skeletal system and the reduction of bone mineral content with age is a major nutritional concern in gerontological nutrition. In recent revisions of the recommendations, evidence for the need for increases in both vitamin D and calcium for older persons has led to changes in the estimates of requirements for these nutrients in later life. Within the DRI system the RDA for vitamin D for males and females over 70 years is 15 mg. The

RDA for adults over 50 years is 10 mg and for young adults is 5 mg. Similar increases in vitamin D intake with age are recommended by the FAO/WHO. With respect to calcium, the recommended levels increase from 1000 mg for younger adults to 1200 mg at age 50 and beyond in the DRI system, and from 1000 to 1300 mg in the FAO/WHO standards. These are justified based on the higher propensity for skeletal fractures after 70 years of age associated with epidemiological evidence of widespread vitamin D deficiency in this age group, and evidence showing a reduction in bone loss with daily calcium intakes exceeding 1000 mg after mid-life.

With respect to chromium it is interesting that the estimation for AI declines with advancing age. The AI for persons over 70 in the DRI is the same as that for individuals between 51 and 70, but it is 5 $\mu g\,day^{-1}$ higher for the 19–50 age range. This reduction is tied to the lower energy demands for individuals over 50 years of age.

The upper tolerable upper intake level (UL) for phosphorus in the DRI system is 3000 $mg\,day^{-1}$ for both men and women over 70 years as compared to 4000 $mg\,day^{-1}$ for adults in the 19–70 age group. This lower tolerance is explained by the greater prevalence of impaired renal function in advanced old age.

Magnesium intake recommendations in the FAO/WHO guidelines decline for individuals over 65 years by 30 $mg\,day^{-1}$ compared to those in the 51–65 years age group. An anomalous finding for the magnesium RDA in the DRI system, which applies to all adult age groups, is that the UL for magnesium has been set at 350 mg. This is only 30 mg higher than the 320 mg daily recommended for older women, and is 70 mg lower than the 420 mg daily intake recommended for older men.

Dietary Guidelines for Health, Function, and Disease Prevention

Concomitant to recommendations for daily nutrient intake based on requirements, guidance and orientation for the pattern of selection of nutrient sources among the food groups have emerged as so-called 'dietary guidelines.' They are often accompanied by an icon or emblem, such as a pyramid in the US, a rainbow in Canada, and a Hindu temple in India, each of which expresses the general tenets of the dietary guidelines in a visual manner. A quantitative prescription, or some notion of balance among foods and food groups, is the basis of dietary guidelines; there is also often a proscription for foods considered to be harmful or noxious.

The additional susceptibility of older persons to chronic degenerative diseases makes adherence to these healthful dietary patterns, throughout the periods in the life span preceding the older years, more relevant. Recent epidemiological research has shown that compliance or behavior concordant with healthy eating guidelines are associated with lower later life incidences of certain cancers, cataracts, diabetes, hypertension, stroke, and cardiovascular diseases, as well as overall survival. There is intense interest in whether and how diet and nutrition influence the maintenance of cognitive function with aging.

Robert Russell and colleagues constructed a food guidelines pyramid, which specifically focused on the health of the elderly. Among the elements and tenets that differed from the standard US pyramid are the following recommendations: to drink additional water and liquid; to increase consumption of dietary fiber; and to consider dietary supplements such as calcium and vitamin E. Otherwise, selecting the same requisite serving portions of the specific food groups, and avoiding excess sugar, salt, and separated fats as indicated by the conventional guidelines emblem is recommended for the older population as well.

It is generally conceded that the major benefits for prevention of nontransmissible disease to be derived from adopting a healthful life style and dietary habits will accumulate over a lifetime; hence, beginning such practices at as early an age as possible will yield the greatest benefits. In this context, the application and emphasis of dietary guidelines specifically for the elderly is controversial and as yet unresolved. One school of opinion, one shared by Russell and coworkers, holds the view that the benefits of adhering to dietary guidelines are continuous, and actively protect from metabolic and neoplastic diseases even in the latter stages of the life span. The alternative proposition suggests that long-term survivorship is a manifestation of a superior genetic constitution resistant to chronic diseases. The very fact of survival to advanced age is a suggestion that the survivor's dietary practice will neither prejudice nor further protect health.

Barriers to Meeting Recommended Nutrient Intakes and Healthful Dietary Intake Patterns by Older Persons

The late Professor Doris Calloway, in the early 1970s, commented: "People eat food, not nutrients." This highlights the paradoxes in considering and enumerating the objectives of dietary intake at the level of the

chemical composition, while most members of the general public are uninformed as to the nutrient composition of the foods and beverages in their diets.

Elderly persons face a number of challenges in meeting their recommended nutrient intakes. In the first instance, they are likely to be those with the least sophisticated or available knowledge of the nutrients required and the food sources to provide them. The social, economic, and physiological changes imposing on the lives of persons surviving to advanced age pose logistical problems for their selecting and purchasing a diet. Economic dependency and the limited incomes of older persons may restrict their access to high-quality foods. Social isolation, depression, and impaired mobility, as well as chewing difficulties may limit the variety of items included in the diet with advancing age. In some circumstances, it may be that free-living and independent elders are relatively less able to optimize their nutrient intake and dietary pattern compared to more dependent individuals served or fed in institutional settings.

The exigencies of consuming a healthful diet for the prevention of chronic diseases, emphasizing a plant-based diet rich in whole grains, fruits and vegetables, limits the nutrient selection that would be obtained from an even wider variety of foods and food-groups. Specific essential fatty acids, and certain minerals (calcium, zinc, selenium) and some vitamins are far less nutrient dense in foods of vegetal origin, setting a dilemma between consuming for nutrient adequacy and prevention of degenerative disease.

The widespread fortification of processed foods with micronutrients by the food industry in industrialized nations could mitigate much of the risk of insufficient nutrient intakes. Fortification of foods with iron may be more disadvantageous than beneficial to older persons whose iron reserves for nutritional purposes should normally be amply filled. A higher intake of iron puts a strain on the intestinal regulatory capacity and would tend toward excessive iron storage with any oxidative consequences for health that may result. National programs of fortification of grain flours and cereal products with folic acid for the prevention of neural tube defects in pregnancies are proliferating; mathematical models of how this policy may in fact increase the masking of macrocytic anemias due to vitamin B_{12} deficiency associated with senile gastric atrophy have been generated for European elderly populations.

Future Considerations

The DRI recommendations are specifically derived for the populations of the US and Canada in North America. The RNIs of the UN System are meant to be universal across the entire world. The slight majority of the living elderly are currently to be found in the low-income, largely tropical regions of the world in which 80% of the global population reside; this shift is due to rise rapidly over the next two decades. A number of caveats apply to the estimation of nutrient intake recommendations for the elderly across the world. If the "applies only to healthy individuals" disclaimer were applied to the developing world, then virtually no older people would qualify as eligible for coverage by any nutrient recommendations system. However, rather than abandon the effort for nutrient intake guidance, an attempt should be made to take into account the influences of life-long climatic issues (heat, humidity) and ecological factors (parasites, recurrent infections) on nutrient needs in later life.

Nature versus nurture issues will also continue to be debated with regard to nutrient requirements, especially in later life. Of course, with the recent assembling of the genetic code, the issues of 'nutrigenomics' and 'nutrigenetics' theoretically could soon be brought to bear on understanding individual variation in needs for and tolerances of essential and nonessential nutrients and dietary bioactive substances. The significance of this potential for the already aged person is likely to be limited for two reasons. First, the accumulative effects of nutrient imbalance will already have been established. Second, the economic and intellectual wherewithal to access and execute such individualized prescriptions for nutrient intakes and dietary patterns will likely escape the majority of older persons with limited financial means. Hence, further refinements in recommended intakes for older persons are likely to remain at the level of this segment of the population as a group and will involve establishing evidence that increased intakes of specific nutrients will have health-protective effects or function-enhancing properties, and that the effective upper tolerable levels/limits for certain nutrients in later life are lower than those for younger members of the adult population.

See also: **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Bone**. **Calcium**. **Chromium**. **Cobalamins**. **Fatty Acids**: Omega-3 Polyunsaturated. **Folic Acid**. **Food Fortification**: Developed Countries. **Older People**: Nutrition-Related Problems; Nutritional Management of Geriatric Patients. **Riboflavin**. **Supplementation**: Role of Micronutrient Supplementation. **Vitamin B_6**. **Vitamin E**: Physiology and Health Effects. **Zinc**: Physiology.

Further Reading

Blumberg J (1997) Nutritional needs of seniors. *Journal of American College of Nutrition* **16**: 517–523.

Drewnowski A and Warren-Mears VA (2001) Does aging change nutrition requirements? *Journal of Nutrition in Health and Aging* **5**: 70–74.

Food and Nutrition Board/World Health Organization (2002) *Recommended Nutrient Intakes*. Geneva: WHO.

Fries JF (2002) Successful aging–an emerging paradigm of gerontology. *Clinics Geriatric Medicine* **18**: 371–382.

Hartz SC, Russell RM, and Rosenberg IH (1992) *Nutrition in the Elderly. The Boston Nutritional Status Survey*. London: Smith-Gordon.

Institute of Medicine, Food and Nutrition Board (2000) *Dietary Reference Intakes. Applications in Dietary Assessment*. Washington, DC: National Academy Press.

Rowe JW and Kahn RL (2000) Successful aging and disease prevention. *Advances in Renal Replacement Therapy* **7**: 70–77.

Russell RM, Rasmussen H, and Lichtenstein AH (1999) Modified Food Guide Pyramid for people over seventy years of age. *Journal of Nutrition* **129**: 751–753.

Russell RM (2000) The aging process as a modifier of metabolism. *American Journal of Clinical Nutrition* **72**(supplement 2): 529S–532S.

Solomons NW (2002) Nutrition and the extremes of life: dilemmas and enigmas of advanced old age. *Asia Pacific Journal of Clinical Nutrition* **11**: 247–250.

World Health Organization/Tufts University School of Nutrition and Policy (2002) *Keep Fit for Life: Meeting Nutritional Needs of Older Persons*. Geneva: WHO.

Nutrition-Related Problems

C P G M de Groot and W A van Staveren,
Wageningen University, Wageningen, The Netherlands

The population older than 55 years of age is increasing rapidly throughout the world. In industrialized countries, the proportion of elderly people will increase by approximately 1% per year; in developing countries an increment of approximately 3% per year is expected. The nutritional needs of this population will require increasingly more attention from professionals working in the food industry as well as in health care.

Aging is defined as all physiological changes that occur from conception until old age and ultimately death. In this article, the term is restricted to changes that occur in adulthood, when growth has stopped.

On the one hand, nutrition is considered one of the key determinants in the process of aging. On the other hand, age-related changes take place in body appearance, in functional capacity and in the body's capacity to adapt to physical stress that affects nutritional needs.

It is difficult to distinguish between changes due to old age *per se* and changes that are the consequences of disease. In this article, the effects of aging on body composition, including energy needs and problems of over- and underweight, bone mass, and water balance are discussed. Physiological functions of the digestive system, malabsorption, nutrient drug interactions, and consequences for nutritional requirements are described, together with the high-risk micronutrients and early warning signs for malnutrition.

Changes in Body Composition and Energy Needs

Fat-Free Mass and Energy Needs

One of the truly age-driven phenomena is the loss of muscle mass and strength, called sarcopenia. It is distinct from muscle loss (cachexia) caused by inflammatory disease or from weight loss and attendant muscle wasting caused by starvation or advanced disease. Regardless of major differences between individuals, aging-related changes in body composition with time are universal. In addition to changes in lean tissue, this also holds for changes in fat mass, body water, and bone mass.

Throughout middle age, body mass tends to increase due to an accumulation of fat, preferentially intra-abdominally. Thereafter, usually after 60 years of age, it declines in association with loss of lean tissue. Diminution of physical activity enhances the changes in body composition occurring with aging, which in turn affect physical function. Ultimately, these processes result in a lower requirement for energy.

The total demand for energy is dominated by the energy needed per day to maintain vital functions, the basal metabolic rate (BMR), representing 60–70% of total energy expenditure. Most of the remainder (approximately 25%) is needed to cover the costs of physical activities. The BMR declines with age by up to 5% per decade. It is the decrease in lean tissue with age that determines this decline. One of the most important preventive measures in this process is the maintenance of physical activity. This helps to maintain lean body mass, physical fitness, and the requirement for energy.

Partly as a response to reduced energy needs, the energy intakes of affluent populations decline with age. This decline in food intake involves a decrease

in meal size and a reduction in between-meal snacks. Morley called this physiologic decline in food intake "anorexia of aging." This type of anorexia may be considered partially as a response to reduced energy needs but also partially as a dysregulation of food intake. Roberts *et al*. studied energy regulation in young and older adults by deliberately overfeeding and underfeeding their subjects. After a period of underfeeding, young people compensated by overeating when fed ad libitum. However, the older adults did not compensate. The same holds following a period of overfeeding; the older adults did not compensate with a reduction in food intake when fed ad libitum.

Anorexia of aging places an increasing number of elderly people at risk for malnutrition because the opportunities for providing an adequate dietary nutrient intake are very limited when total food consumption becomes low (e.g., <6.3 MJ (1500 kcal) (**Figure 1**). Current recommendations for daily energy intake are approximately 9 MJ for elderly men and approximately 8 MJ for elderly women. Institutionalized elderly people or the elderly who are sick are especially likely to fail to achieve such intakes.

For health reasons, it is important that elderly people avoid becoming underweight. Although losing weight may be favorable at younger ages and being overweight is a known health risk in adults, there is evidence that low body weight and loss of body weight in the elderly are more strongly associated with risk of mortality (**Figure 2**). This is clearly shown by data from the Survey in Europe on Nutrition and the Elderly, a Concerted Action (SENECA). Weight loss (>5 kg over 4 years) seemed to be predictive for survival. It is even more important to be slightly overweight than underweight for people older than age 70 years. Therefore, except for those who are obese, elderly people should be encouraged to maintain an adequate energy intake. According to the SENECA study, 20–25% of the relatively healthy participants failed to do so: Approximately 8% lost and 16% gained at least 5 kg of body weight over a period of 4 years. When appetite is reduced, an increase in meal frequency may not only help to promote energy intakes but also prevent blood glucose levels from declining steeply.

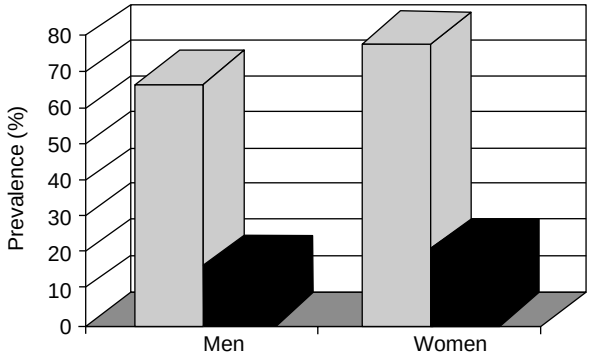


Figure 1 Prevalence of inadequate intake of at least one nutrient among elderly people whose daily energy intake is less than 6.3 MJ (gray bars) and for those whose energy intake exceeds 7 MJ (black bars). (From De Groot CPGM, van Staveren WA, Dirren H *et al.* (eds.) (1996). SENECA, nutrition and the elderly in Europe. Follow-up study and longitudinal analysis. *European Journal of Clinical Nutrition* **50**(supplement 2):127, with permission from Macmillan Press Limited.)

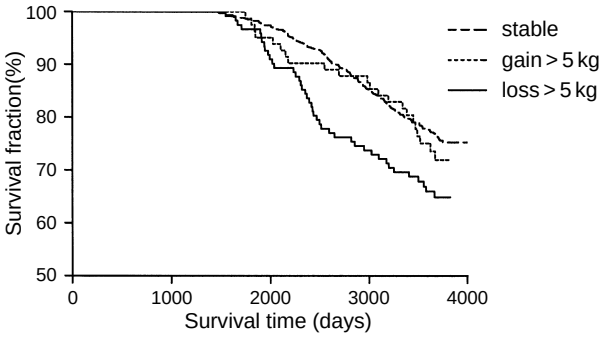


Figure 2 Probability of survival for participants from the SENECA study with and without weight change in the first 4 years. (Reproduced from Thomas D (ed.) (2002) Undernutrition in older adults. *Clinics in Geriatrics* **18**(4), with permission of WB Saunders.)

Body Water, Dehydration, and Medication

Because lean tissue has a high water content, there is a decrease in total body water—especially extracellular water—with advancing age from 80% at birth to 60–70% after age 70 years. In addition, older people experience diminished sensation of thirst, and urinary concentrating ability declines as a function of age. Thus, older people have an increased risk of dehydration, particularly when diuretic or laxative medicines are used or in the presence of some diseases common in old age, such as diarrhea, renal disease, and infection with fever. Because water is essential to all biological functions, fluid intakes during old age should be at least 1700 ml per day. In the body, water acts as a dilutent for water-soluble drugs. Given the decrease in body water with age, older people may need lower dosages of water-soluble drugs than younger adults to achieve the desired therapeutic effect and to avoid drug toxicity.

Bone Mass and Nutritional Factors

Throughout life, bone mass changes, with a maximum (peak bone mass) achieved by age 25–30 years and bone loss occurring after the fourth decade. Higher calcium intakes in childhood and early adulthood result in a 3–8% greater bone mass later in life, thereby improving the key factor in the osteoporotic process and the age-associated risk of fractures. In women, there is a perimenopausal increase in the rate of bone loss that persists after menopause following a decline in oestrogen production (**Figure 3**).

Factors other than age and sex that are associated with low bone mass include low body weight, smoking, alcohol consumption, reduced physical activity, low calcium absorption, and secondary risk factors such as the use of steroids. Although there is still uncertainty about the quantitative role of nutritional factors in the pathogenesis of osteoporosis, preventive measures include adequate calcium intakes (probably even in old age) and exposure to sunlight to ensure vitamin D adequacy and/or dietary supplementation with vitamin D. Restricted sunlight exposure, reduced capacity of the skin to produce vitamin D, and low vitamin D intake make elderly people prone to vitamin D deficiency.

Nutritionally Related Problems and the Digestive System

Taste

The number of taste buds varies widely from person to person but does not decline with age. Taste perception and the perceived flavor of foods decrease, but this is affected by many factors, including diminishing smell, age-related changes in the olfactory system, the integration of the central nervous system, medication, oral hygiene, and nutrition (**Figure 4**). Inadequate intakes of zinc, copper, nickel, and some vitamins have been associated with decreased perception of flavor of food.

Stomach

Atrophic gastritis is common in elderly people, resulting in hypochlorhydria and reduced gastric secretion. There is no consensus whether atrophic gastritis and the decrease in gastric acid secretion are normal processes of aging or a result of *Helicobacter pylori* infection. Independent of the cause, lack of gastric acid may interfere with the optimal absorption of nutrients. In the case of vitamin B_{12}, this may have clinical consequences. The mechanism for the reduced absorption of vitamin B_{12} is not clear, but protein-bound colabamin absorption is reduced in the elderly with atrophic gastritis or hypochlorhydria and this situation can be reversed with free vitamin B_{12} (the crystalline form) administration. Again, the prevalence of vitamin B_{12} deficiency is higher in *H. pylori* infection and eradication of this bacteria may correct vitamin B_{12} levels.

Small Intestine

Transit time is not changed in the elderly; however, their slower gut motility may cause stasis, with

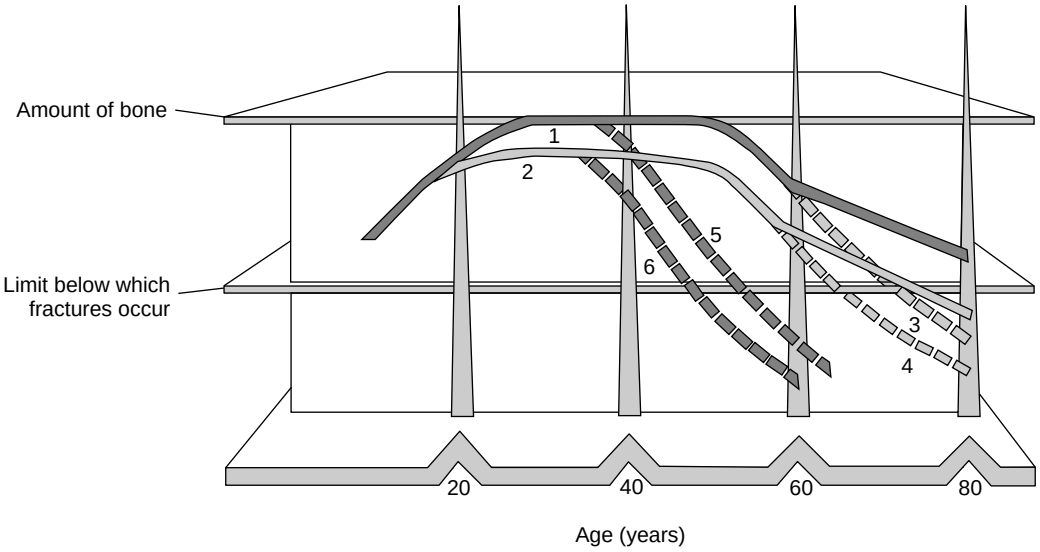


Figure 3 Rate of bone loss: 1, bone mass change in women with a high initial amount of bone and an average loss after menopause; 2, bone mass change in women with a low initial amount of bone and an average loss after menopause; 3 and 4, bone mass change in women with high losses after menopause; 5 and 6, bone mass change in women with an early menopause or after surgical removal of ovaries. First fractures occur approximately 10 years after menopause.

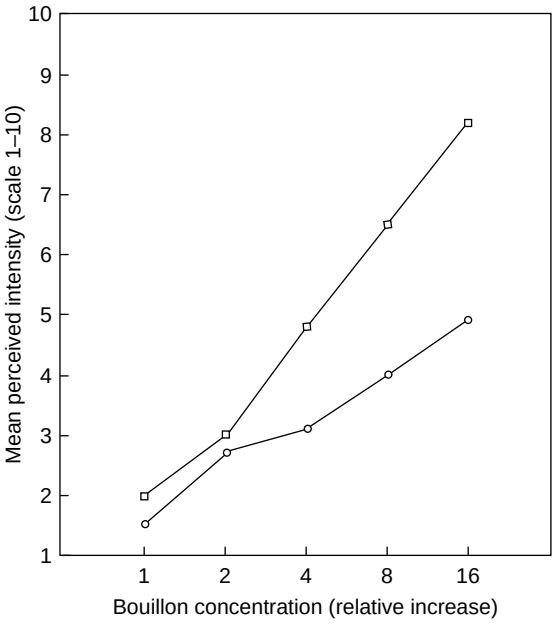


Figure 4 Mean responses of perceived intensity of bouillon flavor judged by a group of 23 elderly subjects (circles) and 32 young subjects (squares). (From Graaf C, Polet P and van Staveren WA (1994) Sensory and pleasantness of food flavors in elderly subjects. *Journal of Gerontology: Psychological Sciences* **94**: 93–99.)

bacterial overgrowth and malabsorption. The latter also may be caused by a reduced mucosal surface due to poor oxygenation of the tips of the villi as a result of a decreased blood supply from a low cardiac output. Decreased absorption is most frequently observed for electrolytes, lactose, vitamin D, and calcium. Absorption of digested food takes place by diffusion and by active transport across membranes. Adequate fluid must be available for absorption to proceed. The dehydrated state of the aged can reduce the capacity of the gut to absorb digested food.

Colon

The mucosal and muscle layers of the colon may atrophy, resulting in weakening of the muscle wall. Reduced motility of the colon allows prolonged exposure of feces to water absorption and drying. Reduced bulk results in further reduction of the stimulus to muscle contraction and will lead to constipation. This may be enhanced by a diet lacking dietary fiber, little physical activity, and poor tone of abdominal muscles.

Liver

Liver reserves of vitamins A, D, and B_{12} are unlikely to be diminished. Protein synthesis and especially the synthesis of vitamin K-dependent factors are reduced. However, it is not clear if this affects vitamin K requirements.

Gastrointestinal System

Elderly people are more prone to nutritional deficiencies that cause significant functionally deleterious consequences. The gastrointestinal system has been identified as an important cause of these problems. However, the mild functional and anatomic age-related changes do not seem to explain the incidence of malnutrition in the elderly. Rather, the lower functional reserve may accelerate nutritional problems under pathological challenge.

Nutrient and Drug Interactions

Many elderly people use drugs. In Europe, 83% of 'apparently healthy' people in the previously mentioned SENECA study use an average of two types of drugs, with antihypertensives (33%), analgesics (31%), diuretics (24%), sleeping pills (18%), and psychotropic drugs (17%) taken most often. Many drugs taken by the elderly can interfere with nutritional status. The possible effects include suppression or stimulation of appetite and impaired nutrient absorption and metabolism. For example, lisdiuretics can have adverse effects on calcium metabolism, salicylates can increase the need for vitamin C, and some types of antihypertensives act as antagonists of vitamin B_6. Negative consequences of laxatives, often taken by the elderly, include interference with nutrient absorption. Dietary interventions may help to reduce the intake of drugs. There is evidence that moderate sodium restriction prevents or delays the development of hypertension. Also, limiting alcohol intake provides protection because approximately 10% of cases of hypertension in men have been attributed to alcohol. Culinary skills become important to ensure that elderly people continue to find eating enjoyable, especially because increases in olfactory and taste thresholds occur with aging (**Figure 4**).

Risks for Malnutrition

Elderly people most at risk of developing malnutrition are those who eat little because of poverty, disability resulting from chronic geriatric disease, or a combination of these factors. Malnutrition is found in elderly people living in their homes if they are indigent, isolated, or homebound because of their own disability or the serious illness of their partner. Ten main risk factors for noninstitutionalized elderly can easily be identified and acted on by nonmedical personnel (**Table 1**). It must be understood that each risk in **Table 1** is only a potential

Table 1 Early warning signs for malnourishment

Medical and physiological factors
Recent unintended weight change of approximately >5% in the preceding month
Disease, polypharmacy or long-term medication
Immobility
High alcohol consumption
Psychological factors
Bereavement and/or depression or loneliness
Mental confusion
Poor nutrional knowledge
Socioenvironmental factors
Lack of sunlight
Low budget for food
Missed meals or snacks

danger sign; each has to be considered in relation to others. It should be stressed, however, that malnutrition is much more common in the elderly in long-term care, especially those who are unable to feed themselves.

Recommended Nutrient Intakes and High-Risk Nutrients in the Elderly

Recommended nutrient intakes for the elderly and very old may be set with different objectives. The values may serve either diagnostic or prescriptive purposes. Today, gerontologists and nutritionists are interested in the amount of nutrient that it takes to prevent a chronic disease from occurring rather than the amount of nutrient it takes to prevent a deficiency state. Most countries have their own set of requirements and age specificity may differ. In **Table 2** values are given as published by the Institute of Medicine (USA and Canada) for the oldest age group, mostly 70 years and older. The values should be accepted cautiously, with the proviso that change may be desirable when new information on nutritional needs of the elderly becomes available; for elderly patients who belong to particular disease groups, including mental diseases; for elderly people using specific drugs; and for elderly

Table 2 Recommended daily allowances (RDA) and observed problems for selected food components

Component	*RDA*[a]	*Problems*
Energy (MJ)		
Men	9–11	Low energy intake (<6.3 MJ) is highly correlated with insufficient micronutrient supply.
Women	8–10	
Protein (g/kg body weight)	0.8	Protein turnover may be lower than in young adults, which indicates lower requirement. However, the efficiency of protein synthesis is decreased.
Vitamin A (μg)		
Men	900	Risk of toxicity from megadoses in supplements.
Women	700	
Vitamin D (μg)	15	Requirement is increased in old age owing to insufficient synthesis with little or no exposure to UV light.
Thiamin (mg)		
Men	1.2	Special attention in those who eat little and elderly with alcoholic problems.
Women	1.1	
Riboflavin mg		
Men	1.3	Those consuming few animal products, especially milk, may be at risk.
Women	1.1	
Vitamin B_6 (mg)		
Men	1.7	Requirement may be higher when using antihypertensive drug hydralazine.
Women	1.5	
Folate (μg)	400	Extra attention for patients with atrophic gastritis and patients using a number of medicines.
Vitamin B_{12} (μg)	2.4	Vegans and patients with hypochlorhydria and atrophic gastritis have high risk; some drugs may interact.
Vitamin C (mg)		
Men	90	Increased requirements for patients using salicytes. Be alert for low vitamin C supply when using cooked meals from catering services and insufficient supply of fresh fruits.
Women	75	
Calcium (mg)	1.2	High-risk groups include elderly people using little or no milk and milk products and patients using lisdiuretica and some other drugs.
Iron (mg)	8	With reduction in lean body mass, iron requirement may be decreased. However, occult blood loss may increase requirement.
Iodine (μg)	150	Supply often inadequate; in some places enriched products (salt) should be used.
Water (ml)	1500–2000	Attention to fluid intake is necessary.

[a]Values derived from recent reports of the Institute of Medicine's Food and Nutrition Board (1999–2002), except RDAs for energy and water. The latter data are derived from the Expert Group Nutrition and the Elderly, The Netherlands (1995).

people using specific diets that may reduce the absorption of some nutrients.

Dietary Guidelines

Dietary surveys do not indicate that dietary guidelines for the elderly should be totally different from those for younger adults. Emphases in the program, however, should be different. Nutrition education programs for the elderly should give priority to drinking habits and to promoting the consumption of foods that are good sources of calcium, zinc, magnesium, potassium, folate, and vitamin B_6. Thus, recommendations should focus on the importance of daily consumption of (green) vegetables, fruit, wholegrain products and fortified cereals, and (low-fat) milk and milk products, and they should emphasize the nutritional value of fish and legumes. Because greater variety is associated with higher nutrient intakes in the elderly, the recommendation to eat a wide variety of foods is also important. Studies have emphasized that a healthy diet as well as other lifestyles, such as being moderately physically active and not smoking, are still important at an older age.

The Use of Dietary Supplements

The few studies on dietary supplementation among elderly people suggest that, as for younger adults, those elderly who need supplements do not use them, whereas the elderly consuming a diet with a high nutrient density use supplements. Food supplements include specially formulated preparations containing vitamins, minerals, and protein or a combination of these and other ingredients. Unnecessary use of supplements should be discouraged because consumption of megadose levels (amounts exceeding 10 times the recommended daily allowance) of various nutrients may cause adverse health effects. However, there are situations in which supplements have a role to play. For example, vitamin D would be indicated for the housebound elderly, and vitamin B_{12}, folate, potassium, or other nutrients may be a necessary supplement in disease conditions or when certain drugs are used that influence nutrient absorption, utilization, or excretion. In addition, suitable supplementation provides a means for improving the nutritional status of malnourished elderly or preventing nutritional deficiencies in people who are at risk.

See also: **Body Composition**. **Bone**. **Colon**: Nutritional Management of Disorders. **Dental Disease**. **Dietary Guidelines, International Perspectives**. **Drug–Nutrient Interactions**. **Energy**: Balance. **Liver Disorders**. **Malabsorption Syndromes**. **Older People**: Physiological Changes; Nutritional Requirements; Nutritional Management of Geriatric Patients.

Further Reading

De Groot CPGM, van Staveren WA, Dirren H *et al.* (eds.) (1996) SENECA, nutrition and the elderly in Europe. Follow-up study and longitudinal analysis. *European Journal of Clinical Nutrition* **50**(supplement 2): 127.

Department of Health (1992) *The Nutrition of Elderly People*, Report on Health and Social Subjects No. 43. London: HMSO.

Expert Group Nutrition and the Elderly (1995) *Nutrition of the Elderly*. The Hague: Netherlands Food and Nutrition Council.

Haveman-Nies A, de Groot CPGM, and van Staveren WA (2003) Relation of dietary quality, physical activity and smoking habits to 10 year changes in health status in older Europeans in the SENECA study. *American Journal of Public Health* **3**: 318–322.

Hazzard WR, Bierman EL, Blass JP *et al.* (1994) *Principles of Geriatric Medicine and Gerontology*, 2nd edn. New York: McGraw-Hill.

Rosenberg IH and Sastre A (2002) *Nutrition and Aging*, Nestle Nutrition Workshop Series, Clinical & Performance Program, vol. 6. Basel: Karger.

Schűrch B and Scrimshaw NS (eds.) (2000) Impact of human aging on energy and protein metabolism and requirements. *European Journal of Clinical Nutrition* **54**(supplement 3): S1–S165.

Thomas D (ed.) (2002) Undernutrition in older adults. *Clinics in Geriatrics* **18**(4).

Nutritional Management of Geriatric Patients

M-M G Wilson and J E Morley, St Louis University, St Louis, MO, USA

Undernutrition

Overwhelming evidence implicates undernutrition as a major index of increased mortality in older adults. Undernourished elders admitted to acute facilities are more likely to develop complications, resulting in increased length of stay and healthcare costs. Rehabilitative efforts are less rewarding as patients often fail to return to baseline functional status and are more likely to require long-term placement or emergency readmission.

Free-living older persons with suboptimal nutritional status are at increased risk of dependence on care givers as a result of compromised activities of daily living. Additionally, convincing evidence exists linking undernutrition with an increased incidence of frailty, gait instability, falls, hip factures, immune dysfunction, delayed wound healing, and decreased cognitive function. Nevertheless, nutritional

assessment and dietary management are often overlooked when health professionals evaluate geriatric patients.

As many as one-third of older adults in the US may be undernourished However, early clinical detection and appropriate intervention occur in less than one-tenth of cases. Health-care providers must remain astutely aware that geriatric health maintenance mandates efficient nutritional evaluation, surveillance, and prompt intervention.

Diagnosis and Evaluation of Undernutrition

Anthropometry

Several anthropometric indices have been proffered for the evaluation of undernutrition in older adults. These include:

- body weight less than 80% of the ideal body weight for height and age;
- weight loss exceeding 10% of baseline weight in the preceding 6 months; or
- body mass index less than 17.

Erroneously, the normative references for most of these criteria are younger subjects.

Within the older population the usefulness of this index is hampered by the lack of age-adjusted reference values. Reference values applicable in younger adults are not suitable for use in older persons as sarcopenia, age-related skin changes, and vertebral osteoporosis with height loss confound such norms. Within the older population, intentional weight loss resulting from dietary restriction should not discourage comprehensive nutritional assessment, as recent evidence indicates that both voluntary and involuntary weight loss in older persons portend similar adverse health outcomes.

Calculation of the body mass index (BMI) is considered to be one of the most objective anthropometric indices, as it permits correction of body weight for height. The BMI, calculated by dividing the weight in kilograms by the height in meters squared, is based on the proven premise that weight in the younger adult increases proportionately with height. However, this concept is false in older persons as height is significantly affected by age-related changes. Loss of height with aging occurs secondary to shortening of the axial skeleton due to age-related osteoporosis, degenerative disc changes, vertebral thinning, and kyphoscoliosis. Furthermore, using height as an anthropometric index is impractical in nonambulant and bed-bound persons. Nevertheless, clinical use of the BMI in the older population has been preserved by the development of adapted nomograms. Such nomograms are based on the determination of BMI using surrogate parameters of height adapted from the appendicular skeleton, which is relatively unaffected by age-related osseous changes. These parameters include total arm length, arm span, erect forearm length, and knee to floor height.

Skin fold thickness measurements are also used as anthropometric indices of total body fat in younger adults. However, the precise relationship between skin fold thickness and total body fat is unpredictable, as is the response of subcutaneous fat to undernutrition. Furthermore, in the older adult, the accuracy of this technique is confounded by age-related qualitative and quantitative changes in body fat. Altered compressibility of body fat has also been shown to occur with aging, rendering skin fold thickness measurements unreliable for use in older adults. Measurement of mid-arm circumference is another frequently used anthropometric index. However, several factors influence muscle bulk including exercise, disease, and genetic factors. In the older person this index is of doubtful clinical utility.

Several factors confound the use of anthropometric indices, underscoring the importance of serial measurements. These allow for quantification of response to intervention and also enhance accuracy of data interpretation by utilizing intrasubject comparison. More accurate methods of body composition analysis are available but are unlikely to be suitable for routine clinical use. These include computerized tomography, bioelectrical impedance, nuclear magnetic resonance imaging, *in vivo* neutron activation analysis, dual energy X-ray absorptiometry (DEXA) and direct photon absorptiometry (DPA). Because of alterations in body water with aging, the value of bioelectrical impedance is questionable. The DEXA technique is excellent but the migration of body fat to the abdomen with aging may result in an underestimation of body fat in older persons. DPA is based on analysis of tissue attenuation of photons transmitted at two different energy levels. This technique permits measurement of different tissue compartments. Both fat mass and fat-free mass can be measured using this technique. Currently, these methods are used almost exclusively for research purposes. Most of these emerging techniques are very expensive and have not been validated for use in clinical settings. Therefore, for practical clinical purposes, the most cost-effective nutritional parameter of proven clinical utility in older adults remains serial body weight measurements.

Biochemistry

Hypoalbuminemia is often erroneously used as an index of undernutrition. However, the diagnostic specificity of this index is poor. Serum albumin levels are determined by a complex interplay between nutritional intake, total body albumin distribution, and several pathological changes that alter the biosynthetic and catabolic rates of albumin. In the acutely ill or stressed older person, cytokine release suppresses albumin and prealbumin synthesis. Additionally, the release of catabolic counter-regulatory hormones in stressful situations reduces albumin synthesis even further. Direct downregulation of albumin gene expression also occurs in situations of acute stress. Paradoxically, undernutrition itself may result in a compensatory reduction in albumin catabolism, yielding inappropriately high albumin levels. Although serum albumin is a poor index of undernutrition, hypoalbuminemia is linked with frailty, excess comorbidity and increased mortality in older adults. Thus, the clinical relevance of hypoalbuminemia lies in the identification of a high-risk subset of older persons in whom early and aggressive nutritional intervention is crucial.

Several other biochemical indices are used as nutritional markers. However, like albumin, they lack diagnostic specificity and have relatively long half-lives, which limit their value in the serial evaluation of undernutrition. Insulin-like growth factor 1 is considered to have the greatest positive predictive value as it has been shown to correlate well with nutritional status even during periods of acute stress. Added advantages of this index are a relatively short half-life of 2–6 h and a rapid response to fasting and refeeding. Nonetheless, routine use of this assay in the evaluation of undernutrition is precluded by cost. Overall, for practical clinical purposes, the use of biochemical markers in routine nutritional geriatric management is costineffective and unreliable.

Hematology

Anemia of chronic disease resulting directly from undernutrition is a recognized clinical entity. Studies have identified reduced erythropoiesis and alterations of erythrocyte function in undernourished persons that respond to nutritional repletion. Iron and folate deficiency anemias may also result from inadequate micronutrient intake in undernourished persons.

Measurement of the total lymphocyte count (TLC) is helpful mainly in the stratification of the severity of undernutrition. A TLC of less than $1200 \times 10^6 1^{-1}$ indicates mild undernutrition while counts less than $800 \times 10^6 1^{-1}$ are usually found in severely undernourished persons.

Recognizing Causative Factors of Undernutrition

Age-related physiological reduction in appetite, 'anorexia of aging,' is well documented. Several factors have been implicated in the genesis of this phenomenon. Evidence suggests that the decrease in lean body mass, energy expenditure, and metabolic rate that occurs with advancing age may partially account for the reduction of food intake in healthy older persons. Age-related reduction in olfactory and gustatory receptor sensitivity may compromise the hedonic qualities of meals, further reducing the desire to eat. Similarly, age-related alterations in hormonal and neurotransmitter-mediated function may also play a role in suppressing food intake. Animal studies suggest that aging results in a reduction in the opioid feeding drive and an increase in the satiating effect of cholecystokinin. This may lead to the ingestion of smaller meals and prolonged periods of satiety between meals. More recently ghrelin, a hunger-inducing peptide hormone, has been shown to decrease with age. Similarly, older hypogonadal men have inappropriately high levels of leptin, a satiation-inducing peptide hormone.

The occurrence of a variety of pathological factors superimposed on the background of age-related physiological changes may further compromise nutritional status in the older adult (**Table 1**). Existing data suggest that as many as one-third of undernourished older persons suffer from untreated depression. Neuro-vegetative symptoms in depressed older persons often result in anorexia, social withdrawal, reduced motivation, and decreased activity, all of which can compromise nutritional intake. The use of appropriate antidepressants very often reverses these symptoms, resulting in an increase in food intake and restoration of adequate nutritional status. Choice of antidepressants is crucial in the management of depressed, older undernourished persons. The popularity of selective serotonin reuptake inhibitors in younger persons has led to their increasing use in the older population. However, in older persons the efficacy of such agents in improving mood may be marred by adverse gastrointestinal effects, such as nausea, vomiting, and diarrhea, which may further compromise nutritional status. Thus, where such agents are used, careful monitoring of nutritional status is mandatory. Mirtazapine is a useful antidepressant that is unrelated to selective serotonin

Table 1 Common and uncommon causes of undernutrition in older persons

Reduced food intake
Anorexia
Ill-fitting dentures
Periodontal disease
Oropharyngeal disease
Orofacial dyskinesias
Psychosocial factors
Depression
Eating disorders
Bereavement
Social isolation
Low financial income
Physical/mental disability
Persistent tremors
Dyskinesia/dyspraxia
Arthritides
Parkinsonism
Cerebrovascular disease
Dementia
Behavioral disorders
Increased nutrient metabolism
Hyperthyroidism
Phaeochromocytoma
Wandering, agitation
Movement disorders
Hemiballismus
Reduced nutrient utilization
Malabsorption syndrome
Chronic inflammatory bowel disease
Gluten enteropathy
Gastroesophageal disease
Inflammatory
Neoplastic
Dysmotility
Multifactorial
Chronic bronchitis, emphysema
Cardiac failure
Malignant disease
Substance abuse

reuptake inhibitors, tricyclics, or monoamine oxidase inhibitors (MAOI). Mirtazapine belongs to the piperazino-azepine group of compounds. Available evidence suggests that Mirtazapine has an additional orexigenic and anti-emetic effect, which may increase energy consumption. Electroconvulsive therapy is a viable option in depressed persons with severe anorexia. Evidence exists in support of the efficacy of this treatment modality in restoring appetite following failure of pharmacological agents.

Minor dysphoric changes may adversely affect nutritional status and warrant intervention. Over 30% of older community-dwelling persons live alone, usually as a result of bereavement or migration of younger family members. Meals are often eaten alone and the lack of social interaction during meal preparation and consumption can compromise the recreational and hedonic aspects of dining. Consequently, such elders are poorly motivated to prepare and eat meals. Particular attention should be paid to the recreational aspects of mealtimes, and older persons should be encouraged to socialize during meals. This can be accomplished in a variety of ways. Participation in dining clubs, where available, should be encouraged. Arrangements can also be made for older persons to dine at senior citizens' centers. Ambulant senior citizens should be encouraged to eat out, if this is preferred.

Effective nutritional intervention mandates due consideration of financial and socioeconomic factors. Approximately one-third of the older population live below the poverty line and many experience difficulty with the purchase of food items necessary to ensure a balanced diet. Inadequate transportation, limited mobility, and poorly accessible shopping facilities may be added limiting factors. Social and community agency services should be considered where relevant, and an attempt should be made to provide appropriate assistance.

A wide variety of prescribed drugs can cause anorexia, nausea, and other symptoms of gastrointestinal distress in older persons, rendering medication review an important component of nutritional management. Digoxin, theophylline, and nonsteroidal anti-inflammatory agents are frequent culprits in this regard. Enquiry must also be made into the use and tolerance of self-prescribed medication. Offending drugs, once identified, must be discontinued. Iatrogenesis also contributes to undernutrition by way of therapeutic diets. Low-cholesterol and low-salt diets are often prescribed to older persons on the basis of data extrapolated from younger persons. There is currently little evidence to suggest that these diets are of any benefit to older persons when used as primary prevention strategies. Available data actually indicate increased mortality in older adults with low-cholesterol levels. Evidence suggests that hypocholesterolemia may reflect increased cytokine expression in acutely ill and frail older adults. Thus, restrictive diets in older persons should be discouraged, as they often reduce palatability and consequently discourage food intake. Health professionals should also make enquiries regarding self-prescribed diets. Studies indicate that the older population is more susceptible to food fads and advertised commercial diets, which are often unbalanced and of dubious benefit. Prolonged ingestion of such diets can result in marked undernutrition.

A wide variety of medical illnesses require focused therapeutic intervention in order to maintain or restore adequate nutritional status. Degenerative and

neurological diseases can significantly impair mobility and physical function. The use of adapted appliances and cutlery in such cases may improve manual dexterity and preserve the ability to self-feed. In older persons with severely impaired function, who are unable to cook, meal delivery services ('meals on wheels') may be an acceptable alternative to home-cooked meals. Tooth loss is another important risk factor for undernutrition. Periodontal disease and edentulism are highly prevalent among the geriatric population and can impair masticatory ability. Older persons who have lost teeth, experience pain on mastication, or receive inadequate dental care should be carefully screened and offered appropriate therapy. The use of dentures may improve food intake. However, where dentures are poorly tolerated, alteration in the consistency of meals is helpful. Dysphagia occurs commonly in older persons with degenerative and vascular neurological conditions such as dementia, Parkinsonism, and cerebrovascular disease. A bedside swallowing evaluation should be an integral component of nutritional evaluation, followed by a modified barium swallow with fluoroscopy in cases where significant dysphagia is identified. In most cases oral food intake will remain possible, with appropriate modifications regarding swallowing technique, feeding precautions, and food consistency.

Health professionals often wrongly assume that older adults possess adequate knowledge of basic dietetic practice and nutritional studies. There is evidence to suggest that the nutritional attitudes and knowledge of undernourished older persons may be inadequate, particularly with regard to food preparation. Dietary education and counseling are crucial components of nutritional intervention in undernourished older persons who retain the responsibility for preparing their own meals. Such counseling should be targeted towards identifying deficits in basic dietary knowledge and the correction of poor nutritional practices.

Nutritional Assessment Tools

Arrays of nutritional screening tools have been developed to facilitate the identification of older persons at risk for undernutrition. The Nutrition Screening Initiative (NSI) in the US stemmed from a collaborative effort between family physicians, dietitians, and the National Council of Aging. This is a three-tiered tool formulated to assist in the detection of older persons at risk for nutritional compromise and subsequent direction of such persons toward the appropriate level of care. The first level of screening is designed to be initiated by the patient or primary care giver. Persons identified to have an increased risk of undernutrition are then referred for evaluation by healthcare or social services personnel. This constitutes the second level of screening. The identification of factors that may warrant medical intervention will prompt referral to a physician for further evaluation. The NSI is of proven value as an epidemiological tool and serves to increase the awareness of patients and care givers to undernutrition. However, its usefulness within orthodox settings may be hampered by the number of personnel and services required, which may constitute a significant drain on available resources. Added drawbacks to the use of this tool for the individual patient are the lack of professional supervision at initiation and reliance on patient compliance in adhering to the specified clinical pathway protocol.

The Mini Nutritional Assessment (MNA) is a comprehensive and simple tool designed to evaluate the nutritional status of older persons. This is the first well-validated nutritional screening instrument and is recommended for use in people aged over 75 years. Cross-validation indicates that nutritional assessment using this tool will accurately evaluate and categorize nutritional status in about 75% of older persons without the need for further biochemical tests or clinical assessment. The MNA scoring system permits the stratification of older adults into three categories: well-nourished, at risk of undernutrition, and undernourished. An advantage of the MNA is that it can easily be used by a wide range of health professionals in a variety of clinical settings that cater for both free-living and institutionalized older persons. Several other tools are of practical value in the clinical setting. Morley has developed a useful screening tool known by the acronym SCALES (**Table 2**). This uses basic biochemical and anthropometric indices to identify older adults at risk of undernutrition, and can be readily incorporated into serial evaluation of the older person in

Table 2 SCALES: screening tool for the early detection of patients at risk of protein-energy undernutrition

Parameter	*Score 1 point*	*Score 2 points*
Sadness	GDS 10–15	GDS >15
Cholesterol	<4.65 mmol l^{-1} (180 mg dl^{-1})	<4.14 mmol l^{-1} (1660 mg dl^{-1})
Albumin	<40 g l^{-1} (4 g dl^{-1})	<35 g l^{-1} (3.5 g dl^{-1})
Loss of weight	<1 kg (2 lb) in 1 month	<2.7 kg (6 lb) in 6 months
Eating problems	Cognitive impairment or physical limitations	Cognitive impairment and physical limitations
Shopping problems	Inability to shop or prepare a meal	

Patients scoring over 3 are at risk. GDS, geriatric depression score.

Table 3 MEALS ON WHEELS: common causes of undernutrition in older persons

Medication (e.g., digoxin, theophylline, psychotropic drugs)
Emotional (depression)
Anorexia, alcoholism
Late-life paranoia
Swallowing disorders
Oral and dental disease
No money (absolute or relative poverty)
Wandering (dementia, behavioral disorders)
Hyperthyroidism, hyperparathyroidism
Entry problems (malabsorption)
Eating problems
Low-salt or low-cholesterol diets
Shopping and food preparation problems

different clinical settings. The simple mnemonic MEALS ON WHEELS, also devised by Morley, may prove useful in prompting consideration of the risk factors and common causes of nutritional compromise (**Table 3**).

More recently, the Council of Nutrition Appetite Questionnaire has been validated for the evaluation of appetite in older adults. A unique feature of this appetite assessment tool is the ability to predict significant weight loss (**Table 4**).

Table 4 The Council of Nutrition Appetite Questionnaire

1. My appetite is:
1. very poor
2. poor
3. average
4. good
5. very good

2. When I eat:
1. I feel full after eating only a few mouthfuls
2. I feel full after eating about a third of a meal
3. I feel full after eating over half a meal
4. I feel full after eating most of the meal
5. I hardly ever feel full

3. I feel hungry:
1. rarely
2. occasionally
3. some of the time
4. most of the time
5. all of the time

4. Food tastes:
1. very bad
2. bad
3. average
4. good
5. very good

5. Compared to when I was younger, food tastes:
1. much worse
2. worse
3. just as good
4. better
5. much better

6. Normally I eat:
1. less than one meal a day
2. one meal a day
3. two meals a day
4. three meals a day
5. more than three meals a day

7. I feel sick or nauseated when I eat:
1. most times
2. often
3. sometimes
4. rarely
5. never

8. Most of the time my mood is:
1. very sad
2. sad
3. neither sad nor happy
4. happy
5. very happy

Instructions: Complete the questionnaire by circling the correct answers and then tally the results based upon the following numerical scale: A = 1, B = 2, C = 3, D = 4, E = 5. The sum of the scores for the individual items constitutes the CNAQ score. Scoring: If the CNAQ score is less than 28, there is an increased risk of significant weight loss over the next 6 months.

Oral Nutritional Repletion

Appropriate treatment of the underlying causes of undernutrition should be accompanied by oral nutritional supplementation in persons who are able to eat. Objective quantitative baseline assessment of food intake is mandatory. This is best achieved by the maintenance of a food diary, in which the patient records all food items consumed over a 72-h period. Review of the food diary also permits evaluation of food preferences and eating patterns. The goal of nutritional supplementation should be the consumption of the recommended daily allowance of macronutrients and micronutrients. Several predictive equations have been derived for the purpose of determining the optimal energy intake for each individual. However, it remains unclear as to what extent corrections have been made for age-related physiological changes in nutritional requirements and energy expenditure. The Benedict-Harris equation is perhaps the best known and most frequently applied. Using this equation, the required daily energy intake in kilocalories is derived as follows:

- Men: $66 + 13.7W + 5H - 6.8A$
- Women: $665 + 9.6W + 1.8H - 4.7A$

where W is the weight in kilograms, H is the height in centimeters, and A is the age in years. Upward adjustment is required by factors ranging from 1 to 1.5, to compensate for increased activity or pathologically stressful conditions.

For practical clinical purposes, a total daily energy intake of $147\,kJ\,kg^{-1}$ ($35\,kcal\,kg^{-1}$) achieves

efficient nutritional repletion. Recent dietary guidelines emphasize an overall healthy and balanced dietary pattern that includes a wide variety of fruits, vegetables, and grain products. Specifically, at least 5 daily servings of fruits and vegetables and 6 daily servings of grain products, including whole grains. Low-fat dairy products, fish, legumes, poultry, and lean meats are encouraged. Guidelines also suggest at least two servings of fish per week.

The current recommended daily allowance for protein is at least $1\,g\,kg^{-1}$ body weight. However, acutely stressful or hypercatabolic conditions mandate an increase in protein intake to about $1.5\,g\,kg^{-1}$. Generally, compliance with these dietary guidelines achieves the dual purpose of ensuring optimal macronutrient and micronutrient intake. This obviates the need for the routine prescription of pharmacological multivitamin preparations in undernourished persons, unless specific signs of micronutrient deficiency are evident.

Nutritional supplementation with regular or fortified natural food items is the ideal mode of nutritional repletion. This possesses the advantages of familiarity, palatability, and cost-effectiveness. Where the patient is reluctant or unable to consume the required total energy intake in natural food items, commercially formulated nutritional supplements are a reasonable alternative. The choice of preparation should be based on palatability and patient preference unless underlying medical conditions such as lactose or gluten intolerance have to be considered. Patients with malabsorption syndromes should be given hydrolyzed preparations to enhance nutrient absorption. Regardless of the preparation used, an attempt should be made to vary flavors, as age-related sensory-specific satiety may limit intake if only one flavor is used. Erroneously, nutritional supplements are often administered with meals. Recent evidence indicates that liquid supplements are more effective in increasing daily energy intake when administered at least 1 h before meals. Data shows that when supplements are administered with meals, a suppressant effect on food consumption is evident. Thus, older adults on nutritional supplements should receive these between meals to maximize net energy intake. Ultimately, in persons with severe undernutrition, the focus should be on energy intake and patient food preference, not on optimal proportions of macronutrient and micronutrient intake. Frequently, efforts to ensure a balanced diet necessitate the use of food items that may compromise palatability and result in a counterproductive reduction in food intake.

Enteral Tube Feeding

Enteral or parenteral modes of nutrient delivery are often used in people who are unable to eat or swallow. In the presence of a functioning gastrointestinal tract, enteral feeding is more appropriate due to the lower incidence of complications, more efficient nutrient utilization, increased cost-effectiveness, and greater ease of administration. Additionally, small bowel hypoplasia and alterations in gastrointestinal secretions may result from prolonged parenteral nutrition. Nasogastric and nasoenteric tubes should be reserved for short-term nutritional support in persons who may be able to resume oral feeding within 14 days, in order to avoid the significant morbidity associated with the use of nasal tubes. In persons in whom prolonged enteral intake is anticipated, gastrostomy or jejunostomy tubes may be considered.

In patients who retain normal gastrointestinal absorptive function, regular meals may be puréed and delivered through large-bore feeding tubes. A variety of polymeric enteral feeding formulas are also available; these are of relatively low viscosity, rendering them particularly suitable for delivery through small-bore tubes, which are usually more comfortable and aesthetically pleasing. In persons with malabsorption, hydrolyzed predigested formulae are available. Specific formulations also exist for people with special nutritional requirements due to diseases such as diabetes mellitus or renal or respiratory failure.

In older people, large volume bolus tube feedings may be associated with a greater risk of aspiration. Thus, where possible, continuous infusions of feeds are preferred. In order to further reduce the risk of aspiration pneumonia, it is recommended that the patient is positioned in a 30° head-up incline during feedings. Feeds may be infused over a 24-h period or over 14–18 h with a nocturnal break. The latter infusion schedule is often advocated on the grounds that it mimics normal eating patterns more closely. In addition, the absence of a nocturnal feed-free period has been shown to obliterate the physiological diurnal variation in insulin, cortisol, and glucagons secretion. Maximal nutrient utilization is also encouraged by daytime feed infusions as gastric emptying occurs more rapidly during the day. Continuous infusion of enteral tube feeds should be initiated at a rate of $30\,ml\,h^{-1}$ using half-strength feeds. If tolerated, full-strength feeds may then be introduced at the same rate and increased by $25\,ml\,h^{-1}$ every 8–12 h until the recommended daily energy intake is achieved. Despite the popularity of enteral tube feeding, emerging evidence

indicates that the medical risks of percutaneous endoscopic gastrostomy (PEG) tube feeding may outweigh the risks. Studies in older adults with dementia fail to demonstrate any reduction in comorbidity or mortality with PEG feeding. Similarly, available data fails to demonstrate any significant improvement in functional status, nutritional status, or quality of life with this method of feeding. Thus, health providers should set realistic goals for patients and family members who opt for PEG feeding. Ultimately, the indications and benefits of PEG tube placement are more likely to be based on personal psychosocial, cultural, or ethical preferences.

Parenteral Nutritional Repletion

In the older person with a nonfunctioning gastrointestinal tract, parenteral nutrition may be unavoidable. All patients receiving parenteral nutrition must be monitored closely for adverse effects. For short-term intravenous nutritional repletion, peripheral parenteral nutrition may be used. Low osmolality nutritional preparations, with a low risk of toxicity to soft tissue, are best suited for this purpose. There is a paucity of data regarding the safety and efficacy of most peripheral parenteral nutritional products for periods exceeding 14 days. Thus, where longer periods of intravenous feeding are required, total parenteral nutrition through a large central vein is indicated. Standard total parenteral formulations comprising 25% dextrose, 5% amino acids, electrolytes, and trace elements in optimal amounts are suitable for use in most patients. During prolonged parenteral nutrition, lipid emulsion supplements should be added to prevent deficiency of essential fatty acids.

Pharmacological Management of Undernutrition

Older patients with a poor response to treatment of underlying causes and nutritional supplementation may benefit from orexigenic agents (**Table 5**).

Table 5 Orexigenic agents

Megesterol acetate
Mirtazapine
Dronabinol (delta-9-tetrahydrocannabinol)
Corticosteroids
Loxiglumide (Cholecystokinin antagonist)
Oxoglutarate
Anabolic agents (testosterone, anadrol)
Oxandrin
Growth hormone
Cyproheptadine

Megestrol acetate is a synthetic progestogen approved for use by the Food and Drug Administration (FDA) as an orexigenic agent in patients with Acquired Immune Deficiency Syndrome (AIDS) and cancer-related anorexia and cachexia. Recent evidence indicates that megesterol acetate is also an effective orexigenic agent in geriatric patients. Thromboembolic disease and adrenal suppression are rare complications, but patients should be monitored closely for these events.

Dronabinol (delta-9-tetrahydrocannabinol), the active ingredient of *Cannabis sativa*, is another FDA-approved orexigenic agent for use in patients with Acquired Immune Deficiency Syndrome (AIDS). Dronabinol is also an effective orexigenic and antiemetic in patients receiving cancer chemotherapy. Additional evidence indicates that dronabinol induces weight gain in persons with dementia, although research has yet to determine whether weight gain in such patients is due to increased energy intake or reduced agitation with improved behavior and consequently decreased energy expenditure. Side effects of dronabinol in older adults include delirium, euphoria, and increased somnolence. The latter two qualities may favor the use of dronabinol as an orexigenic agent in palliative care.

One third of depressed older adults manifest with weight loss. Effective antidepressant therapy should result in weight gain in this subset of patients. Notably, the choice of antidepressant therapy may influence body weight reuptake. Selective serotonin (5-hydroxytryptamine, 5-HT) inhibitors, such as fluoxetine, can cause significant weight loss at the onset of therapy. Evidence in younger adults suggests that this is a transient phenomenon with baseline body weight being restored as treatment progresses. However, age-related changes in energy regulation and adaptation to chronic disease may delay or prevent return to baseline body weight in older patients. Mirtazapine has proved useful in the management of depressed patients with weight loss. Mirtazapine is a well-tolerated and effective antidepressant that inhibits presynaptic $alpha_2$ adrenergic receptors and postsynaptic 5-HT2 and 5-HT3 receptors. Mirtazapine has been shown to induce an earlier increase in appetite and subsequent weight gain in older depressed persons with weight loss.

Several agents previously touted as effective orexigenic agents, such as human growth hormone, have fallen out of favor. The administration of human growth hormone to healthy older adults has been shown to increase muscle bulk. However, significant side effects such as carpal tunnel syndrome,

gynecomastia and hypoglycemia were noted; furthermore, the increase in muscle bulk failed to produce a parallel increase in muscle strength. Inadequate data regarding the safety and efficacy of growth hormone administration precludes routine clinical use. Similarly, the role of insulin-like growth factor (IGF-I) in the management of undernutrition is questionable. Although the data suggest that exogenously administered IGF-I may enhance nitrogen retention, gluconeogenesis, and maintenance of normal gastrointestinal function, evidence-based outcome studies are lacking.

Abundant data exist regarding the role of anabolic steroids in the management of undernutrition. However, current evidence supports the restriction of testosterone therapy as an orexigenic agent to hypogonadal undernourished men. As a general rule, pharmacological treatment should be considered second-line therapy and reserved for patients who have failed to respond to nonpharmacological measures.

Managing Undernutrition in the Community Setting

With increasing emphasis on home healthcare, the number of community-dwelling persons requiring alternative modes of feeding has increased. Special consideration and appropriate modification of therapeutic regimens may be required in such cases to ease the care giver or personal burden.

If enteral tube feeding is provided at home, continuous infusion may limit the patient's mobility and functional independence. This method also has the disadvantage of requiring immediate access to technical support, in the event of mechanical failure of the infusion pump. Thus, care givers and patients may find intermittent bolus feeding a more convenient and less daunting task. To minimize the aspiration risk, intermittent bolus feeds should be administered, where possible, with the patient in a seated position. Patients should also be encouraged to remain seated for at least 1 h after feeds. Some active older people resent the social inconvenience and embarrassment of tube feeding during daytime hours, and may prefer overnight enteral infusions of hypercaloric feeds. Hypercaloric feeds contain twice the amount of equal volumes of regular enteral feeds, thereby permitting the provision of adequate nutritional support over shorter periods.

Parenteral nutrition within the home is fraught with all the hazards of intravenous therapy. Thus, availability of skilled services to monitor such therapy is critical. Additionally, adequate care giver and social support is mandatory for patients receiving this mode of nutritional repletion at home.

Health providers involved in home delivery of enteral and parenteral nutritional therapy will need to develop and implement comprehensive therapeutic programs incorporating skilled nursing and dietary services to ensure safe and effective treatment.

Managing Undernutrition in Long-Term Care Institutions

Therapeutic strategies for managing undernourished institutionalized older adults are similar to those used within the community, though perhaps due to readily available medical supervision, enteral and parenteral modes of feeding are used more often. The comparatively formal structure of the nursing home environment has the added advantage of encouraging closer supervision of therapy and stricter nutritional surveillance.

A major drawback to oral nutritional repletion in institutionalized older persons is the restricted variety of meals. This can usually be circumvented by involving the residents in menu development and, where feasible, granting permission for meals of the residents' choice to be supplied by family or friends. Residents of nursing homes are often less functional than their peers and thus may be more dependent on assistance for their basic activities of daily living. When the ability to self-feed is compromised, it is imperative that all meals are supervised and assistance with feeding rendered where necessary. Many residents are persistent wanderers, and may expend a considerable amount of energy in this exercise. In such patients an appropriate increase in their daily energy intake is required to prevent weight loss. Similar adjustments may be required for residents with persistent involuntary movements or severe agitation.

Long-term care institutions must preserve the social and recreational aspects of meals; all too often, mealtimes are reduced to clinical, sanitized, and isolated events. Within the nursing home environment mealtimes are best managed as a component of recreational therapy. Socialization and the preservation of each resident's dignity should be encouraged during meals. Nursing facilities should also attempt to mimic community resources by making food items available outside scheduled mealtimes, from vending machines and snack carts.

Nutritional surveillance programs are crucial to the success of established intervention strategies within nursing homes. Quality indicators, preferably

employing anthropometric indices, should be defined to monitor the success of intervention strategies. Continuous quality improvement and total quality management programs must also be implemented as critical components of effective nutritional intervention strategies. Finally, the development of nutrition focus groups and the use of interdisciplinary intervention strategies directed at increasing nutritional intake and preventing undernutrition should be encouraged.

Micronutrient Deficiency

In older people at risk of nutritional compromise, micronutrient supplementation deserves special attention, in order to forestall the development of micronutrient deficiency (**Table 6**). The clinical features of established vitamin deficiency are well recognized. The first recourse in the management of micronutrient deficiencies should be the provision of a well-balanced diet. In the presence of a functioning gastrointestinal tract, an adequate diet containing the recommended daily allowance of each micronutrient effectively prevents and corrects deficiency states. However, the failure to consume the required amount of food may warrant the use of oral pharmacological micronutrient supplements. Vitamin B_{12} deficiency may be considered unique in this regard as, traditionally, replacement therapy has been administered parenterally. However, available evidence suggests that food-cobalamin deficiency may be the most common cause of vitamin B_{12} deficiency in older adults. In this condition cobalamin cannot be extracted from ingested food, although free cobalamin is readily absorbed as absorptive function is normal and intrinsic factor is present in adequate quantities. Thus, in persons with vitamin B_{12} deficiency resulting from food-cobalamin deficiency, repletion may be adequately achieved by oral replacement therapy.

There is a rising trend toward dietary supplementation with pharmaceutical preparations containing large doses of vitamins and minerals, based on conclusions drawn from the results of several studies. Available evidence derived from human and animal studies indicates that antioxidant micronutrients, mainly vitamins A, C and E, may play a role in boosting immunity, preventing neoplastic disease, and preventing or retarding the progression of several degenerative diseases, such as atherosclerosis. Vitamins E and C have also been shown to reduce low-density lipoprotein (LDL) cholesterol levels and increase high-density lipoprotein (HDL) levels, in addition to lowering fasting plasma insulin levels and improving insulin efficiency. Epidemiological studies have suggested a protective role for antioxidants such as vitamin C, vitamin E, β-carotene, and glutathione in macular degeneration and cataracts. Nevertheless, evidence derived from other epidemiological studies suggests that antioxidants may lack significant benefit. Studies are ongoing in an attempt to resolve this controversy.

In older adults reduced cutaneous synthesis and enteric absorption of vitamin D increases the risk of vitamin D deficiency. Reduced renal responsiveness to parathormone is an added risk factor. At least $500\,\text{IU}\,\text{day}^{-1}$ of vitamin D are required to prevent significant osteoporosis in postmenopausal women. Institutionalized patients with reduced exposure to sunlight are at higher risk of vitamin D deficiency due to reduced cutaneous synthesis. The role of calcium supplementation in the prevention of osteoporosis is also well accepted. Additional evidence suggests that inadequate dietary calcium consumption may play a role in the genesis of colorectal cancer and hypertension.

Table 6 Vitamins: recommended daily allowances (RDAs) and clinical features of deficiency states

	RDA	*Deficiency states*
Vitamin A	600–700 μg	Decreased immunity to infections, xerophthalmia, night blindness
Niacin	12–16 mg	Pellagra (dermatitis, dementia, diarrhea), glossitis, cheilosis
Pyridoxine	1.6–2 mg	Dermatitis, delirium, peripheral neuropathy, glossitis
Riboflavin	1.1–1.3 mg	Glossitis, cheilosis, normochromic anemia
Thiamin	0.8–0.9 mg	Beriberi, Wernicke's encephalopathy, Korsakoff's psychosis
Cyanocobalamin	5 μg	Megaloblastic anemia, optic atrophy, peripheral neuropathy, subacute combined degeneration of the cord, dementia
Ascorbic acid	40 mg	Hyperkeratosis, petechial hemorrhages, mucosal bleeding, lethargy
Vitamin D	10 μg	Osteomalacia, osteoporosis
Vitamin E	8–10 mg	Peripheral neuropathy, ataxia, hemolytic anemia
Folate	200 μg	Megaloblastic anemia, cognitive dysfunction
Vitamin K	65–80 mg	Spontaneous hemorrhage, hypothrombinemia

NE, niacin equivalent; RE, retinal equivalent.

Currently, the safety of large pharmacological doses of micronutrient supplements in humans remains to be established. In spite of this, a considerable proportion of the older population consumes large doses of these supplements as a primary preventive health measure. The risk of long-term supplementation with high doses of micronutrients, particularly in the presence of age-related changes, cannot be ignored, and few studies have addressed this issue specifically. Due caution must be exercised, even with the use of micronutrients such as vitamin D and calcium where clinical benefits have been clearly established. The complications of overenthusiastic calcium and vitamin D supplementation include hypercalcemia, nephrocalcinosis, milk-alkali syndrome, ectopic calcification, and rebound gastric acidity. Calcium supplementation may also chelate iron compounds and precipitate iron deficiency. With regard to vitamin A, available data have identified an increase in absorption and reduced peripheral clearance of this vitamin in older adults, therapy increasing the risk of vitamin A toxicity. Similarly, older persons on long-term iron therapy, particularly in the absence of proven iron deficiency, are at increased risk for the development of secondary hemochromatosis.

On the basis of existing evidence, the use of pharmacological doses of vitamin and mineral supplements is probably best restricted to low-potency supplements and reserved for persons with established micronutrient deficiency who are unable to eat an adequate diet. Close monitoring of such patients for adverse effects is mandatory.

Obesity

Men aged 55–64 years have the highest prevalence of overweight for males in the US (71.7%). Although the prevalence drops with age, the prevalence of overweight among men and women over 75 years is still considerable (52% and 44%, respectively). At all ages, African Americans have a higher prevalence of overweight and obesity.

With aging there is increasing upper and central body fat distribution. This trend is accelerated in women following menopause. In women aged 55–69 years, central obesity has been demonstrated to be correlated with greater coronary artery disease mortality as well as total mortality. Even with weight loss, the waist to hip ratio remained an important predictor of mortality in elderly women. Leptin is a hormone produced by fat cells. In women, leptin levels rise in middle age in concert with the increase in fat mass and then fall in late old age as fat mass declines. In men, leptin levels increase progressively from 65 years onwards. This may be due to age-related hypogonadism. In older men, testosterone replacement therapy decreased leptin levels.

As food intake declines with aging, obesity in old age is probably due to other factors. All three components of energy output – resting metabolic rate, thermic energy of feeding, and physical activity – decline with aging; thus the pathogenesis of obesity in old age appears to be predominantly due to altered energy output rather than to increased food intake.

While moderate degrees of overweight appear to confer minimal increased mortality in the older population, those above 130% of average body weight have an increased risk of death even at extreme ages. Most of the complications of obesity in older persons are similar to those seen in younger persons. Certain effects of obesity appear more commonly in older persons; for instance, functional decline is more common compared with younger persons. This is often associated with a 'fear of falling.' This syndrome is particularly common in older urban-dwelling adults and may lead to voluntary restriction of physical activity and consequent frailty. The prevalence of diabetes mellitus increases with age, due in part to the increased fat mass in middle age onwards. Obesity markedly increases the prevalence of sleep apnea in older persons. Overweight increases the rate of progression of osteoarthritis and its effects on function. In nursing homes, obesity has been associated with an increase in pressure ulcers. Increasing weight increases claudication in older persons with peripheral vascular disease.

Management of obesity in older persons usually should focus on enhancing functional status and increasing physical activity as opposed to aggressive caloric restriction. Available evidence linking aggressive weight loss in older adults with increased mortality mandates close monitoring during treatment. Surgery for obesity is not appropriate in older adults as the risks of bariatric surgery outweigh the benefits. For similar reasons, the use of thermogenic and anorexic agents should be avoided. Thus, a combination of exercise, healthy eating, and behavior modification is the cornerstone of therapy in older persons. Older obese adults need to be carefully monitored for the development of sarcopenia, visceral protein depletion, and increasing frailty. Due attention should also be given to micronutrient supplementation.

See also: **Antioxidants**: Diet and Antioxidant Defense. **Body Composition**. **Nutritional Assessment**: Anthropometry. **Nutritional Support**: Infants and

Children, Parenteral. **Obesity**: Definition, Etiology and Assessment.

Further Reading

de Groot CP, Enzi G, Matthys C, Moreiras O, Roszkowski W, and Schroll M (2002) Ten-year changes in anthropometric characteristics of elderly Europeans. *Journal of Nutrition, Health and Aging* **6**(1): 4–8.

Glick MR (2000) Rethinking the role of tube feeding in patients with advanced dementia. *New England Journal of Medicine* **342**: 206–210.

Guigoz Y and Vellas B (1997) The Mini-nutritional assessment for grading the nutritional state of elderly patients, presentation of the MNA, history and validation. *Facts, Research and Intervention Geriatric Newsletter: Nutrition* **6**: 2.

Krumholz HM, Seeman T, and Merrill SS (1994) Lack of association between cholesterol and coronary heart disease mortality, morbidity and all-cause mortality in persons older than 70 years. *JAMA* **272**: 1335–1340.

Mitchell S, Kiely DK, and Lipsitz LA (1998) Does artificial enteral nutrition prolong the survival of institutionalized elders with chewing and swallowing problems. *Journal of Gerontology* **53A**(3): M207–M213.

Morley JE and Thomas DR (1999) Anorexia and aging: pathophysiology. *Nutrition* **15**(6): 499–503.

Morley JE, Thomas DR, and Wilson MG (2001) Appetite and orexigenic drugs. Position Paper Council of Nutrition in Long Term Care. *Annals of Long-Term Care* Supplement: 2–12.

Position of the American Dietetic Association (2000) Nutrition, aging and continuum of care. Journal of the American Dietitic Association **100**(5): 580–595.

Reynolds MW, Fredman L, Langenberg P, and Magaziner J (1999) Weight, weight change, mortality in a random sample of older community-dwelling women. Journal of the American Geriatric Society **47**(12): 1409–1414.

Thomas DR, Ashmen W, Morley JE, and Evans WJ (2000) Nutritional management in long-term care: development of a clinical guideline. Council for Nutritional Strategies in Long-Term Care. *Journal of Gerontology and Medical Science* **55**(12): M725–M734.

Wilson MG and Morley JE (2004) Nutritional assessment and support in chronic disease. In: Bales CW and Ritchie CS (eds.) *Handbook of Clinical Nutrition in Aging*, pp. 77–103. Humana Press Inc.

Osteomalacia *see* **Vitamin D**: Rickets and Osteomalacia

OSTEOPOROSIS

K O O'Brien, Johns Hopkins University, Baltimore, MD, USA

Optimal dietary intake is essential for bone health. During childhood and the pubertal growth spurt, nutrients are needed to fully consolidate skeletal mass and to ensure the attainment of a peak bone mass consistent with one's genetic potential. After peak bone mass is obtained, nutrition continues to play an essential role in skeletal health. If intakes of key nutrients are not consumed at required levels, mineral may be lost from bone or essential bone proteins may not be fully functional.

Osteoporosis and osteopenia are substantial public health problems. Low bone mass (osteopenia and osteoporosis) and vitamin D deficiency are currently substantial public health problems. Osteopenia is defined when adult bone mineral density values are 1–2.5 SD below the mean peak value observed in a young adult. If the deficit in bone is more pronounced, and bone mineral density falls 2.5 SD or more below that observed in a young adult, this is defined as osteoporosis. Approximately 200 million people worldwide have osteoporosis. Many more have suboptimal bone mass and are at increased risk of developing this disease. Vitamin D deficiency in adults can also impair bone mineralization and lead to osteomalacia (in adults) or rickets (if evident in pediatric age groups prior to the completion of longitudinal bone growth). Insufficient bone mass and impaired bone mineralization increases the risk of fractures at considerable cost and loss of quality of life. Because bone loss is not fully reversible, the most effective strategies for reducing osteoporosis should focus on prevention, with nutrition playing a key role.

Dietary Intake and Body Mass

A balanced diet is important to promote health and to maintain an appropriate body weight. An

Table 1 Nutritional and lifestyle parameters that may influence bone health

Minerals	*Vitamins/ hormones*	*Lifestyle and environmental factors*	*Dietary components*
Calcium	Vitamin D	Body mass index	Protein
Phosphorus	Vitamin K	Exercise	Soy/ phytoestrogens
Magnesium	Vitamin A	Cigarette smoking	Fatty acids
Sodium	Vitamin C	Alcohol intake	Homocysteine
Zinc	Vitamin B_{12}		
Copper	Vitamin B_6		
Iron	Folate		
Boron			
Manganese			
Fluorine			
Potassium			
Silicon			

individual's body weight is one of the strongest determinants of bone mass because of the skeleton's responsiveness to the load that is placed on it. Individuals with small body frames or those who are excessively thin have an increased risk of osteoporosis due to a lower overall skeletal reserve to draw on for calcium needed to offset the annual loss of bone that occurs later in life. At the extreme end of this spectrum, individuals with anorexia nervosa are at risk of osteoporosis because of alterations in hormonal status and amenorrhea in addition to insufficient dietary intake of nutrients required for bone health.

Although higher body weight is typically associated with a greater skeletal mass, obese individuals may sequester nutrients needed for skeletal health, such as vitamin D, in adipose tissue. Bariatric surgery as a treatment for morbidly obese individuals is becoming more common and leads to a loss of both body weight and bone mass. The long-term impact of this surgery on skeletal health is not yet fully elucidated, and it remains unclear if the amount of bone lost following surgery is solely a response to the decrease in body weight or if it is also associated with other adverse consequences of this surgery on bone health.

Although overall caloric intake impacts body weight, many nutrients and dietary components have been studied in relation to their impact on bone health (**Table 1**). Several of these key nutrients and components of the diet and their roles in bone health and skeletal homeostasis are detailed next.

Calcium

Calcium is the most abundant mineral found in bone and comprises approximately 33% of bone mineral. Optimal calcium intakes are essential across the life cycle to meet the daily intrinsic requirements of calcium required for skeletal growth and to offset urinary, dermal, and endogenous fecal calcium losses. When dietary intakes of calcium are not sufficient to maintain circulating calcium concentrations and/or when the losses of calcium from the body are excessive, bone calcium will be resorbed to maintain calcium homeostasis. Because calcium is essential for the structural integrity of bone, deficiencies or inadequate intakes of this mineral will have a detrimental impact on bone mass and quality.

Skeletal mass peaks at approximately age 20–30 years, with much of this gain occurring during the pubertal growth spurt. Nearly 50% of adult bone mass is accumulated during the pubertal growth spurt. Thus, this period of skeletal accretion can be viewed as a window of opportunity to maximize skeletal mass. Calcium supplementation studies in children have found increased bone mass with supplementation, an effect that is most pronounced when implemented during the prepubertal period. It is not clear to what degree calcium supplementation during the pubertal growth spurt results in a net gain in peak bone mass or if it solely influences the tempo at which peak bone mass is achieved. To account for the importance of this nutrient in bone mineralization, the recommended adequate intake of calcium is highest (1300 mg or 2.5 mmol.day) between the ages of 9 and 18 years.

Calcium supplementation has also been found to have beneficial effects on bone health in adults and may have the greatest impact in individuals whose habitual dietary calcium intakes are less than 400 mg (10 mmol)/day. To account for a decreased efficiency of intestinal absorption coupled with increased losses of calcium in older individuals, recommended calcium intakes increase to 1200 mg (30 mmol)/day in those age 50 or older. As discussed in more detail later, due to the prevalence of vitamin D deficiency in the elderly, oral vitamin D supplements up to 800 IU/day may also be required in order for the impact of calcium supplementation to be evident.

Several epidemiological studies have found significant relationships between an individual's lifelong intake of milk and subsequent risk of fracture. The degree to which this effect is a consequence of increased calcium intake or due to other nutritional components of milk and dairy products requires further study.

Despite the importance of calcium in bone mineral acquisition and maintenance, calcium intakes fall below the recommended level for the

majority of age groups, with intake being particularly low for adolescent girls and adult and elderly women. It is often difficult to increase consumption of calcium in certain age groups due to low intakes of dairy products or to other factors, such as lactose intolerance, dieting, or altered appetite and food consumption patterns in groups such as the elderly. To increase the calcium content of the diet, nonfat milk powder yogurt or cheese, can be added to a number of recipes to increase the calcium content of the food without adversely affecting taste. An increasing variety of calcium-fortified food products are now also available. Individuals with lactose intolerance may improve intake of calcium by use of lactose-free dairy products or lactase pills. Increasing calcium intake from dietary versus supplemental sources also increases the intake of many other nutrients needed for bone health, including protein, magnesium, zinc, phosphorus, and vitamin D. For this reason, dietary approaches to increase calcium intake should be promoted over the use of calcium supplements alone. Despite these benefits, in some instances it may be necessary to utilize calcium supplements to achieve recommended intake levels.

Several forms of calcium supplements are commercially available. Existing supplemental forms differ slightly with respect to their relative calcium content per tablet and their absorbability; however, the magnitude of these differences is minor and may not be biologically significant. Caution should be used when relying on natural sources of calcium (such as those prepared from bone meal, limestone, or oyster shells) because these preparations may also contain heavy metals such as lead. Several calcium supplements also contain additional nutrients required for bone health, including vitamins D and K. Because the fraction of calcium absorbed falls as calcium intake increases, little additional benefit per dose is achieved when taking supplemental calcium sources containing more than 500 mg (12.5 mmol) per dose.

Magnesium

More than half of the magnesium found in the body is located in bone. In addition to its presence in bone, magnesium is important in calcium metabolism and bone health because it is required for parathyroid hormone secretion. Parathyroid hormone (PTH) is integral to bone health because it increases the production of the active form of vitamin D (1,25-dihydroxyvitamin D) and plays a role in the tubular reabsorption of calcium and phosphorus.

Although magnesium deficiency is associated with abnormalities in vitamin D metabolism, hypocalcemia, and impaired PTH secretion, epidemiological studies linking magnesium intakes to measures of skeletal health have produced conflicting results. Some studies report significant associations between dietary magnesium intake and bone mineral density, but others have not supported this finding. Relationships between magnesium status and bone mass may be more challenging to elucidate due to the lack of a highly sensitive indicator of magnesium status.

Studies have indicated that typical magnesium intakes in healthy adolescents may not be sufficient to maintain magnesium balance. Data on the impact of magnesium supplementation on bone mass remain controversial. While some studies have found magnesium supplementation to result in positive effects on bone mass, others have reported no significant benefit. Additional studies are needed to clarify these discrepancies and to assess the net effect of magnesium status and supplementation on bone metabolism. Because dietary intakes fall below recommended levels in several age groups and because of the known relationships between magnesium and hormones integral to bone health, increased attention should be focused on optimal magnesium intakes in relation to bone homeostasis.

Zinc and Copper

Zinc and copper play important roles in bone metabolism and bone health in part due to the roles they play as cofactors for various enzymes required for the synthesis or modification of bone matrix constituents. Zinc is a cofactor for a myriad of enzymes in the body, including alkaline phosphatase. Alkaline phosphatase is synthesized by osteoblasts and is essential for bone mineralization. Zinc also plays a role in the osteoblast via its involvement in aminoacyl-tRNA synthetase. Copper is a necessary cofactor for lysyl oxidase, an enzyme involved in collagen cross-linking. Both copper and zinc are found as components of superoxide dismutase, and they may protect bone from oxidative damage. Genetic defects that cause zinc deficiency (acrodermatitis enteropathica) or copper deficiency (Menkes' disease) result in growth retardation, stunting, and impaired bone growth.

Although more research on the roles of zinc and copper in bone health and fracture risk is clearly needed, the importance of these nutrients in skeletal health should be recognized and optimal intakes should be promoted in relation to skeletal homeostasis.

Vitamin D

Vitamin D is particularly important for bone health because of the role it plays in calcium homeostasis.

The active form of vitamin D stimulates the synthesis of calcium binding protein in the intestine to facilitate calcium transport across the intestine. Vitamin D also plays a regulatory role in renal calcium reabsorption and in calcium release from bone.

Vitamin D can be obtained from the diet (although only vitamin D-fortified milk and fatty fish provide substantial amounts) or is made in the skin following exposure to sunlight. Deficiency of this vitamin is increasingly recognized as an issue of concern across all age groups of the US population from neonates to the elderly. This deficiency is due to a combination of inadequate dietary intake (dairy products provide the largest dietary contribution to vitamin D intake) and to inadequate sunlight exposure. Rickets is increasing among exclusively breastfed minority infants in the United States. This is thought to be due to the low vitamin D content of human milk combined with insufficient endogenous dermal synthesis. Vitamin D deficiency in adults results in osteomalacia and secondary hyperparathyroidism, increasing bone resorption and the risk of osteoporosis.

Lack of sufficient endogenous production of vitamin D in the skin is influenced by geographical location (more northern latitudes have a shorter season during which the wavelength needed for vitamin D synthesis is available), increased use of sunscreen and cosmetics and skin care products containing sunscreen (sunscreens with SPF values of 8 or greater block the dermal production of vitamin D), and lifestyle factors that decrease exposure to sunlight.

Studies suggest that the optimal serum concentration of vitamin D may be markedly higher (>30 ng/ml) than that traditionally used to define vitamin D deficiency (<10–15 ng/ml). If these increased levels are eventually accepted as optimal target concentrations, an even greater fraction of the population will have suboptimal status of this vitamin.

Supplementation with vitamin D and calcium has been found to be effective in decreasing fracture incidence. Several studies in older adults have found significant relationships between vitamin D status (as determined by 25-hydroxyvitamin D concentrations) and both musculoskeletal function and risk of sarcopenia. Combined vitamin D and calcium supplementation in the elderly may also decrease the risk of falling. Individuals with low dairy product intake, those living in northern latitudes, or those with inadequate sunlight exposure may need to rely on supplemental sources of vitamin D to maintain circulating concentrations at optimal levels required to promote bone health.

Vitamin K

Many proteins are dependent on vitamin K for the carboxylation of γ-carboxyglutamyl (Gla) residues. Several of these vitamin K-dependent proteins play integral roles in the bone matrix. Osteocalcin, one of the vitamin K-dependent proteins, is the most abundant noncollagenous protein in bone. Osteocalcin contains three Gla residues that require vitamin K for carboxylation. The ability of osteocalcin to bind to the hydroxyapatite fraction of bone is dependent on its degree of carboxylation. Deficiency of vitamin K increases the fraction of undercarboxylated osteocalcin in the circulation. In addition to osteocalcin, other vitamin K-dependent proteins (including matrix Gla protein and protein S) are found in bone and cartilage. Research is needed to elucidate the impact of vitamin K deficiency on risk of osteoporosis and fracture. Because of the known relationship between vitamin K and several crucial bone proteins, optimal status of this vitamin should be achieved to promote skeletal health.

Phosphorus

Phosphorus is another mineral that functions as an integral component of bone. Bone contains 85% of the phosphorus found in the body, and together calcium and phosphorus comprise the major fraction of bone mineral. Although sufficient phosphorus intakes are necessary to support bone mineralization, phosphorus homeostasis can be maintained across a range of intakes and ratios of calcium to phosphorus in the diet.

There is considerable controversy regarding the potential impact of elevated phosphorus intake from soda on bone health. Although concern has focused on the phosphoric acid and phosphorus content of soda in relation to calcium retention, the major impact of these products on bone health may be their displacement of other more nutritive beverages (such as milk) from the diet. Because increased soda consumption may increase the risk of excess weight gain and displace other more nutritive beverages from the diet, excessive soda intakes should be a cause for concern, especially in children and adolescents during the peak period of bone acquisition.

Sodium

Although many components of the diet play direct roles in bone mineralization, nutrients such as sodium are known to influence the retention of other nutrients required for optimal bone health. Sodium is one of the strongest determinants of

urinary calcium excretion. Increased dietary sodium intake elevates urinary calcium losses, with every 2300 mg (100 mmol) increase in dietary sodium increasing the urinary excretion of calcium by approximately 40 mg (1 mmol). Thus, excessive intakes of sodium (such as those that may occur in individuals who consume large amounts of processed food, salt food heavily, or consume foods high in sodium) increase the obligatory losses of calcium from the body. During the growth phase this could potentially limit the amount of calcium that can be utilized for bone mineralization. The long-term impact of variation in sodium intake on bone mass and fracture risk has been difficult to quantify because of a lack of sufficient information on how dietary effects on urinary sodium loss are counterbalanced and because other dietary components may modify this response.

Protein

Protein is essential for the formation of the organic matrix of bone and optimal intakes are required for normal skeletal development and growth. The importance of protein in bone health is well-known; however, there are conflicting reports on the relative impact of extremes of protein intake on bone health. Many proteins are rich in sulfur amino acids. The resulting protein-induced acid load must be buffered before excretion from the body. Calcium is a positive cation and can be utilized to buffer increased dietary acid loads from high protein intakes. On average, for every 1-g increase in dietary protein intake, urinary calcium excretion increases by approximately 1 mg.

Differences in habitual protein intakes have been related to bone mass and risk of fracture. Many studies have reported positive relationships between increased animal protein intake and bone health. Higher animal protein intakes in the elderly have been associated with reduced bone loss. Other research has supported a positive association between higher animal protein intake and both greater bone mineral density and decreased risk of hip fractures. Studies have found that although urinary calcium excretion increases in response to acute increases in protein intake, intestinal calcium absorption also increases by an amount nearly comparable to that lost in urine. Insufficient intakes of protein can adversely impact muscle mass and function. In addition, low dietary protein intake has been associated with reductions in serum insulin-like growth factor-1 (IGF-1) concentrations. IGF-1 plays an essential role in skeletal health via its impact on osteoblast formation and bone growth.

In contrast to the many studies that have found positive relationships between protein intake and bone health, other data suggest that high protein intakes may have a detrimental impact on bone mass and fracture risk and it is likely that extremes of protein intake, both high and low, may have adverse consequences on bone homeostasis. To clarify these conflicting findings, more research is required to address the relative impact of the quantity and type of protein on skeletal health.

Phytoestrogens

Phytoestrogens are dietary components that have a chemical structure similar to that of endogenous estrogens. The primary phytoestrogens in the diet are obtained from soybean isoflavones (including genistein and daizein). These compounds appear to be able to weakly mediate some the genomic and nongenomic effects of estrogen and may function as agonists or antagonists, depending on the tissue and type of estrogen receptor involved. To date, supplemental sources of these compounds have not been found to decrease fracture risk. Additional clinical trials will assist in determining the long-term impact of phytoestrogens on bone health and fracture risk.

Homocysteine

For some time, it has been known that individuals with a genetic defect in homocysteine metabolism (homocystinuria) have an increased risk of early onset osteoporosis. However, only recently has attention focused on the potential impact of circulating homocysteine concentrations on bone health among the general population. This interest is based on studies that have reported significant relationships between serum homocysteine concentrations and increased risk of fracture in adults. The strength of the relationship observed is substantial and is similar to the relationship found between serum homocysteine concentrations and cardiovascular disease. The mechanisms responsible for the impact of homocysteine concentrations on fracture risk are not known. Increased homocysteine concentrations could possibly interfere with normal collagen production, but studies have not found a significant relationship between serum homocysteine concentrations and bone mineral density, and the impact of elevated homocysteine concentrations on bone health may be indirect. Further research will assist in identifying the mechanisms and relationships between homocysteine and bone health and the degree to which this relationship is influenced by

folate, vitamin B_{12}, and vitamin B_6 status. Because of the other known adverse consequences of elevated serum homocysteine concentrations, additional incentive to monitor and promote reductions in this amino acid in relation to bone health is warranted.

Other Lifestyle Factors

Other lifestyle choices, such as smoking, alcohol abuse, and physical activity, also impact overall bone health. Excessive alcohol intake is a risk factor for low bone mass. This finding may be a consequence of poor dietary quality in chronic alcoholics and may also be related to adverse effects of excessive alcohol intake on osteoblast function. Cigarette smoking also adversely impacts bone health. Smokers may be leaner, and female smokers may experience an earlier menopause and have lower postmenopausal estrogen levels. Smoking may also have adverse effects on bone cells either directly or indirectly through an increase in oxidative stress.

Exercise is known to positively influence bone mass. During exercise, the strain placed on bone stimulates local bone responses to positively influence the balance in bone remodeling. Many studies have found positive associations between exercise and bone mass at a number of sites, especially the hip and the spine. The impact of exercise on bone mass is related to the intensity of the exercise and is associated with the degree to which it increases the habitual physical activity level of the individual. The impact of exercise on bone mass is also influenced by diet and may be most efficacious when calcium intake is optimal. Exercise not only impacts bone mass but also influences muscle strength, muscle mass, balance, and coordination. These improvements in muscle strength may also lead to improvements in posture, balance, flexibility, coordination, and gait stability that decrease the risk of falls.

Nutrient–Gene Interactions

Optimal nutrition is needed to supply the necessary substrates for bone; however, other parameters also influence the impact of a given nutrient on bone health. A substantial amount of bone mineral acquisition (up to 80%) is genetically determined. An individual's ability to utilize a given nutrient intake is influenced by his or her genetic makeup.

Many candidate genes have been associated or linked with the risk of osteoporosis or fracture, including genes coding for hormones (PTH), receptors (including PTH, vitamin D, estrogen, glucocorticoid, and calcitonin receptors), cytokines and growth factors (including IGF-1, transforming growth factor-β, epidermal growth factor, interleukin-4, and interleukin-6), and bone matrix proteins (such as osteocalcin, collagen type 1 (α1 and α2), and collagen type 11 (α1)). Although many of these genes have obvious roles in bone metabolism, other candidate genes (such as those coding for apolipoprotein E and methylenetetrahydrofolate reductase) have less obvious relationships to bone mass.

Several studies have found interactions between genotype, nutrient level, and environmental factors. For instance, the impact of exercise on bone can be influenced by the habitual dietary calcium intake and the individual's genotype (such as the vitamin D receptor genotype). Further research on the genetic control of bone mineral acquisition and loss will be invaluable in targeting groups at risk for low bone mass and may eventually be useful in setting genotype-specific intakes of bone-related nutrients to maximize skeletal health throughout the life cycle.

Best Practices to Prevent Osteoporosis

In summary, several practices can be adopted to assist in the prevention of osteoporosis. From a nutritional standpoint an emphasis should be made on adequate intakes of calcium, vitamin D and a balanced diet that meets the requirements of other essential bone-related minerals and nutrients (detailed in **Table 1**). A healthy body weight should be achieved and maintained throughout the life cycle. Age-appropriate physical activity and exercise programs should be promoted to maintain fitness, muscle strength and weight bearing activities. Lifestyle habits that adversely impact bone health, including smoking and excessive alcohol intake, should be avoided. Individuals with risk factors known to increase the risk of low bone mass should discuss these concerns with their physician to identify the need for bone density screening. Appropriate screening will allow for the initiation of medical interventions to maintain or build existing bone mass and reduce the subsequent risk of fragility related fractures. Attention to bone health and adoption of bone healthy habits should be initiated during childhood and maintained throughout the lifecycle to promote lifelong attainment of skeletal health.

See also: **Bone**. **Calcium**. **Copper**. **Magnesium**. **Phosphorus**. **Protein**: Requirements and Role in Diet. **Sodium**: Physiology. **Vitamin D**: Rickets and Osteomalacia. **Vitamin K**. **Zinc**: Physiology.

Further Reading

Branca F (2003) Dietary phyto-oestrogens and bone health. *Proceedings of the Nutrition Society* **62**: 877–887.

Bugel S (2003) Vitamin K and bone health. *Proceedings of the Nutrition Society* **62**: 839–843.

Heaney RP and Weaver CM (2003) Calcium and vitamin D. *Endocrinology and Metabolism Clinics of North America* **32**(1): 181–194.

Holick MF (2003) Vitamin D: A millennium perspective. *Journal of Cellular Biochemistry* **88**: 296–307.

Holick MF and Dawson-Hughes B (eds.) (2004) *Nutrition and Bone Health*. Totowa, NJ: Humana Press.

Institute of Medicine (1998) *Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride. Dietary Reference Intakes*. Washington, DC: National Academy Press.

Lowe NM, Fraser WD, and Jackson MJ (2002) Is there a potential therapeutic advantage to copper and zinc for osteoporosis? *Proceedings of the Nutrition Society* **61**: 181–185.

Prentice A (2004) Diet, nutrition and the prevention of osteoporosis. *Public Health Nutrition* **7**(1A): 227–243.

Tucker KL (2003) Dietary intake and bone status with aging. *Current Pharmaceutical Design* **9**: 2687–2704.

Wolf RL, Zmuda JM, Stone KL *et al.* (2000) Update on the epidemiology of osteoporosis. *Curr Rheumatol Rep* **2**: 74–86.

Oxidant Damage *see* **Antioxidants**: Observational Studies; Intervention Studies

P

PANTOTHENIC ACID

C J Bates, MRC Human Nutrition Research, Cambridge, UK

Absorption, Transport and Storage, Status Measurement

A considerable proportion of the pantothenic acid (vitamin B_5, see **Figure 1**) that is present in food eaten by animals or humans exists as derivatives such as coenzyme A (CoA) and acyl carrier protein (ACP). Compared with the crystalline vitamin, only about half of the vitamin in food is thought to be absorbed. The pantothenic acid in its derivatives in food is largely released as free pantothenic acid or pantetheine by pancreatic enzymes, and is then absorbed along the entire length of the small intestine by a combination of active transport and passive diffusion, of which the active transport process seems to predominate at physiological intakes. This active transport process is dependent on sodium, energy and pH and is saturable: the K_m is *c.* 17 μM and V_{max} is *c.* 1000 pmol cm^{-2} h^{-1}, with minor variations among species. The transport pathway is shared by biotin in colonic epithelial cells, and it appears to be regulated by an intracellular protein kinase C-mediated pathway. Calmodulin is also implicated in cellular pantothenic acid transport pathways.

In mice, it was found that usual dietary pantothenate levels did not affect the rate of absorption of a standard pantothenate dose, i.e., there was no evidence for feedback adaptation of the absorption pathway to low or high intakes, and it is assumed that the same is true in other species, including humans. However, there is some evidence from rat studies that the extent of secretion of enzymes degrading CoA into the gut lumen may partially limit the availability of pantothenic acid from CoA.

In humans, studies of urinary excretion of pantothenic acid after oral intakes of either free pantothenic acid or of the pantothenic acid present in food have indicated a relative availability of *c.* 50% from the food-borne vitamin. Urinary excretion of pantothenate was *c.* 0.8 mg day^{-1} when a pantothenate-deficient diet was eaten, rising to 40–60 mg day^{-1} at a high daily intake of 100 mg day^{-1}. At intermediate intakes, in the range 2.8–12.8 mg day^{-1}, the urinary excretion rate varied between 4 and 6 mg day^{-1}. Excretion of less than 1 mg day^{-1} is considered low. Urinary excretion rates reflect recent intakes perhaps more closely than most other biochemical indices.

The contribution of the gut flora to the available pantothenate for humans is unknown, but there is some evidence that bacterial synthesis of the vitamin may be important in animals, especially ruminants, since severe deficiency can only be achieved by using antibiotics or antagonists. Clinical conditions such as ulcers or colitis can adversely affect pantothenate status and excretion rates, and dietary fiber may affect its absorption.

After a dose of ^{14}C-labeled pantothenate, about 40% of the dose appears in muscle tissue and about 10% in the liver, with smaller amounts occurring elsewhere. The differential affinities of the various different tissues determines their individual contents of the coenzyme derivatives, CoA and ACP, since there is no other major store of the vitamin anywhere in the body. Most organs, including placenta, exhibit evidence of a unidirectional active transport process for the intracellular accumulation of pantothenate, which is dependent on sodium, energy, and pH. In placenta (and probably elsewhere) this transport process is also shared by biotin and by some of its analogs, which can exhibit competitive inhibition. The only tissues that have been shown to differ with respect to transport mechanisms are red cells and the central nervous system.

The uptake and efflux of pantothenate into and out of red blood cells is unaffected by sodium, energy, or pH. Red cells contain pantothenate, 4-phosphopantothenate, and pantetheine, but they do not contain mitochondria, or carry out CoA-dependent processes. The function of the pantothenate derivatives found in red cells is unknown, but their formation clearly results in higher concentrations of total pantothenate in red cells than in plasma, and red cell (or whole blood) total pantothenate is considered a better status index, and is more predictably related to

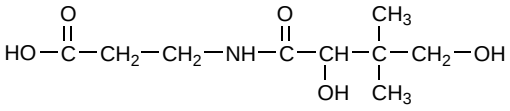


Figure 1 Structure of pantothenic acid.

intake, than is serum or plasma pantothenate. A concentration less than $1\,\mu mol\,l^{-1}$ of pantothenate in whole blood is considered low; the normal range is $1.6–2.7\,\mu mol\,l^{-1}$. Pantothenate in serum appears to be a very short-term marker and it is not well correlated with changes in intake or status.

Concentrations in body fluids are traditionally measured by microbiological assay using *Lactobacillus plantarum*. If CoA is present, enzymatic hydrolysis is needed to liberate free pantothenic acid for the microbiological assay. Other assay methods reported include gas chromatography (after conversion to a volatile derivative), radioimmunoassay (RIA), or enzyme-linked immunoabsorbent assay (ELISA).

Unlike several other B vitamin precursors of cofactors, pantothenate is not entirely converted to coenzyme forms inside the cell, and metabolic 'trapping' is therefore less dominant than it is for some other B vitamins. There is some evidence that the free pantothenate in tissues is more closely related to dietary pantothenate than the coenzyme forms are; the latter are relatively protected during periods of dietary deficiency or of low intakes. Uptake of pantothenate from plasma into most tissues is proportional to the plasma concentration because the active transport process is nowhere near saturated at typical plasma concentrations of *c*. 10^{-6} M.

Pantothenate is required for the hepatic acetylation of drugs by its presence in acetyl CoA, and it has been shown that pantothenate deficiency can impair this process; moreover, 20–60% of human populations are slow acetylators, varying with their ethnic grouping. Whether this function can be used to develop a functional test for pantothenate status is an intriguing but unresolved question.

Metabolism and Turnover

The primary role of pantothenic acid is in acyl group activation for lipid metabolism, involving thiol acylation of CoA or of ACP, both of which contain 4-phosphopantotheine, the active group of which is β-mercaptoethylamine. CoA is essential for oxidation of fatty acids, pyruvate and α-oxogutarate, for metabolism of sterols, and for acetylation of other molecules, so as to modulate their transport characteristics or functions. Acyl carrier protein, which is synthesized from apo-ACP and coenzyme A, is involved specifically in fatty acid synthesis. Its role is to activate acetyl, malonyl, and intermediate chain fatty acyl groups during their anabolism by the biotin-dependent fatty acid synthase complex (i.e., acyl-CoA: malonyl-CoA-acyl transferase (decarboxylating, oxoacyl and enoyl-reducing, and thioester-hydrolyzing), EC 2.3.1.85).

The organ with the highest concentration of pantothenate is liver, followed by adrenal cortex, because of the requirement for steroid hormone metabolism in these tissues. Ninety-five per cent of the CoA within each tissue is found in the mitochondria. However, the initial stages of activation of pantothenate and conversion to CoA occur in the cytosol. It was originally believed that the final stages of CoA synthesis must occur within the mitochondria, but later evidence indicated that transport across the mitochondrial membrane is, after all, possible. β-oxidation within the peroxisomes is also CoA-dependent, and is downregulated by pantothenate deficiency.

The pathways of conversion of pantothenic acid to CoA and to ACP are summarized in **Figure 2**. There are three ATP-requiring reactions and one CTP-requiring reaction in the synthesis of CoA. The rate of CoA synthesis is under close metabolic control by energy-yielding substrates, such as glucose and free fatty acids (via CoA and acyl CoA) at the initial activation step, which is catalyzed by pantothenate kinase (ATP: pantothenate 4-phosphotransferase, EC 2.7.1.33). This feedback control is thought to be a mechanism for conservation of cofactor requirements. There are also direct and indirect effects of insulin, corticosteroids, and glucagon, which result in important changes in tissue distribution, uptake, etc. in persons with diabetes. The mechanisms involved here are complex and not yet fully understood; however insulin represses and glucagon induces the enzyme.

A rare genetic disease, Hallervorden-Spatz syndrome, has recently been shown to result from deficiency of pantothenate kinase, and is now alternatively known as pantothenate kinase-associated neurodegeneration (PKAN). Dystonia, involuntary movements, and spasticity occur, and although there is no cure, some palliative treatment is possible.

In genetically normal people, fasting results in a reduction of fatty acid synthase activity with loss of the coenzyme of ACP, which thus achieves the desired objective of a shift away from fatty acid synthesis, towards breakdown. This interconversion of apo-ACP and holo-ACP is thus a very important process for the short-term regulation of fatty acid synthesis.

Deficiency of sulfur amino acids can result in reduced CoA synthesis; likewise copper overload can (by interfering with sulfur amino acid function) also reduce CoA synthesis.

Excretion of free pantothenate in the urine is the primary excretion route in humans; in other mammals

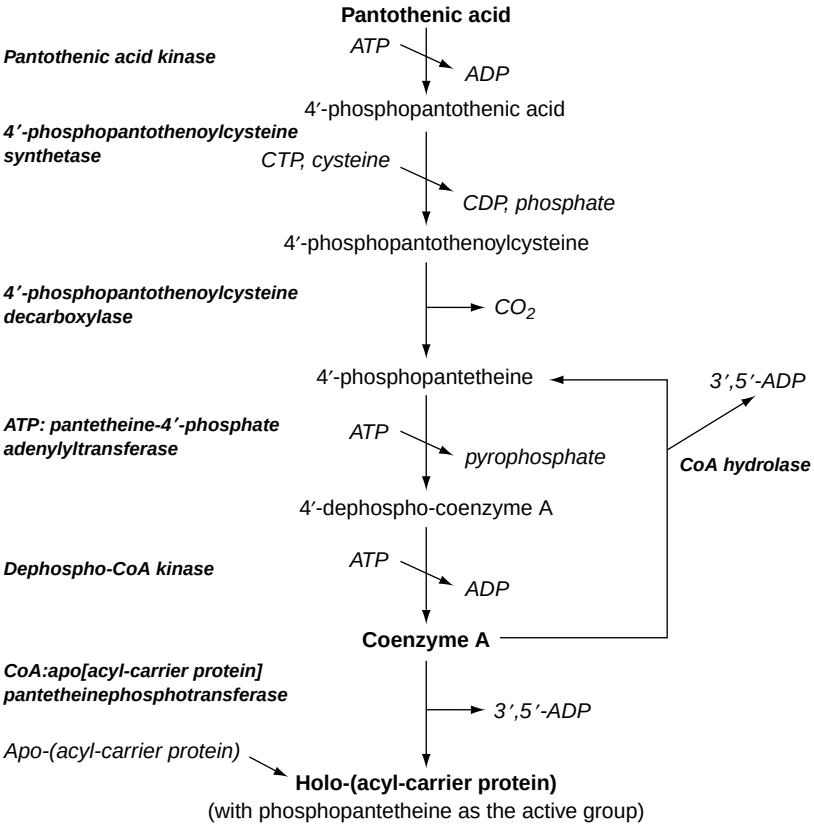


Figure 2 Synthetic pathway between pantothenic acid, coenzyme A, and acyl carrier protein.

the glucuronide or glucoside may be excreted. There is little evidence of degradation to simpler products, and pantothenic acid appears to be very efficiently conserved in animals. Some bacteria can cleave it to yield pantoic acid and β-alanine. A potentially useful breakdown product of CoA is taurine, formed via cysteamine. This amino acid is an essential nutrient for some carnivorous animals such as cats.

When dietary intakes are low, the majority of the circulating vitamin, which is filtered in the kidney tubules, is absorbed by the same type of sodium-dependent active transport process that also occurs at most other sites in the body. Retention of a test dose of pantothenate is, as expected, greater in partially depleted subjects, than in saturated ones. Secretion into breast milk is proportional to intake and to blood levels of the vitamin; therefore, dietary supplements taken by the lactating mother generally increase the breast milk content of the vitamin.

Metabolic Function and Essentiality

As noted above, the biochemical functions, and hence the basis for the dietary requirement of pantothenic acid, arise entirely from its occurrence as an essential component of CoA and of ACP, which cannot be synthesized *de novo* in mammals from simpler precursors.

In addition to the now well-established roles of CoA in the degradation and synthesis of fatty acids, sterols, and other compounds synthesized from isoprenoid precursors, there are also a number of acetylation and long-chain fatty acylation processes which seem to require CoA as part of their essential biological catalytic sites, and which are still being explored today. The acetylation of amino sugars, and some other basic reactions of acetyl-CoA and succinyl-CoA in intermediary metabolism, have been known since the 1980s. However, the addition of acetyl or fatty acyl groups to certain proteins in order to modify and control their specific and essential properties is a more recent discovery. The first category of these modifications comprises the acetylation of the N-terminal amino acid in certain proteins, which occurs in at least half of all the known proteins that are found in higher organisms. The specific amino acids that are recipients of these acetyl groups are most commonly methionine, alanine, or serine. The purposes of this terminal acetylation process are not entirely clear and may be multiple, including modifications of function

(e.g., of hormone function), of binding and site recognition, of tertiary peptide structure, and of eventual susceptibility to degradation. Another possible site of protein acetylation is the side chain of certain internal lysine residues, whose side chain ε-amino group may become acetylated in some proteins, notably the basic histone proteins of the cell nucleus, and the α-tubulin proteins of the cytoplasmic microtubules, which help to determine cell shape and motility. Its essential role in the synthesis of α-tubulin appears to be a particularly important one.

Proteins can also be modified by acylation with certain long-chain fatty acids, notably the 16-carbon saturated fatty acid, palmitic acid, and the 14-carbon saturated fatty acid, myristic acid. Although structurally very similar to each other, these two fatty acids seek entirely different protein locations for acylation and also have quite different functions. They have recently been explored with particular emphasis on viral and yeast proteins, although proteins in higher animals, in organs such as lung and brain, can also become acylated with palmityl moieties. Palmitoyl CoA is also required for the transport of residues through the Golgi apparatus during protein secretion. It is believed that these protein acylations may enable and control specific protein interactions, especially in relation to cell membranes, and proteins that are palmitoylated are generally also found to be associated with the plasma membrane. Signal transduction (e.g., of the human β_2-adrenergic receptor) is one process that appears to be controlled by palmitoylation, and other palmitoylated proteins possess some structural importance, for example in the case of the protein–lipid complex of brain myelin. Clearly, these subtle protein modifications, all of which depend on CoA and hence on pantothenic acid, have a wide-ranging significance for many biological processes, which is still being actively explored.

Pantothenic acid is essential for all mammalian species so far studied, namely humans, bovines, pigs, dogs, cats, and rodents, as well as for poultry and fish. Pantothenate deficiency signs in animals are relatively nonspecific and vary among species. Deficiency in young animals results in impaired growth, and requirement estimates based on maximum growth rates are between 8 and 15 mg per kg diet. Rats that are maintained on a diet low in pantothenate exhibit reduced growth, scaly dermatitis, alopecia, hair discoloration and loss, porphyrin-caked whiskers, sex organ disruption, congenital malformations, and adrenal necrosis. Deficient chicks are affected by abnormal feather development, locomotor and thymus involution, neurological symptoms including convulsions, and hypoglycemia. Pigs exhibit intestinal problems and abnormalities of dorsal root ganglion cells, and several species suffer nerve demyelination. Fish exhibit fused gill lamellae, clumping of mitochondria, and kidney lesions. Signs specific for pantothenate depletion are not well characterized for humans. A syndrome that included 'burning feet' has been described in tropical prisoner-of-war camps during World War II, and it was said to respond to pantothenic acid supplements; however this was likely to have been a more complex deficiency. A competitive analog of pantothenate, ω-methyl pantothenate, interferes with the activation of pantothenic acid; it also produces burning feet symptoms, Reye-like syndrome, cardiac instability, gastrointestinal disturbance, dizziness, paraesthesia, depression, fatigue, insomnia, muscular weakness, loss of immune (antibody) function, insensitivity to adrenocorticotrophic hormone, and increased sensitivity to insulin. Large doses of pantothenate can reverse these changes. One of the earliest functional changes observed in mildly deficient rats was an increase in serum triacylglycerols and free fatty acids, presumably resulting from the impairment in β-oxidation. Paradoxically, CoA levels are relatively resistant to dietary pantothenate deficiency; however there are some inter-organ shifts in pantothenate in certain metabolic states.

As noted above, CoA is required for Golgi function, involved in protein transport. Pantothenate deficiency can therefore cause reductions in the amounts of some secreted proteins. Other metabolic responses to deficiency include a reduction in urinary 17-ketosteroids, a reduction in serum cholesterol, a reduction in drug acetylation, a general reduction in immune response, and an increase in upper respiratory tract infection.

Recently, some studies of wound healing and fibroblast growth have indicated that both pantothenic acid and ascorbic acid are involved in trace element distribution in the skin and scars of experimental animals, and that pantothenic acid can improve skin and colon wound healing in rabbits. It is not yet known whether these observations are relevant to wound healing in humans.

Requirements

In the UK, National Food Survey records suggest that during recent decades mean adult daily pantothenate intakes have been consistently in the range of 4–6 mg. Since there is little evidence for the magnitude of minimum requirements in humans, the UK committee responsible for the revision of dietary reference values in 1991 suggested that intakes in the range 3–7 mg day^{-1} can be considered as adequate (although no specific values for the reference nutrient intake, estimated average

requirement or lower reference nutrient intake for pantothenate were set). The US adequate intake (AI) for pantothenic acid is currently set at 5 mg day^{-1} for adults; 4 mg day^{-1} for children aged 9–13 years; 3 mg day^{-1} for 4–8 years, and 2 mg day^{-1} for 1–3 years. There was insufficient evidence to set an estimated average requirement (EAR), a recommended daily allowance (RDA), or a tolerable upper intake level (UL).

There are few studies in communities where intakes are likely to be low; indeed, pantothenic acid is so widely distributed in human foods that it is unlikely that any natural diets with a very low content will be encountered. Some variations in status among communities have been described, but these do not define requirements. In a group of adolescents in the USA, daily pantothenate intakes were around 4 mg; total blood pantothenate was in the 'normal' range of *c.* 350–400 ng ml^{-1}, and intakes were correlated with red cell pantothenate ($r = 0.38$) and with urinary pantothenate ($r = 0.60$), both $P < 0.001$. In adults, these correlations were less strong.

During pregnancy and lactation there is some evidence that requirements may increase. As for most water-soluble vitamins, maternal blood levels do decrease significantly on normal diets during pregnancy, and the mean daily output of the vitamin in breast milk in the US is of the order of 2–6 mg. The adequate intake (AI) in the USA is 6 mg day^{-1} during pregnancy and 7 mg day^{-1} during lactation. It has been suggested that infant formulas should contain at least 2 mg pantothenate per liter and the AI for infants is 1.7 mg day^{-1} from birth to 6 months and 1.8 mg day^{-1} from 7 to 12 months of age.

Dietary Sources and High Intakes

Pantothenate is widely distributed in food; rich sources include animal tissues, especially liver, and yeast, with moderate amounts occurring in whole grain cereals and legumes (see **Table 1**). It is fairly stable during cooking and storage, although some destruction occurs at high temperatures and at pH values below 5 or above 7. Highly processed foods have lower contents than fresh foods. Commercial vitamin supplements containing pantothenate usually contain the calcium salt, which is crystalline and more stable than the acid.

Synthesis by gut flora in humans is suspected but not yet proven; the rarity of diet-induced deficiency has been attributed to contributions from gut flora sources.

There is some evidence that pantothenic acid supplements may be beneficial for treatment of rheumatoid arthritis and for enhancement of athletic performance, specifically in running. Pantethine, the disulfide dimer of pantetheine, may have cholesterol-lowering properties. The mechanisms of these reported effects are unclear and they require further investigation and verification. A homolog of pantothenate, pantoyl γ-aminobutyrate (hopanthenate), which can act as a pantothenate antagonist, has been used to enhance cognitive function, especially in Alzheimer's disease. It acts on GABA receptors to enhance acetylcholine release and cholinergic function at key sites in the brain.

There is little or no evidence for any toxicity at high intakes: at daily intakes around 10 g there may be mild diarrhea and gastrointestinal disturbance, but no other symptoms have been described. Pantothenate has been prescribed for various chronic disorders, but is not known to be useful in high doses.

Table 1 Pantothenate content of selected foods

Food	*mg per 100 g wet wt*	*mg per MJ*
Meat, offal, and fish		
Stewed minced beef	0.36	0.41
Grilled pork chop	1.22	1.58
Calf liver, fried	4.1	5.59
Lamb's kidney, fried	4.6	5.87
Cod, grilled	0.34	0.85
Dairy products		
Cow's milk, full cream	0.58	2.12
Cheese, cheddar	0.50	0.29
Yogurt (whole milk, plain)	0.50	1.50
Boiled chicken's egg	1.3	2.12
Human milk	0.25	0.87
Fruits		
Apples, eating, flesh and skin	trace	trace
Oranges, flesh	0.37	2.34
Pears, flesh and skin	0.07	0.41
Strawberries, raw	0.34	3.01
Dried mixed fruit	0.09	0.08
Vegetables		
Potatoes, boiled, new	0.38	1.18
Carrots, boiled, young	0.18	1.94
Brussel sprouts, boiled	0.28	1.83
Cauliflower, boiled	0.42	3.59
Onions, fried	0.12	0.18
Grains, grain products, nuts		
White bread	0.40	0.43
Wholemeal bread	0.60	0.65
Rice, boiled, white	0.10	0.17
Comflakes	0.30	0.19
Baked beans in tomato sauce	0.18	0.51
Peanuts, plain	2.66	1.14

Compiled from Food Standard Agency (2002) McCance and Widdowson's *The Composition of Foods*, 6th Sixth Summary edn. Cambridge: Royal Society of Chemistry, © Crown copyright material is reproduced with the permission of the Controller of HMSO and Queen's Printer for Scotland.

See also: **Cofactors**: Organic. **Energy**: Metabolism. **Fatty Acids**: Metabolism. **Lactation**: Dietary

Requirements. **Lipids**: Chemistry and Classification. **Nutritional Assessment**: Biochemical Indices.

Further Reading

Bender DA (1992) Pantothenic acid. *Nutritional Biochemistry of the Vitamins*, ch. 12, pp. 341–359. Cambridge: Cambridge University Press.

Bender DA (1999) Optimum nutrition: thiamin, biotin and pantothenate. *Proceedings of the Nutrition Society* **58**: 427–433.

Institute of Medicine (2000) *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin and Choline*, pp. 357–373. Washington, DC: National Academy Press.

Miller JW, Rogers LM, and Rucker RB (2001) Pantothenic acid. In: Bowman BA and Russell RM (eds.) *Present Knowledge in Nutrition*, 8th edn., ch. 24. pp. 253–260. Washington, DC: ILSI Press.

Plesofsky NS (2001) Pantothenic acid. In: Rucker RB, Suttie JW, McCormick DB, and Machlin LJ (eds.) *Handbook of Vitamins*, 3rd edn., ch. 9. pp. 317–337. New York: Marcel Dekker Inc.

Plesofsky-Vig N and Brambl R (1988) Pantothenic acid and coenzyme A in cellular modification of proteins. *Annual Review of Nutrition* **8**: 461–482.

Plesovsky-Vig (2000) Pantothenic acid. In: Shils ME, Olson JA, Shike M, and Ross AC (eds.) *Modern Nutrition in Health and Disease*, 9th edn., ch. 25. pp. 423–432. Baltimore: Williams & Wilkins.

Smith CM and Song WO (1996) Comparative nutrition of pantothenic acid. *Journal of Nutritional Biochemistry* **7**: 312–321.

Swaiman KF (2001) Hallervorden-Spatz syndrome. *Pediatric Neurology* **25**: 102–108.

Tahiliani AG and Beinlich CJ (1991) Pantothenic acid in health and disease. *Vitamins and Hormones* **46**: 165–227.

van den Berg H (1997) Bioavailability of pantothenic acid. *European Journal of Clinical Nutrition* **51**: S62–63.

PARASITISM

P G Lunn, University of Cambridge, Cambridge, UK

Introduction

In common with all other animals, human beings are susceptible to a range of parasitic organisms. The most important and commonest of these have been with man for countless years and have become so well adapted that in most cases man is their major if not only host. Although parasitic infections occur throughout the world, it is in the wet tropics and subtropics where they are found at their greatest prevalence and intensity. Most developing countries are also located in these areas and the consequent poverty, poor hygiene, and inadequate sanitation augment the favorable environmental conditions to enhance proliferation of these organisms. Only those that are known to interfere with host nutritional status will be discussed in this article.

Parasitic infections of the gastrointestinal tract are among the commonest diseases in the world (**Table 1**) and in most developing countries there has been little improvement in prevalence rates for many years. Indeed in some cases, e.g., schistosomiasis, local prevalence has been increasing with expanding irrigation schemes. Their association with poverty ensures that these diseases occur in areas where poor child growth and malnutrition are common and where there are persistent health problems. While there is no doubt that severe infections of any parasite can result in severe illness or even death of the host, such cases are rare even in areas of high prevalence and the norm is for low to moderate parasite numbers, which result in few, if any overt clinical symptoms. Nevertheless, by causing subtle reductions in appetite, digestion and absorption; by increasing chronic inflammation, and by inducing

Table 1 Estimated world prevalence of parasites important to human nutrition

	Approx. prevalence (millions)
Helminth parasites	
Ascaris lumbricoides (roundworm)	1500
Necator americanus and *Ancyclostoma duodenale* (hookworms)	1300
Trichuris trichiura (whipworm)	1100
Schistosoma haematobium, S. japonicum, and *S. mansoni*	200
Strongyloides stercoralis and *S. fülleborni*	200
Protozoal parasites	
Giardia intestinalis	200 symptomatic cases, total much higher
Entamoeba histolytica	400 but may be much higher
Cryptosporidium spp.	?

Data from Crompton, DWT (1999) How much human helminthiasis is there in the world? *Journal of Parasitology* **85**: 397–403; Olsen BE, Olson ME, and Wallis PM (2002) *Giardia: The Cosmopolitan Parasite*. Wallingford, UK: CABI; Haque R *et al.* (2003) Current concepts: amebiasis. *New England Journal of Medicine* **348**: 1565–1573.

nutrient loss, particularly of iron and protein, it is believed that such low-level but long-term infections contribute to the persistently poor nutritional state of many, especially children, in the developing world.

The most important parasites of man are from two main groups: the helminth worms and protozoans. Although several hundred different species have been described, the vast majority of infections are caused by relatively few.

Mechanisms of Parasite–Host Nutrition Interactions

Gastrointestinal parasites interfere with the nutrition of their host by one or more of the following mechanisms (**Figure 1**).

Loss of Appetite, Anorexia

Loss of appetite is a common feature in many illnesses and not only those involving the gastrointestinal tract. It is now thought that much of the appetite loss in disease is mediated by one or more cytokines released by lymphocytes as part of the body's response to tissue damage or invasion. Additionally, however, parasitized individuals often complain of symptoms such as nausea, abdominal pain, flatulence, and distension and discomfort, while the protozoal infections are associated with vomiting, diarrhea, or dysentery, all of which can be expected to reduce appetite.

Maldigestion and Malabsorption

Several GI parasites are well placed to interfere with these processes so it is not surprising that maldigestion and/or malabsorption of fat, protein, and carbohydrate as well as many of the micronutrients has been reported during infection. Structural damage to the mucosa of the intestine, such as the flattening or thickening of villi, or villus atrophy, will reduce the absorptive surface area. Damage to the cells diminishes their absorptive properties and limits active transport processes, while accelerated replacement of damaged cells may result in immature mucosal cells with reduced enzymatic and transport capacity. Food that is not fully digested and absorbed in the small intestine will enter the large bowel where excessive colonic fermentation may result in diarrhea.

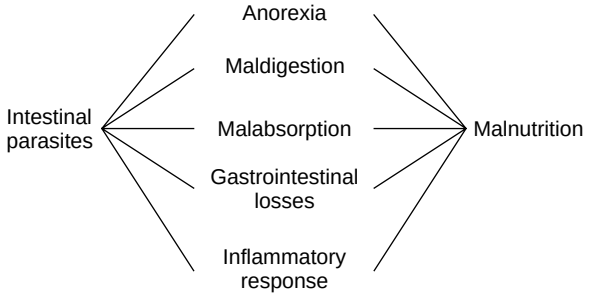


Figure 1 Mechanisms of parasite–host nutrition interactions.

Nutrient Losses

Accelerated loss of nutrients from the body is probably the most important mechanism by which parasitic infections compromise the nutritional status of their host. Nutrient losses arise both directly and indirectly.

Direct losses occur during the feeding of the blood-sucking and tissue-invading parasites. Blood and tissue ingested by the worms forms part of the loss but the lesions caused by feeding and burrowing activity continue to ooze blood and tissue fluids after the parasites have moved on. Similarly, the passage of schistosome eggs through the tissues of the bladder or intestine is often accompanied by tissue damage and blood loss. Increased turnover and accelerated shedding of parasite-damaged enterocytes into the lumen of the GI tract is another mechanism of increased nutrient loss. Even though some of the nutrients lost into the lumen may be reabsorbed, the process is far from complete. Vomiting or diarrhea causes loss of electrolytes and important trace elements such as zinc.

Indirect losses arise from stimulation of the host's immunological and inflammatory mechanisms that are mobilized to combat the infection and repair tissue damage. Localized inflammation at the site of the parasite activity, often accompanied by lymphocytic infiltration of tissues cause further damage to the mucosa, augmenting maldigestion, malabsorption, and nutrient losses as more damaged cells are shed. Activation of the systemic inflammatory system, i.e., the acute phase response, is a general reaction of the body to pathogen invasion or tissue damage. It results in a widespread cytokine-mediated catabolic response. Growth slows or ceases, muscle tissue is broken down to provide substrates for gluconeogenesis and the repair of damaged cells, and a negative nitrogen balance ensues. Anorexia occurs and there are increased losses of amino acids, minerals, and vitamins in the urine and feces.

Competition for Nutrients

Competition for nutrients is generally unlikely owing to the considerable difference in biomass of the host and parasite. However, the tapeworm, *Diphyllobothrium latum*, does compete for vitamin B_{12} taken in the diet. The worm concentrates large amounts of this vitamin in its own tissues, depriving the host and in some cases leading to megaloblastic anemia.

Parasite Epidemiology and Impact on Host Nutrition

Clinical Studies

Much of our knowledge of the impact of parasitism on host nutrition (**Table 2**) comes from hospital studies of heavily parasitized patients. Irrespective of the organism involved, nutritional status and anthropometric indices of such severely ill patients are invariably poor on admission but quickly improve following treatment. Such data must however be interpreted with caution. In developing countries, malnourished individuals admitted to hospital rarely suffer from a single parasitic infection; viral and bacterial pathogens and other parasitoses are frequently present as are frank dietary deficiencies. Patients are routinely dosed with wide-range antibiotics and anthelmintics and given high-quality rehabilitation diets, so, in general, neither the cause of their symptoms nor the basis for recovery can be established with certainty.

Helminth Parasites

Ascaris lumbricoides (**roundworm**) About 73% of all infections by this worm are estimated to occur in Asia with many countries having prevalence rates greater than 50%. In some rural areas over 90%

Table 2 Parasite interference with host nutrition

Parasite	*Symptom*	*Nutritional effect*
Ascaris lumbricoides	Anorexia and abdominal pain	Growth retardation, weight loss
	Malabsorption syndrome	Reduced fat and nitrogen uptake
		Reduced vitamin A status
	Lactose intolerance	Growth retardation, weight loss
	Acute-phase response	Growth retardation, weight loss
Hookworm	Anorexia and abdominal pain	Growth retardation, weight loss
	Diarrhea	Growth retardation, weight loss
	Blood loss	Iron deficiency, anemia
	Protein-losing enteropathy	Hypoalbuminemia, edema
Schistosoma spp.	Anorexia	Growth retardation, weight loss
	Diarrhea	Growth retardation, weight loss
	Blood loss	Iron deficiency, anemia
	Plasma protein loss	Hypoalbuminemia
	Acute-phase response	Growth retardation, weight loss
Trichuris trichiura	Anorexia	Growth retardation, weight loss
	Abdominal pain and vomiting	Growth retardation, weight loss
	Diarrhea and dysentery	Loss of trace elements, e.g., zinc
		Growth retardation, weight loss
	Blood loss	Iron deficiency, anemia
	Plasma protein loss	Hypoalbuminemia, edema
	Acute phase response	Growth retardation, weight loss
Strongyloides spp.	Anorexia	Growth retardation, weight loss
	Abdominal pain and vomiting	Growth retardation, weight loss
	Malabsorption syndrome	Reduced fat absorption
	Protein-losing enteropathy	Hypoalbuminemia, edema
	Acute-phase response	Growth retardation, weight loss
Giardia intestinalis	Anorexia	Growth retardation, weight loss
	Diarrhea and vomiting	Loss of trace elements
	Malabsorption syndrome	Reduced fat absorption
		Reduced vitamin A status
	Mucosal disruption	Lowered disaccharidase activity
		General maldigestion
	Acute-phase response	Growth retardation, weight loss
Cryptosporidium spp.	Anorexia	Growth retardation, weight loss
	Abdominal pain	Growth retardation, weight loss
	Diarrhea and vomiting	Loss of trace elements
	Mucosal disruption	Lowered disaccharidase activity
		General malabsorption
	Acute-phase response	Growth retardation, weight loss
Entamoeba histolytica	Diarrhea and dysentery	Fluid and electrolyte loss
		Electrolyte imbalance
		Loss of trace elements
	Acute-phase response	Growth retardation, weight loss

of children harbor the infection. It is less prevalent in Africa (about 12% of all cases) and in central and southern America (about 8% of all cases). It is uncommon but still present in some rural areas of Europe and southeastern parts of the US. Adult *A. lumbricoides* live in the lumen of the upper part, i.e., the jejunum, of the small intestine. The worms live for some 12–20 months and females grow to 20–35 cm in length and 3–6 mm in diameter. An adult female discharges 200 000–240 000 eggs per day into the lumen and these pass out of the body in the feces. Infection occurs by oral ingestion of eggs from fecally contaminated food, water, hands, kitchen utensils, or play things. Both the prevalence and intensity of infection with *A. lumbricoides* increase rapidly during early childhood and although prevalence often remains high throughout life, intensity of infection tends to peak in the 5–15 years age range.

Despite the large size of these worms, mild to moderate infections are generally well tolerated with few, if any, overt symptoms. Clinical studies give inconsistent results. Although anorexia, abnormal mucosal histology, decreased absorption of fat and carbohydrate, reduced lactase activity, decreased transit time, reduced nitrogen retention, reduced vitamin A absorption, and lower vitamin A status have all been reported, they are by no means present in all cases. These abnormalities are in keeping with a stimulation of the host's immune and inflammatory mechanisms and it seems likely that occurrence of these symptoms depends on whether such mechanisms have been initiated. Why the immune and inflammatory response should be initiated in some cases of *A. lumbricoides* infection but not others is not known, but it may be at least partly due to genetics.

Hookworms Although 13 different human hookworm parasites have been listed, only two species, *Necator americanus* and *Ancyclostoma duodenale*, are responsible for virtually all cases of hookworm disease in humans. The two worms are similar in appearance, feeding pattern, and life history. Man is their only known host. *Necator americanus* is the only species seen in North America and it predominates in central and southern America, central Africa, southern India, Indonesia, and the South Pacific. *Ancyclostoma duodenale* is found in Mediterranean Europe, the Middle East, North Africa, Pakistan, Iran, and northern India. Both species occur in parts of Brazil, India and Africa, throughout Southeast Asia, Indonesia and the Pacific islands.

Adult worms live in the upper part of the small intestine and eggs are discharged into the lumen. Up to 10 000 (*N. americanus*) or 25 000 (*A. duodenale*) eggs per day can be produced and are passed out in the feces. Eggs hatch within 48 h and the larvae are free living for 2–3 weeks but then must reach a host or die. Adult female *A. duodenale* are 10–13 mm in length, *N. americanus* 9–11 mm, and the males about 2 mm shorter. *Necator americanus* can live for up to 5 years.

Both prevalence and intensity of infection increase with age in childhood up to about 10–15 years, and then remain constant during adulthood. High prevalence is associated with inadequate or unhygienic disposal of feces, which contaminates the soil. Lack of footwear, a common state in developing countries, allows feet to come in contact with infective larvae.

Loss of blood, particularly of its iron content, is the most important pathological feature of hookworm infection. Iron deficiency anemia is one of the commonest deficiency diseases in the world and there is no doubt that hookworms contribute significantly to the estimated two billion individuals who suffer from this problem. Through its feeding activity, each *N. americanus* worm causes the loss of about 0.03 ml of blood per day, while the larger *A. duodenale* accounts for approximately 0.15 ml per day. Part of this loss is blood ingested by the worm, but each time the worm moves to a new site, perhaps up to six times per day, the lesions continue to ooze blood into the lumen.

Daily blood loss from an individual passing 2000 eggs per gram of feces has been estimated at 4.3 ml (containing 2.0 mg of iron) and 8.9 ml (4.2 mg of iron) for *N. americanus* and *A. duodenale*, respectively. Although the intestine will reabsorb approximately 35% of this iron, daily losses will be 1.3 and 2.7 mg of iron, respectively. Assuming that only 10% of dietary iron is absorbed, an increased dietary intake of 13 mg and 27 mg, respectively, is required to make good these losses. As most diets contain only 15–20 mg of iron per day and some 10–15 mg of this is needed to cover daily metabolic requirements, intake would need to at least double to replace the loss from even this moderate hookworm load. In developing countries this is rarely possible, so without iron supplements, iron stores are soon depleted and iron deficiency anemia ensues. In lighter infections, subclinical iron deficiency is shown by low plasma ferritin and iron concentration, low transferrin saturation, and elevated erythrocyte protoporhyrin content.

Protein is also lost into the lumen of the small intestine during hookworm disease. Estimates of plasma loss vary considerably and values over 100 ml (containing 6–7 g of protein) per day have

been recorded, although much of this may be reabsorbed. Nevertheless, moderate to heavy hookworm infections are associated with hypoalbuminemia, hypoproteinemia, edema, and kwashiorkor, especially in areas where the protein content of the diet is low.

Schistosomes The three commonest species responsible for disease are *Schistosoma haematobium, S. mansoni*, and *S. japonica*, with some individuals in Africa harboring two species. Urinary schistosomiasis, found mainly in Africa and some eastern Mediterranean countries, is caused by *S. haematobium*. Infection with either *S. mansoni* (found in Africa, the Middle East, parts of South America, and the Caribbean) or *S. japonica* (occurs in China, the Philippines, and Indonesia) results in intestinal schistosomiasis. These worms live in blood vessels: *S. haematobium* in the vesicle venules of the urinary bladder with the other two species infecting the mesenteric veins adjacent to the intestines. Adults live in male/female pairs and damage is caused by passage of eggs through the tissues into either the bladder (*S. haematobium*) or the gut lumen. Eggs leave the body in the urine or feces. If they reach fresh water, they hatch to produce miracidia, which must find a suitable snail host. After entering the snail, the parasites multiply by asexual reproduction, eventually producing free-swimming cercaria that are infective to man. Infection is by skin penetration during contact with fresh water containing cercaria. Egg production starts some 2–3 weeks after infection. The parasite lives for 3–8 years. The prevalence of this parasite in many developing countries is increasing as irrigation schemes allow the intermediary snail hosts to extend their range.

Iron deficiency anemia associated with blood loss occurs in both urinary and intestinal schistosomiasis. Although blood loss can be severe in heavy *S. haematobium* infection, in a study of nonhospitalized children with low to moderate infection, iron losses ranged from 120 to 500 $\mu g\,day^{-1}$, increasing with rising egg count. This is less than losses due to hookworm, but dietary iron consumption would need to increase by about a third to compensate. In areas where iron status is poor, the extra burden due to *S. haematobium* will undoubtedly contribute to the onset of anemia. Intestinal schistosomiasis caused by *S. mansoni* can also result in iron deficiency but it is generally less severe than that seen in hookworm disease. Little data is available for *S. japonicum* infection, but its effect appears to be similar to *S. mansoni*.

The poor nutritional status of infected individuals may be related to anorexia, diarrhea, and activation of the inflammatory mechanisms of the host. Blood cytokine concentrations are raised in schistosomiasis causing growth faltering and weight loss.

Trichuris *trichiuris* (whipworm) This helminth is widespread throughout the tropics and subtropics. Most cases of infection (63% of the worldwide total) occur in Asia, with 11% in Africa and 14% in the Americas; however, a few cases are still seen in the US, Western Europe, and Japan.

Man is the principal host of the parasite, which lives in the large intestine. Adults are 3–5 cm in length and are whip shaped; the long thin anterior end is embedded in the mucosa, with the thicker posterior end in the lumen. Worms feed on mucosal cells but may also ingest red and white blood cells. Eggs leave the host in feces and embryonate in the soil. Infection occurs by oral ingestion of embryonated eggs on fecally contaminated food, hands, or utensils. They hatch in the small intestine and larvae develop in the villi before moving down to the large intestine. Egg production starts 30–90 days after ingestion.

In some rural areas the prevalence can exceed 90% and although prevalence remains high throughout life, peak intensity usually occurs between the ages of 5 and 15 years.

This helminth causes loss of blood and iron from the large intestine of its host by its burrowing and feeding activities. More than 3000 worms have been found in heavy *T. trichiuris* infection and such individuals do have marked iron deficiency anemia. However, in the majority, where worm counts rarely exceed 100, the infection is usually asymptomatic. Plasma protein loss can also be substantial in heavy infections but although plasma albumin values are frequently reduced, hypoproteinemic edema is rare.

Heavy infections are characterized by persistent dysentery, abdominal pain, nausea, vomiting, and tenesmus leading to rectal prolapse. Appetite is reduced and raised plasma cytokines and acute-phase proteins indicate activation of host immune and inflammatory mechanisms. Loss of nutrients including zinc and other trace elements in the persistent dysentery and vomiting may further lower nutritional status.

Strongyloides stercoralis This worm has a worldwide distribution but is found predominantly in the tropics. Prevalence rates are uncertain as detection of the larvae by direct fecal examination (the method usually employed) gives a considerable underestimate. Prevalence rates of up to 85% have been reported but are uncommon. In parts of Africa,

a closely related worm, *Strongyloides fülleborni* is often more common than *S. stercoralis*.

Adult worms are about 2.7 mm in length and are usually found in the duodenum and upper jejunum. Eggs are passed into the lumen but most hatch while still in the GI tract. Although the majority of larvae pass out in the feces, some penetrate the wall of the intestine and reinfect the host, a situation known as autoinfection. Larvae passed with the feces live in the soil and grow into adults of both sexes. Eggs are laid and larvae hatch within 1–2 weeks. They metamorphose to an infective stage when they must either locate a host or die. Infection is usually by skin penetration. Because of the autoinfection process, infection with this parasite can last indefinitely and severe disease can suddenly appear many years after an individual has left an endemic area.

The impact of this worm on nutritional status has not been clearly defined. Heavily infected subjects have a severe small intestinal illness with anorexia, abdominal pain, nausea, diarrhea, and vomiting. There is some evidence for a malabsorption syndrome, steatorrhea is often present, but it is not seen in all cases. A substantial protein-losing enteropathy can occur, resulting in severe hypoalbuminemia and kwashiorkor-like oedema. Protein loss arises from a combination of the burrowing activity of the worms and a local inflammatory reaction from the host. The little information on *S. fülleborni* infection suggests it has a similar impact on host nutrition.

Special features of helminth parasites Helminth infections all exhibit certain characteristic features by which they differ from most other infective organisms:

1. In contrast to most infective organisms, most helminths cannot reproduce within the host; each worm has to gain individual access to the host, usually by ingestion or skin penetration.
2. Intensity of infection shows an overdisperse distribution; it is usual for 20% of an infected population to harbor 80% of the parasites. Thus, a large majority of individuals will have only light infections and show few if any symptoms.
3. Some individuals appear to be predisposed to have heavy worm burdens; they quickly reacquire a heavy load after eradication of their original infection. Household and family clustering of high parasite loads is also seen. Clearly such differences might be explained on the basis of increased exposure of individuals and family groups due to particularly unhygienic living conditions or greater occupational risk. However, increased host genetic susceptibility has recently been demonstrated to account for between 21 and 44% of the observed variance in infection intensity.
4. Infection by several different parasites at the same time (polyparasitism) is extremely common in many areas.
5. Re-infection following deworming occurs very quickly because of considerable contamination of the environment by the large numbers of eggs produced by the parasites. In a study in Myanmar, preinfection prevalence of *A. lumbricoides* was reached only 6–8 months after deworming.

Protozoal Parasites

***Giardia intestinalis* (= *lamblia*)** This organism is a common parasite of the human gastrointestinal tract and is found in all parts of the world. Although its prevalence is greatest in developing countries where hygiene facilities are poor, outbreaks of giardiasis continue to occur in many developed countries. It has a simple life history. The trophozoite (the active form in the intestine) lives in the duodenum and jejunum of the host where it attaches to the enterocytes by means of a ventral disk. It reproduces rapidly by mitotic division and in heavy infections can cover large areas of the mucosa. Some trophozoites encyst; a protective wall forms around the organism, and the cysts pass out in the feces. Cysts are directly infective and after ingestion by a new host, the organisms emerge to establish a new infection. Disease can follow the ingestion of as few as 10 cysts which, given moist conditions, are viable for several months.

In developed countries, most infections can be traced to contaminated water, but direct person-to-person transmission has been documented. In developing countries, poverty-related unsanitary conditions and inadequate disposal of feces promote oro-fecal spread of the parasite, but contaminated water is also likely to be important. The large number of cyst-producing individuals with asymptomatic infection constitutes a reservoir of *G. intestinalis*. In addition, some animals are known to harbor *Giardia* and may be a source of human giardiasis.

Infection with *G. intestinalis* can be associated with a wide range of symptoms: from mild, self-limiting watery diarrhea to persistent foul-smelling diarrhea with vomiting, abdominal pain and distension, and a severe malabsorption syndrome. However, many infected individuals (from 20 to 84% of infected cases) remain asymptomatic. It is not clear why the parasite can cause such a range of degrees of illness.

The nutritional impact varies with both the severity and duration of the symptoms. In the early stages, anorexia is of major importance, but if the disease persists, intestinal aspects compound the situation. In at least 50% of symptomatic patients there is malabsorption of fat, carbohydrates, protein, and micronutrients (particularly vitamin A) associated with structural and functional abnormalities in the small intestine. Damage to the mucosa can range from little to subtotal villus atrophy, but most subjects have mild villus shortening and increased crypt depth. The abnormalities are associated with a reduction in disaccharidases, notably lactase activity and in lowered intraluminal concentrations of the hydrolytic enzymes trypsin, chymotrypsin, and lipase. The small intestinal barrier function is compromised, allowing translocation of potentially antigenic macromolecules into the body with consequent stimulation of the immune and inflammatory mechanisms resulting in growth retardation. Little is known about the nutritional effects of nonsymptomatic giardiasis.

***Cryptosporidium parvum and* other *Cryptosporidium* species** These organisms have only been recognized as human parasites since 1976. They have a worldwide distribution but in developed countries they generally causes a self-limiting disease, which occurs most commonly in child institutions and in people working with animals. However, water-borne outbreaks have occurred in which large numbers of people have become infected. Cryptosporidiosis is much more prevalent in developing countries where it is mainly a disease of children. The parasites live in the upper part of the small intestine, attached to the mucosal cells from which they feed.

Both sexual and asexual reproduction occurs in the host and cysts are produced, most of which pass out in the feces. However, some excyst while passing through the gastrointestinal (GI) tract resulting in autoinfection that can prolong the disease long after the original source of infection has been eliminated. Infection is by ingestion of cysts in fecally contaminated food, water, or utensils, or from unhygienic contact with infected persons. Continued exposure is facilitated by the many infected individuals who remain asymptomatic while passing cysts. *Cryptosporidium* spp., including *parvum*, also occur in many animals and can be transmitted to humans.

Most infected individuals remain asymptomatic, but in others, acute or chronic diarrhea associated with vomiting, abdominal pain, dehydration, and fever can occur. Immunocompromised and previously malnourished cases tend to have more severe and prolonged disease.

The nutritional impact of the infection depends on the severity and duration of the infection but growth retardation and lowered nutritional indices occur in asymptomatic cases as well as those with symptoms. Structural damage to the mucosa of the small intestine is seen, with shortened and fused villi and lengthening of the crypts due to accelerated cell division to replace damaged cells. Surface area is greatly reduced and the immature enterocytes have lower enzymatic and transport activity than mature cells. The resulting maldigestion, malabsorption, and stimulation of the host immune and inflammatory mechanisms are likely to account for the adverse nutritional effects. In children, growth remains poor for many months after infection has resolved. The nutritional status of asymptomatic individuals appears to be compromised by less extreme expression of these same mechanisms.

Entamoeba histolytica This ameba has a very wide distribution but is most commonly found in developing countries where lack of hygienic facilities exacerbate fecal contamination of water, food, and hands. The organism is exclusive to humans; there are no animal hosts. These parasites generally infect the large intestine where they can cause severe disease by invading mucosal tissues.

The life cycle is simple: adult amebae reproduce asexually forming substantial colonies and in some cases cause ulcerative lesions in the mucosa. Some organisms encyst and pass out with the feces. Following ingestion of the cysts by another host, the amebae emerge when the cyst reaches the large bowel. The organism can also invade other organs, notably the liver, resulting in a life-threatening illness.

Although this parasite can cause life-threatening diarrhea and dysentery in some, most infected individuals remain free of symptoms. In others, persistent diarrhea can continue for months, interspersed with periods of apparently normal bowel function. As the parasite is most commonly found in the large intestine, there is little interference with food digestion and absorption and its main effect on nutrition seems to be due to loss of trace elements and electrolytes in watery stools. In more severe cases, blood is also lost in this way but amounts are small. Infection is associated with inflammation of the large bowel (colitis) indicating that host immune and inflammatory mechanisms have been stimulated and this may account for reports of hypoalbuminemia.

Community and Intervention Studies

Iron Deficiency and Iron Deficiency Anemia

A close relationship between the level of hookworm infection and severity of anemia has been observed in many cross-sectional field studies. Similar, though generally less severe, levels of anemia have been associated with intensity of schistosome species and *T. trichiura* disease. The cause and effect relationship suggested by this data has been confirmed by longitudinal investigations of iron status following anthelmintic administration. Community studies in Kenya, India, and Papua New Guinea have recorded substantial increases (up to $6\,g l^{-1}$) in hemoglobin concentration between 4 and 8 weeks after treatment for hookworm. These marked improvements were seen even when parasite loads were not completely eliminated. Effective treatment of severe *Schistosoma and T. trichiura* infections also results in much improved iron status.

Growth and Protein-Energy Malnutrition

Although clinical studies confirm that these parasites have the potential to interfere with growth and nutritional status, evidence that they are a major cause of the widespread stunting and protein-energy malnutrition seen in developing countries is not as conclusive as may be expected. This may be because most infected individuals in a community will have only low to moderate parasite loads and whether a particular disease is important in precipitating malnutrition on a community or public health scale will depend on whether or not such low level infections impact on nutritional status. Information has come from two types of study: (1) cross-sectional surveys; and (2) longitudinal, placebo controlled intervention studies in which nutritional improvements are sought following the use of antiparasite drugs.

A large number of cross-sectional community studies have associated parasitic infection with growth deficits and poor anthropometric indices. Schistosomiasis has long been associated with poor growth, and an extreme condition, schistosomiasis dwarfism, in which physical and sexual development were severely retarded was reported to be quite common in China until the 1950s. Most recent studies of mild to moderate infection with all three schistosome species confirm an association with poor nutritional status that is more marked in girls, but the degree of impairment is variable between different regions and at best can only explain a small part of the total nutritional deficit of the subjects. Hookworm infection is similarly associated with poor appetite, slower growth, and lowered nutritional indices, all of which become more marked with increasing severity of iron deficiency anemia. Iron supplementation of hookworm-infected children has been reported to improve appetite and growth performance as well as iron status, suggesting that the lowered nutritional status may be secondary to iron deficiency rather than a direct effect of the parasite. Growth retardation seen in moderate to heavy *T. trichiura* infection may be similarly explained, although heavier burdens of this worm frequently cause dysentery, which can result in loss of essential trace elements such as zinc.

The impact of the protozoal parasites *Giardia* and *Crytosporidium* on nutritional status has been less well studied, but infection appears to be associated with persistent diarrheal disease and prolonged growth faltering even after apparent elimination of the parasites. Moreover, these parasites, unlike the helminths, are very common in children during the first 2 years of life when growth is at its greatest. Growth-retarding infections at this time of life, particularly in developing countries, appear to compromise growth throughout the whole growth period, thus the impact of these parasites on nutritional status may be far greater than currently appreciated. This is certainly an area requiring more research.

The results of these cross-sectional studies have been reinforced by longitudinal community-wide studies of nutritional improvement following reduction or eradication of parasite burden with anthelmintic drugs. The results of such studies have, however, been less than convincing. Successful treatment of heavily poly-parasitized Kenyan children harboring hookworm, *A. lumbricoides* and *T. trichiura,* with albendazole resulted in improvements in weight, arm circumference, and skinfold thickness and was associated with increased appetite and fitness. Statistical analysis of this data implicated hookworm as being the most important in compromising nutritional status. Weight gain above placebo-treated counterparts averaged 1.3 kg per 6 months, and added about 3% points to a weight-for-age of approximately 80%. However, similar studies in many parts of the world in subjects with lower intestinal helminth burdens have reported only small improvements, whereas others found no change at all in nutritional status indices following successful deworming. Treatment of schistosomiasis in Kenyan, Brazilian, and Filipino children showed only small improvements in nutritional status, e.g., in Kenya, the per cent weight-for-age only increased from 72.9 to 74.9% following eradication with praziquantel. A recent meta-analysis of these studies concluded that deworming did improve nutritional status, but that the effect was small.

Overall, both community and intervention studies do suggest that elimination of GI parasites would improve growth and anthropometric status of children in developing countries but that such improvement would be limited. This contrasts with the very substantial improvement in iron status and iron deficiency anemia that follows effective treatment of organisms causing blood loss.

Treatment and Prognosis

Table 3 shows the drugs most commonly used in treatment of these parasitic infections. Anthelmintic drugs have improved dramatically during the last 20 years and are now highly effective; in most cases a single course of treatment will result in parasite eradication. However, immunocompromised hosts, including malnourished children, may require more extensive courses of therapy to completely eliminate the infection. This is particularly the case in the treatment of cryptosporidiosis. Iron supplements are usually provided where blood loss has resulted in iron deficiency anemia.

Recovery from infection is usually complete and rapid as most parasites do not cause lasting damage to their host. Schistosomiasis is the exception and can result in permanent granuloma formation in several tissues, particularly the liver and spleen, which may become life threatening.

Prevention

Although drugs are now available to eradicate infections, unless the home environment changes, most individuals will soon become reinfected. The transmission of all the parasites discussed occurs most commonly through close contact between the host and infected human feces, either orally or by skin penetration. The basic requirement for prevention is an efficient and hygienic mode of disposal of feces, improved facilities in the home, for example clean running water, concrete floor to the home, plus a knowledge of basic hygiene. Use of footwear and avoidance of contact with water likely to contain schistosome cercaria would help. For the foreseeable future, however, such control measures are quite unrealistic in most developing countries and the alternative may be the large-scale, nation-wide use of anthelmintics to regularly deworm all individuals in endemic areas. School-based regular treatment programmes can be effective. Safe, effective, and relatively cheap drugs are now available and their use in this way could substantially reduce the level of helminth disease throughout the developing world. Such programs can be expected to result in a marked reduction in the prevalence and severity of iron deficiency anemia but in most situations, to have a relatively small impact on child growth, stunting, and incidence of protein-energy malnutrition.

Table 3 Drugs of choice for parasitic infections

Infection	*Drug*
Ascariasis	Mebendazole, albendazole, pyrantel pamoate
Hookworm infection	Mebendazole, albendazole
Schistosomiasis	Praziquantel, metrifonate, niridazole, oltipraz
Trichuriasis	Mebendazole, albendazole
Strongyloidiasis	Thiobendazole, ivermectin
Giardiasis	Metronidazole, tinidazole, secnidazole, furazolidone, albendazole
Cryptosporidiosis	Nitazoxanide, spiramycin, clindamycin,
Amebiasis	Metronidazole, secnidazole, paromomycin, nitazoxanide

See also: **Anemia**: Iron-Deficiency Anemia. **Cytokines**. **Diarrheal Diseases**. **Infection**: Nutritional Interactions. **Iron**. **Zinc**: Physiology.

Further Reading

Cooper ES, Whyte-Alleng CAM, Finzi-Smith JS, and MacDonald TT (1992) Intestinal nematode infections in children: the physiological price paid. *Parasitology* **104**: S91–S103.

Crompton DWT (1999) How much human helminthiasis is there in the world? *Journal of Parasitology* **85**: 397–403.

Crompton DWT (2000) The public health importance of hookworm disease. *Parasitology* **121**: S39–S50.

Crompton DWT and Nesheim MC (2002) Nutritional impact of intestinal helminthiasis during the human life cycle. *Annual Review of Nutrition* **22**: 35–59.

Dickson R, Awasthi S, Williams P, Demellweek C, and Gamer P (2000) Effects of treatment for intestinal helminth infection on growth and cognitive performance in children: systematic review of randomised trials. *British Medical Journal* **320**: 1697–1701.

Grove DI (1996) Human strongylodiosis. *Advances in Parasitology* **38**: 251–309.

Haque R, Huston CD, Hughes M *et al.* (2003) Current concepts: amebiasis. *New England Journal of Medicine* **348**: 1565–1573.

Lunn PG and Northrop-Clewes CA (1993) The impact of gastrointestinal parasites on protein-energy malnutrition in man. *Proceedings of the Nutrition Society* **52**: 101–111.

O'Lorcain P and Holland CV (2000) The public health importance of *Ascaris lumbricoides*. *Parasitology* **121**: S51–S71.

Olson BE, Olson ME, and Wallis PM (2002) *Giardia: The Cosmopolitan Parasite* Wallingford, UK: CABI.

Solomons NW (1993) Pathways to the impairment of human nutritional status by gastrointestinal pathogens. *Parasitology* **107**: S19–S35.

Stephenson LS (1993) The impact of schistosomiasis on human nutrition. *Parasitology* **107**: S107–S123.
Stephensen LS, Holland CV, and Cooper ES (2000) The public health significance of *Trichuris trichiura*. *Parasitology* **121**: S73–S95.
Stephenson LS, Latham MC, and Ottesen EA (2000) Malnutrition and parasitic helminth infections. *Parasitology* **121**: S23–S38.
Tzipori S (2002) Cryptosporidiosis: current trends and challenges. *Microbes and Infection* **4**: 1045–1080.

Pathogens *see* **Infection**: Nutritional Interactions; Nutritional Management in Adults

PELLAGRA

C J Bates, MRC Human Nutrition Research, Cambridge, UK

History of Pellagra: Recognition, Causes, and Treatment

The following historical summary is based mainly on Carpenter's excellent compendium of key pellagra-related publications published in 1981. Pellagra (meaning 'rough' or 'raw' skin) was common in western Europe (e.g., France and Italy) and especially the southern United States until the early decades of the twentieth century. It has been estimated that it claimed approximately half a million lives between the early eighteenth century and 1930, with as many as 10 000 deaths in the United States in 1929 alone. The typical signs and symptoms of human pellagra are summarized in **Table 1**.

The characteristic signs and symptoms of pellagra were given the name 'mal de la rosa' in 1720 by doctors working in the Asturia region of Spain, and it was very common in Italy at the end of the eighteenth century. As early as 1810, one European description concluded that the disease was neither contagious nor hereditary but was probably caused by a poor diet, especially diets in which grain such as corn (i.e., sweet corn) was the principal staple. The concept of a 'protein' deficiency, as distinct from the characteristic body-wasting calorie deficiency of common famines, was proposed in approximately 1850, and a good hospital diet was shown to have positive curative effects. As early as 1860, however, one observer commented that poor Mexican peasants whose diet was mainly corn based did not exhibit pellagra, and he attributed this to their practices of roasting the corn with lime and of preventing mould growth.

Table 1 Signs and symptoms of niacin deficiency in man and animals

Human deficiency[a]
Loss of appetite and weight
Dermatosis (hyperpigmentation, hyperkeratosis, desquamation of the epidermis, especially where frequently exposed to strong sunlight)
Anorexia
Achlorhydria
Angular stomatitis, cheilosis, magenta tongue
Diarrhea
Anemia
Neuropathy (headache, dizziness, tremor, neurosis, apathy)
Death in severe and prolonged cases
Blacktongue in dogs and cats
Pustules in mouth and excessive salivation, darkening and necrosis of tongue
Diarrhea
Pigs
Neurological lesions affecting ganglion cells; histopathology of nerves
Anemia
Degeneration of intestinal mucosa and diarrhea
Rats
Reduced growth rate
Alopecia
Damage to peripheral nerves (cells and axons)
Birds (e.g., chickens and ducks)[b]
Inflammation of the upper gastrointestinal tract
Dermatitis
Diarrhea
Poor growth of feathers; bowed and weakened legs

[a]All animal species lose appetite and weight when deficient, but the characteristic skin lesions that are observed in human pellagra are rarely seen in other species. Some deficiency symptoms have been produced in other primates (e.g., monkeys).
[b]Ruminants are usually resistant to pellagra, except when forced to produce high quantities of milk or lean tissue.

For the next approximately 50 years, the 'toxin' theory of pellagra causation held sway, and there were government-backed campaigns in Europe to prevent mould growth (in Italy) and to reduce the population's reliance on corn as a staple (in France). In Europe, the prevalence of pellagra was clearly declining sharply (by the beginning of the twentieth century) just as it was beginning to emerge as a major new scourge in the southern United States. A plethora of conflicting hypotheses in the United States included poor sanitation, infection, insect-borne disease, and toxins from bacteria or moulds, but the largely correct 'poor diet' hypothesis was completely ignored until Joseph Goldberger, during the period from 1914 until his death in 1929, carried out classical and definitive controlled feeding studies, both in human convicts and in an animal model, causing 'blacktongue,' a corn diet-induced condition that could be induced in dogs (**Table 2**). Although the exact relationship of the diseases that Goldberger described, and then produced in his animal model, to classical human pellagra remains controversial, the most important outcome of his studies was that pellagra was now seen not as an infectious disease but as primarily a diet-related one, and it was recognized that it and similar diseases could unequivocally be induced by monotonous, poor-quality diets. Families who kept a cow were relatively protected.

Funk's newly formulated hypothesis about essential 'vitamines' and Gowland Hopkins' concept of 'accessory food factors', both of which were emerging at about the same time as Goldberger's studies, set the scene for a focused hunt for a specific organic substance that would be present in the 'curative' diets and which might thereby be identified as the elusive 'pellagra-preventive' (PP) that was thought to be present in the preventative and curative foods. Goldberger classified a range of foods according to their PP properties and found that dried yeast and a water-soluble extract from yeast were both curative, even in small quantities. During the 1930s, following the elucidation of the role of the pyridine nucleotides in food energy metabolism and release, the central roles of nicotinic acid and nicotinamide were elucidated and were equated with the curative PP factor present in the curative food extracts. The term 'niacin' was then coined because 'nicotinic acid' appeared to be etymologically associated with tobacco and was therefore considered to be unsuitable as the name for an essential dietary factor.

Complex Causation

Although pellagra in dogs and humans usually responds well to supplements of pure niacin, there are several further strands to the story that complicate the idea that all the characteristics and manifestations of pellagra can be explained as the result of a simple dietary deficiency of a single water-soluble factor (i.e., vitamin) identified as the molecule niacin. First, it soon became clear that the total niacin content of different foods, as measured by chemical analysis, was not necessarily a good guide to their pellagra-producing or preventing properties. During the 1940s, it was shown that in rats (which respond to pellagragenic diets by a reduced growth rate but not by skin lesions) high dietary tryptophan levels could substantially reduce the requirement for dietary niacin. Tryptophan was then shown to be equally effective in humans in reducing the pellagragenic properties of poor diets.

Table 2 History of the recognition of pellagra, and its probable causes, in human populations

1. A poorly understood disease (dermatitis, gastrointestinal and mental signs/symptoms) appeared first in Europe in the eighteenth and nineteenth centuries and then in the southern United States in the 1910s and was named 'pellagra' (= raw/rough skin).
2. Favored causal hypotheses (USA) initially included infection, mouldy grain and insects.
3. In 1914–1916, Joseph Goldberger disproved the infection hypothesis by self-experimentation and then producing pellagra in prisoners fed mainly corn diets.
4. In the 1920s, he developed the "blacktongue" model of pellagra in dogs, with corn diets.
5. In the 1930s, nicotinic acid was isolated as a pure water-soluble compound ('vitamin') of known structure from yeast and liver extracts, able to cure pellagra and blacktongue.
6. In the mid-twentieth century, niacin was shown to be bound as an unavailable part of the large chemical complex 'niacytin' in corn. Heating in alkaline environment (e.g., Mexican tortillas) can liberate this bound niacin.
7. Intermediary metabolism studies revealed that niacin can be produced from tryptophan in the body. Pellagragenic diets are therefore low in tryptophan as well as niacin. The concept of 'niacin equivalents,' usually [mg niacin plus one-sixtieth of mg tryptophan] in food, developed. Human requirements are estimated.
8. The Indian cereal 'jowar' shown to be pellagragenic. Some, but not all, studies have implicated its high leucine content.
9. Niacin and riboflavin deficiencies often coexist; therefore, the signs and symptoms of 'pellagra-like' disease are often attributable to multiple vitamin deficiencies, of which niacin and riboflavin are usually the most important.
10. Certain inborn errors of metabolism (genetic defects) or iatrogenic effects of drugs can mimic pellagra signs, symptoms, and metabolic defects.

Soon after this, the complex metabolic pathway linking tryptophan to niacin and to the pyridine nucleotide coenzymes was elucidated. Between 34 and 86 mg tryptophan in human diets is now considered to be equivalent to 1 mg niacin, with a mean conversion ratio of 60 mg tryptophan per 1 mg niacin used universally to calculate the niacin equivalents (NE) value of any diet. This is recognised to be a much better index of the anti-pellagra potency of a diet than its niacin content *per se*.

However, this was only one part of the complex etiology of pellagra. The causes of the signs of pellagra in human populations, and indeed also in Goldberger's experimental studies, were and are likely to be a complex mixture of B vitamin deficiencies, of which niacin and tryptophan content are indeed the dominant effectors, but riboflavin is also an important component, followed sometimes by thiamin, vitamin B_6, and possibly vitamin B_{12} and some other nutrients including zinc and iron. The antivitamin effects of certain toxins, especially mould toxins, cannot be ruled out, and several bacterial, fungal, and other toxins have been shown to be capable of depleting cellular levels of NAD(P). In addition, and perhaps most important for corn and the other grain diets, the niacin present in corn and other grains is often chemically bound into a macromolecular complex that is sometimes called niacytin, from which the niacin cannot readily be released by digestive enzymes in the gastrointestinal tract but which requires heat and alkali treatment during food preparation (as in the preparation of Mexican tortillas, which involves lime and heat treatment) so as to make it adequately bioavailable.

In India, the millet-type cereal called 'jowar' is frequently associated with pellagra signs and symptoms, even though it is apparently a reasonably good source of available niacin and tryptophan. Studies have suggested an association with its high leucine content, which may impair the conversion of tryptophan to niacin coenzymes. However, the aetiology of this association remains controversial and unresolved. Balance studies in humans have failed to show a consistent effect of either leucine or vitamin B_6 supplementation on the excretion of the metabolites of tryptophan, which is a sensitive test for imbalanced tryptophan conversion pathways.

Oestrogenic hormones can affect the conversion of tryptophan to niacin coenzymes, and this is thought to be the causal basis for the observation that women (except during pregnancy) seem to be considerably more susceptible to pellagra than men.

Several commonly used drugs also have anti-niacin (iatrogenic) effects in man. Perhaps the most important is isoniazid, which is commonly used in the treatment of tuberculosis. It inhibits kynureninase activity (an enzyme in the tryptophan conversion pathway) by inactivating the enzyme's essential cofactor, pyridoxal phosphate, derived from vitamin B_6. There are several other metabolic interconnections between niacin and vitamin B_6, such that any interference with vitamin B_6 metabolism is likely to affect niacin economy as well. Since up to 60% of Asian Indians are genetically slow acetylators (i.e., deactivators) of isoniazid, the use of this drug in Indians is especially apt to cause pellagra symptoms. The anti-Parkinsonism drugs Carbidopa and Benseride can also cause pellagra symptoms, and in people taking these, there is a reduced rate of excretion of the niacin metabolite *N*-methylnicotinamide. Another niacin antagonist is *N*-acetyl pyridine, which can cause neurological symptoms and histological damage to the hippocampus in some animals.

There are also several inborn errors of metabolism that can result in pellagra-like symptoms in man. Although none of these are very common, the best known is Hartnup's disease, an autosomal recessive condition in which the cellular transport of tryptophan (and other neutral amino acids) is impaired so that tryptophan is rapidly lost in the urine through failure of renal tubular reabsorption. Patients respond well to supplementation with niacin or with tryptophan peptides but not to free tryptophan, which cannot be well absorbed and retained. Other inborn errors of tryptophan economy that can result in pellagra include the vitamin B_6-responsive condition xanthurenic aciduria, hydroxykynurenuria, tryptophanuria (i.e., tryptophan dioxygenase deficiency), and another linked to an increased activity of the enzyme picolinate carboxylase. All these conditions are rare but informative inborn errors of metabolism affecting the tryptophan–niacin metabolic pathway. Tumors of the enterochromaffin cells, which synthesise excessive amounts of 5-hydroxytryptophan and 5-hydroxytryptamine, can also result in pellagra since hyperactivity of this pathway can result in the diversion of tryptophan away from the alternative pathway that converts it to the niacin coenzymes.

High-Risk Groups in Present-Day Society

Today, the most high-risk group for development of pellagra signs and symptoms in Western society is chronic alcoholics, whose diets are often poor, and in addition are subject to liver damage from alcohol abuse and its cellular toxicity. Certain forms of psychosis, including depression and schizophrenia, are associated with abnormalities of the tryptophan metabolism pathways, including those involved in the formation of 5-hydroxytryptamine (serotonin)

and 5-hydroxytryptophan in the central nervous system. Some of these may benefit from modulation of these pathways by drugs and/or supplements. People with AIDS may exhibit some impairment of NAD production, which may in turn respond to niacin supplements as a support treatment, and high-dose nicotinic has been used as one of many alternative treatments for people with cardiovascular disease. In the developing world, pellagra is most commonly encountered in certain African countries (e.g., refugees in Malawi) and areas of India and China.

Biochemical Status Assays: Recommended Intakes

The detection of subclinical niacin deficiency and the confirmation of a clinical deficiency require the objective measurement of biochemical status (and, if possible, of dietary intake, which is usually a more time-consuming task) in order to provide confirmatory evidence, especially because the typical clinical signs and symptoms of pellagra are not entirely specific and pathognomic. Biochemical status estimates can be used to characterize a population, particularly any high-risk subgroups, and to monitor the efficacy of any anti-pellagra interventions. For most micronutrients, robust, specific, and sensitive blood component status assays have been developed. However, for niacin the only promising blood-based assay, namely of intracellular pyridine nucleotide (NAD(P)) concentrations, has not been developed into a definitive and generally accepted biochemical status assay with well-defined normal ranges and a demonstrated association between low concentrations and clinical deficiency signs and symptoms. It has been suggested that a ratio of NAD to NADP below 1.0 in erythrocytes may provide evidence of niacin deficiency, but this requires confirmation.

The practical measurement of niacin status has mainly depended on urinary assays of the excretory products of niacin metabolism, namely N^1-methyl nicotinamide (NMN), N^1-methyl-2-pyridone-5-carboxamide (2-pyridone), and N^1-methyl-4-pyridone-3-carboxamide (4-pyridone), which can be quantitatively estimated by high-performance liquid chromatography separation followed by UV absorption–detection. The Interdepartmental Committee on Nutrition for National Defense has selected as the preferred principal index of niacin status an NMN excretion rate of 5.8 μmol (0.8 mg) per day in 24 h urine samples as defining the junction between biochemical deficiency and sufficiency. If only casual (spot) urine samples are available, then the ratio of NMN to 2-pyridone may provide a useful alternative index, and one study suggested that <8.8 μmol of the combined excretion of NMN plus 2-pyridone can be considered as defining borderline adequacy, corresponding to a niacin intake in the region of 6 mg NE/day. The average adult NE requirement has been estimated from depletion–repletion studies to be approximately 5.5 mg NE/1000 kcal food energy/day, and thus with a 20% allowance for individual variation to cover the needs of the majority of healthy individuals, an RNI (UK) of approx. 6.6 mg NE/1000 kcal (4200 kjoule) food energy/day translates into the broad ranges of UK RNI values that are shown in **Table 3**. In the USA, the basis for the calculation is

Table 3 Reference and recommended intakes of niacin equivalents[a]

Age group	*United Kingdom*[b]		*United States*[c]
	LRNI (mg niacin equivalents/day)	*RNI*	*RDA (mg niacin equivalents/day)*
0–6 months	2	3	2
6–12 months	3–4	4–5	4
12 months–13 years	5–10	8–15	6–12
Adult	8–12	13–18	14–16
Lactation	10	15	17

[a]One niacin equivalent (mg NE) is equivalent to 1 mg niacin or one-sixtieth of the tryptophan consumed. Since the mean energy intake increases with age, and differs between the sexes after puberty, there is a corresponding difference in the absolute values for each population group. This table provides only a simplified summary of the published values.

[b]UK values are calculated on the basis of an LRNI (Lower Reference Nutrient Intake) of 4.4 mg NE/1000 kcal food energy and an RNI (Reference Nutrient Intake) of 6.6 mg NE/1000 kcal, both of which are constant for all population groups. The LRNI is intended to cover the needs of the lower 2.5% of a healthy population, whereas the RNI is intended to cover 97.5% of a healthy population.
Source: Department of Health (1991) *Report on Health and Social Subjects No. 41. Dietary Reference Values for Food Energy and Nutrients for the United Kingdom*. London: HMSO.

[c]The US RDA values are intended to cover the needs of 97.5% of a healthy population and are set 30% above the Estimated Average Requirements. Source: Food and Nutrition Board (1998) *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin and Choline*. Washington, DC: National Academy Press.

now somewhat different, but nevertheless, the US RDA ranges shown in **Table 3** are generally similar to the UK RNIs.

Glossary

Niacin Combination of nicotinamide and nicotinic acid.

Niacin equivalents (NE) mg niacin plus one-sixtieth of mg tryptophan in a defined quantity (e.g., mg/100 g or mg/d) of food or diet.

Niacytin A macromolecular complex of niacin found in some cereals, notably maize (sweet corn), which is poorly bioavailable but liberates niacin on heating with alkali. Mexican tortillas are cooked by heating maize flour with lime, which liberates the niacin in a bioavailable form.

Jowar An Indian cereal (staple) that is associated with high pellagra risk and is rich in the amino acid leucine, but which contains only a marginally adequate amount of niacin + tryptophan.

Isoniazid A drug used in the treatment of tuberculosis that antagonizes vitamin B_6 and is associated with pellagra symptoms, especially in Asian Indians.

Carbidopa, Benseride Anti-Parkinsonism drugs that can cause pellagra-like symptoms.

***N*-acetyl-pyridine** A niacin antagonist, studied in animals.

NAD, NADH, NADP, NADPH The pyridine nucleotide coenzymes containing niacin that are involved in many essential hydrogen transfer reactions of intermediary metabolism, namely nicotinamide adenine dinucleotide (oxidized and reduced forms) and nicotinamide adenine dinucleotide phosphate (oxidized and reduced forms).

NMN N^1-methyl nicotinamide, a degradation product of niacin found in the urine.
2-pyridone, N^1-methyl-2-pyridone-4-carboxamide, another degradation product of niacin, is also found in the urine.

RDA Recommended Dietary Amount of a named nutrient, being the daily amount needed to cover the needs of the majority, usually 97.5% (i.e., mean plus 2 standard deviations) of the members of a healthy population of each defined age group and sex in the USA. RNI (Reference Nutrient Intake) is the UK equivalent of the RDA. (EAR: Estimated Average Requirement; LRNI: Lower Reference Nutrient Intake (UK only), which is the amount need to cover only that 2.5% of the population with the lowest requirements for the named nutrient).

Pellagra A human disease that is often equated with a clinical niacin deficiency but may have more complex causes.

Hartnup's disease A genetic disease of humans that exhibits some features similar to those of pellagra but that results from an impaired membrane transport of tryptophan.

Blacktongue Induced niacin deficiency disease in dogs, an animal model for human pellagra.

Units Niacin or niacin equivalents in food are usually expressed as milligrams per 100 g, usually of wet weight of food. Concentrations of analytes in tissues, blood, or urine are usually expressed in SI units (e.g., μmol/l or $\mu mol\,l^{-1}$).

See also: **Cereal Grains**. **Drug–Nutrient Interactions**. **Niacin**. **Riboflavin**. **Thiamin**: Physiology. **Vitamin B_6**.

Further Reading

Anonymous (1987) Pellagra treated with tryptophan. *Nutrition Reviews* **45**: 142–151.

Bender DA (1992) Niacin. In *Nutritional Biochemistry of the Vitamins*, pp. 184–222. Cambridge: CUP.

Carpenter KJ (ed.) (1981) *Pellagra: Benchmark papers in History of Biochemistry*, vol. 2. Stroudsberg, PA: Dowden, Hutchinson & Ross.

Carpenter KJ and Lewin WJ (1985) A reexamination of the composition of diets associated with pellagra. *American Journal of Clinical Nutrition* **115**: 543–552.

Combs GF Jr (1998) Niacin. In *The Vitamins. Fundamental Aspects in Nutrition and Health*, 2nd edn., pp. 312–331. New York: Academic Press.

Cook NE and Carpenter KJ (1987) Leucine excess and niacin status in rats. *Journal of Nutrition* **117**: 519–526.

Fu CS, Swendseid ME, Jacob RA, and McKee RW (1989) Biochemical markers for assessment of niacin status in young men: Levels of erythrocyte niacin coenzymes and plasma tryptophan. *Journal of Nutrition* **119**: 1949–1955.

Henderson LM (1983) Niacin. *Annual Review of Biochemistry* **3**: 289–307.

Horwitt MK, Harvey CC, Rothwell WS, Cutler JL, and Haffron D (1956) Tryptophan–niacin relationship in man. *Journal of Nutrition* **60**(supplement 1): 1–43.

Jacob RA (2001) Niacin. In *Present Knowledge in Nutrition*, 8th edn., pp. 199–206. Washington, DC: ILSI Press.

Multiauthor symposium (1981) Pellagra. *Federation Proceedings* **40**: 1519–1537.

Sauberlich HE (1999) Niacin (nicotinic acid, nicotinamide). In *Laboratory Tests for the Assessment of Nutritional Status*, 2nd edn., pp. 161–174. Boca Raton, FL: CRC Press.

Van Eys J (1991) Nicotinic acid. In: Machlin LJ (ed.) *Handbook of Vitamins*, vol. 2, pp. 311–340. New York: Marcel Dekker.

Pesticides *see* **Food Safety**: Pesticides

Phenylketonuria *see* **Inborn Errors of Metabolism**: Nutritional Management of Phenylketonuria

Phosphate *see* **Small Intestine**: Structure and Function

PHOSPHORUS

J J B Anderson, University of North Carolina, Chapel Hill, NC, USA

The consumption of a diet sufficient in phosphorus, in the form of phosphate salts or organophosphate molecules, is critical for the support of human metabolic functions. Too much phosphorus, in relation to too little dietary calcium, may contribute to bone loss, and too little phosphorus along with too little dietary calcium may not adequately maintain bone mass, especially in the elderly. Therefore, under normal dietary conditions, dietary phosphorus is used for numerous functions without any concern; it is only when too much or too little phosphorus is ingested that skeletal problems may arise. Certainly, elderly subjects need to consume sufficient amounts of phosphorus, like calcium, to maintain bone mass and density, but too much phosphorus may contribute to inappropriate elevations of parathyroid hormone (PTH) and bone loss. It is not clear where most elderly subjects fall along this continuum of intake patterns. This article discusses the mechanisms by which phosphate ions impact on calcium and also on bone tissue.

Calcium–Phosphate Interrelationships

Although phosphorus in the form of phosphate ions is essential for numerous body functions, its metabolism is intricately linked to that of calcium because of the actions of calcium-regulating hormones, such as PTH and 1,25-dihydroxyvitamin D, on bone, the gut, and the kidneys. Adequate phosphorus and calcium intakes are needed not only for skeletal growth and maintenance but also for many cellular roles, such as energy production (i.e., adenosine triphosphate (ATP)). Phosphate ions are incorporated in many organic molecules, including phospholipids, creatine phosphate, nucleotides, nucleic acids, and ATP.

Dietary Sources of Phosphorus

Animal products, including meats, fish, poultry, eggs, milk, cheese, and yogurt, are especially rich in phosphorus, as phosphates, but good amounts of phosphorus can be obtained from cereal grains and many vegetables, including legumes. Because of the abundance of phosphorus in the food supply, deficiency is highly unlikely except perhaps late in life when some elderly individuals consume little food. An extremely rare deficiency disease, phosphate rickets, in infants has been reported to result from inadequate phosphorus intake.

In the United States, mean phosphorus intakes approximate 1200–1500 mg per day in adult males and 900–1200 in adult females. In addition, phosphate additives used in food processing and cola beverages are also consumed, but the quantities are not required by food labeling laws to be given on the label so that the actual additional amounts consumed can only be estimated. Phosphate additives used by the food industry may be found in baked goods, meats, cheeses, and other dairy products.

Table 1 Calcium and phosphorus composition of common foods

Food category	*Phosphorus mg/serving*	*Calcium mg/serving*	*Ca:P ratio (wt:wt)*
Milk, eggs, and dairy			
Cheddar cheese, 1 oz.	145	204	1.4
Mozzarella cheese-part skim, 1 oz.	131	183	1.4
Vanilla ice milk, 1 cup	161	218	1.4
Lowfat yogurt, 1 cup	353	448	1.3
Skim milk, 8 oz.	247	301	1.2
Skim milk-Lactose reduced, 8 oz.	247	302	1.2
Vanilla ice cream, 1 cup	139	169	1.2
Vanilla soft-serve ice cream, 1 cup	199	225	1.1
Egg substitute, frozen, 1/4 cup	43	44	1.1
Chocolate pudding, 5 oz.	114	128	1.1
Processed American cheese, 1 oz.	211	175	0.8
Lowfat cottage cheese, 1 cup	300	200	0.7
Processed cheese spread, 1 oz.	257	129	0.5
Instant chocolate pudding, 5 oz.	340	147	0.4
Soy milk, 8 oz.	120	10	0.1

A conservative estimate is that most adults in the United States consume an extra 200–350 mg of phosphorus each day from these sources and cola beverages. Therefore, the total phosphorus intakes for men and women are increased accordingly. Because the typical daily calcium intake of males is 600–800 mg and that of females is 500–650 mg, the Ca:P ratios decrease from approximately 0.5–0.6 to less than 0.5 when the additive phosphates are included. As shown later, a chronically low Ca:P dietary ratio may contribute to a modest nutritional secondary hyperparathyroidism, which is considered less important in humans than in cats. **Table 1** provides representative values of calcium and phosphorus in selected foods and the calculated Ca:P ratios. Only dairy foods (except eggs), a few fruits, and a few vegetables have Ca:P ratios that exceed 1.0.

Recommended intakes of phosphorus have been set for adults in the United States at 900 mg per day for men and 700 mg per day for women.

Intestinal Absorption of Phosphates

Because phosphate ions are readily absorbed by the small intestine (i.e., at efficiencies of 65–75% in adults and even higher in children), a prompt increase in serum inorganic phosphate (Pi) concentration follows within an hour after ingestion of a meal begins. (Calcium ions or Ca^{2+} are much more slowly absorbed.) The increased serum Pi ($HPO_4^{=}$) concentration then depresses the serum calcium ion concentration, which in turn stimulates the parathyroid glands to synthesize and secrete PTH. PTH acts on bone and the kidneys to correct the modest decline in Ca^{2+} and homeostatically return it to the set level. Reports suggest that an elevation of serum Pi ionic concentration directly influences PTH secretion independently of hypocalcemia. These meal-associated fluctuations in Pi and Ca^{2+} are part of normal physiological adjustments that occur typically three or more times a day.

Pi ions are thought to be absorbed primarily by transcellular mechanisms that involve cotransport with cations, especially sodium (Na^{+}). These rapid mechanisms account for the uptake of Pi ions in blood within 1 h after ingestion of a meal. The blood concentration of Pi is less tightly regulated than the serum calcium concentration. Wider fluctuations in serum Pi concentrations reflect both dietary intakes and cellular releases of inorganic phosphates.

Most Pi absorption by the small intestine occurs independently of the hormonal form of vitamin D. The reported role of 1,25-dihydroxyvitamin D in intestinal Pi transcellular absorption is somewhat unclear because of the normally rapid influx of Pi ions after a meal, but this hormone may enhance the late or slower uptake of Pi ions. Paracellular passive absorption of Pi ions may also occur, but the evidence for this is limited.

Phosphate Homeostatic Mechanisms

The blood concentrations of Pi ions are higher early in life and then decline gradually until late life. The normal range for adults is 2.7–4.5 mg/dl (0.87–1.45 mmol/l). The percentage distributions of the blood fractions of phosphorus compared to those

Table 2 Approximate percentage (%) distributions of calcium and phosphate in blood

Serum fraction	*Calcium*	*Phosphate*
Ionic	50–55%	55–60%
Protein-Bound	45–50	10–13
Complexed	0.3–0.6	30–35

of calcium are given in **Table 2**. The homeostatic control of this narrow concentration range of Pi is maintained by several hormones, including PTH, $1,25(OH)_2$ vitamin D, calcitonin, insulin, glucagon, and others, but the control is never as rigorous as that of serum calcium. In contrast to calcium balance, which is primarily regulated in the small intestine by $1,25(OH)_2$ vitamin D, Pi balance is mainly regulated by the phosphaturic effect of PTH on the kidney, primarily the proximal convoluted tubule. In this sense, Pi regulation is less critical than that of calcium, which may result from the presence of multiple stores of this ion distributed throughout the body (i.e., bone, blood, and intracellular compartments).

A major regulator of Pi is PTH, whose role has been fairly well uncovered. PTH increases bone resorption of Pi (and calcium ions), it blocks renal tubular Pi reabsorption following glomerular filtration (whereas PTH favors calcium reabsorption), and it enhances intestinal Pi absorption (and calcium absorption) via the vitamin D hormone, $1,25(OH)_2$ vitamin D. Other hormones have more modest effects on serum Pi concentration.

Functional Roles of Phosphates

Several major roles of Pi ions have been briefly noted (i.e., intracellular phosphate groups for cellular energetics and biochemical molecules as well as for the skeleton and teeth (structures)). Other important functions also exist. For example, in bone tissue phosphates are critical components of hydroxyapatite crystals, and they are also considered triggers for mineralization after phosphorylation of type 1 collagen in forming bone. Serum phosphates, $HPO_4^{=}$ and $H_2PO_4^{-}$, also provide buffering capacity that helps regulate blood pH and also cellular pH.

Considerable cellular regulation occurs through the phosphorylation or dephosphorylation of Pi ions under the control of phosphatase enzymes, including protein kinases. These cell regulatory roles of Pi ions coexist with regulatory functions involving calcium ions, but Pi ions are much more widely distributed within cells and cell organelles than Ca ions.

Insulin affects Pi ions by increasing their intracellular uptake, although temporarily, for the prompt phosphorylation of glucose. Insulin may also influence the use of Pi ions when insulin-like growth factor-1 acts to increase tissue growth or other functions. Because of the broad uses of Pi ions in structural components, energetics, nucleic acids, cell regulation, and buffering, there is an overall generalization that these versatile yet critical ions support life.

Phosphate in Health and Disease

Phosphate balance in adults is almost always zero, in contrast to calcium balance, which is usually negative, because of the effective action of PTH on renal tubules to block Pi reabsorption. In late life, however, intestinal phosphate absorption decreases and the serum phosphate concentration declines. These physiological decrements may contribute to disease, especially to increased bone loss and osteopenia or more severe osteoporosis. Typically, these changes in Pi balance are also accompanied by similar changes in calcium balance. Too little dietary phosphorus and too little dietary calcium may be determinants of low bone mass and density and, hence, increased bone fragility. The usual scenario invoked to explain osteoporosis in old age, however, is that too little dietary calcium in the presence of adequate dietary phosphorus stimulates PTH release and bone loss (**Figure 1**).

Three human conditions that involve abnormal Pi homeostasis need explanation.

Aging and Renal Function

The serum concentration of Pi increases with a physiological decline in renal function associated with aging (but not renal disease per se). Healthy individuals excrete approximately 67% of their absorbed phosphate via the urine and the remainder via the gut as endogenous secretions. As the glomerular filtration capacity of the kidneys declines, the serum Pi concentration increases and more Pi is retained by the body. PTH secretions increase but the typical serum PTH concentrations, although elevated, remain within the upper limits of the normal range, at least for a decade or so. Thereafter, however, serum Pi and PTH both continue to climb as renal function declines and increased rates of bone turnover lead to measurable bone loss. This situation probably affects millions in the United States each year as they enter the 50s and proceed into the 60s; many of these individuals are overweight or obese and have the metabolic syndrome, which

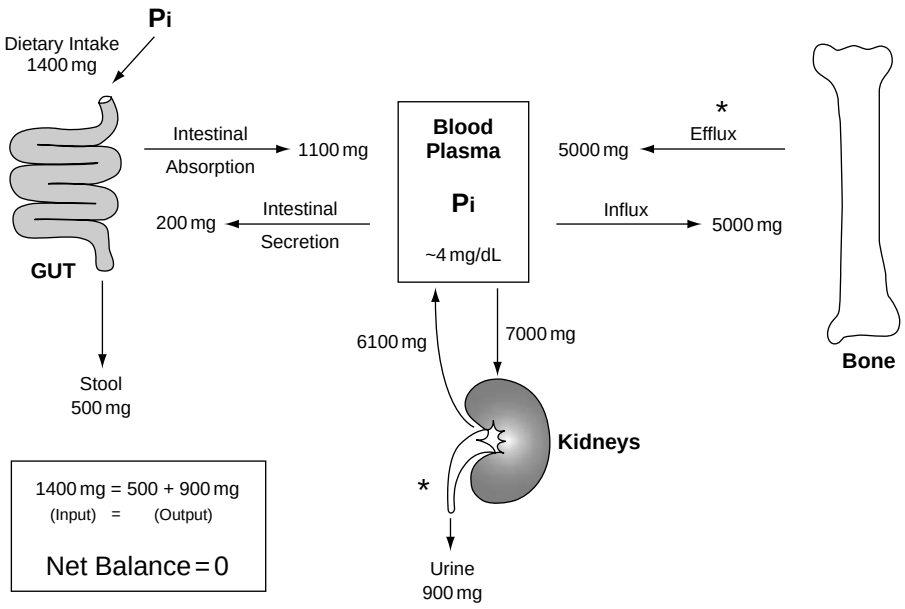


Figure 1 Phosphorus homeostasis and balance. The intestine, kidneys, and bone are organs involved in phosphate homeostasis. Fluxes of phosphate ions between blood and these organs are shown. Note the high fluxes in and out of bone each day. To convert phosphorus values from g to mmol, multiply by 32.29; from mg/dl to mmol/l, multiply by 0.3229. *Steps enhanced by parathyroid hormone. (Adapted with permission from Anderson JJB, Sell ML, Garner SC, and Calvo MS (2001) Phosphorus. In: Bowman BA and Russell R (eds.) *Present Knowledge in Nutrition*, 8th edn, p. 282. Washington, DC: International Life Sciences Institute Press.)

may negatively impact renal function. As the syndrome worsens, many of these individuals will progress to chronic renal failure and renal secondary hyperparathyroidism.

Nutritional Secondary Hyperparathyroidism

This mild condition has not been fully assessed in any longitudinal studies lasting as long as 1 year. The initiating event is a chronic low-calcium and high-phosphorus intake (low Ca:high P ratio) that leads to a chronic elevation of serum PTH. Elevations in PTH stimulate osteoclastic bone resorption and declines in bone mass and density. This condition has only been studied experimentally using human subjects for 28 days, but the chronic increases in PTH and vitamin D hormone suggest that even a lowering of the Ca:Pi ratio below 0.5—in this study to ~0.25—resulted in adverse effects. Longer term studies are needed to determine if bone losses occur under this chronic dietary regimen.

Renal Secondary Hyperparathyroidism

The true secondary hyperparathyroidism of chronic renal failure (CRF) has been extremely difficult to treat by clinicians because of high Pi and PTH concentrations in this condition. Traditional treatment includes the use of binders (chemical) to prevent Pi absorption from the small intestine. In recent years, a calcium-sensing receptor (CaR) in the parathyroid glands has been identified and drugs are being developed that will trick the CaR into thinking that serum calcium is normal rather than depressed, thereby reducing PTH secretion. A reduction in PTH then helps in the conservation of bone tissue since bone loss is such a severe problem in CRF patients.

Conclusions

The general view of dietary phosphorus, supplied in foods as phosphates, is that too much relative to calcium skews the Ca:P ratio to much less than 0.5. Another view, however, has been emerging that suggests that many elderly subjects, especially women, have very low phosphorus intakes in addition to low calcium intakes and that they may benefit from increased consumption of both calcium and phosphate from foods and supplements. In dietary trials designed to reduce fractures of elderly women and men, especially nonvertebral fractures, calcium plus vitamin D has been the treatment, but at least one trial that used calcium phosphate plus vitamin D has shown significant reduction in fractures over 18 and 36 months of follow-up. Further studies are needed to target the role of phosphate ions in reducing fractures among the elderly.

See also: **Aging**. **Bone**. **Calcium**.

Further Reading

Anderson JJB, Sell ML, Garner SC, and Calvo MS (2001) Phosphorus. In: Bowman BA and Russell R (eds.) *Present Knowledge in Nutrition*, 8th edn. Washington, DC: International Life Sciences Institute Press.

Baker SS, Cochran WJ, Flores CA *et al.* (1999) American Pediatrics Committee on Nutrition. Calcium requirements of infants, children, and adolescents. *Pediatrics* ***104: 1152–1157.***

Brot C, Jorgensen N, Jensen LB, and Sorensen OH (1999) Relationships between bone mineral density, serum vitamin D metabolites and calcium:phosphorus intake in healthy perimenopausal women. *Journal of Internal Medicine* **245**: 509–516.

Calvo MS, Kumar R, and Heath HH III (1990) Persistently elevated parathyroid hormone secretion and action in young women after four weeks of ingesting high phosphorus, low calcium diets. *Journal of Clinical Endocrinology and Metabolism* **70**: 1340–1344.

Calvo MS and Park YM (1996) Changing phosphorus content of the US diet: Potential for adverse effects on bone. *Journal of Nutrition* **126**: 1168S–1180S.

Chapuy MC, Arlot ME, Duboeuf F *et al.* (1992) Vitamin D_3 and calcium to prevent hip fractures in elderly women. *New England Journal of Medicine* **327**: 1637–1642.

Garner SC (1996) Parathyroid hormone. In: Anderson JJB and Garner SC (eds.) *Calcium and Phosphorus in Health and Disease*, pp. 157–175. Boca Raton, FL: CRC Press.

Goulding A, Cannan R, Williams SM *et al.* (1998) Bone mineral density in girls with forearm fractures. *Journal of Bone and Mineral Research* **13**: 1143–1148.

Harnack L, Stang J, and Story M (1999) Soft drink consumption among US children and adolescents: Nutritional consequences. *Journal of the American Dietetic Association* **99**: 436–441.

Institute of Medicine, Food and Nutrition Board (1997) *Dietary Reference Intakes: Calcium, Phosphorus, Magnesium, Vitamin D and Fluoride*. Washington, DC: National Academy Press.

Khosla S, Melton LJ III, Dekutoski MB *et al.* (2003) Incidence of childhood distal forearm fractures over 30 years: A population-based study. *Journal of the American Medical Association* **290**: 1479–1485.

Ritter CS, Martin DR, Lu Y *et al.* (2002) Reversal of secondary hyperparathyroidism by phosphate restriction restores parathyroid calcium-sensing receptor expression and function. *Journal of Bone and Mineral Research* **17**: 2206–2213.

Shea B, Wells G, Cranney A *et al.* (2002) VII. Meta-analysis of calcium supplementation for the prevention of postmenopausal osteoporosis. *Endocrine Reviews* **23**: 552–559.

Slatopolsky E, Dusso A, and Brown A (1999) The role of phosphorus in the development of secondary hyperparathyroidism and parathyroid cell proliferation in chronic renal failure. *American Journal of Medical Sciences* **317**: 370–376.

Uribarri J and Calvo MS (2003) Hidden sources of phosphorus in the typical American diet: Does it matter in nephrology? *Seminars in Dialysis* **16**: 186–188.

Vinther-Paulsen N (1953) Calcium and phosphorus intake in senile osteoporosis. *Geriatrics* **9**: 76–79.

Wyshak G (2000) Teenaged girls, carbonated beverage consumption, and bone fractures. *Archives of Pediatric and Adolescent Medicine* **154**: 610–613.

Wyshak G and Frisch RE (1994) Carbonated beverages, dietary calcium, the dietary calcium/phosphate ratio, and bone fractures in girls and boys. *Journal of Adolescent Health* **15**: 210–215.

Physical Activity *see* **Exercise**: Beneficial Effects; Diet and Exercise

PHYTOCHEMICALS

Contents

Classification and Occurrence

A Cassidy, School of Medicine, University of East Anglia, Norwich, UK

There is a considerable body of evidence to suggest that populations that consume diets rich in fruits and vegetables, whole-grain cereals, and complex carbohydrates have a reduced risk of a range of chronic diseases. This has led to the suggestion that the diversity of substances found in food, particularly plant-derived or plant-based foods, may underlie the protective effects that are attributed to diets high in fruits and vegetables and other plant foods. Although fruits and vegetables are rich sources of micronutrients and dietary fiber, they

also contain a wide variety of secondary metabolites, which provide the plant with color, flavor, and antimicrobial and insecticide properties. Many of these substances have been attributed a wide array of properties but have yet to be recognized as nutrients in the conventional sense. Many of these potentially protective plant compounds, termed phytochemicals, are receiving increasing attention. Phytochemicals, also known as phytonutrients, are plant-based compounds that exert numerous physiological functions in mammalian systems. Many of them are ubiquitous throughout the plant and as a result are present in our daily diet. Among the most important classes are the flavonoids, which are classified based on their chemical and structural characteristics. This article focuses on the different classes of phytochemicals and their relationships to human diseases.

Phytochemicals: General

Plants synthesize a wide array of compounds that play key roles in protecting plants against herbivores and microbial infection and as attractants for pollinators and seed-dispersing animals, allelopathic agents, UV protectants, and signal molecules in the formation of nitrogen-fixing root nodules in legumes. Although they have long been ignored from a nutritional perspective, the function of these compounds and their relative importance to human health are gaining significant interest.

Phytochemicals comprise a wide group of structurally diverse plant compounds, which are predominantly associated with the cell wall and widely dispersed throughout the plant kingdom. They are secondary plant metabolites, characterized by having at least one aromatic ring with one or more hydroxyl groups attached. The nature and distribution of these compounds can vary depending on the plant tissue, but they are mainly synthesized from carbohydrates via the shikimate and phenylpropanoid pathways. They range in chemical complexity from simple phenolic acids, such as caffeic acid, to complex high-molecular-weight compounds, such as the tannins, and they can be classified according to the number and arrangement of their carbon atoms. In plants, they are commonly found conjugated to sugars and organic acids and can be classified into two groups, flavonoids and nonflavonoids. The most researched group of compounds to date is the flavonoids, and this article focuses on this group.

Flavonoids

Flavonoids constitute a large class of phytochemicals that are widely distributed in the plant kingdom, are present in high concentrations in the epidermis of leaves and skin of fruits, and have important and varied roles as secondary metabolites. More than 8000 varieties of flavonoids have been identified, many of which are responsible for the colors of fruits and flower. They are found in fruits, vegetables, tea, wine, grains, roots, stems, and flowers and are thus regularly consumed by humans. Although it has been widely known for centuries that derivatives of plant origin possess a broad spectrum of biological activities, it was first suggested that flavonoids may be important for human health in the 1930s when it was observed that a fraction from lemon juice could decrease the permeability of arteries and partially prevent symptoms in scorbutic pigs. At the time, it was suggested that these compounds should be defined as a new class of vitamins, vitamin P, and the substance responsible for the effects was identified as the flavonoid rutin. However, the data were not generally accepted and the term vitamin P was abandoned in the 1950s. There was renewed interest in flavonoids when a potentially protective role for flavonoids in relation to heart disease in humans was reported. Since that time, there has been a surge of interest in the potential role of flavonoids in human health, with research suggesting antioxidant effects, hormonal actions, antiinfectious actions, cancer-preventative effects, the ability to induce chemical defense enzymes, and actions on blood clotting and the vascular system. However, concrete evidence that they positively influence human health is lacking, and adverse effects have also been reported for some polyphenols. The main subclasses of flavonoids are flavones, flavonols, flavan-3-ols, isoflavones, flavanones, and anthocyanidins (**Figure 1** and **Table 1**).

Other flavonoid groups that are thought to be less important from a dietary perspective are the dihydroflavones, flavan-3,4-diols, coumarins, chalcones, dihydrochalcones, and aurones. The basic flavonoid skeleton can have numerous constituents; hydroxyl groups are usually present at the 4-, 5-, and 7- positions. Sugars are very common, and the majority of flavonoids exist naturally as glycosides. The presence of both sugars and hydroxyl groups increases water solubility, but other constituents, such as methyl or isopentyl groups, render flavonoids lipophilic.

Although many thousands of different flavonoids exist, they can be classified into different subclasses. The main subclasses that are important from a human health perspective are the flavones,

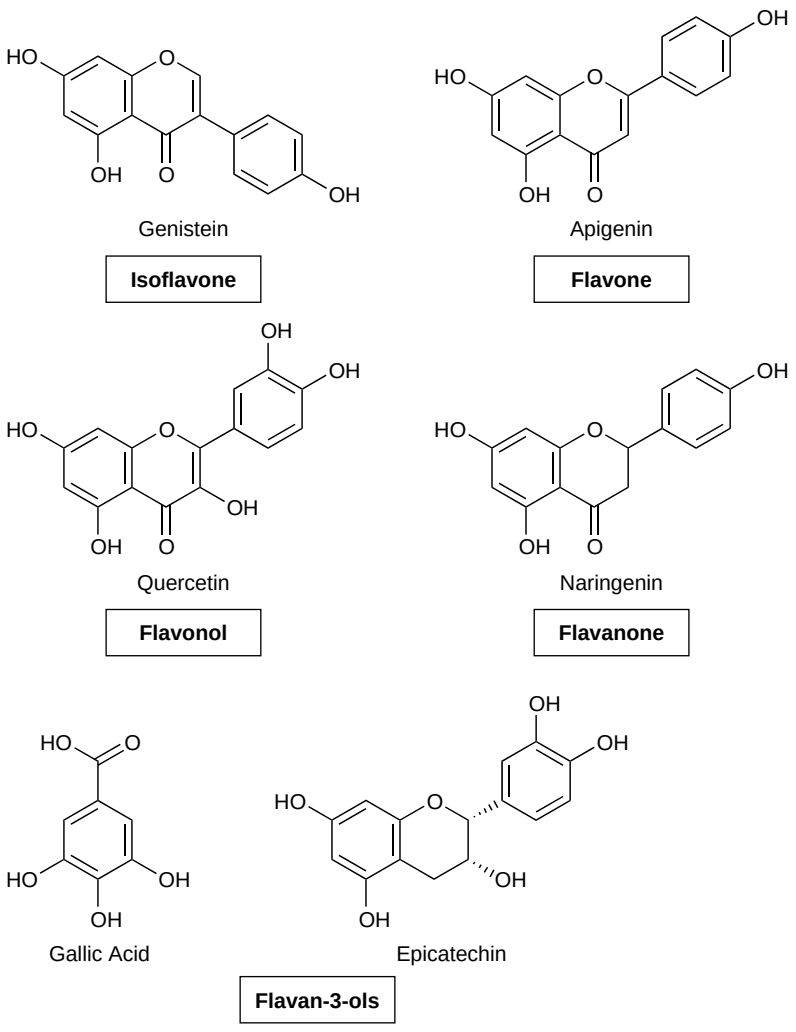


Figure 1 Structures of the major subclasses of flavonoids.

flavonols, flavan-3-ols, isoflavones, flavanones, and anthocyanidins (**Figure 1**).

Flavonols

These are arguably the most widespread of the flavonoids because they are dispersed throughout the plant kingdom. The distribution and structural variations of flavonols are extensive and have been well documented. Extensive information on the different flavonols present in commonly consumed fruits, vegetables, and drinks is available; however, there is wide variability in the levels present in specific foods, in part due to seasonal changes and varietal differences. The most common flavonols are kaempferol, quercetin, isorhamnetin, and myricetin.

Flavones

Flavones have a close structural relationship to the flavonols, but unlike flavonols they are not widely distributed in plants. The only significant occurrences in plants are in celery, parsley, and a few other herbs, and they predominantly occur as 7-*O*-glycosides (e.g., luteolin and apigenin). In addition, polymethoxylated flavones have been found in citrus fruits (e.g., nobiletin and tangeretin).

Flavan-3-ols

Flavan-3-ols, often referred to as flavanols, are the most complex class of the flavonoids because they range from simple monomers (catechin and its isomer epicatechin) to the oligomeric and polymeric proanthocyanidins, which are also known as condensed tannins. Proanthocyanidins can occur as polymers of up to 50 units, and when hydroxylated they can form gallocatechins or undergo esterification to form gallic acid. Red wine contains oligometric proanthocyanidins derived mainly from the seeds of black grapes. Green tea is also a rich source of

Table 1 Principal dietary sources of flavonoids

Flavonoid	*Compound*	*Food source*
Flavonol	Quercetin, kempferol, myricetin	Onion, apple, broccoli, tea, olives, kale, cranberry, lettuce, beans (green, yellow)
Flavone	Luteolin, apigenin	Olives, celery
Flavan-3-ol	Catechin, epicatechin	Tea, red wine, apple
Flavanone	Naringenin, hesperidin	Citrus fruit
Anthocyanidins	Cyanidin, delphinidin, malvidin, petunidin	Grapes, cherries
Chalcones, dihydrochalcones		Heavily hopped beer, tomatoes (with skins), cider, apple juice
Isoflavone	Genistein, daidzein	Soy

Information from Hollman PC, Katan MB (1997) Absorption, metabolism and health effects of dietary flavonoids in man. Biomed Pharmacother **51**(8): 305–10.
Scalbert A, Williamson G (2000) Dietary intake and bioavailability of Polyphenols. *J Nutr* **130** (8S Suppl): 2073S–85S.

flavan-3-ols, principally epigallocatechin, epigallocatechin gallate, and epicatechin gallate. However, during fermentation of tea leaves the levels of catechins decline and thus the main components of black tea are high-molecular-weight thearubigins, whose structures are derived from flavonoids but are unknown. The catechins are widespread, but the main sources in the diet come from tea, wine, and chocolate.

Anthocyanins

Anthocyanins are widespread in nature, predominantly in fruits and flower tissues, in which they are responsible for the red, blue, and purple colors. They are also found in leaves, stems, seeds, and root tissue. In plants, they protect against excessive light by shading leaf mesophyll cells. Additionally, they play an important role in attracting pollinating insects. The most common anthocyanins are pelargonidin, cyanidin, delphinidin, peonidin, petunidin, and malvidin, which are predominantly present in plants as sugar conjugates.

Flavanones

The flavanones are the first flavonoid products of the flavonoid biosynthetic pathway. They are characterized by the presence of a chiral center at C2 and the absence of the C2–C3 bond. The flavanone structure is highly reactive, and they have been reported to undergo hydroxylation, glycosylation, and O-methylation reactions. Flavanones are present in high levels in citrus fruits, with the most common glycoside known as hesperidin (hesperetin-7-*o*-rutinoside), which is present in citrus peel. Interestingly, flavanone rutinosides are tasteless, whereas the flavanone neohesperidoside conjugates (e.g., neohesperidin) from bitter orange and naringenin (naringenin-7-*o*-neohesperidoside) from grapefruit peel have an intensely bitter taste.

Isoflavones

Isoflavones are flavonoids, but they are also called phytoestrogens because of their oestrogenic activity. Structurally, they exhibit a similarity to mammalian oestrogens and bind to oestrogen receptors α and β. Apart from basic structural similarities, the key to their estrogenic effect is the presence of the hydroxyl groups on the A and B rings. They are classified as oestrogen agonists but also as oestrogen antagonists since they compete with oestrogen for their receptor. They have also been demonstrated to exert effects that are independent of the oestrogen receptor.

Current Estimates of Intake

Diets rich in plant-derived foods can provide more than 1 g of phenolic compounds per day, although there are major international and interindividual differences in exposure. Flavonols, flavones, and flavan-3-ols constitute the three major subclasses of flavonoids, and a significant amount of information on the content of selected flavonoids from these subclasses in fruits and vegetables has been obtained using high-performance liquid chromatography techniques. The other subclasses are flavanones, anthocyanidins, and isoflavones.

Given the differences in dietary intake, particularly for fruits and vegetables, between populations, it is not surprising that the relationships between the predominant flavonoids and their sources will vary between populations, nor is it unexpected that there will be wide inter- and intraindividual variations in intake of the individual subclasses of the flavonoids. Flavonol intake was estimated to be highest in a Japanese population group (64 mg/day) and lowest in Finland (6 mg/day). International comparisons of dietary sources also reflect this variation, but only a few sources of flavonoids are responsible for most of the intake. Red wine was the main source of the flavonol quercetin in Italy, tea was the main source in Japan and The Netherlands, and onions were the most significant contributor to intake in Greece, the United States, and the former Yugoslavia (**Table 1**).

Table 2 Estimated dietary intake of flavonoid subclasses in different countries

Flavonoid subgroup	*Estimated intake (mg/day)*			
	Denmark	*Holland*	*Finland*	*Japan*
Flavonol	1.5–8.6	1–17	1.1–7	16.4
Flavone	1–2	2	No data	0.3
Flavan-3-ol	45	50	8.3	40
Flavanone	7.1–9.3	No data	8.3–28.3	No data
Isoflavone	<1	No data	No data	50
Anthocyanidins	6–60	No data	No data	No data

The estimated daily intake of flavonoids, including catechins and anthocyanins, is >50 mg for all the countries presented in **Table 2**, and realistically intake is probably higher than 100 mg/day if data on all flavonoid groups were available. If this intake is compare to daily intakes of other dietary antioxidants, such as vitamin C (80 mg/day), vitamin E (8.5 mg/day), and β-carotene (1.9 mg/day), it is clear that flavonoid intakes exceed or are at least comparable to those of other established antioxidants, indicating that these compounds constitute an important part of dietary intake of antioxidants.

Absorption and Metabolism of Flavonoids

The flavonols and flavones are generally present in plants in the form of glycosides and as such are water-soluble. Thus, some of the flavonol glycosides may be absorbed intact in the small intestine or hydrolyzed by mucosal enzymes and absorbed as aglycones. However, those that pass through the small intestine unabsorbed or reenter the gut from the bile become available for bacterial metabolism in the colon.

The colon contains numerous microorganisms and as a result it has significant capacity for catalytic and hydrolytic reactions. These colonic bacteria produce enzymes that are capable of stripping flavonoid conjugates of their sugar moieties, enabling free aglycones to be absorbed. The enzymes produced by colonic bacteria can also break down the flavonoids into simple compounds, resulting in the production of a range of derivatives, some of which may be more biologically active than the parent compound. This is an important area for future research because the metabolism of flavonoids is influenced by intestinal microflora and these metabolic reactions may result in deactivation of bioactive compounds or activation of previously inactive compounds. It is therefore critical to identify the bacteria involved in these transformation reactions and define their relative importance and occurrence in the human gut to gain a better understanding of the transformation processes.

Other key body compartments that are important in defining the metabolism of flavonoids are the liver and, to a lesser extent, the small intestine and kidney, in which the biotransformation enzymes are located. Flavonols and flavan-3-ols are primarily metabolized in the colon and liver.

The evidence for absorption of intact flavonoid glycosides is weak. Recent data showing β-glucosidase activity in the small intestine, together with the absence of intact glycosides in plasma and urine, strongly suggest that only free flavonoid aglycones are being absorbed. In addition, data also indicate that there is a more rapid and efficient absorption of flavonoids originating from glucosides than from other glycosides or free aglycones. This suggests that dietary sources containing high levels of glucose-bound flavonoids are more likely to have potential health benefits than foods containing other flavonoid glycosides.

Bioavailability of Flavonoids

Critical to a food's 'nutritional' value is whether the 'nutrient' or compound is provided in a bioavailable form from the food. Flavonoids therefore may have to be absorbed from the large intestine if they are to exert a potential health effect. Early data from animal studies suggested that flavonoids were only absorbed to a limited degree because gut microflora preferentially destroyed the heterocyclic rings of the compounds before absorption occurred in the small intestine. However, an increasing number of studies suggest that the bioavailability of flavonoids is greater than was previously recognized, although increases in the concentrations of flavonoids and its associated metabolites in plasma and urine do not necessarily mean that they have significant effects *in vivo*. There are few data on their intracellular location and mechanism of action. Thus, a key area for future research will be to clarify the absorption, bioavailability, and metabolism of a range of flavonoid compounds.

Potential Mechanisms of Action

The effect of flavonoids on enzymatic, biological, and physiological processes has been extensively studied, but few studies have attempted to determine the actual compound or metabolite responsible for the observed effects. Much of the *in vitro* data assume that the biological activity originates

from the flavonoid ingested, without taking into consideration the biotransformations that may occur following ingestion and metabolism, as it is well established that following ingestion they are transformed into a range of structurally distinct compounds.

In interpreting the mechanistic data, it is also important to remember that little attention has been paid to the physiological relevance of the concentration used in the *in vitro* model systems. Thus, in some instances biological effects have been shown at concentrations that are unachievable *in vivo*; therefore, the biological relevance of these mechanisms to humans is questionable.

Since flavonoids are complex groups of compounds with variable structures and activities, it is unlikely that they exert their biological effects by common mechanisms. However, since it is also now established that the pathophysiological processes leading to the development of cardiovascular disease and cancer are complex, this means that there are many potential sites and stages at which bioactive plant compounds present in food could act to potentially reduce the formation of cancerous cells or the atherosclerotic plaque in cardiovascular disease. Elucidating the underlying mechanisms of how flavonoids work is a key aim for nutrition research.

In vitro experimental systems suggest that flavonoids can scavenge oxygen-derived free radicals; exert antiinflammatory, antiallergic, and antiviral effects; and have anticarcinogenic properties.

Potential Health Effects

There is substantial epidemiological evidence that populations that consume diets rich in plant foods have a reduced risk of cardiovascular disease and various cancers, and the potential role of bioactive compounds in plants in this association is gaining significant attention within nutrition research. Identification of the role of flavonoids in the primary mechanisms that may protect against cellular damage may yield clues to slowing aspects of the aging process and postpone age-related diseases.

Most research on flavonoids and health has focused on quercetin due to its antioxidant potency and potential role in cardiovascular disease. However, the diverse and broad nature of flavonoids means that subclasses other than the flavonols may be more important to human health since they appear to be more bioavailable and thus have a greater potential to protect against the various mechanisms involved in aging and disease development.

Cardiovascular Health

The stimulus for much of the research on the role of flavonoids in human health was derived from epidemiological studies, particularly a study suggesting that dietary flavonoids may protect against cardiovascular disease. During the past decade, a significant amount of research has examined the effect of flavonoids in foods and pure flavonoid compounds at various stages in the atherosclerosis process.

A significant proportion of the research on flavonoids has concentrated on their antioxidant actions, and their capacity to act as antioxidants remains their best described biological property to date. Their antioxidant ability is well established *in vitro*, and *in vivo* animal data also suggest that consumption of compounds such as rutin or red wine extracts, tea, or fruit juice lowers oxidative products such as protein carbonyls, DNA damage markers, and malonaldehyde levels in blood and a range of tissues.

The flavones and catechins appear to be the most powerful flavonoids at protecting the body against reactive oxygen species. Although the mechanisms and sequence of events by which free radicals interfere with cellular functions are not fully understood, one of the most important events may be lipid peroxidation, which results in cellular damage. Flavonoids may prevent such cellular damage by several different mechanisms, including direct scavenging of free radicals such as superoxides and peroxynitrite, inhibition of nitric oxide, or antiinflammatory effects.

Cancer

The specific mechanisms by which individual dietary components can alter the cancer process remain poorly understood. However, mechanisms underlying the carcinogenesis process are understood sufficiently so that model systems to evaluate the ability of a specific compound to inhibit or promote processes that may prevent or delay cancer development can be predicted. Phytochemcials can act at a variety of sites relevant to the development of the cancer cells. They may inhibit carcinogen activation, induce hepatic detoxification pathways, exert antioxidant effects/metal chelation properties, enhance immune response, induce apoptosis, and alter hormonal environment.

From a mechanistic perspective, evidence suggests that flavonoids have the potential to alter the cancer

development process by several different mechanisms. These include inhibition of the metabolic activation of carcinogens by modifying the expression of specific phase I and II enzymes, acting as antioxidants, inhibiting protein kinase C, interfering with expression of the mutated *ras* oncogene, and influencing other redox-regulated aspects of cell proliferation.

In addition to *in vitro* data, it is also well established that certain flavonoids can protect against chemically induced and spontaneously formed tumors in animal models. However, despite the significant amount of experimental evidence indicating that specific flavonoids have potent anticarcinogenic effects, the available epidemiological data are contradictory. Some ecological, cohort, and case–control studies suggest that tea consumption lowers the risk of developing cancer, whereas other investigations have failed to find such an association. The inconclusive nature may relate to poor information on dietary intake of flavonoids.

Safety

Although flavonoids may have potential health effects, the function of many of these compounds in the plant is to discourage attack by fungal parasites, herbivores, and pathogens. As a result, it is not surprising than many are toxic and mutagenic in cell culture systems, and excessive consumption by animals or humans may cause adverse metabolic reactions. However, the concentrations used in cell culture experiments in general tend to exceed the levels that are achievable *in vivo* following dietary consumption. Results of recent studies using β-carotene supplements should reinforce the need to proceed with caution in using flavonoid supplements, where levels could easily exceed doses obtained from normal dietary intake. For the majority of the identified phytochemicals, there are limited data on the 'safe level' of intake or optimal level of intake for health benefits, and it is critical that these margins be more clearly defined in future research.

Conclusions

There is increasing evidence that flavonoids may be protective against a number of age-related disorders. Data suggest that diets high in flavonoids may not only reduce the risk of cardiovascular disease and cancer but also, by protecting against cellular damage, may slow aspects of the aging process and improve quality of life by postponing age-related diseases. There is still much to be uncovered about their bioavailability, metabolism, mode of action, and optimal doses or, indeed, the actual compounds responsible for the health effect. Research has focused on foods as well as individual components of food to help us further our knowledge. Given the limited information to date, there are no recommended dietary intakes for phytochemicals, but people should consume a wide variety of foods that incorporate the various phytochemicals to maximize disease prevention. Further research is required to define optimal doses for potential health effects and to define safe levels of intakes for many of these phytochemicals. Many of these compounds should be viewed as pharmaceutical compounds because although they occur naturally, they still require the same levels of proof of efficacy and safety in use as synthetic pharmaceutical agents.

See also: **Antioxidants**: Diet and Antioxidant Defense; Observational Studies; Intervention Studies. **Cancer**: Epidemiology and Associations Between Diet and Cancer; Effects on Nutritional Status. **Coronary Heart Disease**: Prevention. **Fruits and Vegetables**. **Phytochemicals**: Epidemiological Factors. **Tea**. **Whole Grains**.

Further Reading

BNF (2003) Plants: Diet and health. In: Goldberg G (ed.) *The Report of the BNF Task Force*. Oxford: Blackwell. [ISBN 0-632-05962-1].

Bravo L (1998) Polyphenols: Chemistry, dietary sources, metabolism, and nutritional significance. *Nutrition Reviews* **56**: 317–333.

Cassidy A, DePascual S, and Rimbach GH (2003) Molecular mechanisms by which isoflavones potentially prevent atherosclerosis. *Expert Reviews in Molecular Medicine* **5**: 1–9.

Day AJ and Williamson G (2001) Biomarkers for exposure to dietary flavonoids: A review of the current evidence for identification of quercetin glycosides in plasma. *British Journal of Nutrition* **86**(supplement 1): S105–S110.

Dragsted LO (2003) Antioxidant actions of polyphenols in humans. *International Journal for Vitamin and Nutrition Research* **73**: 112–119.

Duthie G and Crozier A (2000) Plant-derived phenolic antioxidants. *Current Opinion in Clinical Nutrition and Metabolic Care* **3**: 447–451.

Hasler CM and Blumberg JB (1999) Phytochemicals: Biochemistry and physiology. Introduction. *Journal of Nutrition* **129**: 756S–757S.

Hertog MG, Kromhout D, Aravanis C *et al.* (1995) Flavonoid intake and long-term risk of coronary heart disease and cancer in the seven countries study. *Archives of Internal Medicine* **155**: 381–386.

Hollman PC and Arts ICW (2000) Flavonols, flavones and flavonols—Nature, occurrence and dietary burden. *Journal of the Science of Food and Agriculture* **80**: 1081–1093.

Hollman PC and Katan MB (1999) Health effects and bioavailability of dietary flavonols. *Free Radical Research* **31**(supplement): S75–S80.

Nielsen SE, Freese R, Kleemola P *et al.* (2002) Flavonoids in human urine as biomarkers for intake of fruits and

vegetables. *Cancer Epidemiology Biomarkers and Prevention* **11**: 459–466.
Nijveldt RJ, van Nood E, Van Hoorn DE *et al.* (2001) Flavonoids: A review of probable mechanisms of action and potential applications. *American Journal of Clinical Nutrition* **74**: 418–425.
Setchell KD and Cassidy A (1999) Dietary isoflavones: Biological effects and relevance to human health. *Journal of Nutrition* **129**: 758S–767S.

Epidemiological Factors

H Wiseman, King's College London, London, UK

There is considerable interest in the role that dietary phytochemicals may play in the protection of human health. This article considers the epidemiological evidence for the health protective effects of phytochemicals such as flavonoids, phytoestrogens, glucosinolates, and their derivatives and allium organosulfur compounds, particularly against cancer and heart disease. Possible health benefits of the soya isoflavone phytoestrogens to brain (especially cognitive function) and bone health are also considered together with the importance of their metabolism by the gut microflora (conversion of daidzein to equol) (**Table 1**). The possible mechanisms of action of these phytochemicals and others of related interest are also considered.

Epidemiological Sources of Evidence Indicating Potential Health Benefits of Phytochemicals

Flavonoids

Flavonoids are a group of more than 4000 polyphenolic compounds found in many plant foods. This group includes the flavonols such as quercetin, flavanols (or catechins, including catechin, epicatechin, epigallocatechin, and epigallocatechin gallate), flavones such as apigenin, and flavanones and anthocyanadins.

Until recently, the extent of absorption and bioavailability of flavonoids was somewhat unclear. Studies with ileostomy patients have shown that humans can absorb significant amounts of quercetin and that glycosides can be absorbed from the small intestine. Absorption of quercetin glucosides was 52%, absorption of pure quercetin was 24%, and that of quercetin rutinoside was 17%. This shows that not only can the glycone form of quercetin be absorbed but also absorption of the glucoside was greater than that of both the aglycone and the rutinoside, showing absorption to be enhanced by conjugation with glucose.

Epidemiological evidence suggests that dietary flavonoids, such as the quercetin, kaempferol, myricetin, apigenin, and luteolin found in tea, apples, onions, and red wine (usually as glycoside derivatives of the parent aglycones), may help to protect against coronary heart disease (CHD). The main epidemiological evidence comes from the Zutphen Elderly study and the Seven Countries Study. In the Zutphen Elderly study (805 men aged 65–84 years), the mean baseline flavonoid intake was 25.9 mg daily and the major sources of intake were tea (61%), onions (13%), and apples (10%). Flavonoid intake, which was analyzed in tertiles, was significantly inversely associated with mortality from CHD, and the relative risk of CHD in the highest versus lowest tertile of flavonoid intake ($\geq$28.6 vs $<$18.3 mg/day) was 0.42 (95% confidence interval, 0.20–0.88).

The Zutphen Elderly study thus suggests that regular flavonoid consumption, as part of the food matrix, may reduce the risk of death from CHD in elderly men. This study also provides evidence for flavonoid-mediated protection against stroke.

Table 1 Overview of epidemiological data relating to the role of soybean products in breast cancer risk

Study	*Soybean product*	*Findings*	*Estimate of relative risk*
Case–control[a]	Soybean protein	↓ Risk	0.43
	Soybean:total protein	↓ Risk	0.29
Case–control	Soybean	Not significant	[b]
Prospective	Miso soup	↓ Risk	0.46
Prospective	Miso soup	↓ Risk[c]	[b]
	Tofu	↓ Risk[c]	[b]

[a]Premenopausal women only.
[b]Could not be calculated.
[c]Decreased risk was only found to be significant for the baseline period 1971–1975.
Adapted from Messina MJ, Persky V, Setchell KDR and Barnes S (1994) Soy intake and cancer risk: A review of the *in vitro* and *in vivo* data. *Nutrition and Cancer* **21**: 113–131.

Dietary flavonoids (particularly quercetin) were inversely associated with stroke incidence. The relative risk of the highest versus the lowest quartile of flavonoid was 0.27 (95% confidence interval, 0.11–0.70). Black tea contributed approximately 70% to flavonoid intake and the relative risk for a daily consumption of 4.7 cups or more of tea versus less than 2.6 cups of tea was 0.31 (95% confidence interval, 0.12–0.84). This study also found that intake of catechins, whether from tea or other sources (e.g., chocolate), may reduce the risk of ischemic heart disease but not stroke. In the Rotterdam study (a large population-based study of men and women aged 55 or older), an inverse association was found between tea and flavonoid (quercetin, kaempferol, and myricetin) intakes and the incidence of myocardial infarction.

In 16 cohorts of the Seven Countries Study, the average long-term intake of flavonoids was inversely associated with mortality from CHD (**Table 2**). Surprisingly, flavonoid intake, did not appear to be an important determinant of cancer mortality in this study. This is in contrast to the anticarcinogenic effects observed in animal models and in human cancer cells *in vitro*. An inverse association between tea consumption and the incidence of some cancers has been reported in a prospective cohort study of 35 369 postmenopausal women. Inverse associations with increasing frequency of tea drinking were seen for cancers of the digestive tract and the urinary tract. The relative risk for women who reported drinking $\geq$2 cups (474 ml) of tea per day compared to those who never or only occasionally drank tea was 0.68 (95% confidence interval, 0.47–0.98) for digestive tract cancers and 0.4 (95% confidence interval, 0.16–0.98) for urinary tract cancers. Another epidemiological study reported a reduced risk of gastric cancer from drinking 10 cups or more daily of green tea. Tea, especially green tea, is particularly rich in catechins, such as epicatechin, epigallocatechin, and epigallocatechin gallate, in addition to flavonols such as quercetin.

The association between flavonoid intake and chronic diseases has been studied in Finland in 10 054 men and women. The incidence of cerebrovascular disease was lower at higher kaempferol, naringinin, and hesperetin intakes. Asthma incidence was lower at higher quercetin, naringinin, and hesperetin intakes. Men with high quercetin intakes had a lower lung cancer incidence, and men with higher myricetin intakes had a lower prostate cancer risk.

Flavonol and flavone intakes have been studied in the United States in health professionals (37 886 men and 78 886 women) using a semiquantitiative food frequency questionnaire. Of the flavonols and flavones investigated, quercetin contributed 76% in men and 73% in women. The mean flavonol and flavone intake was 20–22 mg/day, and onions, tea, and apples contributed the greatest amounts of flavonols and flavones. This information should prove useful in the investigation of the role of flavonoids in disease prevention.

Phytoestrogens

Phytoestrogens are phytochemicals found in a number of edible plants. The highest levels of dietary intakes of phytoestrogens are found in countries with a low incidence of hormone-dependent cancers. The main phytoestrogens in the human diet are the isoflavonoids and the lignans. Isoflavonoids include the isoflavones genistein, daidzein, and glycitein and occur mainly (as glycosides of the parent aglycone) in soybeans (*Glycine max*), a wide range of soy products, and to a lesser extent in other legumes. The main source of plant lignans are various seeds, such as linseed (secoisolariciresinol), sesame seed

Table 2 Data from the Seven Countries Study: Flavonoid (flavonol and flavone) intakes of middle-aged men in various countries in approximately 1960 and contribution of different foods to total flavonoid intake

Country	*Flavonol and flavone intake (mg/day)*	*Quercetin intake (mg/day)*	*Tea (%)*	*Fruit and vegetables (%)*	*Red wine (%)*
The Netherlands	33	13	64	36	0
Japan	64	31	90	10	0
United States	13	11	20	80	0
Finland	6	6	0	100	0
Croatia	49	30	0	82	18
Serbia	12	10	0	98	2
Greece	16	15	0	97	3
Italy	27	21	0	54	46

Adapted from Hertog MGL and Hollman PCH (1996) Potential health effects of the dietary flavonol quercetin. *European Journal of Clinical Nutrition* **50**: 63–71.

(matairesinol), and various grains (matairesinol and secisolariciresinol).

The incidence of breast and prostate cancer is much higher in Western countries than in Far Eastern ones, where there is an abundance of dietary phytoestrogens. Populations in the Far East have been consuming soyabean for centuries. In contrast, Western cultures and diets have only started to adopt soy foods much more recently. Western-style soy foods are produced by modern processing techniques in large soybean-processing plants. Traditional soy foods, made from soybeans, include both nonfermented and fermented foods. The nonfermented soy foods include soy milk and the soy milk product tofu and also whole-fat soy flour, soy nuts, whole dry beans, and fresh green soybeans. Traditional fermented soy foods include soy sauce, tempeh, natto, miso, and fermented tofu and soy milk products. Soy milk is the name given to the aqueous extract derived from whole soybeans. A cup of soy milk is thought to contain approximately 40 mg of isoflavones. In soybeans, textured vegetable protein, and tofu (soybean curd), there are high levels of the conjugated isoflavones called daidzin and genistin. In contrast, in the fermented soybean products such as miso, nearly all the isoflavones are present in their unconjugated forms called genistein and daidzein.

After ingestion, the glycones daidzin and genistin are hydrolyzed by gut bacterial glucosidases and by mammalian intestinal lactase phlorizin hydrolase to release the aglycones genistein and daidzein. These may be absorbed or further metabolized. Although most studies suggest that the bioavailabilities of genistein and daidzein are similar, some indicate greater bioavailability for genistein. Daidzein can be metabolized by the gut microflora to form the isoflavan equol (oestrogenic and more potent antioxidant than daidzein) or *O*-desmethylangolensin (*O*-DMA; nonoestrogenic), whereas genistein is metabolized to the nonoestrogenic *p*-ethyl phenol. In studies, only approximately 35% of subjects are able to convert daidzein to equol. Interindividual variation in the ability to metabolize daidzein to equol could thus influence the potential health protective effects of soya isoflavones. Equol is produced in greater amounts by subjects who consume diets that are low in fat and high in carbohydrate and fiber. Developmental changes in isoflavone metabolism occur, and although isoflavone absorption and the ability to convert daidzein to *O*-DMA develop early in infancy, equol production appears much later.

The lignan phytoestrogen precursors matairesinol and secisolariciresinol are present in foods as glycosides and are converted by gut bacteria to the two main mammalian lignans enterolactone and enterodiol, respectively, which are weakly oestrogenic. Matairesinol undergoes dehydroxylation and demethylation directly to enterolactone, whereas secisolariciresinol is converted to enterodiol, which can then be oxidized to enterolactone. After absorption, enterolactone and enterodiol are converted to their β-glucuronides and eventually excreted in urine.

In humans, omniverous subjects usually have quite low levels of isoflavonoid excretion. The Japanese (males and females) have the highest levels of isoflavonoid excretion in subjects following macrobiotic, vegan, and lactovegetarian diets. Urinary lignan excretion is higher in Finland compared to the United States and Japan. In assessing exposure to the protective effects of phytoestrogens, urinary excretion rates should be considered in combination with actual plasma levels. In some Japanese men, the plasma biologically active sulfate + free lignan fraction was similar or even higher than in Finnish men.

Urinary excretion of phytoestrogens can be used as a measure of intake and thus possible exposure and possible protection against cancer. Low urinary excretion of enterolactone in breast cancer patients was found in an epidemiological case–control study in Australia. Prospective studies from Finland and Sweden have shown low plasma concentrations of enterolactone to be associated with a high risk of breast cancer. However, the Swedish study also found a greatly increased risk of breast cancer in the highest quintile of enterolactone concentrations. A plasma enterolactone concentration of 30–80 nmol/l is therefore probably protective against breast cancer. Production of equol is associated with a decreased risk of breast cancer, and production of large amounts of equol is associated with an increased ratio of 2-hydroxyestrone to 16α-hydroxestrone in urine and this has been suggested to decrease breast cancer risk.

Japanese women and women of Japanese origin living in Hawaii but who consume a diet similar to the traditional Japanese diet (rich in soy products) have a low breast cancer incidence and mortality. Women in the Far East who have low rates of breast cancer are thought to consume approximately 30–50 times more soy products than women in the United States. A case–control study in Singapore found that premenopausal women who consumed 55 g of soy per day had a 50% reduced risk of breast cancer compared to women who infrequently consumed soy foods. A high intake of miso soup has been associated with a reduced risk of breast cancer in Japanese women. In prospective trials, a trend toward an inverse association between intake of tofu and

subsequent risk of breast cancer and an inverse association between intake of miso soup and development of breast cancer have been found. However, a large prospective study in Japan did not show any effect of soy consumption on breast cancer risk, although this may be because dietary intake was studied in adult women rather than in children or adolescents. A number of studies in rodents have indicated that a protective effect of a soy isoflavone-rich diet may occur only if soya is consumed before puberty or during adolescence. Soy, if consumed throughout life, appears to protect against breast cancer, particularly if consumed before and during adolescence. Soy isoflavones may decrease breast cancer risk by influencing the menstrual cycle and endogenous sex hormone concentrations. In some but not all studies, increased concentrations of sex hormone binding globulin leading to lower free sex hormone concentrations and a longer menstrual cycle were observed.

A trend toward protective effects against prostate cancer of tofu but not miso has been shown in a large group (approximately 8000) of men of Japanese ancestry in Hawaii followed for 20 years. The latency period for prostate cancer appears to be lengthened in these men, who have a low mortality from prostate cancer. However, the incidence of *in situ* prostate cancer in autopsy studies is similar to that of men in Western countries. The consumption of soy isoflavones by these men may be responsible for this long latency period. This probably means that they die of other causes, including old age, before the prostate cancer can develop to a life-threatening stage. The three most recent studies all suggest that soy intake does protect against prostate cancer. Two studies showed that reduced risk is related to consumption of soy foods and one was a prospective study that showed that consumption of soy milk more than once a day was protective against prostate cancer.

Although soy and isoflavonoids appear not to protect against colon cancer, lignans or lignan-rich foods can protect against colon cancer development in animal models. There is also increasing evidence for cardioprotective effects, bone protective effects, and possibly cognitive benefits of phytoestrogens, and these are under investigation. A lower incidence of heart disease has been reported in populations consuming large amounts of soy products, often in combination with oily fish consumption, which also has cardioprotective benefits. Increased bone mineral density has been found in epidemiological studies in women with high dietary intakes of soy isoflavones. The incidence of dementia has been reported to be lower in Asian countries, particularly Japan, where consumption of soy isoflavones is high. Although one epidemiological study found an association between high intakes of tofu and cognitive impairment, other factors, including age and education, may explain this possible increased risk among tofu consumers.

Brassica Glucosinolates and Their Derivatives

Glucosinolates (previously known as thioglucosides) are sulfur-containing phytochemicals found in cruciferous or brassica vegetables, such as broccoli, cabbage, kale, cauliflower, and Brussels sprouts. Although approximately 100 different glucosinolates are found in the plant kingdom, only approximately 10 are found in brassica vegetables. They are also found in other plant foods. Degradation products of glucosinolates include other organosulfur compounds, such as the isothiocyanates and dithiothiols. Glucosinolate degradation products also include indoles.

Epidemiological data suggest that the relatively high content of glucosinolates and related compounds may be responsible for the observed protective effects of brassica vegetables in the majority of the 87 case–control studies and 7 cohort studies that have been carried out on the association between brassica consumption and cancer risk (**Tables 3–5**). In the case–control studies, 67% of studies showed an inverse association between consumption of brassica vegetables and risk of cancer at various sites. If individual brassica vegetables are considered, then the values for the number of studies that showed an inverse association between consumption of brassica vegetables and risk of cancer at various sites are as follows: broccoli, 56%; Brussels sprouts, 29%; cabbage, 70%; and cauliflower, 67%. The cohort studies showed inverse associations between broccoli consumption and the risk of all types of cancer taken together; between the consumption of brassicas and risk of stomach cancer and the occurrence of second primary cancers; and between the consumption of cabbage, cauliflower, and broccoli and the risk of lung cancer. Overall, it appears that a high consumption of brassica vegetables is associated with a decreased risk of cancer. The associations were most consistent for stomach, lung, rectal, and colon cancer. The epidemiological literature also provides some support for the hypothesis that high intakes of brassica vegetables can reduce risk of prostate cancer. Further epidemiological research is required to separate the cancer protective effects of brassica vegetables from those of vegetables in general.

Allium Organosulfur Compounds

There is increasing epidemiological evidence that other organosulfur compounds in addition to those derived from glucosinolates can protect against

Table 3 Case–control studies of stomach, colon, and rectal cancer showing inverse, null, or positive associations for the consumption of different types of phytochemical-rich fruit and vegetables

Fruit or vegetable type	*No. of studies*								
	Stomach cancer[a]			*Colon cancer*[b]			*Rectal cancer*[c]		
	Inverse	*Null*	*Positive*	*Inverse*	*Null*	*Positive*	*Inverse*	*Null*	*Positive*
Fruit	14	3	0	5	2	1	3	0	1
Citrus fruit	11	1	0	2	1	3	4	1	0
Tomatoes	9	1	1	4	0	2	3	2	1
Vegetables	11	0	0	8	0	1	2	0	2
Raw vegetables	10	0	0	3	0	1	—	—	—
Allium vegetables	9	1	1	4	1	1	2	0	1
Cruciferous vegetables	—	—	—	8	3	1	5	0	0
Green vegetables	8	0	0	4	1	0	—	—	—
Legumes	7	0	2	1	2	2	—	—	—
Carrots	7	1	1	—	—	—	4	0	1

[a]Data summarize the results from 31 studies (both statistically significant and nonsignificant results included).
[b]Data summarize the results from 21 studies (both statistically significant and nonsignificant results included).
[c]Data summarize the results from 13 studies (both statistically significant and nonsignificant results included).
Adapted from Steinmetz KA and Potter JD (1996) Vegetables, fruit and cancer prevention: A review. *Journal of the American Dietetic Association* **96**:1027–1039.

Table 4 Case–control studies of lung, breast, and pancreatic cancer showing inverse, null, or positive associations for the consumption of different types of phytochemical-rich fruit and vegetables

Fruit or vegetable type	*No. of studies*								
	Lung cancer[a]			*Breast cancer*[b]			*Pancreatic cancer*[c]		
	Inverse	*Null*	*Positive*	*Inverse*	*Null*	*Positive*	*Inverse*	*Null*	*Positive*
Fruit	8	0	0	3	0	1	6	1	0
Citrus fruit	—	—	—	1	0	2	1	2	0
Tomatoes	4	0	0	—	—	—	—	—	—
Vegetables	7	0	0	—	—	—	5	1	0
Raw vegetables	—	—	—	—	—	—	2	1	0
Green vegetables	9	0	0	5	1	0	—	—	—

[a]Data summarize the results from 13 studies (both statistically significant and nonsignificant results included).
[b]Data summarize the results from 13 studies (both statistically significant and nonsignificant results included).
[c]Data summarize the results from nine studies (both statistically significant and nonsignificant results included).
Adapted from Steinmetz KA and Potter JD (1996) Vegetables, fruit and cancer prevention: A review. *Journal of the American Dietetic Association* **96**: 1027–1039.

cancer. Allium species such as garlic (*Allium sativum*) and onions (*Allium cepa*) are a rich source of organosulfur compounds, such as the diallyl sulfides. There is epidemiological evidence from the Netherlands Cohort Study (120 852 men and women 55–69 years of age) for a strong inverse association between onion consumption and incidence of stomach carcinoma. However, the consumption of leeks and the use of garlic supplements were not associated with stomach carcinoma risk. The relative risk for stomach carcinoma in the highest onion consumption category ($\geq$0.5 onions/day) was 0.50 (95% confidence interval, 0.26–0.95) compared to the lowest consumption category (no onions/day). However, this study did not support an inverse association between the consumption of onions and leeks and the use of garlic supplements and the incidence of male and female colon and rectal carcinoma. There is only limited epidemiological evidence concerning the beneficial influence of garlic organosulfur compounds on cardiovascular disease.

Potential Importance of Flavonoids to Human Health: Molecular Mechanisms of Action

Flavonoids possess a broad spectrum of biological actions ranging from anticarcinogenic to antiinflammatory, cardioprotective, immune-modulatory, and

Table 5 Cohort and case–control studies of all types of cancer showing inverse, null, or positive associations for the consumption of different types of phytochemical-rich fruit and vegetables

Fruit or vegetable type	*All types of cancer*[a]		
	Inverse	*Null*	*Positive*
Fruit	29	12	5
Citrus fruit	26	8	6
Tomatoes	35	5	10
Vegetables	55	4	9
Raw vegetables	33	4	2
Allium vegetables	27	3	4
Cruciferous vegetables	38	8	8
Green vegetables	61	5	13
Legumes	14	6	16
Carrots	50	7	7

[a]Data summarize the results from 194 studies (both statistically significant and nonsignificant results included).
Adapted from Steinmetz KA and Potter JD (1996) Vegetables, fruit and cancer prevention: A review. *Journal of the American Dietetic Association* **96**: 1027–1039.

antiviral. The mechanisms by which flavonoids cause these effects may include induction of the activity of some important enzymes while inhibiting the activity of others. Modulation of membrane function, including the activity of membrane-bound enzymes, through a protective membrane antioxidant action is likely to be of prime importance.

Membrane function is understood to be of vital importance to many cellular processes, including the role of membrane enzymes and receptors in cell growth and signalling. Membrane function may be influenced by dietary components directly by altering membrane fluidity or indirectly by protection against the free radical-mediated process of membrane lipid peroxidation. This can arise from oxidative stress and result in oxidative membrane damage. Flavonoids such as quercetin and myrecetin have been widely found to inhibit membrane lipid peroxidation. Flavonoids inhibit lipid peroxidation *in vitro* by acting as chain-breaking antioxidants: They donate a hydrogen atom to lipid radicals, thus terminating the chain reaction of lipid peroxidation. Additionally, flavonoids can act as metal chelating agents. Furthermore, kaempferol-3-*O*-galactoside protected mice against bromobenzene-induced hepatic lipid peroxidation. The relative potencies of flavonoids as antioxidants is governed by a set of structure–function relationships: In general, optimum antioxidant activity is associated with multiple phenolic groups, a double bond in C2–C3 of the C ring, a carbonyl group at C4 of the C ring, and free C3 (C ring) and C5 (A ring) hydroxy groups. It is of related interest that consumption of 300 ml of either black or green tea greatly increased plasma antioxidant capacity in 10 volunteers. This suggests that normal levels of tea consumption could provide sufficient flavonoids to achieve a potentially health protective effect.

There is increasing evidence for the role of free radicals in the oxidative DNA damage implicated in carcinogenesis. The ability of flavonoids to act as antioxidants may contribute to the anticancer effects observed in animal models and human cells in culture *in vitro*, which could potentially be important to human health despite the current lack of epidemiological evidence and the finding that consumption of flavonoids in onions and black tea (providing 91 mg/day of quercetin for 2 weeks) by young healthy male and female subjects had no effect on oxidative DNA base damage in leucocytes. Quercetin has been shown to have growth inhibitory effects *in vitro* on breast cancer cells, colon cancer cells, squamous cell carcinoma cell lines, acute lymphoid and myeloid leukemia cell lines, and a lymphoblastoid cell line. These effects appear to be mediated *via* binding to cellular type 2 oestrogen binding sites. Furthermore, when the ability of two citrus flavonoids, hesperetin and naringenin (found in grapefruit mainly as its glycosylated form naringin), and three noncitrus flavonoids to inhibit the proliferation and growth of a human breast cancer cell line was investigated, the concentrations required to achieve 50% inhibition ranged from 5.9 to 56 μg/ml. The effectiveness of the citrus flavonoids was enhanced by using them in combination with quercetin, which is widely distributed in other foods. Quercetin fed to rats in the diet at levels of 2% or 5% inhibited the incidence and multiplicity of chemical carcinogen-induced mammary tumors. Mammary tumorigenesis in rats was delayed in the groups given orange juice (rich in citrus flavonoids together with other phytochemicals and nutrients) or fed the naringin-supplemented diet compared with the other groups. A number of the phenolic compounds of green tea, including the catechins, have been shown to inhibit tumour formation in rats induced by *N*-methyl-*N*′-nitro-*N*-nitrosoguanidine and also mutation induced by aflatoxin and benz(*a*)pyrene.

Quercetin has been shown to inhibit the activity of two enzymes that play an important role in mammary cell growth and development, tyrosine protein kinase activity and phosphoinositide phosphorylation, and it also inhibits protein kinase C, which is vital in the regulation of cellular proliferation. Blockade of the tyrosine kinase activity of the EGR receptor leading to growth inhibition and apoptosis in pancreatic tumor cells have been reported for quercetin and luteolin. Furthermore, inhibition of

tumor growth through cell cycle arrest and induction of apoptosis by quercetin are thought to be functionally related to activation of the tumor supressor protein p53. In addition, quercetin has been shown to regulate the growth of endometrial cancer cells (Ishikawa cell line) via suppression of EGF and the cell cycle protein, cyclin D1. A further mechanism for the antiproliferative action of quercetin may be via perturbation of microtubule functions such as polymerization through the binding of quercetin to tubulin, which induces conformational changes.

A number of mechanisms have been proposed for the protection by flavonoids against CHD, including antioxidant activity. Oxidative damage to low-density lipoprotein (LDL) (particularly to the apoprotein B molecule) is considered to be an important stage in the development of atherosclerosis: It is a prerequisite for macrophage uptake and cellular accumulation of cholesterol leading to the formation of the atheromal fatty streak. Flavonoids such as quercetin are effective inhibitors of *in vitro* oxidative modification of LDL by macrophages or copper ions. Although consumption of flavonoids in onions and black tea (providing 91 mg/day of quercetin for 2 weeks) by young healthy male and female subjects had no effect on plasma F_2-isoprostane concentrations (a biomarker of *in vivo* lipid peroxidation) or on resistance of LDL to copper–ion-induced oxidation, flavonoids in red wine have been reported to protect LDL against oxidative damage. The antioxidant properties of flavonoids may contribute to the reduced risk of CHD in wine drinkers, the so-called French paradox. Resveratrol, another phenolic phytochemical found in wine, has been shown to protect LDL against oxidative damage and appears to protect against cancer in animal models. Further studies on this interesting compound are clearly warranted.

Quercetin displays potent antithrombotic effects: It inhibits thrombin and ADP-induced platelet aggregation *in vitro*, and this may be through inhibition of phospholipase C activity rather than through inhibition of thromboxane synthesis. Flavonoid binding to platelet membranes may inhibit the interaction of activated platelets with vascular endothelium. In addition, quercetin elicits coronary vasorelaxation that is endothelium independent. The antioxidant activity of flavonoids may also prevent the damaging action of lipid peroxides generated by activated platelets on endothelial nitric oxide and prostacyclin, which both inhibit platelet aggregation and have vasodilatory activity.

The activity of flavonoids as inhibitors of the viral enzyme reverse transcriptase also suggests that they may be beneficial in the control of retroviral infections such as AIDS.

Possible adverse effects on human health should also be considered. Quercetin was reported to induce bladder cancer in rats when administered in the diet at a level of 2%. These results were not confirmed in another study, however, which used quercetin at levels reaching 10%. It should be noted that under certain *in vitro* conditions flavonoids and other phenols can act as prooxidants and cause DNA damage. However, phenols have complex pro- and antioxidant effects *in vitro*, depending on the assay system used, and it is often difficult to predict their net effect *in vivo*. For example, many synthetic and dietary polyphenols (including quercetin, catechin, gallic acid ester, and caffeic acid ester) can protect mammalian cells from the cytotoxicity induced by peroxides such as hydrogen peroxide. Although tea is a good source of flavonoids, phenolic compounds including tannins and also polyphenols and phenol monomers are good inhibitors of iron absorption, which could contribute to the nutritional problem of iron deficiency. In general, it is unlikely that sufficiently toxic quantities of any particular flavonoid could be consumed from the diet, which contains many diverse varieties of flavonoids in varying quantities.

Potential Importance of Phytoestrogens to Human Health: Molecular Mechanisms of Action

The probable beneficial effects of phytoestrogens against breast cancer are likely to be mediated via numerous mechanisms. However, it has not been fully established whether the protective effects of soya and cereals result from their phytoestrogen content or from some other effect.

Many studies utilising breast cells in culture such as the oestrogen-sensitive MCF-7 cell line show that phytoestrogens (genistein was used in most of studies) stimulate tumor growth at low concentrations while inhibiting growth at higher concentrations. Genistein is a potent and specific *in vitro* inhibitor of tyrosine kinase action in the autophosphorylation of the epidermal growth factor (EGF) receptor and is thus frequently used as a pharmacological tool. The EGF receptor is overexpressed in many cancers, particularly those with the greatest ability for metastasis, and it has therefore often been assumed that some of the anticancer effects of genistein are mediated *via* inhibition of tyrosine kinase activity. However, this is likely to be an oversimplification of the true *in vivo* situation.

Although genistein is a much better ligand for oestrogen receptor β (ERβ) than for the ERα (20-fold higher binding affinity), it can also act as an oestrogen agonist via both ERα and ERβ in some test systems. Mechanisms other than those involving oestrogen receptors are likely to be involved in the inhibition of cell proliferation by genistein because genistein inhibits both the EGF-stimulated and the 17β-oestradiol-stimulated growth of MCF-7 cells. Although studies have shown that exposure to genistein can reduce the tyrosine phosphorylation of cell proteins in whole cell lysates, studies using cultured human breast and prostate cancer cells have not confirmed that genistein has a direct effect on the autophosphorylation of the EGF receptor. Many other mechanisms of anticancer action for isoflavones and genistein in particular have been suggested, including inhibition of DNA topoisomerases, cell cycle progression, angiogenesis, tumor invasiveness, and enzymes involved in oestrogen biosynthesis. They also include effects on the expression of DNA transcription factors c-*fos* and c-*jun*, on reactive oxygen species, on oxidative membrane damage and oxidative damage *in vivo*, and on the negative growth factor, transforming growth factor-β (TGF-β).

Although cholesterol lowering is probably the best documented cardioprotective effect of soya, vascular protection is also likely to contribute and may be mediated via a number of mechanisms. Soya isoflavones are likely to contribute to the cardioprotective benefits of soya.

ERβ is the predominant ER isoform expressed in the rat, mouse, and human vascular wall. In the rat carotid injury model, following endothelial denudation of rat carotid artery, ERα is expressed at a low level, whereas the expression of ERβ increases by greater that 40-fold and treatment of ovariectomized female rats with genistein provides a similar dose-dependent vasculoprotective effect in this model to that observed with 17β-oestradiol. However, studies in ERβ knockout mice have shown that ERβ is not required for oestrogen-mediated inhibition of the response to vascular injury and suggest that either of the two known oestrogen receptors (or another unidentified one) is sufficient to protect against vascular injury.

Vascular protection could also be conferred by the ability of genistein to inhibit proliferation of vascular endothelial cells and smooth muscle cells and to increase levels of TGF-β. TGF-β helps maintain normal vessel wall structure and promotes smooth muscle cell differentiation while preventing their migration and proliferation. Genistein has been shown to increase TGF-β secretion by cells in culture, and increased TGF-β production may be a mediator of some of the cardioprotective effects of soya isoflavones.

Antioxidant action is one of the mechanisms that may contribute to the vascular protective effects of soya isoflavones. Antioxidant properties have been reported for isoflavones both *in vitro* and *in vivo*. In a randomized crossover study of young healthy male and female subjects consuming diets that were rich in soy that was high (56 mg total isoflavones/day: 35 mg genistein and 21 mg daidzein) or low in isoflavones (2 mg total isoflavones/day), each for 2 weeks, plasma F_2-isoprostane concentrations were significantly lower after the high-isoflavone dietary treatment than after the low-isoflavone dietary treatment. The lag time for copper–ion-induced LDL oxidation was significantly longer.

Increased resistance to LDL oxidation has also been reported in a 12-week single open-group dietary intervention with soy foods (60 mg total isoflavones/day) in normal postmenopausal women. A randomized crossover study in hyperlipidemic male and female subjects consuming soya-based breakfast cereals (168 mg total isoflavones/day) and control breakfast cereals, each for 3 weeks, reported decreased oxidized LDL (total conjugated diene content) following consumption of the soy-based breakfast cereal compared to the control.

Effects of soya isoflavones on arterial function, including flow-mediated endothelium-dependent vasodilation (reflecting endothelial function) and systemic arterial compliance (reflecting arterial elasticity), may contribute to vascular protection and these have been measured in a number of studies. A randomized double-blind study administering either soy protein isolate (118 mg total isoflavones/day) or cesin placebo for 3 months to healthy male and postmenopausal subjects (50–75 years of age) showed a significant improvement in peripheral pulse wave velocity (reflecting peripheral vascular resistance and one component, together with systemic arterial compliance, of vascular function) but worsened flow-mediated vasodilation in men and had no significant effect on the flow-mediated vasodilation in postmenopausal women.

Some beneficial effects following dietary intervention with soy isoflavones have been observed on bone health, and the mechanism is likely to be via an oestrogenic action, particularly because ERβ is highly expressed in bone, although this requires further investigation. Consumption by postmenopausal women (6-month parallel group design) of soy protein (40 g/day providing either 56 mg isoflavones/day or 90 mg isoflavones/day) compared to caesin and nonfat dry milk (40 g/day) produced significant increases in bone mineral content (BMC)

and bone mineral density (BMD) in the lumbar spine (but not in any other parts of the body) only in the higher isoflavone (90 mg/day) group compared to the control group. In a long-term study, consumption by postmenopausal women (2-year parallel group design) of isoflavone-rich soy milk (500 ml/day providing 76 mg isoflavones/day) compared to isoflavone-poor soy milk control (providing 1 mg isoflavones/day) resulted in no decline in BMC and BMD in the treatment group compared to significant losses in the control group. The ability to produce equol was associated with a better response to the treatment.

Some beneficial effects following dietary intervention with soy isoflavones have been observed on the cognitive function aspect of brain health, and the mechanism is likely to be via an oestrogenic action, particularly because ERβ, in addition to ERα, is expressed in brain. Although other mechanisms may contribute, they remain to be elucidated. Consumption by young healthy male and female subjects (parallel group design) of a high-soy diet (100 mg isoflavones/day for 10 weeks) compared to a low-soy diet (0.5 mg isoflavones/day) resulted in improved cognitive function, including significantly improved short-term and long-term memory and mental flexibility. These improvements were found in males and females. Consumption by postmenopausal women (parallel group design, placebo controlled) of a dietary supplement (soy extract containing 60 mg isoflavones/day for 12 weeks) resulted in improved cognitive function, particularly improved long-term memory.

Phytoestrogens can cause infertility in some animals and thus concerns have been raised over their consumption by human infants. The isoflavones found in a subterranean clover species (in Western Australia) have been identified as the agents responsible for an infertility syndrome in sheep. No reproductive abnormalities have been found in peripubertal rhesus monkeys or in people living in countries where soy consumption is high. Indeed, the finding that dietary isoflavones are excreted into breast milk by soy-consuming mothers suggests that in cultures in which consumption of soy products is the norm, breast-fed infants are exposed to high levels without any adverse effects. Isoflavone exposure soon after birth at a critical developmental period through breast feeding may protect against cancer and may be more important to the observation of lower cancer rates in populations in the Far East than adult dietary exposure to isoflavones. Although some controversy exists as to whether soy-based infant formulas containing isoflavones pose a health risk, a review of studies on the use of soy milk in infants suggests that there is no real basis for concern. Toxicity from isoflavones may arise from their action as alternative substrates for the enzyme thyroid peroxidase, and people in Southeast Asia would be protected by the dietary inclusion of iodine-rich seaweed products.

Potential Importance of Glucosinolate Derivatives and Related Compounds to Human Health: Molecular Mechanisms of Action

There may be some important health protective effects of glucosinolate derivatives and related compounds. The hydrolytic products of some glucosinolates have been shown to display anticancer properties. Glucosinolates are hydrolyzed following exposure to the endogenous plant enzyme myrosinase (also found in the gut microflora) to form isothiocyanates. Isothiocyanates are biologically active compounds with anticancer properties and are more bioavailable than glucosinolates.

A metabolite of glucobrassicin (3-indoylmethylglucosinolate), indole-3-carbinol has been shown to inhibit the growth of human tumors of the breast and ovary. Furthermore, indole-3-carbinol may modulate the oestrogen hydroxylation pathway such that a less potent form of oestradiol is produced, thus conferring protection against oestrogen-related cancers.

Consumption of Brussels sprouts (300 g/day of cooked sprouts) for 1 week has been shown to increase rectal glutathione *S*-transferase -α and -π isoenzyme levels. Enhanced levels of these detoxification enzymes may partly explain the epidemiological association between a high intake of glucosinolates in cruciferous vegetables and a decreased risk of colorectal cancer. It is likely that genetic polymorphisms and associated functional variations in biotransformation enzymes, particularly in glutathione *S*-transferases, will alter the cancer preventative effects of cruciferous vegetables.

Compounds including the isolated glucosinolate sinigrin and aqueous extracts of cooked and autolyzed Brussels sprouts (rich in glucosinolate degradation products) decreased hydrogen peroxide-induced DNA stand breaks in human lymphocytes and thus exerted a DNA-protective effect. Oral adminiatration of sinigrin has been shown to induce apoptosis and suppress aberant crypt foci in the colonic mucosa of rats treated with 1,2-dimethylhydrazine. Similar effects were observed with oral administration of freshly prepared Brussels sprout juice, rich in glucosinolate breakdown products including isothiocyanates.

Isothiocyanates can prevent the formation of chemical carcinogen-induced tumors of the liver, lung, mammary gland, stomach, and oesophagus in animal models. The anticarcinogenic effects of isothiocyanates may be mediated by a combination of mechanisms, including inhibition of carcinogen activation by cytochromes P450: This could be achieved by both direct inhibition of enzyme catalytic activity and downregulation of enzyme levels and induction of phase 2 enzymes such as glutathione transferases and NAD(P)H:quinone reductase (these detoxify any remaining DNA-attacking electrophilic metabolites generated by phase 1 enzymes). Dietary glucosinolates and their breakdown products have been tested as anticarcinogens in terms of their ability to induce the anticarcinogenic phase 2 enzyme marker quinone reductase in murine Hep a1c1c7 cells, and the relative activities observed were found to be dependent on the nature of the side chain of the parent glucosinolate.

Phenethyl isothiocyanate protects mice against nitrosoamine-induced lung tumorigenesis. It also modulates the activity of phase 1 and phase 2 xenobiotic-metabolizing enzymes, resulting in the inhibition of the oxidative activation of a number of chemical carcinogens.

The isothiocyanate sulforophane is a particularly potent inducer of detoxification enzymes. A novel isothiocyanate-enriched broccoli has been developed that has an enhanced ability to induce phase 2 detoxification enzymes in mammalian cells compared to standard commercial broccoli.

Undesirable goitrogenic effects have been identified for isothiocyanates and other hydrolytic products of glucosinolates. Furthermore, in contrast to the anticancer effects of brassica vegetables discussed previously, a number of genotoxic effects have also been demonstrated in bacterial and mammalian cells. In bacterial assays (induction of point mutations in *Salmonella* TA98 and TA100 and repairable DNA damage in *Escherichia coli* K-12), juices from eight brassica vegetables tested caused genotoxic effects in the absence of metabolic activation. The order of potency was Brussels sprouts > white cabbage > cauliflower > green cabbage > kohlrabi > broccoli > turnip > black raddish. In mammalian cells, structural chromosome aberrations were observed with some of the juices, with the most potent being Brussels sprouts and white cabbage, and genotoxic effects were accompanied by decreased cell viability. The isothiocyanate-containing fraction (and other breakdown products of glucosinolates) of these brassica juices was found to contain 70–80% of the total genotoxic activity of the juices. The flavonoid- and other phenolic-containing fraction had a much weaker effect. In related studies, the isothiocyanates, allyl isothiocyanate and phenethyl isothiocyanate, were found to be more than 1000-fold more cytotoxic in a Chinese hamster ovary cell line than their parent glucosinolates (sinigrin and gluconasturtiin, respectively). Phenethyl isothiocyanate also induced genotoxic effects (chromosome aberrations and sister chromatid exchanges).

More data are required before an overall recommendation can be made regarding the likely beneficial or otherwise influences of glucosinolates (and their derivatives) on human health.

S-Methyl Cysteine Sulfoxide

S-methyl cysteine sulfoxide is another sulfur-containing phytochemical found in all brassica vegetables, in addition to glucosinolates. Both *S*-methyl cysteine sulfoxide and methyl methane thiosulfinate (its main metabolite) can block genotoxicity, induced by chemicals, in mice. *S*-methyl cysteine sulfoxide is thus likely to contribute to the observed ability of brassica vegetables to protect against cancer in both human and animal studies. It is of interest that a hydrolytic product of *S*-methyl cysteine sulfoxide was linked in the 1960s to the severe hemolytic anemia or kale poisoning observed in cattle in Europe in the 1930s.

Potential Importance of Other Phytochemicals to Human Health: Molecular Mechanisms of Action

Allium Organosulfur Compounds

Allium organosulfur compounds may be phyotchemicals of importance to human health by acting as antioxidants, thus protecting against free radical-mediated damage to important cellular targets such as DNA and membranes implicated in cancer and neurodegenerative diseases and aging. Protection against oxidative damage to LDL and cellular membranes could also protect against cardiovascular disease. Aged garlic extract (AGE) inhibits lipid peroxidation and the oxidative modification of LDL, reduces ischemic/reperfusion injury, and enhances the activity of the cellular antioxidant enzymes superoxide dismutase, catalase, and glutathione peroxidase. AGE also inhibits the activation of the oxidant-induced transcription factor NF-κB. Investigation of the major organosulfur compounds in AGE identified highly bioavailable water-soluble organosulfur compounds with antioxidant activity, such as *S*-allylcysteine and *S*-allylmeracptocysteine.

Organosulfur compounds such as diallyl sulfide may also protect against cancer by modulation of carcinogen metabolism, and this may involve altered

ratios of phase 1 and phase 2 drug-metabolizing enzymes. Various garlic preparations including aged garlic extract have been shown to inhibit the formation of nitrosamine-type carcinogens in the stomach, enhance the excretion of carcinogen metabolites, and inhibit the activation of polyarene carcinogens. Inhibitory effects of organosulfur compounds on the growth of cancer cells *in vitro*, including human breast cancer cells and melanoma cells, have been observed. Modulation of cancer cell surface antigens, associated with cancer cell invasiveness, has been observed, and in some cases cancer cell differentiation can be induced. AGE can reduce the appearance of mammary tumors in rats treated with the powerful carcinogen dimethyl benz(*a*)anthracene (DMBA), which is activated by oxidation by cytochromes P450 to form the DNA binding form of DMBA diol epoxide, resulting in DNA legions and cancer initiation. The antibacterial activity of these allium compounds may also prevent bacterial conversion of nitrate to nitrite in the stomach. This may reduce the amount of nitrite available for reacting with secondary amines to form the nitrosamines likely to be carcinogenic particularly in the stomach.

Allium organosulfur compounds appear to possess a range of potentially cardioprotective effects. In one study, 432 cardiac patients were divided into a control group (210) and a garlic-supplemented group (222), and garlic feeding was found to reduce mortality by 50% in the second year and by approximately 66% in the third year. Furthermore, the rate of reinfarction was reduced by 30 and 60% in the second and third year, respectively. It should be noted that only a small number of patients in both groups experienced the end event of death or myocardial infarction, and a much larger scale study is needed. AGE lowers cholesterol and triglycerides in laboratory animals and can reduce blood clotting tendencies. It has been suggested that garlic supplementation at a level of 10–15 g of cooked garlic daily could lower serum cholesterol by 5–8% in hypercholestrolemic individuals. However, there may be more important cardioprotective effects of garlic. In animal studies, AGE suppressed the levels of plasma thromboxane B_2 and platelet factor levels, which are important factors in platelet aggregation and thrombosis. In rats, frequent low doses (50 mg/kg) of aqueous extracts of garlic or onions (onion was less potent) produced significant antithromotic activity (lowering of thromboxane B_2) without toxic side effects.

Aqueous extracts of raw garlic also inhibited cyclooxygenase activity in rabbit platelets, again contributing to an antithrombotic effect. In addition, AGE and *S*-allyl cysteine and *S*-allyl mercaptocysteine have antiplatelet adhesion effects. Platelet adhesion to the endothelial surface is involved in atherosclerosis initiation. Furthermore, *S*-allyl mercaptocysteine inhibits the proliferation of rat aortal smooth muscle cells, another important atherosclerotic process. Indeed, this antiproliferative effect on smooth muscle cells may be indicative of a possible antiangiogenic ability in relation to prevention of tumor growth and metastasis.

Saponins

Saponins are another steroidal phytochemical of interest that may, in addition to isoflavone phytoestrogens, contribute to the health protective effects of soya products. Soyabeans have a high saponin content and soyabean saponins have been shown to have a growth inhibitory effect on human carcinoma cell *in vitro*, probably by interacting with the cell membrane and increasing membrane permeability. The proposed anticarcinogenic mechanisms of saponins include normalization of carcinogen-induced cell proliferation, direct cytotoxicity, bile acid binding, and immune-modulating effects. Of particular interest is the finding that saponins actively interact with cell membrane components: They possess surface active characteristics because of the amphiphillic nature of their chemical structure. Thus, they can act to alter cell membrane permeability and cellular function. Soybean saponins have been reported to inhibit hydrogen peroxide damage to mouse fibroblast cells and thus may protect human health through antioxidant-mediated mechanisms.

Saponins from ginseng root (*Panax ginseng C.A. Mey.*) may also be important. Antioxidant effects have been reported for total ginseng saponins and its individual saponins (ginsenosides Rb1, Rb2, Rc, and Rd; others include Re and Rg1). Furthermore, ginsenosides Rb1 and Rb2 protected cultured rat myocardiocytes against superoxide radicals, and the mechanism for this may involve induction of genes responsible for antioxidant defences rather than radical scavenging. Ginsenosides stimulate endogenous production of nitric oxide in rat kidney, and this may contribute to the observed antinephritic action of these compounds and suggest a protective role in the kidney. Furthermore, it has been suggested that the observed cardioprotective effects of ginsenosides in animal models may be mediated by nitric oxide release. In addition, ginsenoside enhanced release of nitric oxide from endothelial cells, particularly from perivascular nitric oxidergic nerves in the corpus cavernosum of animal models, may partly account for the reported aphrodisiac effects of ginseng. Also, ginsenosides have been shown to have beneficial effects on inferior human

sperm motility and progression. It is of interest that regulation of lipid metabolism by ginseng has been reported, and although the mechanism of action remains unclear, it is likely that the peroxisome proliferator-activated receptor-α is involved.

Other Phytochemicals of Interest

A wide range of other phytochemicals may have important beneficial effects on human health if consumed in sufficient amount to be efficacious. In many cases, their full spectrum of molecular actions remains to be elucidated. Nevertheless, the following phytochemicals and their main botanical sources are deemed worthy of mention.

The phytochemicals dihydrophthalic acid, ligustilide, butylidene, phthalide, and *n*-valerophenone-*O*-carboxylic acid have been isolated from Angelica root (*Angelica sinensis*). They are likely to contribute to the observed circulatory modulating effects of Angelica root, including increasing coronary flow, modulation of myocardial muscular contraction, and antithrombotic effects.

Phytochemicals extracted from licorice (*Glycyrrhiza glabra L.*) include glycyrrhetic acid, glycyrrhizic acid (the sweet principle of licorice), and an active saponin glycyrrhizin (a 3-*O*-diglucuronide of glycyrrhetic acid). In rats, dietary supplementation with 3% licorice elevated liver glutathione transferase activity, suggesting a potential detoxification and anticancer effect of these phytochemicals because glutathione transferase catalyses the formation of glutathione conjugates of toxic substances for elimination from the body. Antibacterial, antiviral, antioxidant, and antiinflammatory effects have also been reported for these compounds. Indeed, glycyrrhizin has been reported to inhibit HIV replication in cultures of peripheral blood mononuclear cells taken from HIV-seropositive patients.

Phytochemicals found in ginkgo (*G. biloba*) leaves, including ginkgolic acid, hydroginkgolic acid, ginkgol, bilobol, ginon, ginkgotoxin, ginkgolides (A–C), and a number of flavonoids common to other plants, such as kaempferol, quercetin, and rutin, are currently attracting attention for their possible effects on circulation, particularly cerebral circulation, and this may improve brain function and cognition. Indeed, ginkgo, ginseng, and a combination of the two extracts have been found to improve different aspects of cognition in healthy young volunteers. A number of studies have reported that extracts of ginkgo leaves enhanced brain circulation, increased the tolerance of the brain to hypoxia, and improved cerebral hemodynamics. It has been suggested that these effects are mediated via calcium ion flux over smooth cell membranes and via stimulation of catecholamine release. In addition, protection against free radical-mediated retinal injury has been reported; thus, other antioxidant-mediated protective effects on human health are also possible. Damage to mitochondrial DNA could play a role in neurodegenerative diseases such as Alzheimer's disease and Parkinson's disease. There is limited evidence for significant improvements in CHD patients following treatment with a daily dose equivalent to 12 mg total ginkgetin. Ginkgolide B-activated inhibition of glucocorticoid production has been reported and is likely to result from specific transcriptional suppression of the adrenal peripheral-type benzodiazepine receptor gene in rats. This suggests that ginkgolide B may be useful pharmacologically to control excess glucocorticoid formation.

See also: **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Cereal Grains**. **Coronary Heart Disease**: Prevention. **Fruits and Vegetables**. **Phytochemicals**: Epidemiological Factors. **Tea**.

Further Reading

Adlercreutz CHT (2002) Phyto-oestrogens and cancer. *Lancet Oncology* **3**: 32–41.

Arts IC, Hollman PC, Feskens EJ, Bueno de Mesquita HB, and Kromhout D (2001) Catechin intake might explain the inverse relationship between tea consumption and ischemic heart disease: The Zutphen Elderly Study. *American Journal of Clinical Nutrition* **74**: 227–232.

Beatty ER, O'Reilly JD, England TG *et al.* (2000) Effect of dietary quercetin on oxidative DNA damage in healthy human subjects. *British Journal of Nutrition* **84**: 919–925.

File SE, Jarrett N, Fluck E *et al.* (2001) Eating soya improves human memory. *Psychopharmacology* **157**: 430–436.

Gupta K and Panda D (2002) Perturbation of microtubule polymerization by quercetin through tubilin binding: A novel mechanism of its antiproliferative activity. *Biochemistry* **41**: 13029–13038.

Kim H, Xu J, Su Y *et al.* (2001) Actions of the soy phytoestrogen genistein in models of human chronic: Potential involvement of transforming growth factor β. *Biochemical Society Transactions* **29**: 216–222.

Knekt P, Kumpulainen J, Jarvinen R *et al.* (2002) Flavonoid intake and risk of chronic diseases. *American Journal of Clinical Nutrition* **76**: 560–568.

Mithen R, Faulkner K, Magrath R *et al.* (2003) Development of isothiocyanate-enriched broccoli and its enhanced ability to induce phase 2 detoxification enzymes in mammalian cells. *Theoretical Applied Genetics* **106**: 727–734.

O'Reilly JD, Mallet AI, McAnlis GT *et al.* (2001) Consumption of flavonoids in onions and black tea: Lack of effect on F_2-isoprostanes and autoantibodies to oxidized LDL in healthy humans. *American Journal of Clinical Nutrition* **73**: 1040–1044.

Rowland IR, Wiseman H, Sanders TAB, Adlercreutz H, and Bowey EA (2000) Interindividual variation in metabolism of soy isoflavones and lignans: Influence of habitual diet on equol production by the gut microflora. *Nutrition and Cancer* **36**: 27–32.

Shapiro TA, Fahey JW, Wade KL, Stephenson KK, and Talalay P (2001) Chemoprotective glucosinolates and isothiocyanates of

broccoli sprouts: Metabolism and excretion in humans. *Cancer Epidemiology Biomarkers and Prevention* **10**: 501–508.

Thomson M and Ali M (2003) Garlic [allium sativum]: A review of its potential use as an anticancer agent. *Current Cancer Drug Targets* **3**: 67–81.

Wiseman H (2000) The therapeutic potential of phytoestrogens. *Expert Opinion in Investigational Drugs* **9**: 1829–1840.

Wiseman H, Goldfarb P, Ridgway T, and Wiseman A (2000) *Biomolecular Free Radical Toxicity: Causes and Prevention.* Chichester, UK: John Wiley.

Wiseman H, O'Reilly JD, Adlercreutz H *et al.* (2000) Isoflavone phytoestrogens consumed in soy decrease F_2-isoprostane concentrations and increase resistance of low-density lipoprotein to oxidation in humans. *American Journal of Clinical Nutrition* **72**: 395–400.

Phyto-estrogens *see* **Phytochemicals**: Classification and Occurrence; Epidemiological Factors

Polyunsaturated Fatty Acids *see* **Fatty Acids**: Omega-3 Polyunsaturated; Omega-6 Polyunsaturated

POTASSIUM

L J Appel, Johns Hopkins University, Baltimore, MD, USA

The major intracellular cation in the body is potassium, which is maintained at a concentration of approximately 145 mmol/l of intracellular fluid but at much lower concentrations in the plasma and interstitial fluid (3.8–5 mmol/l of extracellular fluid). The high intracellular concentration of potassium is maintained via the activity of the Na^+/K^+-ATPase pump. Because this enzyme is stimulated by insulin, alterations in the plasma concentration of insulin can affect cellular influx of potassium and thus plasma concentration of potassium. Relatively small changes in the concentration of extracellular potassium greatly affect the extracellular/intracellular potassium ratio and thereby affect nerve transmission, muscle contraction, and vascular tone.

In unprocessed foods, potassium occurs mainly in association with bicarbonate-generating precursors such as citrate and, to a lesser extent, with phosphate. In processed foods to which potassium is added and in supplements, the form of potassium is potassium chloride. In healthy people, approximately 85% of dietary potassium is absorbed. Most potassium (approximately 77–90%) is excreted in urine, whereas the remainder is excreted mainly in feces, with much smaller amounts excreted in sweat. Because most potassium that is filtered by the glomerulus of the kidney is reabsorbed (70–80%) in the proximal tubule, only a small amount of filtered potassium reaches the distal tubule. The majority of potassium in urine results from secretion of potassium into the cortical collecting duct, a secretion regulated by a number of factors including the hormone aldosterone. An elevated plasma concentration of potassium stimulates the adrenal cortex to release aldosterone, which in turn increases secretion of potassium in the cortical collecting duct.

Acid–Base Considerations

A diet rich in potassium from fruits and vegetables favorably affects acid–base metabolism because these foods are also rich in precursors of bicarbonate. Acting as a buffer, the bicarbonate-yielding organic anions found in fruits and vegetables neutralize noncarbonic acids generated from meats and other high-protein foods. In the setting of an inadequate intake of bicarbonate precursors, excess acid in the blood titrates bone buffer. As a result, bone becomes demineralized and calcium is released. Urinary calcium excretion increases. This state has been termed a 'low-grade metabolic acidosis.' Increased bone breakdown and

calcium-containing kidney stones are adverse clinical consequences of excess diet-derived acids. Diets rich in potassium with its bicarbonate precursors might prevent kidney stones and bone loss. In processed foods to which potassium is added and in potassium supplements, the conjugate anion is typically chloride, which cannot act as a buffer.

Adverse Effects of Insufficient Potassium

Severe potassium deficiency, which most commonly results from diuretic-induced potassium losses, is characterized by a serum potassium concentration of less than 3.5 mmol/l. The adverse consequences of hypokalemia are cardiac arrhythmias, muscle weakness, and glucose intolerance. Moderate potassium deficiency, which commonly results from an inadequate dietary intake of potassium, occurs without hypokalemia and is characterized by increased blood pressure, increased salt sensitivity, an increased risk of kidney stones, and increased bone turnover. An inadequate intake of dietary potassium may also increase the risk of stroke and perhaps other cardiovascular diseases.

Kidney Stones and Bone Demineralization

Because of its effects on acid–base balance, an increased dietary potassium intake might have favorable effects on kidney stone formation. In one large observational study of women (**Figure 1**), there was a progressive inverse relationship between greater intake of potassium and incident kidney stones. At a median potassium intake of 4.7 g/day (119 mmol/day), the risk of developing a kidney stone was 35% less compared to that for women with an intake of <2.0 g/day (52 mmol/day). In the one available trial, an intake of approximately 3.6–4.7 g/day (92–120 mmol/day) of potassium in the form of potassium citrate reduced the risk of recurrent kidney stones.

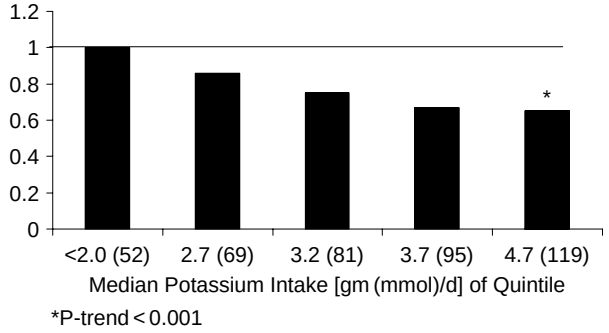


Figure 1 Relative risk of kidney stones during 12 years of follow-up by quintile of potassium intake in 91 731 women. (Data from Curhan GC, Willett WC, Speizer FE, Spiegelman D, and Stampfer MJ (1997) Comparison of dietary calcium with supplemental calcium and other nutrients as factors affecting the risk of kidney stones in women. *Annals of Internal Medicine* **126**: 497–504.)

Epidemiologic studies have consistently documented that increased potassium intake is associated with greater bone mineral density. In trials, supplemental potassium bicarbonate reduced bone turnover as manifest by less urinary calcium excretion and by biochemical evidence of greater bone formation and reduced bone resorption. However, no trial has tested the effect of increased potassium or diets rich in potassium on bone mineral density or clinical outcomes related to osteoporosis.

Elevated Blood Pressure

High levels of potassium intake are associated with reduced blood pressure. Observational data have been reasonably consistent in documenting this inverse relationship, whereas data from individual trials have been less consistent. However, three meta-analyses of these trials have each documented a significant inverse relationship between potassium intake and blood pressure in nonhypertensive and hypertensive individuals. In one meta-analysis, average net systolic/diastolic blood pressure reductions associated with a net increase in urinary potassium excretion of 2 g/day (50 mmol/day) were 4.4/2.4 mmHg. Typically, greater blood pressure reductions from potassium occur in African Americans compared to non-African Americans. Most of the trials that tested the effects of potassium on blood pressure used pill supplements, typically potassium chloride.

A high potassium intake has been shown to blunt the rise in blood pressure in response to increased salt intake. The term 'salt-sensitive blood pressure' applies to those individuals or subgroups who experience the greatest reduction in blood pressure when salt intake is reduced. One metabolic study of 38 healthy, nonhypertensive men (24 African Americans and 14 non-African Americans) investigated the effect of potassium supplementation on the pressor effect of salt loading (5.7 g/day of sodium (250 mmol)). Before potassium was supplemented, 79% of the African American men and 26% of the non-African American men were termed 'salt sensitive,' as defined by a salt-induced increase in mean arterial pressure of at least 3 mmHg. There was a progressive reduction in the frequency of salt sensitivity as the dose of potassium was increased. In the African Americans with severe salt sensitivity, increasing dietary potassium to 4.7 g/day (120 mmol/day) reduced the frequency of salt sensitivity to 20%, the same percentage as that observed in non-African American subjects when their potassium intake was increased to only 2.7 g/day (70 mmol/day).

Other studies indicate that potassium has greater blood pressure lowering in the context of a higher salt intake and lesser blood pressure reduction in the setting of a lower salt intake. Conversely, the blood pressure reduction from a reduced salt intake is greatest when potassium intake is low. These data are consistent with subadditive effects of reduced salt intake and increased potassium intake on blood pressure.

Cardiovascular Disease

The beneficial effects of potassium on blood pressure should reduce the occurrence of blood pressure-related cardiovascular disease. Potassium may also have protective effects that are independent of blood pressure reduction. This possibility has been tested in experimental studies conducted in rodents. In a series of animal models, the addition of either potassium chloride or potassium citrate markedly reduced mortality from stroke. Interestingly, these reductions occurred when blood pressure was held constant. Such data indicate that potassium has both blood pressure-dependent and blood pressure-independent properties that are cardioprotective.

In many, but not all, epidemiologic studies, an inverse relationship between dietary potassium intake and subsequent stroke-associated morbidity and mortality has been noted. A few observational studies have also shown an inverse association between potassium intake and coronary heart disease. In a 12-year follow-up of 859 men and women enrolled in the Rancho Bernardo Study, a significant inverse relationship between potassium intake and subsequent risk of stroke-related mortality was documented. Similarly, during the course of 8 years of follow-up in 43 738 US men in the Health Professionals Follow-Up Study, there was a significant inverse relationship between baseline potassium intake and stroke after adjustment for established cardiovascular disease risk factors, including blood pressure and caloric intake (**Figure 2**). In this study, a median potassium intake of 4.3 g/day (110 mmol/day) was associated with a 41% reduced risk of stroke in comparison to those with a median intake of 2.4 g/day (61 mmol/day). Consistent with these studies are other observational studies that have repeatedly documented a reduced risk of stroke from an increased intake of fruits and vegetables.

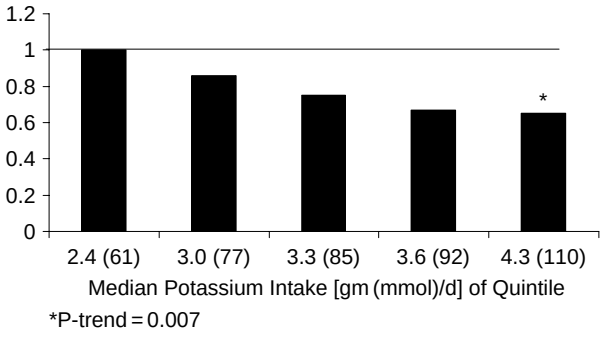


Figure 2 Relative risk of ischemic stroke by quintile of potassium intake in 43 738 men. (Data, from Ascherio A, Rimm EB, Hernan MA *et al.* (1998) Intake of potassium, magnesium, calcium, and fiber and risk of stroke among U.S. men. *Circulation* **98**: 1198–1204.)

Adverse Effects of Excess Potassium Intake

In the generally healthy population with normal kidney function, a high potassium intake from foods poses no risk because excess potassium is readily excreted in the urine. In contrast, supplemental potassium can lead to acute toxicity in healthy individuals. Also, in individuals whose urinary potassium excretion is impaire a potassium intake less than 4.7 g/day (120 mmol/day) is appropriate because of adverse cardiac effects (arrhythmias) from hyperkalemia. Drugs that commonly impair potassium excretion are angiotensin converting enzyme inhibitors, angiotensin receptor blockers, and potassium-sparing diuretics. Common medical conditions associated with impaired potassium excretion are diabetes, chronic renal insufficiency, end stage renal disease, severe heart failure, and adrenal insufficiency. Elderly individuals are at increased risk of hyperkalemia because they often have one or more of these conditions or take one or more of the medications that impair potassium excretion.

Recommended Potassium Intake, Current Intake, and Dietary Sources

On the basis of available data, an Institute of Medicine committee set an Adequate Intake for potassium at 4.7 g/day (120 mmol/day) for adults. This level of dietary intake should maintain lower blood pressure levels, reduce the adverse effects of salt on blood pressure, reduce the risk of kidney stones, and possibly decrease bone loss. Current dietary intake of potassium is considerably lower than this level.

Humans evolved on a diet that was rich in potassium and bicarbonate precursors and low in salt. However, contemporary Western-style diets have the opposite pattern—that is, relatively low content of potassium and high content of salt. Based on intake data from the Third National Health and Nutrition Examination Survey (NHANES-III,1988–1994), the

percentage of men and women who consumed equal to or more than 4.7 g/day (120 mmol/day) was less than 10 and 1%, respectively. Median intake of potassium in the United States ranged from 2.8 to 3.3 g/day (72 to 84 mmol/day) for adult men and 2.2 to 2.4 g/day (56 to 61 mmol/day) for adult women. The median potassium intake of non-African Americans exceeded that of African Americans. Because African Americans have a relatively low intake of potassium and a high prevalence of elevated blood pressure and salt sensitivity, this subgroup would especially benefit from an increased potassium intake.

Dietary intake surveys typically do not include estimates from salt substitutes and supplements. However, less than 10% of those surveyed in NHANES-III reported using salt substitutes or a reduced-sodium salt. Because a high dietary intake of potassium can be achieved through diet rather than pills and because potassium derived from foods also comes with bicarbonate precursors, as well as a variety of other nutrients, the preferred strategy to achieve the recommended potassium intake is to consume foods rather than supplements.

Dietary sources of potassium, as well as bicarbonate precursors, are fresh fruits, fruit juices, dried fruits, and vegetables. Although meat, milk, and cereal products contain potassium, their content of bicarbonate precursors does not sufficiently balance the amount of acid-forming precursors, such as sulfur amino acids, found in higher protein foods. The typical content of potassium-rich foods is displayed in **Table 1**. Salt substitutes currently available in the marketplace range from 0.4 to 2.8 g/teaspoon (11–72 mmol/teaspoon) of potassium, all as potassium chloride.

Table 1 Foods rich in potassium

Food	*Portion size*	*Potassium content, g (meq)*
Beans		
Cooked dried beans	1/2 cup	0.4 (10.7)
Lima beans	5/8 cup	0.4 (10.8)
Fruit		
Apple	1 medium	0.1 (2.8)
Apricots	3 medium	0.3 (7.2)
Banana	6 in.	0.4 (9.5)
Cantaloupe	1/4 medium	0.3 (6.4)
Dates	10 pitted	0.6 (16.6)
Orange	1 small	0.3 (7.7)
Peach	1 medium	0.2 (5.2)
Prunes, dried	10 medium	0.7 (17.8)
Raisins	1 tablespoon	0.1 (2.0)
Watermelon	1 slice	0.6 (15.4)
Fruit juices		
Grapefruit	1 cup	0.4 (10.4)
Orange	1 cup	0.5 (12.4)
Pineapple	1 cup	0.4 (9.2)
Tomato	1 cup	0.5 (13.7)
Vegetables		
Corn	1 ear	0.2 (5.0)
Potato		
– White	1 boiled	0.3 (7.3)
– Sweet	1 boiled	0.3 (7.7)
Tomato	1 medium	0.4 (9.4)
Squash, winter	1/2 cup boiled	0.5 (11.9)
Meats		
Hamburger	1 patty	0.4 (9.8)
Rib roast	2 slices	0.4 (11.2)
Fish (e.g., haddock)	1 medium fillet	0.3 (8.0)
Milk		
Skim milk	8 oz.	0.3 (8.5)
Whole milk	8 oz.	0.4 (9.0)

See also: **Bone**. **Electrolytes**: Acid-Base Balance. **Hypertension**: Etiology; Dietary Factors; Nutritional Management. **Osteoporosis**.

Conclusion

Potassium is an essential nutrient that is required for normal cellular function. Although humans evolved on diets rich in potassium, contemporary diets are quite low in potassium. An increased intake of potassium from foods should prevent many of the adverse effects of inadequate potassium intake, which are higher blood pressure levels, greater salt sensitivity, increased risk of kidney stones, and possibly increased bone loss. An inadequate potassium level may also increase the risk of stroke. In view of the high prevalence of elevated blood pressure, stroke, and conditions related to bone demineralization (i.e., osteoporosis and kidney stones) in the general population, individuals should strive to increase their consumption of potassium-rich foods, particularly fruits and vegetables.

Further Reading

Ascherio A, Rimm EB, Hernan MA *et al.* (1998) Intake of potassium, magnesium, calcium, and fiber and risk of stroke among U.S. men. *Circulation* **98**: 1198–1204.

Bazzano LA, Serdula MK, and Liu S (2003) Dietary intake of fruits and vegetables and risk of cardiovascular disease. *Current Atherosclerosis Reports* **5**: 492–499.

Curhan GC, Willett WC, Speizer FE, Spiegelman D, and Stampfer MJ (1997) Comparison of dietary calcium with supplemental calcium and other nutrients as factors affecting the risk of kidney stones in women. *Annals of Internal Medicine* **126**: 497–504.

He FJ and MacGregor GA (2001) Beneficial effects of potassium. *British Medical Journal* **323**: 497–501.

Institute of Medicine (2004) *Dietary Reference Intakes for Water, Potassium, Sodium, Chloride and Sulfate*. Washington, DC: National Academy of Sciences.

Lemann J, Bushinsky D, and Hamm LL (2003) Bone buffering of acid and base in humans. *American Journal of Physiology* **285**: F811–F832.

Morris RC Jr, Sebastian A, Forman A, Tanaka M, and Schmidlin O (1999) Normotensive salt-sensitivity: Effects of race and dietary potassium. *Hypertension* **33**: 18–23.

Whelton PK, He J, Cutler JA *et al.* (1997) Effects of oral potassium on blood pressure. Meta-analysis of randomized controlled clinical trials. *Journal of the American Medical Association* **277**: 1624–1632.

Poultry *see* **Meat, Poultry and Meat Products**

PREGNANCY

Contents

Role of Placenta in Nutrient Transfer

P Haggarty, Rowett Research Institute, Aberdeen, UK

Introduction

The main nutritional role of the placenta is to provide the correct mix of nutrients in sufficient quantities to support fetal growth and development throughout pregnancy. It has to do this whilst coping with wide variations in maternal nutrient intake between pregnancies and temporal variations within a pregnancy. The effective barrier to nutrients within the placenta is a single layer of cells called the syncytiotrophoblast, which close to term is around 4 μm thick with a total exchangeable surface area of around 10–15 m^2. This membrane represents a nutritional 'bottleneck' where competition for nutrient transporters and metabolic selectivity allows the placenta to regulate the nutrient mix within the fetal circulation to one best suited to fetal development. In addition to its role as a simple nutrient transporter the placenta also acts as an extra fetal organ for many metabolic transformations with the feto-placental unit working as a metabolic whole. Adaptive mechanisms within the placenta allow the mother to meet the nutrient demands of the growing fetus whist consuming apparently poor diets during pregnancy. The fetus itself plays an active role in regulating key aspects of placental metabolism and nutrient transfer function to meet its own nutrient requirements.

Under normal circumstances the nutrient transfer capacity of the human placenta exceeds the fetal requirement and a considerable proportion of transport function would have to be lost before it became limiting for fetal growth. Although relatively rare, intrauterine growth restriction resulting from uteroplacental insufficiency does occur. However, this is a complicated syndrome in which almost all aspects of placental and fetal metabolism are altered and it may not simply be due to a limitation of placental nutrient transfer capacity.

Fetal Nutrient Requirements

Prenatal development can usefully be divided into two periods: the embryonic period, which covers the

first 8 weeks of life, and the fetal period, which lasts from the 9th week of gestation until term. During the latter period the fetus is entirely dependent on the placenta for its supply of nutrients. The fetus has an absolute requirement for the same essential nutrients as the adult but the adequacy of supply is particularly critical during *in utero* life when all the structures of the body are being established. In addition, because of the particularly high demand for some strictly nonessential nutrients these may be considered as 'conditionally essential' if the rate of utilization exceeds the fetal capacity for *de novo* synthesis.

The placenta has to maintain the supply of all nutrients at a rate adequate to allow unrestricted fetal growth. It also has to provide an appropriate mix of nutrients to meet the needs of the fetus at the different stages of pregnancy. For example, in the first two-thirds of pregnancy the fetus deposits mainly protein, while in late gestation fat takes over as the dominant form of deposition (**Figure 1**).

The availability of individual nutrients to the fetus depends not only on the maternal dietary intake but also on the function of the placenta and the many physiological and biochemical adaptations that occur during pregnancy (**Figure 2**). An understanding of placental function and its interaction with diet is essential to the setting of appropriate dietary guidelines for pregnancy.

The Human Placenta

The human placenta is a hemochorial, villous type where the maternal blood enters the intervillous space via the spiral arteries and flows directly around the terminal villi of the fetal circulation without any intervening maternal vessel wall. The surface area available for exchange gradually increases throughout pregnancy until it reaches around 10–15 m^2 in the last trimester (**Figure 1**). The nature of the exchangeable surface of the placenta also changes throughout gestation with the mature intermediate villi appearing towards the end of the second trimester and the terminal villi, which represent the main site of feto-maternal exchange, appearing a few weeks later. The rate of fetal blood delivery to the placenta (umbilical flow) also changes markedly during pregnancy and is approximately linearly related to fetal weight, and hence the fetal nutrient requirement, throughout gestation (**Figure 1**).

Anatomically, the human placenta is a large structure typically weighing around half a kilogram. However, its physical bulk belies the flimsy nature of the separation between the maternal and fetal

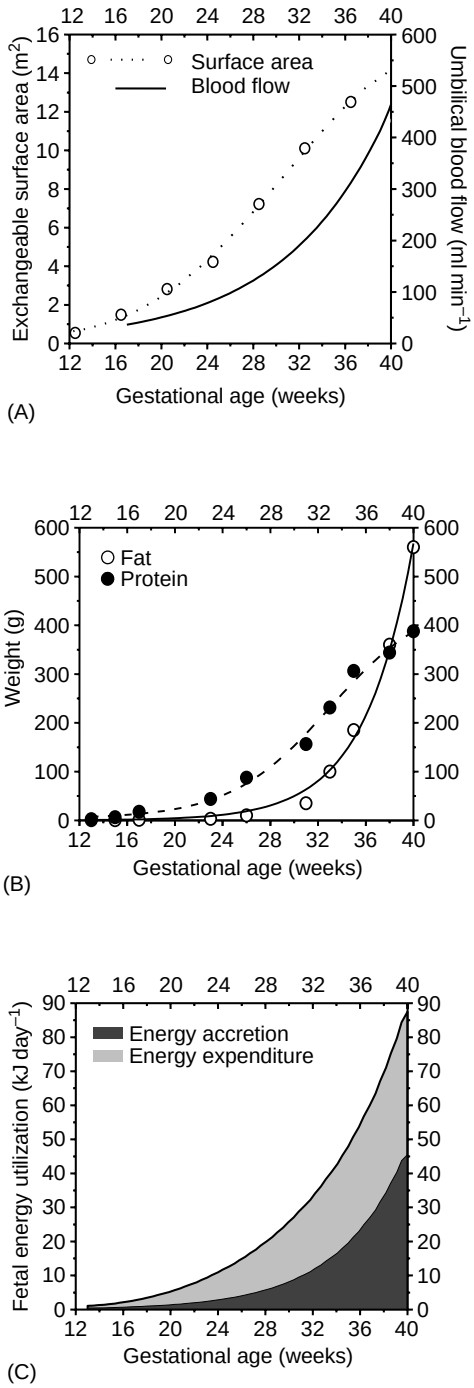


Figure 1 Changes with gestational age in placental exchangeable surface area and umbilical blood flow (A), accretion of fat and protein in the fetus (B), and the components of fetal energy requirements (C). (Reproduced with permission from: Sutton MS, Theard MA, Bhatia SJ, Plappert T, Saltzman DH, and Doubilet P (1990) Changes in placental blood flow in the normal human fetus with gestational age. *Pediatric Research* **28**: 383–387;Widdowson EM (1968) Growth and composition of the fetus and newborn. In: Assali NS (ed.) *The Biology of Gestation*, pp. 1–49 New York: Academic Press; Sparks JW (1984) Human intrauterine growth and nutrient accretion. *Seminars in Perinatology* **8**: 74–93.)

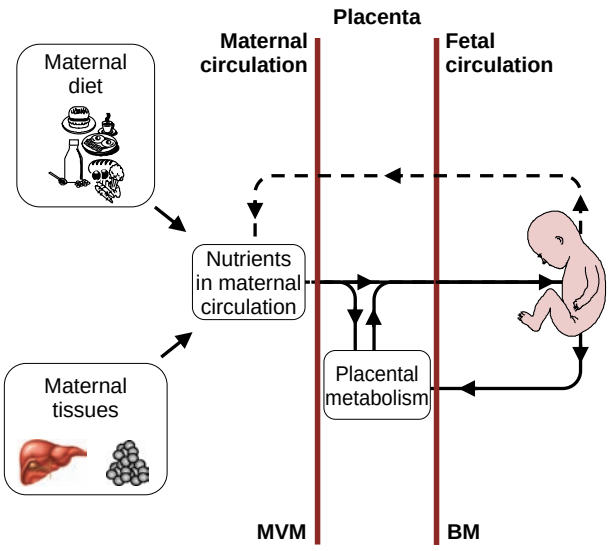


Figure 2 Nutrient exchanges between the maternal circulation, placenta, and fetus.

circulations, which consists of only two cell layers; the syncytiotrophoblast and the capillary endothelium. The endothelium allows the passage of nutrients through pores within the interendothelial cleft and therefore is not a significant barrier to nutrient passage. The effective barrier between the maternal and fetal circulation is provided by a thin trophoblastic cover in the form of a syncytium (a tissue in which the cytoplasm of constituent cells is continuous), known as the syncytiotrophoblast. Between 10 weeks and term the thickness of the villous trophoblast falls from around 10 μm to 4 μm and the overall maternofetal diffusion distance from 40 μm to 5 μm. Any substance crossing between the maternal and fetal circulation has to pass though this barrier, which consists of two membranes: the micovillous membrane (MVM) facing the maternal blood and the basal membrane (BM) facing the fetal blood. The surface area of the maternal-facing MVM is around 5–6 times that of the fetal-facing BM. There are other cell types and structures within the placenta, such as maternal myometrium and decidua, connective tissue, Hofbauer cells, and persisting cytotrophoblast cells, which contribute to the metabolic activity and nutrient requirements of the placenta but which are not thought to be significant barriers to transport.

Methods Used to Study Placental Function

Direct measurement of placental nutrient transport function in human pregnancy is practically and ethically extremely difficult to achieve. All of the available techniques have drawbacks and involve a trade-off between physiological relevance and the quality of the information derived. There is a very small number of reports of studies where stable isotope-labeled amino acids and fatty acids have been administered to the mother and their appearance measured in the cord blood. These studies have the potential to provide information on dynamic placental nutrient transfer rates *in vivo* but their interpretation is severely constrained by the number of sequential cord blood samples that can be taken, and the conclusions have therefore been necessarily tentative. Placental function is often inferred by measurements of concentration differences in the maternal and fetal circulations. The most sophisticated of these involve measurements of arterio-venous differences across the umbilical cord at Caesarean section before the cord is cut but such studies are therefore only carried out in very late gestation. Cord blood levels may also be measured following delivery or, more informatively, at earlier stages of development using the invasive method of cordocentesis. However, an important disadvantage of any 'snapshot' of cord blood nutrient concentrations is that these are the net result of both placental delivery and fetal utilization.

Because of the problems with interpretation of results from *in vivo* studies a number of *in vitro* approaches have been developed. These include the dually perfused placenta, which retains the cellular structure and metabolic activity of the syncytiotrophoblast and the placental vascular structure but allows the nutrient composition of the maternal and fetal circulation to be controlled and transfer rates to be measured dynamically using isotopic tracers. The problems with this *ex vivo* technique are that the placenta tends to be very mature, the efficiency of perfusion cannot be assumed to exactly mimic the *in vivo* situation, and the composition of the maternal and fetal perfusates are not truly physiological. More detailed but less physiologically relevant to absolute rates of transfer are vesicles formed from the syncytiotrophoblast, which are particularly well suited to the study of nutrient transport mechanisms under highly controlled conditions. The most reductionist methodology involved the identification and characterization of individual transport proteins.

The Mechanisms of Placental Nutrient Transport

The transport of individual nutrients across the placenta generally depends on the same principles

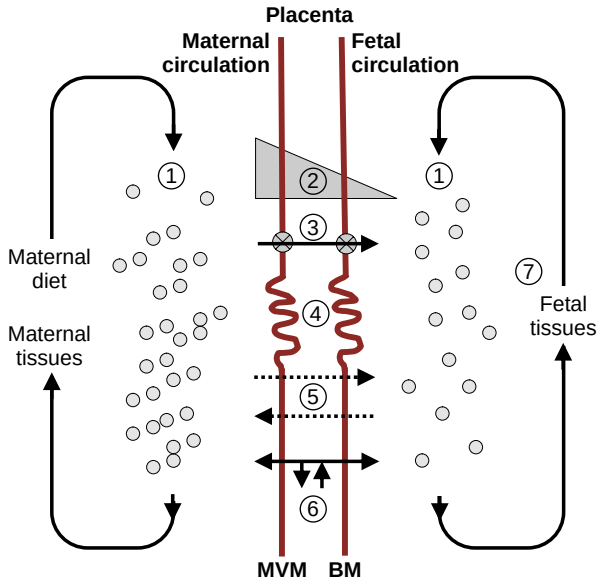


Figure 3 Factors affecting nutrient transfer across the syncytiotrophoblast. These include: (1) maternal and fetal blood flow; (2) the nutrient concentration gradient across the syncytiotrophoblast; (3) the concentration of transport proteins to facilitate or actively transport nutrients; (4) the exchangeable surface area; (5) the rate of diffusion of some nutrients across membranes without the intervention of transport proteins; (6) metabolism (utilization and *de novo* synthesis) within the placenta; and (7) the rate of nutrient utilization by the fetal tissues.

and the presence of the same or similar transport systems to those in the tissues and organs of the adult, although there are some additional factors specific to the placenta (**Figure 3**). In particular, unlike most tissues in the adult where either uptake or export dominate at any given time, the syncytiotrophoblast whose primary function is transport has to do both simultaneously.

The placental transport systems for the macronutrients (carbohydrate, fat, and protein) have been extensively studied. Glucose transport within the placenta appears to be mediated exclusively by the GLUT1 transporter, which has been located on both the MVM and BM. GLUT3 and GLUT4 are also present in the placenta but not in the syncytiotrophoblast itself. They are located on the vascular endothelium and the intravillous stromal cells, respectively. The syncytiotrophoblast also contains a wide range of amino acid transporters: system A, ASC, Asc, B^0, $b^{0,+}$, L, N, Gly, y^+, y^+L and X_{AG} and β. A number of fatty acid-binding proteins are also found in the placenta. Of these proteins FAT/CD36 and FATP have been located to both the MVM and BM but there is also a placenta-specific protein (p-FABPpm), which has been located exclusively on the MVM. This p-FABPpm is similar in size (~40 kDa) to the ubiquitous FABPpm found in most mammalian cells but it has a different amino acid composition.

The driving force that results in the net transfer of nutrients to the fetus is different for different nutrients and this is reflected in their transplacental gradients (**Figure 4**). Where the nutrient concentrations are lower in the cord than maternal blood this has been cited as a reason to supplement the mother but in many cases it is precisely this gradient that drives placental nutrient transfer. Glucose is thought to flow down a concentration gradient from the mother to the fetus and this process of 'facilitated diffusion' is mediated by GLUT1. Unlike glucose the concentration of most amino acids in the fetal circulation is greater than that in the maternal circulation suggesting some form of active transport. For many amino acids the concentration is even higher within the placenta than the fetal circulation and the key gradient generating step for amino acids is the active transport across the MVM. The amino acids can then diffuse down a concentration gradient into the fetal circulation, and to some extent back to the mother. The concentration of water-soluble vitamins and lactate in the fetal circulation also exceeds that in the maternal circulation.

Like glucose, the fats and fat-soluble vitamins also flow down a concentration gradient from the mother to the fetus mediated by the various fatty acid transport proteins. However, unlike glucose or the amino acids, fat-soluble compounds can also cross the syncytiotrophoblast, and all other membranes for that matter, by simple diffusion and partition without the intervention of a carrier protein. The role of the fatty acid-binding proteins appears to be to improve the efficiency of this process. The key factor in understanding the driving force for the placental transfer of fat-soluble nutrients is that these compounds are only sparingly soluble in water (13 μM for C18:0 at 37 °C) and have to be transported in the plasma in hydrophobic binding sites on carrier proteins. The partition of fats between the maternal and fetal circulations is largely determined by the relative abundance of available hydrophobic binding sites within those compartments. Since only NEFA are thought to cross membranes it is the NEFA concentration gradient that is most relevant to the transplacental flow of fatty acids. The concentration of NEFA in the maternal plasma at term is around 3 times that in the fetal circulation but the concentration of its primary carrier protein, albumin, is actually 10–20% higher in the fetal circulation. This results in a ratio of NEFA to albumin on the fetal side of the placenta of around a quarter of that on the

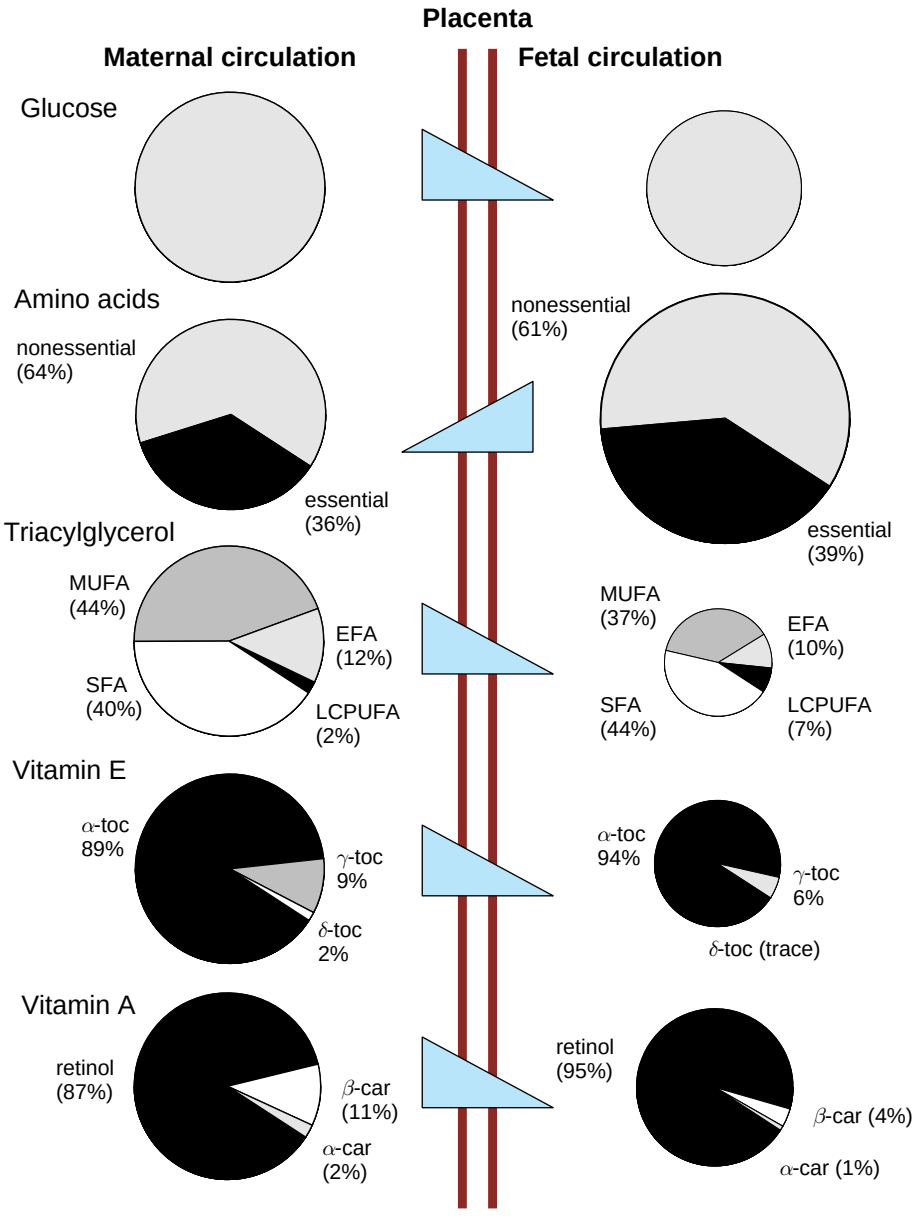


Figure 4 The relative concentration of nutrients in the maternal and fetal circulations. The concentration differences for each nutrient class are represented by the area of the circle in the fetal circulation relative to the maternal circulation. Apart from glucose the relative concentrations of individual nutrients within the nutrient groups are shown as segments of the circle. For triglyceride the fractions are saturated (SFA) monounsaturated (MUFA), essential (EFA), and long-chain polyunsaturated fatty acids (LCPUFA). For vitamin E the fractions are α-tocopherol (α-toc), γ-tocopherol (γ-toc), and δ-tocopherol (δ-toc). For vitamin A the abbreviated fractions are β-carotene (β-car) and α-carotene (α-car). (Reproduced with permission from: Berghaus TM, Demmelmair H, and Koletzko B (1998) Fatty acid composition of lipid classes in maternal and cord plasma at birth. *European Journal of Pediatrics* **157**: 763–768; Kiely M, Cogan PF, Kearney PJ and Morrissey PA (1999) Concentrations of tocopherols and carotenoids in maternal and cord blood plasma. *European Journal of Clinical Nutrition* **53**: 711–715; Cetin I., Marconi AM, Bozzetti P, Sereni LP, Corbetta C, Pardi G, and Battaglia FC (1988) Umbilical amino acid concentrations in appropriate and small for gestational age infants: a biochemical difference present *in utero*. *American Journal of Obstetrics and Gynecology* **158**: 120–126; Bozzetti P, Ferrari MM, Marconi AM, Ferrazzi E, Pardi G, Makowski EL, and Battaglia FC (1988) The relationship of maternal and fetal glucose concentrations in the human from midgestation until term. *Metabolism* **37**: 358–363.)

maternal side at term. The fat-soluble vitamins (A, E, and D) are also present in the fetal circulation in lower concentrations than in the maternal circulation. These materno-fetal concentration differences for the macronutrients develop gradually throughout gestation.

It is less easy to generalize about the transplacental gradient for minerals as some are at a

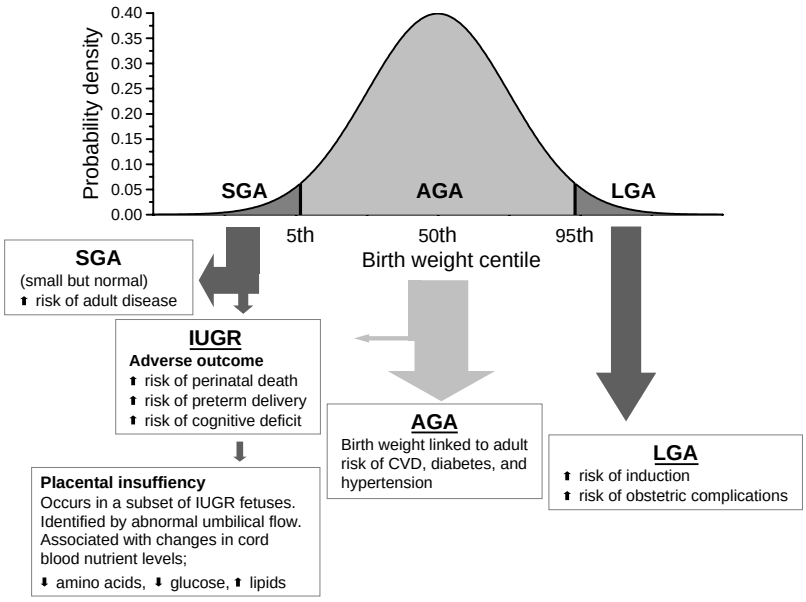


Figure 5 The normal distribution of birth weights and relative risks associated with babies that are small for gestational age (SGA), appropriate for gestational age (AGA), large for gestational age (LGA) and those subjected to intrauterine growth retardation (IUGR), and the relationship to placental insufficiency.

lower concentration in the fetal circulation (Se, Cu, Ba), some are higher (Ca, Zn, Be, Rb), and some are about the same (Co, Mg, Mo, Sn, Bi, Cd, Cs, La, Li, Pb). Iron is particularly important during pregnancy and its concentration in the fetal venous blood leaving the placenta is almost 3 times that of the maternal serum. Iron is transported in the serum on the transport protein transferrin and, like the fats and fat-soluble vitamins, its rate of transfer may be influenced by the availability of free binding sites.

Placental Selectivity

One of the key functions of placental nutrient transport is to maintain the most appropriate balance of nutrients in the fetal circulation and the balance of nutrients transferred by the placenta may be as important as the overall transfer capacity in influencing the pattern of fetal growth. Nutrients such as the fatty acids and amino acids occur in many forms yet they are translocated across membranes by a relatively small number of transporter molecules. This nutritional 'bottleneck' results in competition for transfer and the possibility of placental selectivity. An example of the resulting change in nutrient quality can be seen in the increase in the relative proportion of the essential to nonessential amino acids in the fetal circulation compared to the maternal circulation (**Figure 4**). The same is true of the fat-soluble vitamins where the relative concentration of the most biologically active form is increased in the fetal circulation. In the case of the fatty acids it is the long-chain polyunsaturated fatty acids (LCPUFA) such as arachidonic acid (20:4 n-6; AA) and docosahexaenoic acid (22:6 n-3; DHA) that perform most of the essential functions in the fetus. Although the overall concentration of the lipid classes are greatly reduced in the fetal circulation the critical LCPUFA make up a greater proportion of total fatty acid in the fetal circulation. In the case of the fatty acids the placenta has multiple mechanisms including preferential binding of LCPUFA by p-FABPpm, selective uptake by the syncytiotrophoblast, intracellular metabolic channeling of individual fatty acids, and selective export to the fetal circulation, which allow it to preferentially deliver DHA and AA to the fetal circulation.

Placental Metabolic Activity

Although the barrier between the maternal and fetal circulation is effectively only one cell thick, the placenta is a substantial organ made up of many cell types. It is extremely active metabolically and has its own requirement for nutrients and this is consistent with the observations that the surface area of the maternal-facing membrane (MVM) is around 5 times greater than that of the fetal-facing membrane (BM), that the concentration of expression of GLUT1 is greater on the MVM, that the MVM contains additional fatty acid binding

proteins that are not present on the BM, and that the amino acid transporters act to produce the maximum amino acid gradient across the MVM. The metabolic transformations within the placenta are intimately linked to fetal metabolism and represent another way in which the placenta can regulate nutrient transport availability within the fetal circulation.

In late pregnancy the overall contribution of fat to whole body oxidation is reduced and this is thought to result from the preferential utilization of carbohydrate and amino acids such as glutamate as an energy source in the feto-placental unit and the sparing of fatty acids to maximize fetal accretion of the critical LCPUFA in particular. The interrelationships between the placenta and fetus are particularly complex for the amino acids. The placenta is a net user of serine, glutamate, leucine, isoleucine, and valine and there is significant interconversion of alanine, pyruvate, and lactate between the placenta and fetal tissues. The concentration of lactate in the fetal circulation is considerably greater than that in the maternal circulation and a considerable proportion of the glucose taken up by the placenta is converted into lactate prior to export into the fetal circulation for use by the fetus. The placenta takes up serine from both the maternal and fetal circulation, converting this into glycine and exporting it into the fetal circulation for oxidation by the fetal liver and there is significant cycling of glutamate and glutamine between the placenta and fetal liver. This partition of the various segments of metabolic pathways between the placenta and fetal tissues is a general phenomenon and in many respects the feto-placental unit can be considered as a metabolic whole with the placenta acting as an extra fetal organ in addition to its role as a simple nutrient transporter. Metabolic activity in the feto-placental unit is also responsive to nutrient supply and fetal demand. For example, AA is an important precursor of the prostacyclins, prostaglandins, thromboxanes, and leukotrienes, which play key roles in pregnancy. When the maternal circulation of AA is low there is net uptake from the fetal circulation to maintain placental synthesis of these compounds.

Placental Buffering of Maternal Dietary Intake

In cases where the increased demand for nutrients during pregnancy is not met by the diet alone the shortfall may be made up from the maternal stores and the placenta may play a role in orchestrating some of the maternal nutritional adaptations in pregnancy. For example, placentally derived leptin is a potent stimulator of lipolysis and there is evidence that the rate of export into the maternal circulation is controlled to allow the placenta to modulate its own substrate supply in response to the fetal demand for fats. The various homeostatic mechanisms within the placenta and their interaction with maternal physiological adaptations during pregnancy act to ensure a constant supply of substrate to the fetus, free of large diurnal fluctuations corresponding to the timing of maternal meals, and to protect the fetus against a transiently poor intake during critical periods of fetal growth. These adaptations help the mother to meet the full fetal requirement for nutrients such as LCPUFA and iron whilst consuming apparently poor diets.

Placental Insufficiency and Fetal Growth

Potentially the most important public health issue relating to pregnancy is the epidemiological association between birth weight and adult disease susceptibility (cardiovascular disease, diabetes, and hypertension). The highest risk is associated with the lowest birth weight but, because of the nature of the normal distribution, in terms of the numbers potentially affected in adult life, it is the small variations in the normal birth weight range that have the largest public health implications. A causal connection between birth weight and adult disease has been proposed in the 'fetal origins' hypothesis, which is that fetal undernutrition in middle to late gestation leads to disproportionate fetal growth and programs later disease susceptibility. The close association between birth weight and placental weight has led to speculation that the placenta may limit fetal growth within the normal weight range. However, the available evidence suggests that the capacity of the human placenta to transport macronutrients exceeds the fetal requirement and that a considerable proportion of transport function would have to be lost before it became limiting for fetal growth.

True intrauterine growth restriction (IUGR) resulting from utero-placental insufficiency is a serious pathology that is associated with a greatly increased risk of adverse outcomes including perinatal mortality and morbidity, impaired mental, visual and aural development, autism, and cerebral palsy. IUGR is often detected indirectly by measuring abnormal umbilical artery flow velocity waveforms and/or abnormal fetal heart rate. The abnormal waveforms are thought to result from increased vascular resistance associated with abnormal arteriolar tree and villi branching and a reduction in the

villous capillary tree. Pregnancies in which these abnormalities are observed are also associated with fetal hypoxia and reduced concentrations of glucose and amino acids in the fetal circulation and reduced activity of the system A amino acid transporter within the placenta. However, *in vitro* studies have shown that the hypoglycemia observed in some IUGR fetuses is not caused by a decreased glucose transport capacity within the placenta (expression and activity of GLUT1) and IUGR fetuses are actually hypertriglyceridemic compared to their appropriately grown counterparts. The fetal blood concentrations of the trace elements are also either normal or elevated in IUGR. Thus, while it is possible that the placenta from IUGR fetuses may limit the supply of amino acids there is no evidence that placental delivery is the first limiting factor in the supply of glucose, lipids, or trace elements. IUGR is a complicated syndrome in which almost all aspects of placental and fetal metabolism are altered and many researchers have emphasized the primary importance of the fetal hypoxia and its effects on fetal metabolism rather than a simple limitation of placental nutrient transfer capacity.

There is considerable uncertainty about the magnitude of the problem of IUGR. The lowest 5% of weight-for-gestational-age babies (defined according to well-nourished fetal growth centile charts) are referred to as small for gestational age (SGA) but babies in this range need not be growth retarded but may be naturally small and have no increased risk of adverse outcome. A further complication is that a baby born within the normal birth weight range could have suffered growth retardation *in utero* if its genetic potential was for a higher birth weight. The true incidence of IUGR resulting from uteroplacental insufficiency is therefore unknown but if it is defined in relation to umbilical flow or fetal heart rate abnormalities then it is only a fraction of even those in the lowest 5% of weight-for-gestational-age that are affected by uteroplacental insufficiency. At the other end of the spectrum babies that are large-for-gestational-age (LGA) are at higher risk of adverse obstetric outcomes and early developmental problems but there is no evidence that LGA or macrosomic babies are produced as a result of a primary alteration in the placenta.

The Role of the Fetus

The nutrient composition of the human diet varies enormously among populations yet the healthy human newborn is essentially the same the world over. The available evidence points to extensive homeostatic mechanisms at work within the placenta to ameliorate some of the variation in the quality of the maternal diet by regulating the mix of nutrients to the developing fetus. However, these mechanisms can only operate on the nutrients already available in the maternal circulation. The maternal diet and maternal circulating concentrations of many nutrients are major determinants of the concentrations in the fetal circulation and the fetus clearly has the ability to cope with relatively large variations in nutrient availability in the cord blood. The fetus also plays an active role in regulating placental nutrient transfer. The rate of placental nutrient transport is directly influenced by the transplacental concentration gradient, which is in turn largely determined by the rate of uptake by the fetal tissues. Another major determinant of placental nutrient transfer is the umbilical blood flow, which is approximately linearly related to the fetal weight, and hence the fetal nutrient requirement, throughout gestation. Finally, the most intimate connection between the fetus and the placenta is the way in which different parts of metabolic pathways and cycles are distributed between the placenta and fetal tissues, mainly the fetal liver. Thus, whilst the placenta has to provide the correct mix of nutrients in sufficient quantities to support fetal growth and development throughout pregnancy it is the fetus itself that ultimately regulates many key aspects of placental nutrient transfer function.

Acknowledgements

The author acknowledges the support of SEERAD.

See also: **Early Origins of Disease**: Fetal. **Low Birthweight and Preterm Infants**: Causes, Prevalence and Prevention. **Pregnancy**: Nutrient Requirements; Energy Requirements and Metabolic Adaptations; Safe Diet for Pregnancy.

Further Reading

Cetin I (2003) Placental transport of amino acids in normal and growth-restricted pregnancies. *European Journal of Obstetrics, Gynecology, and Reproductive Biology*, **110**(supplement 1): S50–S54.

Dutta-Roy AK (2000) Transport mechanisms for long-chain polyunsaturated fatty acids in the human placenta. *American Journal of Clinical Nutrition* **71**: 315S–322S.

Gagnon R (2003) Placental insufficiency and its consequences. *European Journal of Obstetrics, Gynecology, and Reproductive Biology* **110**: S99–S107.

Haggarty P (2002) Placental regulation of fatty acid delivery and its effect on fetal growth – a review. *Placenta* **23**(supplement A): S28–S38.

Hay WW Jr (1995) Metabolic interrelationships of placenta and fetus. *Placenta* **16**: 19–30.
Illsley NP (2000) Glucose transporters in the human placenta. *Placenta* **21**(1): 14–22.
Jansson T, Ylven K, Wennergren M, and Powell TL (2002) Glucose transport and system A activity in syncytiotro phoblast microvillous and basal plasma membranes in intrauterine growth restriction. *Placenta* **23**(5): 392–399.
Kaufmann P and Scheffen I (1998) Placental development. In: Polin RA and Fox WW (eds.) *Fetal and Neonatal Physiology*, pp. 59–70. W.B. Philadelphia: Saunders Company.
Marconi AM, Paolini C, Buscaglia M, Zerbe G, Battaglia FC, and Pardi G (1996) The impact of gestational age and fetal growth on the maternal-fetal glucose concentration difference. *Obstetrics and Gynecology* **87**: 937–942.
Pardi G, Marconi AM, and Cetin I (2002) Placental-fetal interrelationship in IUGR fetuses – a review. *Placenta* **23**(supplement A): S136–141.
Sparks JW (1984) Human intrauterine growth and nutrient accretion. *Seminars in Perinatology* **8**: 74–93.
Regnault TR, de Vrijer B, and Battaglia FC (2002) Transport and metabolism of amino acids in placenta. *Endocrine* **19**(1): 23–41.

Nutrient Requirements

L H Allen, University of California at Davis, Davis, CA, USA

Providing pregnant women with their nutrient needs is a public health priority in both wealthier and poorer countries, although the local resources to attain this objective may vary widely. The inability to meet nutrient requirements during pregnancy can have serious and often long-term adverse effects on development during the fetal and postpartum period and on maternal health. Most of the research that provides information on nutrient requirements during pregnancy has been conducted in industrialized countries, although trials in developing countries have been important in revealing the adverse effects of maternal nutrition and the benefits of nutrient interventions. In general, even in wealthier countries there is an unacceptably high rate of pregnancy complications that may be prevented by improved maternal nutrition, including anemia, low birth weight, birth defects, and preeclampsia. The situation is far worse in poorer regions of the world, however.

The most recent and best described recommended intakes of nutrients during pregnancy are those of the Institute of Medicine, developed for the United States and Canada, and these are the main set presented in this article (**Table 1**). The recommendations for the United Kingdom were published in 1992 and are discussed here when recommendations differ substantially from those of the Institute of Medicine. Many other countries have their own sets of recommendations, as do organizations such as the Food and Agriculture Organization/World Health Organization and the European Economic Community.

The set of Dietary Reference Intake (DRI) recommendations developed by the Institute of Medicine includes several values. The Estimated Average Requirement (EAR) is the intake required to meet the nutrient needs of 50% of a population group (e.g., pregnant women). It is an important value for two reasons. First, it is the value used to estimate the prevalence of inadequate intakes of a nutrient in a population group; the percentage of a group consuming less than the EAR of a nutrient is the percentage with an inadequate intake. For energy, the Estimated Energy Requirement is equivalent to the EAR. Second, the Recommended Dietary Allowance (RDA) is calculated by adding two standard deviations (usually unknown but assumed to be 20%) to the EAR. The RDA should meet the requirements of 97.5% of a population group. The Tolerable Upper Level (UL) for a nutrient is the intake above which there is a risk of adverse effects.

Table 1 shows the RDA for nonpregnant women and the EAR, RDA, and UL for pregnant women.

Energy

Maternal energy requirements increase during pregnancy due to higher basal energy expenditure as well as energy deposition in maternal and fetal tissues. Basal metabolism of the mother is higher due to the increased work by the lungs and heart and because of the metabolism of the fetus and uterus. A longitudinal study by Butte *et al.* found that basal metabolic rate increased by 10.7 ± 5.4 kcal per week of gestation, mostly in the second and third trimesters. On average, the fetus requires approximately 168 kcal/day. The substantial variability in basal energy expenditure among individual women is caused mainly by differences in fat-free mass (including maternal skeletal muscle mass and fetal tissue). The cumulative increase in basal energy expenditure during pregnancy is positively correlated with maternal fatness and weight gain. Energy requirements for the thermic effect of feeding are not different from those of nonpregnant women, nor is there much change in the total energy cost of activity. Although the increasing body weight of the mother means that the energy cost of each activity is higher, the net effect is cancelled out by the fact that after approximately 25 weeks of gestation

Table 1 Recommended Dietary Allowances (RDAs) for nonpregnant and pregnant women and Estimated Average Requirements (EARs) and Upper Limits of nutrients for pregnant women[a]

	AI/RDA,[b] adult woman	*EAR, pregnancy*	*AI/RDA,[b] pregnancy*	*Upper Limit, pregnancy*
Energy (kcal)	2200	2500	—	—
Energy (MJ)	9.2	10.5	—	—
Protein	50	+21	+25	None
Vitamins				
Vitamin A (μg retinol activity equivalents)	800	550	770	3000
Vitamin D (μg)	5	—	5	50
Vitamin E (mg α-tocopherol)	15	12	15	1000
Vitamin K (μg)	90	—	90	None
Vitamin C (mg)	75	70	85	2000
Folate (μg dietary folate equivalents)	400	520	600	1000 from fortified food + supplements
Thiamin (mg)	1.1	1.2	1.4	None
Riboflavin (mg)	1.1	1.2	1.4	None
Vitamin B_6 (mg)	1.3	1.6	1.9	100 as pyridoxine
Niacin (mg NE)	14	14	18	35
Vitamin B_{12} (μg)	2.4	2.2	2.6	None
Pantothenic acid (mg)	5	—	6	None
Biotin (μg)	30	—	30	None
Choline (mg)	425	—	450	3500
Minerals and trace elements				
Calcium (mg)	1000	—	1000	2500
Phosphorus (mg)	700	580	700	3500
Magnesium (mg)	320	300	360	+350 as supplement
Iron (mg)	18	22	27	45
Zinc (mg)	8	9.5	11	40
Iodine (μg)	150	160	220	1100
Copper (μg)	900	800	1000	10 000
Selenium (μg)	55	49	60	400
Chromium (μg)	25	—	30	None
Fluoride (mg)	3	—	3	10
Manganese (mg)	1.8	—	2	11
Molybdenum (μg)	34	40	50	2000

[a]Dietary Reference Intakes published by the Institute of Medicine, National Academy Press, for the United States and Canada (http://www.nap.edu).
[b]Values are RDAs except for vitamins D and K, pantothenic acid, biotin, choline, calcium, chromium, fluoride, and manganese, where value is an Adequate Intake (AI).

women tend to become less active. The longitudinal study by Butte *et al*. suggests that energy expenditure in physical activity decreases by approximately 100–200 kcal/day in women with a low or normal body mass index prior to pregnancy and by an average of more than 400 kcal/day in those with a high body mass index ($>26\,kg/m^2$).

In deriving the recommendations for the United States and Canada, the Estimated Energy Requirement (EER) during pregnancy is accepted to be the sum of the Total Energy Expenditure (TEE) of the nonpregnant woman, measured using a doubly labeled water technique, plus an estimated median change in TEE of 8 kcal/week, plus 180 kcal/day to cover energy deposited in maternal and fetal tissues. In the first trimester of pregnancy, TEE changes little and weight gain is small, so the energy requirement is increased only during the second and third trimesters. There is no RDA or UL because energy intakes greater than the EER would lead to undesirable weight gain.

The EER for pregnancy is as follows:

Trimester 1: nonpregnant EER + 0 kcal

Trimester 2: nonpregnant EER + 160 kcal (based on 8 kcal/week × 20 weeks) + 180 kcal

Trimester 3: nonpregnant EER + 272 kcal (based on 8 kcal/week × 34 weeks) + 180 kcal

Note that these formulae present average requirements in trimesters 1 and 2. If a more precise estimate of requirements is needed at a specific stage of

gestation, instead of the mean increment of 160 kcal in trimester 1 and 272 kcal in trimester 2, the actual weeks of gestation can be multiplied by 8 kcal per week.

The UK recommendation is for an additional 200 kcal (0.8 MJ)/day above the prepregnant EAR but only in the last trimester. The recommendation is lower than that in the United States and Canada, in part because of the observation that the actual increase in energy intake during pregnancy is usually small.

Protein

The turnover of body protein is higher after approximately 13 weeks of pregnancy, and the mother adjusts by losing less nitrogen as urea even during the first trimester. A woman who gains 12.5 kg of body weight has deposited 925 g of protein, the fetus gains 440 g, the uterus 166 g, expanded maternal blood volume contains 81 g, the placenta 100 g, and the increment in extracellular fluid 135 g. The mother probably stores some additional protein in her body, presumably in muscle. The EAR for all age groups is 0.88 g/kg/day protein or 21 g of additional protein/day. The RDA is 1.1 g protein/kg/day or 25 g/day.

One-third of the 925 g total protein deposition during the 40 weeks of pregnancy occurs in the second trimester and two-thirds in the third trimester. By the end of the third trimester, the US–Canada recommendations assume that an additional consumption of 17 g protein/day is required to meet the needs for protein deposition, and since about half of this occurs during the second trimester this amounts to 8 g/day. It is also assumed that no additional protein is needed in trimester 1, but for the last two trimesters consumption of an additional 21 g/day (a total of 1.1 g/kg/day) is recommended. Recommended protein intakes for UK women are that an additional 6 g should be consumed during all three trimesters.

No UL has been set for protein, including for pregnancy, in the US–Canada recommendations due to lack of data on harmful effects. However, some earlier studies noted adverse pregnancy outcomes when high-protein supplements were given to relatively well-nourished pregnant women, so caution in this regard is certainly warranted.

Vitamins

Folic Acid

Maternal folate requirements increase markedly during pregnancy due to the utilization of the vitamin in cell division in the mother and fetus, single-carbon transfer reactions, and deposition in the fetus. Approximately a decade ago, research including randomized controlled trials finally proved that the risk of women giving birth to an infant with a neural tube defect (NTD) was significantly reduced if they consumed folic acid supplements prior to conception through approximately the first 4–6 weeks of pregnancy—during the time of neural tube closure. Some women are at greater risk of producing an infant with this birth defect, especially when their folate intake is rather low. Because such women are unaware of this risk unless they have had a previous NTD delivery, the recommendation is that all women who are capable of becoming pregnant consume at least 400 μg of folic acid daily from supplements, fortified food, or both in addition to consuming food folate from a varied diet.

In pregnancy, the recommendation is for all women to consume an additional 200 μg dietary folate equivalents daily (approximately 100 μg of folic acid as a supplement, which is more than twice as bioavailable as folate in food) in addition to the RDA for the nonpregnant woman of 400 μg/day. This amount was shown to prevent plasma homocysteine from becoming elevated during pregnancy and to maintain normal folate concentration in red blood cells. The UL of 1000 μg/day, the same as for nonpregnant women, is set to avoid potential exacerbation of vitamin B_{12} deficiency.

In the United Kingdom, the recommendation is substantially lower—an intake of 100 μg folate daily in addition to the recommendation of 200 μg/day for the nonpregnant, nonlactating woman. The UK committee's recommendation was based on the assumption that 100 μg/day will maintain plasma and erythrocyte folate concentrations at least at the level of those of nonpregnant women. Prevention of NTDs was not discussed, probably in part because the results of folic acid intervention trials were not clear at the time the recommendations were set.

In addition to its importance for lowering risk of NTDs in the periconceptional period, there is evidence that adequate folate status, which is important for maintaining normal plasma homocysteine concentrations, lowers the risk of other delivery problems and birth defects, including preeclampsia, preterm delivery, very low birth weight, club foot, and placental abruption. In the United States, Canada, and many other countries (more than 20 in Latin America alone), wheat flour is fortified with folic acid to ensure adequate folate status for pregnant women.

Other B Vitamins

Several B vitamin deficiencies cause homocysteinemia, notably folic acid, vitamin B_{12}, riboflavin, and vitamin B_6. Importantly, homocysteinemia is associated with adverse pregnancy outcomes. In a large retrospective study in Norway, for example, women in the highest 25% of plasma homocysteine concentrations had significantly more placental abruption, stillbirths, very low-birth-weight and preterm infants, preeclampsia, club foot, and NTDs in their offspring compared to women with values in the lowest 25%. Supplementation with folic acid up to 500–600 μg/day lowers plasma homocysteine, but few studies have been done on the other B vitamins. Of these, it is most difficult for poor women to obtain their dietary vitamin B_{12} requirement because this vitamin in found only in animal source foods, such as meat and dairy products.

The recommended intakes of most B vitamins and choline are increased above nonpregnant values as shown in **Table 1** The increases are based on evidence for higher maternal requirements (in the case of thiamin, riboflavin, niacin, and vitamin B_6) and for fetal and placental deposition of the vitamin (thiamin, riboflavin, niacin, vitamin B_6, vitamin B_{12}, and choline). UL values, the same as for nonpregnant women, have been set for niacin when consumed as nicotinic acid in supplements based on a 'flushing' reaction and for choline based on cholinergic reactions and a fishy body odor.

Vitamin A

The increment in vitamin A requirements during pregnancy is based on the amount of the vitamin that is found in fetal liver at birth. The liver content is assumed to be 36 μg, mostly accumulated during the last 3 months of gestation. Using an estimated 70% absorption of the vitamin from the maternal diet, the EAR is 50 μg above the requirement for the nonpregnant woman, whereas the RDA is 20% higher (60 μg).

In wealthier regions of the world, vitamin A deficiency during pregnancy is rare. Rather, there is more concern about the potentially adverse effects of consuming excessive amounts of the vitamin. Based on the potential for retinol excess to cause birth defects (malformations), especially if high doses are consumed early in pregnancy, a UL of 3000 μg/day is set for all women who may become pregnant as well as those who are pregnant. This intake is unlikely to be achieved with natural food, although it would be possible if large amounts of liver, foods fortified with the vitamin, or supplements were consumed. One situation in which this restrictive UL becomes important is in the context of developing countries where high-dose vitamin A supplements are provided to postpartum women and their infants as part of the Expanded Program on Immunization, National Vitamin A Days, or similar programs. It is accepted that it is only safe to provide these high-dose supplements to the mother during the first 6 weeks postpartum, in case she becomes pregnant again.

Nevertheless, it is important to provide pregnant women with their recommended intake of the vitamin because one major study in Nepal showed a 40% reduction in infection-related maternal mortality by supplementing the women with approximately their RDA as retinol per week. Supplementation with β-carotene reduced mortality by 49%, and it is a nontoxic alternative. Additional trials to confirm the benefits of maternal supplementation with the vitamin in deficient populations are ongoing.

Vitamin D

In the form of 25(OH) cholecalciferol, vitamin D is transferred from the mother to the fetus in relatively small amounts that do not appear to cause maternal depletion. Those women who obtain adequate exposure to ultraviolet light do not need higher amounts during pregnancy. However, if usual intake declines below 150 IU (3.8 μg)/day at high latitudes (where there is little ultraviolet radiation in the winter, such as in France), evidence of low maternal 25(OH) cholecalciferol and infant depletion has been observed at delivery.

The recommendation for both adolescent and adult women is to continue to consume the amount recommended as adequate (the Adequate Intake (AI)) for nonpregnant women, 5 μg (200 IU/day). The UL of 50 μg (2000 IU/day) is the same as before pregnancy, based on prevention of high serum calcium concentrations. In the United Kingdom, the recommended intake is higher at 10 μg/day, which is probably appropriate based on its generally more northern latitude (thus less ultraviolet radiation) and lower synthesis of the vitamin in skin.

Vitamin C (Ascorbic Acid)

The EAR for nonpregnant women is based on the intake that attains the maximum neutrophil concentration of ascorbic acid. Maternal plasma vitamin C concentrations decline during pregnancy, probably as a result of normal hemodilution. Oxidized ascorbic acid is transferred from the maternal circulation to the fetus, where it is retained in the reduced form. Although vitamin C deficiency in pregnancy is rare

in most situations, it has been associated with premature rupture of the membranes, increased risk of infections, preterm birth, and eclampsia. Smokers have lower levels of ascorbic acid in their serum and amniotic fluid. Based on the amount known to prevent infants from developing scurvy, the EAR is increased by 10 mg/day to 66 mg/day for those 14–18 years old and to 70 mg/day for adult women, and the RDA is 80 and 85 mg/day for these groups, respectively. The recommended intake is also increased by 10 mg/day in the United Kingdom. Women who smoke more than 20 cigarettes per day and regular aspirin users may require twice as much, as may heavy users of alcohol and street drugs. The UL of 2000 mg/day is based on prevention of diarrhea and gastrointestinal disturbances that occur with high intakes.

Vitamin E

There is no increase in the recommended intake of vitamin E during pregnancy, so the RDA remains at 15 mg of α-tocopherol/day for all ages. There have been no reports of deficiency of vitamin E during pregnancy nor any evidence of benefit from maternal supplementation. The UL is 1000 μg/day of any form of the vitamin taken as a supplement, extrapolated from data showing that high levels cause hemorrhaging in rats.

Minerals

Calcium

It has become recognized relatively recently that changes in maternal calciotropic hormones and calcium metabolism (i.e., increased intestinal absorption and reduced urinary excretion) enable the fetus to be supplied with adequate amounts of this mineral, and that little change in maternal intake is needed. There is no correlation between the number of pregnancies a woman has and her risk of bone fracture, so the maternal skeleton does not serve as the calcium reservoir for the fetus. Thus, for the United States and Canada there is no increase in recommended calcium intakes for pregnancy and the AI recommendation remains at 1300 mg/day for women aged 14–18 years and 1000 mg/day for the 18- to 51-year-old group. In the United Kingdom, the recommendation is also that no increase in intake is required during pregnancy, although the level of intake for nonpregnant, nonlactating women is considerably lower at 700 mg/day.

In a series of 14 randomized, controlled calcium intervention studies in different countries, increasing calcium intake in the range of 375–2000 mg/day reduced maternal blood pressure and the risk of pregnancy-induced hypertension and preeclampsia by 30–40%, with a greater effect in populations that consumed diets relatively low in calcium. The multicenter Calcium for Preeclampsia Prevention trial on 4589 pregnant women in the United States found no such benefits of a 2000 mg/day supplement, presumably because of reasonably high usual intakes of the mineral. It is possible that women at higher risk of pregnancy-induced hypertension, such as those with very low calcium intakes or adolescents, may benefit from calcium supplementation.

The UL for calcium in pregnancy is the same as that for the nonpregnant woman, 2500 mg/day. This safe level is set based on documented cases of 'milk-alkali syndrome,' in which there is high blood calcium, renal failure, and sometimes metabolic alkalosis as a result of chronic consumption of high calcium intakes.

Phosphorus

The efficiency of phosphorus absorption increases by 15% during pregnancy. The term infant contains approximately 17 g of phosphorus at birth, mostly in bone and water. The physiological adaptations of the mother that increase calcium retention also help to supply the fetus with more phosphorus. There is no evidence that the EAR needs to increase over that recommended for the nonpregnant women, so the RDA for women aged 14–18 years is 1250 mg/day and for those aged 19–50 years it is 700 mg/day. Based on the need to avoid high serum phosphorus concentrations, and the fact that phosphorus absorption is more efficient in pregnancy, the UL is set at 3500 mg/day, slightly lower than the 4000 mg/day for nonpregnant women.

Magnesium

It is assumed that the gain in fat-free mass in pregnancy (7.5 kg) is associated with a greater deposition of magnesium. If this tissue contains 470 mg/kg, after adjustment for a bioavailability of 40%, the EAR is an increase of 35 mg/day for pregnant women of all ages, and the RDA is 10% higher than this; for women aged 14–18 years, the EAR and RDA respectively are 335 and 400 mg; for those aged 19–30 years, these values are 290 and 350 mg; and for those 31–50 years, they are 300 and 360 mg. In the United Kingdom, there is no increment for magnesium in pregnancy based on the assumption that phosphorus metabolism becomes more efficient to meet fetal needs.

The UL for magnesium in pregnancy is set at 350 mg/day taken as a supplement, based on the potential for higher doses of magnesium salts to cause an osmotic diarrhea.

Iron

Incremental iron requirements for the mother and fetus are relatively well established, although how these requirements should be met is more controversial. It is generally accepted that the mother needs to absorb an additional 6 mg/day to supply the amount retained by the fetus (300 mg) and placenta (60 mg) and that used to synthesize additional maternal erythrocytes (450 mg) and replace blood loss during delivery (200 mg). Some iron is saved by the lack of menstruation in pregnancy. The fetus obtains iron from the placenta in a process that involves iron transfer from maternal transferrin to transferrin receptors on the placenta, endocytosis of holotransferrin, and release of iron into the fetal circulation. Maternal iron absorption and transfer to the fetus increases during the second and third trimesters. This process is upregulated if the mother is iron deficient, although in recent years it has become apparent that maternal iron deficiency does reduce the amount of fetal iron stored at birth and available to the fetus during the first months of life.

The EAR for pregnancy is set at 23 mg/day for adolescents and 22 mg/day for adult women, and the RDA is 27 mg/day for both groups. Although the requirement is mainly in the last trimester, it is important to build iron stores early and to avoid high doses later, so the higher intake recommendation is distributed throughout pregnancy. The UL is the same as that for the nonpregnant woman and is based on the need to avoid gastrointestinal distress.

It has been calculated that the maternal diet can supply enough iron to meet these increased needs during pregnancy, especially if maternal iron stores are adequate at conception. For this reason, the United Kingdom does not recommend that iron intake be increased during pregnancy, except when there is evidence of iron deficiency anemia. Iron deficiency anemia is a relatively common occurrence during pregnancy, especially in the following situations: Maternal iron status is poor at conception, and maternal diet is low in absorbable iron including heme iron from meat, fish, and poultry. The World Health Organization estimates that approximately 18% of women in industrialized countries and 35–75% of those in developing countries develop iron deficiency anemia during pregnancy. In the United States, the Centers for Disease Control and Prevention reports that anemia affects 10% of low-income women in the first trimester, 14% in the second, and 33% in the last, with a much higher proportion of women becoming iron depleted by term. Accepted cut points for adequate hemoglobin concentration are 110 g/l in trimesters 1 and 3 and 105 g/l in trimester 2 due to midpregnancy hemodilution.

In most countries, iron supplements are recommended routinely for all pregnant women. Benefits clearly include reduction of anemia risk, improved maternal and iron status that can persist through the early postpartum period, and possibly some protection against low birth weight. The amount recommended has been reduced from former levels of 60–120 mg to 30 mg for nonanemic women and 60 mg for anemic women. The World Health Organization recommends 60 mg/day plus 400 μg folic acid, starting as soon as pregnancy is confirmed, but recognizes that 30 mg/day may be as effective as 60 mg/day. The folic acid recommendation was originally set based on older studies showing development of folate deficiency anemia in women. Although the risk of this anemia is probably low on a global scale, folic acid supplementation is recognized to have other potential benefits. Some countries still recommend iron supplementation only when pregnant women become anemic. There has also been considerable controversy concerning the best time to start supplementation.

Zinc

The estimated additional zinc required for pregnancy is approximately 100 mg, equivalent to 5–7% of the mother's body zinc, part of which is obtained through more efficient intestinal zinc absorption. Approximately half of this is deposited in the fetus. The EAR for pregnant women is based on an additional requirement of 2.7 mg/day during the last 10 weeks of gestation. The UL is based on evidence of impaired copper status at high intakes, as for nonpregnant women. No increment is recommended for pregnancy in the UK report, based on the assumption that needs can be met through adjustments in maternal zinc metabolism.

Zinc plays critical roles in cell division, hormone metabolism, protein and carbohydrate metabolism, and immunocompetence. Because zinc deficiency in pregnant animals causes birth defects and fetal growth retardation, there has been considerable effort to determine the effects of human zinc status on pregnancy outcome, especially in developing countries, where zinc intakes

are often inadequate. In an analysis of 12 randomized, controlled intervention trials, only 2 (1 in India and 1 in the United States) found that zinc supplementation increased birth weight and reduced preterm delivery risk, whereas 6 found no effect. In the United States study, a positive effect was found in low-income, obese African American women with below average plasma zinc concentrations. Trials in Peru and Bangladesh showed no such benefits. In general, however, meeting recommended zinc intakes is more difficult but more critical for women whose diets are low in animal source foods and higher in fiber. High intakes (supplements) of iron and calcium may also impair zinc absorption and therefore increase requirements.

Iodine

In the many countries with endemic iodine deficiency, which include parts of the United States, Canada, and substantial areas of Europe and many other industrialized and developing countries, there is clear potential for the harmful effects of this deficiency to emerge during pregnancy. The most damaging effect of iodine deficiency is on the brain of the fetus since iodine is required for thyroid hormone, which in turn affects myelination and function of the developing central nervous system. The clinical expression of severe maternal iodine deficiency during pregnancy is cretinism, including severe mental retardation, deaf mutism, short stature, and spasticity. Injections of iodized oil before midpregnancy have markedly reduced cretinism and neonatal mortality in areas of severe iodine deficiency. In most countries, Universal Salt Iodization has reduced the prevalence of cretinism substantially, but milder indications of maternal deficiency persist even in Western Europe, including countries such as Belgium.

The EAR for pregnancy is set at 150 μg/day and the RDA at 160 μg/day for the United States and Canada based on the amount needed to prevent increased thyroid size in previously deficient women. The UL is 1100 μg/day, the same as for nonpregnant, nonlacting women, and it is based on the need to avoid elevated thyroid-stimulating hormone concentrations.

Trace Elements: Copper, Selenium, Chromium, Fluoride, Manganese, and Molybdenum

Copper is required for the function of many enzymes, primarily oxidases. In pregnancy, an increased intake of this mineral is recommended to cover deposition of approximately 18 mg/day, most of which is in fetal liver. The UL (10 000 μg/day) is the same as for nonpregnant women, based on the need to prevent the liver damage that occurs with high intakes.

Recommended intakes of selenium for adults are based on the criterion of maximizing plasma glutathione peroxidase activity. Based on an estimated selenium content of the fetus of 1000 μg, across pregnancy this would require that an additional 4 μg/day be consumed. The EAR is therefore increased from 45 to 49 μg/day and the RDA from 55 to 60 μg/day. The UL is determined on the basis of hair loss and brittle nails, which occur at higher levels of intake, and is the same as that set for nonpregnant women. An intake of 60 mg/day is also recommended throughout pregnancy in the United Kingdom, which is the same as the prepregnancy value for that population.

Chromium is required for normal insulin metabolism. There are no data from which to derive a recommendation for pregnancy, so an increase of 5 μg/day is recommended (as an AI) based on the additional weight and tissue chromium gained in pregnancy. No UL was set due to lack of documented adverse effects in humans.

For fluoride, there is no evidence that increasing the AI in pregnancy above that for the nonpregnant woman would benefit fetal tooth or bone content or afford protection against later tooth decay in the child. The UL is set at 10 mg/day to avoid fluorosis (discoloration of tooth enamel, joint pain, and skeletal abnormalities).

Manganese is required for bone formation and the normal metabolism of amino acids, lipids, and carbohydrates. The AI for pregnancy, estimated from the manganese content of maternal weight gain, is 2 mg/day. The UL is based on avoidance of elevated blood manganese and neurotoxicity, and it is not increased for pregnancy.

Recommended molybdenum intakes, based on the mineral's role as a cofactor for several enzymes, increase by 16 mg/day in pregnancy to cover the increment in fetal and maternal weight. The UL is derived from adverse reproductive effects seen in animals.

Water and Electrolytes

The US–Canada recommended intake of water for pregnant women is based on median intake from a large national survey in the United States. The AI of 3 l/day is anticipated to come from foods (0.7 l) and beverages (2.3 l). No UL was set because individuals stop drinking once their intake is adequate.

The AI for sodium in pregnancy is 1500 mg/day based on an intake level to cover daily losses, provide adequate intakes of other nutrients, and maintain normal function. The UL of 2300 mg/day is based on the adverse effects of higher intakes on blood pressure in susceptible members of the population.

The AI for potassium in pregnancy (4.7 g/day) is set at a level that will lower blood pressure, reduce the extent of salt sensitivity, and minimize the risk of kidney stones. There is no evidence that adverse effects of potassium are seen with high intakes from food and no UL was set, but potassium supplements can cause high blood potassium in some chronic diseases, such as renal disease and type 1 diabetes.

Summary

In the US–Canada recommendations, the recommended intakes are increased for most, but not all, nutrients during pregnancy. However, the recommendations are often based on less than ideal experimental data, in part due to the difficulty of conducting experiments on pregnant women.

For most nutrients, it is likely that some population groups may have higher requirements than those recommended in **Table 1**, notably women bearing more than one fetus or adolescents (see the Institute of Medicine volumes for specific recommendations for this age group). In order to meet the recommended nutrient increases, dietary quality often needs to be improved during pregnancy. It is often advised that pregnant women should also take iron supplements and/or a multiple vitamin–mineral supplement. The specific benefits of supplementation in pregnancy, optimal timing, and optimal doses are still somewhat controversial and the subject of ongoing research. Currently, some countries recommend routine supplementation for all pregnant women, whereas others recommend supplementation only when there is evidence of anemia, other nutritional deficiencies, a poor diet, or other problems such as drug or alcohol abuse.

See also: **Anemia**: Iron-Deficiency Anemia. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Calcium**. **Choline and Phosphatidylcholine**. **Chromium**. **Cobalamins**. **Copper**. **Folic Acid**. **Iodine**: Physiology, Dietary Sources and Requirements. **Iron**. **Magnesium**. **Manganese**. **Phosphorus**. **Potassium**. **Pregnancy**: Energy Requirements and Metabolic Adaptations; Safe Diet for Pregnancy. **Protein**: Requirements and Role in Diet. **Sodium**: Physiology. **Vitamin A**: Physiology; Biochemistry and Physiological Role. **Vitamin B_6**. **Vitamin E**: Metabolism and Requirements. **Zinc**: Physiology.

Further Reading

Allen LH (2000) Anemia and iron deficiency: Effects on pregnancy outcome. *American Journal of Clinical Nutrition* **71**(supplement): 1280S–1284S.

Allen LH (2001) Pregnancy and lactation. In: Bowman BA and Russell RM (eds.) *Present Knowledge in Nutrition*, 8th edn, pp. 403–415. Washington, DC: ILSI Press.

Berry RJ, Li Z, Erickson JD *et al.* (1999) Prevention of neural-tube defects with folic acid in China. China–U.S. Collaborative Project for Neural Tube Defect Prevention. *New England Journal of Medicine* **341**: 1485–1490.

Butte NF, Wong WW, Treuth MS, Ellis KJ, and O'Brian Smith E (2004) Energy requirements during pregnancy based on total energy expenditure and energy deposition. *American Journal of Clinical Nutrition* **79**: 1078–1087.

Institute of Medicine, six volumes on Dietary Reference Intakes. http://nap.edu

Kaiser LL and Allen LH (2002) Position of the American Dietetic Association: Nutrition and lifestyle for a healthy pregnancy outcome. *Journal of the American Dietetic Association* **102**: 1479–1490.

Murphy MM, Scott JM, Arija V, Molloy AM, and Fernandez-Ballart JD (2004) Maternal homocysteine before conception and throughout pregnancy predicts fetal homocysteine and birth weight. *Clinical Chemistry* **50**: 1406–1412.

Prentice A (2000) Maternal calcium metabolism and bone mineral status. *American Journal of Clinical Nutrition* **71**: 1312S–1316S.

Scholl TO, Hediger ML, Bendich A *et al.* (1997) Use of multivitamin/mineral prenatal supplements: Influence on the outcome of pregnancy. *American Journal of Epidemiology* **146**: 134–141.

Energy Requirements and Metabolic Adaptations

G R Goldberg, MRC Human Nutrition Research, Cambridge, UK

The subject of energy metabolism in human pregnancy has received extensive consideration for more than 60 years, dating back to early work that assessed the contribution of fetal metabolism to the overall energy costs of pregnancy. Since then, the emphasis of much work has been on separating and quantifying the different components of gestational energy needs and on establishing appropriate recommendations for the energy requirements of pregnant women, with the intention to quantify average amounts. Deviations from average values were mostly regarded as undesirable biological or measurement noise that needed to be

Table 1 Protein and fat deposition during pregnancy for a reference woman[a]

Site	Protein		Fat		Water (kg)	Total	
	kg	MJ (kcals)	kg	MJ (kcals)		kg	MJ (kcals)
Fetus	0.44	12.76 (3050)	0.44	20.24 (4840)	2.41	3.29	33.00 (7890)
Placenta	0.10	2.90 (690)	0.04	0.18 (43)	0.54	0.64	3.08 (740)
Amniotic fluid	0.003	0.09 (21)	0.00	0.00	0.79	0.79	0.09 (21)
Uterus	0.17	4.81 (1150)	0.04	0.18 (43)	0.80	0.97	5.00 (1200)
Breasts	0.08	2.35 (560)	0.12	0.55 (130)	0.30	0.40	2.90 (690)
Blood	0.14	3.92 (940)	0.02	0.92 (220)	1.29	1.44	4.84 (1157)
Water	0.00	0.00	0.00	0.00	1.50	1.50	0.00
Subtotal	**0.93**	**26.83 (6400)**	**0.48**	**22.08 (5280)**	**7.63**	**9.04**	**48.9 (11 700)**
Fat stores	0.07	1.94 (460)	2.68	123.10 (29 400)	0.60	3.35	125.04 (29 900)
Total	**0.99**	**28.77 (6900)**	**3.16**	**145.18 (34 700)**	**8.24**	**12.38**	**173.94 (41 600)**

[a]Adapted from Prentice AM, Spaaij CJK, Goldberg GR *et al.* (1996) Energy requirements of pregnant and lactating women. *European Journal of Clinical Nutrition* **50**(supplement 1): S82–S111.

overcome by studying large samples of women to get a more precise estimate of the mean values. These inter-individual variations in the metabolic responses to pregnancy are increasingly recognized as biologically significant 'plasticity' that has true adaptive value in enabling women to carry a pregnancy to term under a wide range of nutritional conditions. The shorter and longer term consequences of such adaptations are being explored as part of fetal and infant origins of adult disease hypotheses.

Extra Energy Costs of Pregnancy

The question of how much extra dietary energy a pregnant woman needs is closely linked to the question of the amount of weight she should gain during pregnancy. This in turn is linked to her age and to her prepregnant body mass index as a proxy for energy status.

Hytten and Leitch's theoretical estimations of the overall energy costs of human pregnancy published more than 30 years ago have subsequently been experimentally validated as reasonable average values, and they have been adopted by many national and international bodies as a partial basis for developing recommended energy intakes in pregnancy. The costs can be divided into three main components: the energy deposited as new tissue in the conceptus, the energy deposited as fat, and the energy required to maintain this new tissue.

Tissue Deposition

Weight gain during pregnancy consists of the fetus, placenta, and amniotic fluid (the products of conception) and the extra growth of several maternal tissues. The deposition of fat in pregnancy is presumed to help meet the extra energy demands of lactation. The total energy deposited as new tissue, excluding maternal fat, averages approximately 49 MJ (11 700 kcal). If an average maternal fat gain of 2.6 kg is assumed, then the estimate of the total energy deposited as new tissue during an average pregnancy is approximately 174 MJ (41 600 kcal) (**Table 1**).

Maintenance Energy Costs of Pregnancy

Because of the increase in tissue mass, the body's oxygen consumption also increases during pregnancy. Estimates suggest that the increase in oxygen consumption is equivalent to an extra 187 (45), 414 (100), 620 (148), and 951 (230) kJ/day (kcal) at 0–10, 10–20, 20–30, and 30–40 weeks of gestation, respectively. The total maintenance cost for an average human pregnancy is approximately 150 MJ (35 800 kcal) (**Table 2**).

Table 2 Increases in oxygen consumption during pregnancy[a]

	ml/min			
	10 weeks	20 weeks	30 weeks	40 weeks
Cardiac output	4.5	6.8	6.8	6.8
Respiration	0.8	1.5	2.3	3.0
Kidneys	7.0	7.0	7.0	7.0
Breasts	0.1	0.6	1.2	1.4
Uterus	0.5	1.2	2.2	3.6
Placenta	0	0.5	2.2	3.7
Fetus	0	1.1	5.5	12.4

[a]Adapted from Hytten FE (1991) Nutrition; Weight gain in pregnancy. In: Hytten F and Chamberlain G (eds.) *Clinical Physiology in Obstetrics*, 2nd edn. Oxford: Blackwell Scientific.

Theoretical Total Metabolic Costs of Pregnancy

Compared to many other mammals, humans have a relatively small and usually single infant, which develops during a long gestation period. The energy stress to the mother is therefore low per unit time. The 49 MJ of energy deposited as the products of conception represents only 4 or 5 days of food intake for the mother. Humans also differ from most other mammals because their large fat stores can help meet some of these costs. The theoretical total metabolic costs (i.e., due to extra tissue and increased metabolism) of pregnancy are approximately 335 MJ (80 000 kcal), or 1.25 MJ/day (300 kcal). This value does not make any allowance for changes (increases or decreases) in energy expended on physical activity. It is assumed that the majority of the energy costs of human pregnancy are met by behavioral adjustments in energy metabolism rather than increased energy intake. This assumption has formed the basis for energy intake recommendations, some of which are summarized in **Table 3**. It should be noted that the 1985 estimates used by WHO/FAO/UNU are under revision. Future recommendations may separate the obligatory costs (e.g., by fixed increments for basal metabolic rate (BMR) and tissue deposition) and differences in physical activity (based on PAL values).

Longitudinal Studies of the Energy Costs of Pregnancy

Fat Deposition

The increase in maternal fat stores is by far the largest contributor to the energy cost of tissue deposition. It is also the most variable. Although the average increase for a well-nourished woman who has an uncomplicated pregnancy and healthy infant is approximately 3 kg, a large number of studies have reported ranges of −2 to 8 kg and standard deviations of 2–4 kg. There is also a wide range in fat deposition between different populations, particularly when those from developed and developing countries are compared. Fat is very energy dense and therefore changes in body fat stores have a large impact on the energy costs of pregnancy. A loss of 2 kg saves approximately 78 MJ (18 600 kcal), whilst a gain of 8 kg costs approximately 312 MJ (74 600 kcal). Women most likely to need an energy reserve to help meet the costs of lactation are often those who are least able to deposit spare energy as fat in pregnancy. Conversely, women who store large amounts of fat during pregnancy are least likely to need to use it during lactation. They are often able to increase food intake and/or decrease physical activity instead.

Table 3 Examples of current recommendations for energy intakes during pregnancy

	Trimester(s)	*Increment, MJ/day (kcal/day)*	*Total for pregnancy, MJ (kcal)*	*Qualifying comments*
FAO/WHO/UNU (1985)	All	1.20 (300)	336 (80 300)	
	All	0.84 (200)	235 (56 150)	For healthy women who reduce activity Energy and protein requirements are undergoing revision (interim report published 2004)
United Kingdom (1991)	3rd	0.80 (190)	74 (17 000)	Underweight women and those not reducing activity may need more
United States and Canada (2002)	1st	Adult EER + 0		For women aged 19–50 years
	2nd	Adult EER + 160 kcal (8 kcal/week × 20 weeks) + 180 kcal		EERs for pregnant adolescents are based on EER for 14- to 18-year-olds
	3rd	Adult EER + 272 kcal (8 kcal/week × 34 weeks) + 180 kcal		EER is based on total energy expenditure in the nonpregnant state; increments for pregnancy are 8 kcal/week for total energy expenditure and 180 kcal/day for tissue deposition

EER, estimated energy requirement.

Studies have shown that excess energy intake during pregnancy results in excess maternal weight (and fat) gain. Postpartum retention of excess fat has implications for the development of obesity and its comorbidities such as type 2 diabetes.

Basal Metabolic Rate

The cumulative increase in BMR can comprise a large part of the total energy costs of pregnancy. Although 150 MJ is a good estimate of the average energy cost of maintenance for a well-nourished woman, there is a very wide range. This has an important influence on the extra daily requirements for individual women. Studies in which BMR has been measured every 6 weeks from prepregnancy to 36 weeks of pregnancy have shown very marked differences. In some women, there is the expected response to pregnancy—an immediate and progressive increase in BMR. In other women, BMR actually decreases or increases only slightly in the early stages of pregnancy and does not increase substantially until late gestation. This offsets the later increase in BMR such that there is actually a slight net saving of energy over the entire gestation period in some of these 'energy-sparing' women. The total net cost of maintenance, estimated as the cumulative area under the curve represented by the rise in a mother's BMR above the prepregnancy baseline metabolic rate, is negative or only very small. Data indicate that this between-subject variability is found in women from both well-nourished and marginally-nourished populations. However, 'energy-sparing' and 'energy-profligate' responses dominate in marginally and well-nourished women, respectively. There is a more than 5-fold range between the most energy-profligate and the most energy-sparing women.

In addition to the wide variability in changes in BMR between individual women, there are also wide variations between different populations. Well-nourished affluent women from developed countries tend to show an energy-profligate increase in BMR. In marginally nourished thinner women from developing countries the increase in BMR is delayed and/or preceded by a decline in early pregnancy. The total maintenance costs of pregnancy in these studies range from +210 MJ (+50 000 kcal) to −45 MJ (−11 000 kcal).

Diet-Induced Thermogenesis

A reduction in diet-induced thermogenesis (DIT) may be a mechanism by which energy is saved during pregnancy. However, when expressed as a proportion of energy intake, DIT remains essentially unaltered during pregnancy and any changes are small and unlikely to be biologically significant.

Energy Cost of Activities

Results from a number of longitudinal studies have shown that the cost of non-weight-bearing activity changes little until very late pregnancy. From approximately 35 weeks, the gross costs (which include changes in BMR) increase by approximately 11% and net costs by approximately 6%. The gross and net costs of weight-bearing exercise (treadmill walking and standardized step testing) remain fairly constant during the first half of pregnancy and then increase progressively by approximately 15–20% at term.

Behavioral Changes in Physical Activity

It has frequently been assumed that a behavioral reduction in the energy expended on physical activity helps to counteract the increases in expenditure due to increased body weight, and in some women this leads to saving of energy that largely meets the costs of pregnancy. However, although relatively small changes in activity patterns can potentially result in significant energy savings, there is little evidence that this occurs to a large extent. A possible reason for this is that affluent women are habitually so sedentary that there is little scope for further reduction. In contrast, in developing countries habitual levels of physical activity are high and there is therefore more potential for behavioral reductions. However, many women are likely to be unable to reduce their physical activity because of the constraints imposed by a subsistence livelihood, where farm work is obligatory for survival.

This topic has been one of considerable debate in recent years, particularly since longitudinal studies that have measured total energy expenditure with doubly labeled water have shown that many women increase the energy expended on physical activity during pregnancy, and that any decreases are not sufficient to counterbalance the energy costs of pregnancy due to tissue (fat) deposition and maintenance energy metabolism. It has been recommended that the data used by the World Health Organization should be revised to take account of changes in energy expended on physical activity and to separate these energy costs from those of maintenance and tissue deposition. The Dietary Reference Intakes for the United States and Canada have already incorporated these changes (**Table 3**).

Between-Country Comparison of the Metabolic Costs of Pregnancy

The average costs across different populations result in a wide range of energy needs from −30 MJ (−7000 kcal) to 523 MJ (125 000 kcal). Studies found that the average costs in the well-nourished groups were similar to the current international assumption of 336 MJ (80 000 kcal). These studies have also shown that the amount of prepregnancy body fat is strongly correlated with both the maintenance costs and the total metabolic costs of pregnancy. The combined costs of maintenance, fat deposition, and conceptus across studies from different countries drawn from emerging and affluent nations show that the energy cost of fat deposition also varies according to the state of affluence and is positively correlated with variations in maintenance requirements.

This flexibility in energy metabolism acts in a protective manner, with undernourished women showing significant energy-sparing adaptive strategies that tend to normalize energy balance. Body fat content is one of the measures of fitness for reproduction; fertility is suppressed in undernourished women. However, future unfavorable conditions cannot be anticipated and pre- or early pregnant fatness may be indicative of overall nutritional status and energy balance during pregnancy.

These relationships suggested the existence of a mechanism that can monitor the mother's prepregnancy energy status and adjust the homeorrhetic changes in maternal metabolism accordingly. The discovery of leptin provides a plausible mechanism by which peripheral energy status can be centrally monitored and may coordinate the metabolic responses to pregnancy. It is clear that in addition to its role in the regulation of adipose tissue, appetite, and metabolic rate, leptin plays a significant role in several components of the reproductive axis. Evidence suggests that it plays a key role in pregnancy, including the modulation of fetal growth.

Individual Variability in the Total Energy Costs of Pregnancy

Because of the marked differences between individuals in the different components of the energy costs of pregnancy (changes in BMR, body fat, and energy expended on physical activity), the total energy costs, and therefore energy requirements, are also variable. Studies of well-nourished women indicate that the total extra energy costs of pregnancy average 418 MJ (100 000 kcal), considerably higher than the estimates in **Table 3**, and there is a large range from 34 to 1200 MJ (8000–287 000 kcal). These values are probably representative of many women in developed countries. They show that it is impossible to prescribe energy intakes for individual women since it cannot be predicted how they will respond metabolically (BMR and fat) or behaviorally (physical activity and food intake) to pregnancy.

Implications of Energy-Sparing Adaptations for Mother and Infant

Human energy metabolism is particularly adaptable during pregnancy, with early/prepregnancy body 'fatness' being a major determinant. The adaptive strategies that maintain energy balance seem to be a coordinated biological system in which energy-sensitive modulations in metabolism help to sustain human pregnancies and protect fetal growth in highly marginal environmental circumstances. However, the existence of such mechanisms should not be misinterpreted as suggesting that maintenance of optimal nutritional status in pregnant women is not a priority because the adaptive mechanisms of the women will cope. It cannot be assumed that pregnant women will have energy-sparing alterations in metabolism and/or that physical activity decreases. Any adaptations that do occur should not be overinterpreted as suggesting that this is the case. The possible long-term detrimental effects must also be considered. The biochemical and physiological processes that are downregulated in the mother causing the suppression in BMR are unknown and there may be long-term consequences to her health and that of her infant.

The associations between maintenance needs, pregnancy weight gain, and prepregnant fatness indicate that a target weight gain of 12.5 kg is associated with maintenance costs of approximately 160 MJ (38 000 kcal). Although individual women or populations may have lower maintenance requirements, these may be associated with inadequate weight gain and low-birth-weight infants. A major determinant of birth weight is maternal weight gain, and the single most important determinant of infant survival is birth weight. Although birth weight is relatively well preserved at different planes of nutrition, weight alone is an inadequate measure of an infant's overall condition at birth. Even subtle nutritional influences on the fetal environment may have long-term consequences.

As mentioned previously, pregnancy weight gain is a critical component of the overall energy costs of pregnancy. The issue of whether pregnancy weight

gain drives, or is driven by, the metabolic changes is interesting, but it is clear that women who consume marginal diets have small weight gains and that women from poorer countries have much lower percentage weight gains despite having lower initial body weights. Extremes of weight gains during pregnancy may have several consequences, which may or may not be mediated directly through an effect on birth weight. Other effects of weight gain may be more subtle and may be mediated through qualitative effects on fetal growth and development at different stages of intrauterine growth. There is a considerable body of evidence that suggests that many chronic adult diseases have their origins in fetal and infant nutrition, which has refocused attention on early life as a critical period in human development.

See also: **Energy**: Metabolism; Balance; Requirements. **Energy Expenditure**: Indirect Calorimetry; Doubly Labeled Water. **Pregnancy**: Nutrient Requirements; Weight Gain.

Further Reading

Barash IA, Cheung CC, Weigle DS *et al.* (1996) Leptin is a metabolic signal to the reproductive system. *Endocrinology* **137**: 3144–3147.

Butte N, Hopkinson J, and Nicholson M (1997) Leptin in human reproduction: Serum leptin levels in pregnant and lactating women. *Journal of Clinical Endocrinology and Metabolism* **82**: 585–589.

Butte N, Treuth M, Mehta N *et al.* (2003) Energy requirements of women of reproductive age. *American Journal of Clinical Nutrition* **77**: 630–638.

Department of Health (1991) *Dietary Reference Values for Food Energy and Nutrients for the United Kingdom*, Report on Health and Social Subjects No. 41. London: HMSO.

FAO/WHO/UNU (1985) *Report of a Joint Expert Consultation: Energy and Protein Requirements*, Technical Report Series 724. Geneva: WHO.

Forsum E (2004) Energy requirements during pregnancy: Old questions and new findings. *American Journal of Clinical Nutrition* **79**: 933–934.

Goldberg GR (1997) Reproduction: A global nutritional challenge. *Proceedings of the Nutrition Society* **56**: 319–333.

Goldberg GR, Prentice AM, Coward WA *et al.* (1993) Longitudinal assessment of energy expenditure in pregnancy by the doubly-labelled water method. *American Journal of Clinical Nutrition* **57**: 494–505.

Harigaya A, Nagashima K, Nako Y *et al.* (1997) Relationship between concentration of serum leptin and fetal growth. *Journal of Clinical Endocrinology and Metabolism* **82**: 3281–3284.

Hytten FE (1991) Nutrition; Weight gain in pregnancy. In: Hytten F and Chamberlain G (eds.) *Clinical Physiology in Obstetrics*, 2nd edn. Oxford: Blackwell Scientific.

Kopp-Hoolihan LE, Loan MV, Wong WW *et al.* (1999) Longitudinal assessment of energy balance in well nourished, pregnant women. *American Journal of Clinical Nutrition* **69**: 697–704.

National Academy of Sciences (2002) *Dietary Reference Intakes for Energy, Carbohydrates, Fiber, Fat, Protein and Amino Acids (Macronutrients)*. Washington, DC: National Academy Press.

Poppitt SD, Prentice AM, Goldberg GR *et al.* (1994) Energy-sparing strategies to protect human fetal growth. *American Journal of Obstetrics and Gynecology* **171**: 118–125.

Poppitt SD, Prentice AM, Jequier E *et al.* (1993) Evidence of energy-sparing in Gambian women during pregnancy: A longitudinal study using whole-body calorimetry. *American Journal of Clinical Nutrition* **57**: 353–364.

Prentice AM and Goldberg GR (2000) Energy adaptations in human pregnancy: Limits and long-term consequences. *American Journal of Clinical Nutrition* **71**(supplement): 1226S–1232S.

Prentice AM, Spaaij CJK, Goldberg GR *et al.* (1996) Energy requirements of pregnant and lactating women. *European Journal of Clinical Nutrition* **50**(supplement 1): S82–S111.

Scrimshaw NS, Waterlow JC, and Schurch B (1996) Energy and protein requirements. Proceedings of an IDECG Workshop. *European Journal of Clinical Nutrition* **50**(supplement 1): S1–S197.

Weight Gain

L H Allen and J M Graham, University of California at Davis, Davis, CA, USA

During the past 40 years, there have been dramatic changes in the recommendations for optimal maternal weight gain during pregnancy. In the past, it was thought that it was necessary to restrict the diet of many pregnant women in order to reduce the perceived risks associated with higher weight gains. The fetus was thought to be relatively unaffected by this advice. In contrast, the current recommendations in the United States are based on weight changes in pregnancy that have been taken from records and known to be compatible with a healthy pregnancy outcome. The US recommendations have been widely accepted by many other Western countries. Because several maternal factors influence the amount of weight gained in pregnancy, these factors have to be taken into consideration when basing recommendations on actual weight gain. The result has been the development of more realistic weight gain guidelines that are based to some extent on the characteristics of the mother. Additional experience has been gained since these guidelines were developed that encompasses a variety of subpopulations including different ethnic groups and overweight women, and the knowledge gained broadens the scope of these recommendations among modern diverse populations. However, there is still much to be learned about the

determinants of, and variability in, energy requirements and balance of pregnant women.

Pregnancy Weight Gain Recommendations

In 1970, the US National Academy of Sciences published guidelines for weight gain during pregnancy in the report, Maternal Nutrition and the Course of Pregnancy. The recommended pregnancy gain was 24 lb (10.9 kg), with a range of 10–25 lb (9.1–11.4 kg). The report advised health care providers and pregnant women not to restrict weight gain—a practice that had been fairly widespread during the previous decade in order to reduce the perceived risks of labor complications, preeclampsia, and excess weight retention postpartum. In fact, many obstetricians had been recommending gains of only 15–20 lb (6.8–9.1 kg).

Even with the more generous recommendations set in 1970, by the 1980s it had become clear that average gains of women in the United States far exceeded these guidelines. An analysis of data from the National Natality Survey in 1980 showed the average pregnancy weight gain to be 29 lb (13.2 kg), and by the time of the National Maternal Infant Health Survey in 1988 the average had increased to 32 lb (14.5 kg). The range of gain was very wide, from no gain to more than 75 lb (34.1 kg).

Based on this realization, in 1990 the weight gain recommendations were revised completely by a committee established by the Institute of Medicine (IOM) of the National Academy of Sciences. Existing data from a national survey were analyzed to determine the weight gain that was compatible with a normal pregnancy outcome. The latter was defined as the infant being born full term and of normal birth weight and the absence of pregnancy or delivery complications. It became apparent from these analyses that maternal weight-for-height at conception, expressed as body mass index (BMI; weight in kilograms and height in meters squared), was an important predictor of actual weight gain. Thin women (with a low BMI) gained more weight than fatter women. Different weight gain recommendations were therefore developed for women entering pregnancy with different BMIs (**Table 1**). For thinner women (BMI <19.8 or <90% of ideal body weight), recommended gains are 28–40 lb (12.7–18.2 kg); for women with a normal BMI (19.8–25.9), gain should be 25–35 lb (11.4–15.9 kg) or 1 lb (0.45 kg) per week; and for overweight women (BMI >29.0 or >135% ideal body weight), gain should be at least 15 lb (6.8 kg) or 0.7 lb (0.32 kg) per week. New weight gain grids were constructed that showed the recommended gains over the course of pregnancy for each BMI group (**Figure 1**), enabling the adequacy of weight gain to be tracked for individual women. To use the chart, women's height and weight should be measured as near to the time of conception as possible (because pregnancy causes a temporary reduction in height) and used to obtain their BMI from a table. The US recommendations are deemed to be appropriate for women in developed countries worldwide.

Table 1 Recommendations for pregnancy weight gain by body mass index (BMI) at conception

BMI category	*Recommended total gain*	
	Kilograms	*Pounds*
Low (BMI <19.8)	12.8–18.0	28–40
Normal (BMI >19.8–26.0)	11.5–16.0	25–35
High (BMI >26.0–29.0)	7.0–11.5	15–25
Obese (BMI >29.0)	≤6.0	≤13

Modified from Institute of Medicine, Committee on Nutritional Status during Pregnancy and Lactation (1990) Nutrition During Pregnancy. Weight Gain. Nutrient Supplements. Food and Nutrition Board. Washington, DC: National Academy Press.

Pattern of Weight Gain

Relatively little (1–2.5 kg) of the total weight gain during pregnancy occurs during the first trimester, whereas gain in the last two trimesters is relatively linear. Nevertheless, it is important to pay attention to the quality of pregnant women's diets during the first trimester and to ensure that they do not restrict their intake during this time, when there is the

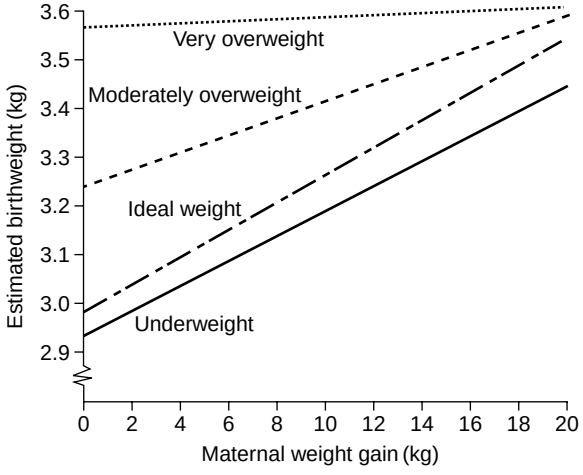


Figure 1 The relationship between maternal pregnancy weight gain and birth weight. (Reproduced with permission from the Institute of Medicine, Committee on Nutritional Status during Pregnancy and Lactation (1990) *Nutrition during Pregnancy. Weight Gain. Nutrient Supplements. Food and Nutrition Board.* Washington, DC: National Academy Press.)

strongest risk of nutrition-related birth defects and spontaneous abortions. In some studies, an association has been noted between low weight gain in the first trimester and increased risk of spontaneous preterm delivery.

Variability in Weight Gain

The BMI-specific target ranges for pregnancy weight gain are relatively narrow, but a very wide range of gain actually occurs. In a California study, for example, only 50% of the mothers who had an uncomplicated pregnancy with a normal birthweight infant gained the recommended 12.5–18 kg, with the remainder gaining more or less. Since a substantial amount of the variation in weight gain is due to physiological variability and prepregnancy BMI, deviation from the recommended range may not necessarily be cause for concern. However, it is especially important to assess the dietary patterns and other behaviors of women whose weight gain is unexpectedly high or low. The IOM *Implementation Guide* for weight gain recommendations provides helpful information on the assessments that should be used.

Maternal Weight Gain and Birth Weight

Inadequate weight gain is associated with poor fetal growth even when the contribution of fetal weight and factors such as length of gestation are taken into consideration. Birth weight is an important determinant of child health and survival; low-birth-weight (<2.5 kg) infants are 40 times more likely to die in the neonatal period. Low weight-for-length at birth may be a risk factor for chronic disease in later life. It has been estimated that in women with a normal prepregnancy BMI, each kilogram of total pregnancy weight gain has an average effect on birth weight of 20 g. In California, women with pregnancy weight gains below recommendations had a 78% higher risk of the infant being born small, whereas women who gained in excess of recommendations were twice as likely to give birth to a large infant.

As noted previously, maternal BMI at conception is strongly inversely related to expected pregnancy weight gain. Nevertheless, heavier women still tend to deliver heavier infants (**Figure 2**) and thinner women tend to have smaller infants. In thinner women, birth weight is more strongly related to pregnancy weight gain. Thus, as is evident from

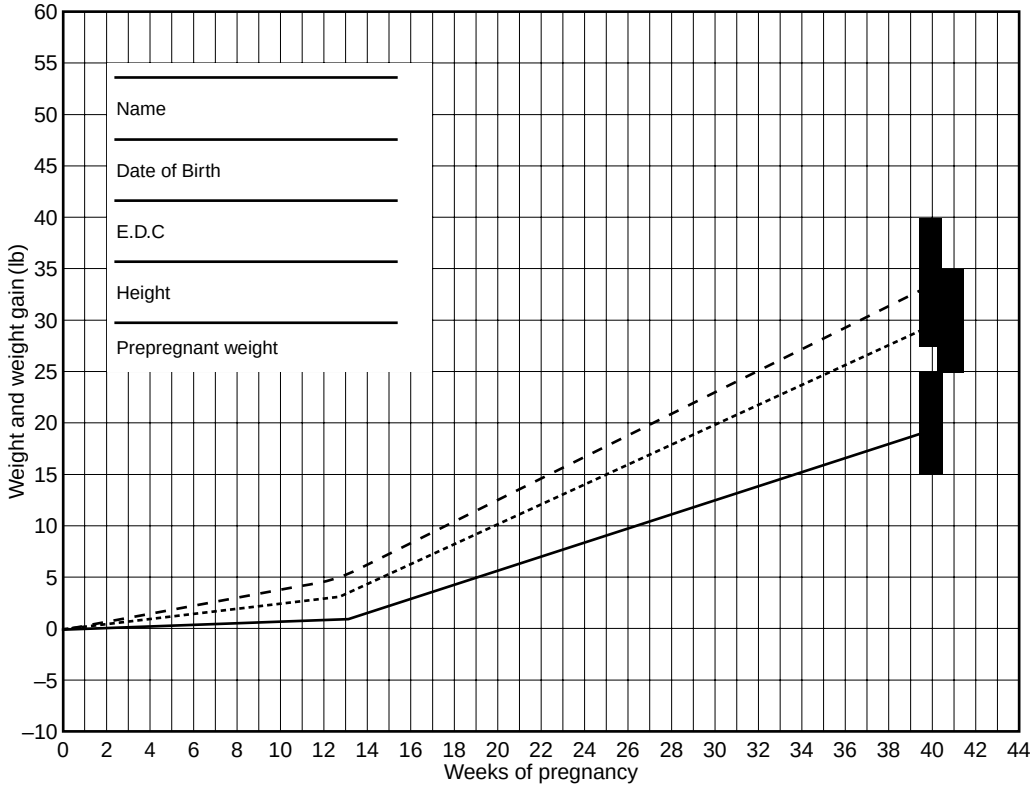


Figure 2 Recommended pregnancy weight gains based on body mass index (BMI) at conception. Dotted line, prepregnancy BMI <19.8 lbs; dashed line, prepregnancy BMI 19.8–26.0 lbs (normal); solid line, prepregnancy BMI >26.0 lbs. (Reproduced with permission from the Institute of Medicine, Committee on Nutritional Status during Pregnancy and Lactation (1992) *Nutrition during Pregnancy and Lactation. An Implementation Guide. Food and Nutrition Board*. Washington, DC: National Academy Press.)

Figure 2, the greatest risk of low birth weight is for thin women with a low pregnancy weight gain. It is crucial that thin women gain adequate amounts of weight. These associations are not explained by other risk factors associated with thinness, such as smoking.

Changes in Body Composition and Maternal Energy Status

It used to be assumed that maternal energy intake during pregnancy was the main determinant of the amount of weight gained. Although our knowledge of this relationship is still inadequate, newer information indicates that other maternal factors, and especially body composition, are more important predictors.

The weight gained during pregnancy can be roughly divided into the weight of the fetus, placenta, and amniotic fluid (a total of approximately 5 kg), maternal gain in the uterus, breasts, blood, and fluid (approximately 4 kg), and maternal fat. The latter component is the most variable, accounting for approximately 70% of the variability in pregnancy weight gain. Although average fat gain in different studies is approximately 2–5 kg, values for individual women range from a loss of several kilograms to a gain of approximately 12 kg. Even in a group of women with normal BMIs at conception, the range of fat gain was 0.5–9.5 kg. Fatter women at conception gained less fat during pregnancy, as would be expected from their lower weight gains. The greater fat gain of thinner women is a potential energy store for the fetus and would afford some protection against maternal malnutrition in late pregnancy—a situation that is not uncommon in some economically disadvantaged countries.

Maternal BMI at conception influences not only the amount of maternal weight and fat gained during pregnancy but also changes in maternal basal metabolic rate (BMR). In studies of well-nourished pregnant women, BMR has been reported to increase by approximately 20–30%. For undernourished women, however, the increment in BMR may be only 20% of that seen in those who are well nourished. In contrast, in a group of well-nourished Californian women weighing 55–116 kg, the BMR of those with higher BMIs was almost twice that of the thinnest women in the group.

Overall, it is clear that heavier women gain less weight and fat during pregnancy and have a larger increase in BMR. It has not been determined how these changes translate into energy requirements for women in the different BMI groups used to predict weight gains. Therefore, a single value for energy requirements is used for all pregnant women regardless of their BMI at conception.

Weight Gain for Special Population Groups

Adolescents

Well-nourished adolescents tend to gain at least as much, if not more, weight than adult women. The relationship between BMI, pregnancy weight gain, and birth weight is probably no different in this group, but to ensure adequate nutrition for those who are still growing, weight gain in the upper range of BMI-specific recommendations is advised. The effects of this recommendation on weight retention postpartum have not been evaluated adequately.

Short Women

Women who are less than 157 cm tall tend to give birth to infants who are large relative to maternal pelvic size, with a subsequently slightly greater risk of a more difficult delivery. These women are therefore advised to gain near the lower end of the weight gain range that is compatible with their BMI.

Ethnic Groups

Black women in the United States tend to gain less weight in pregnancy and to produce lower birthweight infants. The reasons for this are not known, but it could not be explained by differences in gestational age or other factors that were measured. Adequate weight gain in this group is known to be especially important for the prevention of fetal growth retardation. In one study, 18% of nonobese black women who gained less than the IOM recommendations gave birth to low-birth-weight infants compared to 10% whose gain was in the ideal range and 4% who gained more than the recommendations. In obese black women, the low birth weight prevalence was approximately six times higher than that for those who gained less than the recommendations.

Most surveys indicate that Hispanics seem to gain approximately the same amount of weight as Anglo women. In the 1980 National Natality Survey, Hispanic and non-Hispanic white women gained a similar amount of pregnancy weight, but the risk of low birth weight was twice as high in Hispanics. Surveillance of a predominantly Hispanic population indicated that half of the underweight women and one-third of the normal weight women gained the recommended amount of weight, whereas more

than half and three-fourths of overweight and obese women, respectively, had excessive gains. Inadequate weight gain during the third trimester was predictive of preterm birth. Underweight Hispanic women had nearly twice the risk of premature delivery.

The maternal weight gain recommendations have been evaluated in a group of Chinese women with good pregnancy outcomes ($N = 504$) to assess the need for an ethnic-specific recommendation for this group. The BMI categories were used at different levels. The recommended total pregnancy weight gain ranges according to BMI for Chinese women were 13–16.7 kg for BMI <19, 11–16.4 kg for BMI 19–23.5, and 7.1–14.4 kg for BMI >23.5. Women with weight gain in the lowest quartile had twice the risk of having a low-birth-weight infant, and those with excessive weight gain were in need of assisted delivery (either vaginal or cesarean delivery).

Substance Abusers

Cigarette smokers tend to gain less weight during pregnancy and to produce smaller infants. This effect is not explained by a lower food intake of smokers. Alcohol and drug use have similar effects. Simply gaining more weight during pregnancy will not compensate for the adverse effects of these practices on fetal outcome or pregnancy complications.

Multiple Births

Relatively few data are available from national surveys on which to base weight gain recommendations for women with twins. A weight gain of 15.9–20.5 kg or 0.7 kg per week in the second and third trimesters is usually consistent with a healthy pregnancy outcome for these women. No recommendations are available for women carrying more than two fetuses, but it is reasonable to expect that they will increase by 3.5 kg for each additional infant.

Obese and Overweight Women

Obesity during pregnancy is associated with higher morbidity for both the mother and the child. Higher prepregnancy weights have been shown to increase the risk of late (>28 weeks of gestation) fetal deaths. In addition, the prevalence of gestational hypertension increases 3-fold and there is a 3–4 times greater risk of gestational diabetes in obese pregnant women.

Exercising Women

Women who are physically fit at conception appear to be able to continue to exercise during pregnancy without harm to themselves or the fetuses, as long as the activity is not too strenuous or prolonged. In several studies it was observed that exercising women gained 2 or 3 kg less than those who were more sedentary.

Pregnancy Weight Gain and Postpartum Risk of Obesity

On average, well-nourished women retain relatively little weight approximately 1 year postpartum (approximately 0.5–1.5 kg). Delivery is followed by a rapid loss of weight in the subsequent 2 weeks due to fluid loss. This is followed by a slower rate of loss for the next 6 months, so a complete return to preconception weight should not be expected in less time than this. In general, weight still retained at 1 year postpartum is unlikely to be lost without lowering intake and/or increasing physical activity. If weight retention is substantial, it can add to the risk of obesity in the longer term, and obesity is a major public health concern in many countries.

The relatively low average weight retention postpartum obscures the fact that many women do retain an excessive amount of weight. Those who retain most are likely to have gained large amounts of weight during pregnancy. At 10–18 months postpartum, weight retention was 2.5 kg for women who gained more than the IOM recommendation compared to 0.7 kg for white women and 3.2 kg for black women who gained the advised amount. These large racial differences in weight retention have not been explained and certainly may be a risk factor for the higher prevalence of later obesity in this group.

Most women breast-feed their infants exclusively or partially for a relatively short time. There is little difference in weight loss between women who breast-feed and those who do not for periods up to 6 months postpartum. This is presumably due to the greater appetite and energy intake of women who are breast-feeding and perhaps to dieting on the part of non-breast-feeders. One study of women who breast-fed until 12 months postpartum did report a 2-kg greater weight loss compared to women who stopped breast feeding before 3 months. Even more weight was lost by those who breast-fed more often and gave longer feeds.

Women with a high BMI at conception tend to either lose or gain more weight postpartum than those with a normal BMI; approximately one-third end up weighing less than at conception, and one-third weigh substantially more. The reasons for the highly variable weight retention in this group are not known.

Although inadequate intake of nutrients during lactation can lead to maternal nutrient depletion and lower breast milk content of some nutrients and

especially vitamins, breast feeding women who choose to lose weight can do so by exercising and/or reasonable restriction of energy intake. Exercising by jogging, biking, and aerobics for 45 minutes, four or five times per week for 12 weeks did not affect well-nourished mothers' ability to lactate or influence their milk composition. However, it is possible that severe energy deficit in lactation, especially of thinner women, will reduce breast milk volume.

Impact of Supplementation

Numerous investigators have explored the benefits of energy and/or protein supplementation for pregnancy weight gain and other outcomes. However, relatively few trials have randomly assigned these supplements and used control diets. A statistical analysis was conducted of the 10 such studies that met this criterion in 1995. Most, but not all, of these studies were performed in developing countries. A 5-year controlled trial in The Gambia provided daily prenatal dietary supplements (two biscuits) that contained 4250 kJ energy and 22 g protein. This supplement increased pregnancy weight gain and birth weight during the hungry and harvest seasons. There was a significant but very small increase in head circumference and a significant reduction in perinatal mortality.

It was originally thought that timing of supplementation during later gestation would be most likely to increase birth weight. This hypothesis was supported by data from the Dutch famine, during which women in their third trimester had infants with the lowest birth weights. An increase in low birth weight prevalence was also observed in The Gambia when third-trimester gestation overlapped with the hungry season. Nonetheless, research suggests nutrition interventions initiated earlier in pregnancy will have the strongest effect on birth weight. There are enduring advantages to continued supplementation postpartum (during lactation) and into the ensuing pregnancy. A longitudinal study in Guatemala reported a significant increase (approximately 350 g) in birth weight in the second pregnancy when the mother was supplemented during the previous pregnancy and throughout subsequent lactation and the second pregnancy compared to those who were not supplemented during the prior pregnancy. Overall, it is appropriate for supplementation to begin as early in the pregnancy as possible so that both mother and fetus receive the maximum benefits for optimal health and development. However, this advice is tempered by concerns that supplementation of short Asian women may increase their offspring's risk of diabetes in later life.

See also: **Adolescents**: Nutritional Requirements. **Breast Feeding**. **Lactation**: Physiology; Dietary Requirements. **Obesity**: Complications. **Pregnancy**: Role of Placenta in Nutrient Transfer; Nutrient Requirements; Energy Requirements and Metabolic Adaptations; Safe Diet for Pregnancy; Dietary Guidelines and Safe Supplement Use; Prevention of Neural Tube Defects; Pre-eclampsia and Diet.

Further Reading

Ceesay SM, Prentice AM, Cole TJ *et al.* (1997) Effects on birth weight and perinatal mortality of maternal dietary supplements in rural Gambia: 5 year randomised controlled trial. *British Medical Journal* **315**: 786–790.

Cnattingius S, Bergstrom R, Lipworth L, and Kramer MS (1998) Prepregnancy weight and the risk of adverse pregnancy outcomes. *New England Journal of Medicine* **338**: 147–152.

Dewey KG and McCrory M (1994) Effects of dieting and physical activity on pregnancy and lactation. *American Journal of Clinical Nutrition* **59**(supplement): 439–445.

Hickey C, Cliver S, Goldenberg R, Kohatsu J, and Hoffman H (1993) Prenatal weight gain, term birth weight, and fetal growth retardation among high risk multiparous black and white women. *Obstetrics and Gynecology* **81**: 529–535.

Institute of Medicine, Committee on Nutritional Status during Pregnancy and Lactation (1990) *Nutrition during Pregnancy. Weight Gain. Nutrient Supplements. Food and Nutrition Board*. Washington, DC: National Academy Press.

Institute of Medicine, Committee on Nutritional Status during Pregnancy and Lactation (1992) *Nutrition during Pregnancy and Lactation. An Implementation Guide. Food and Nutrition Board*. Washington, DC: National Academy Press.

Keppel K and Taffel S (1993) Pregnancy-related weight gain and retention: Implications of the 1990 Institute of Medicine Guidelines. *American Journal of Public Health* **83**: 1100–1103.

King JC, Butte NF, Bronstein MN, Kopp LE, and Lindquist SA (1994) Energy metabolism during pregnancy: Influence of maternal energy status. *American Journal of Clinical Nutrition* **59**(supplement): 439S–445S.

Kramer M (1993) Effects of energy and protein intakes on pregnancy outcome: An overview of the research evidence from controlled clinical trials. *American Journal of Clinical Nutrition* **58**: 627–635.

Luke B, Minogue J, Witter F, Keith LG, and Johnson TRB (1993) The ideal twin pregnancy: Patterns of weight gain, discordancy, and length of gestation. *American Journal of Obstetrics and Gynecology* **169**: 588–597.

Parker J and Abrams B (1992) Prenatal weight gain advice: An examination of the recent prenatal weight gain recommendation of the Institute of Medicine. *Obstetrics and Gynecology* **79**: 664–669.

Siega-Riz AM, Adair LS, and Hobel CJ (1994) Institute of Medicine maternal weight gain recommendations and pregnancy outcome in a predominantly Hispanic population. *Obstetrics and Gynecology* **84**: 565–573.

Wong W, Tang NL, Lau TK, and Wong TW (2000) A new recommendation for maternal weight gain in Chinese women. *Journal of the American Dietetic Association* **100**: 791–796.

Safe Diet for Pregnancy

S Stanner, British Nutrition Foundation, London, UK

A balanced diet that contains adequate amounts of all the nutrients needed by a mother and her growing fetus is essential for a healthy pregnancy. Pregnant women also need to be advised about how to reduce their risk of exposure to substances that may be toxic to the fetus during development (teratogenic) and therefore associated with the production of physical defects in the developing embryo (e.g., alcohol and excess vitamin A), as well as other dietary and lifestyle behaviors that could optimize maternal health and reduce the risk of health problems in their children.

The aim of this article is to describe evidence relating to food safety issues during pregnancy, including potential risks to the fetus as a result of prenatal exposure to food pathogens or toxic food components (e.g., heavy metals and dioxins) and the potentially harmful effects of high doses of alcohol, caffeine, and vitamin A.

Food-Borne Infections during Pregnancy

For many years it has been recognized that food-borne antenatal infections may cause death or serious fetal damage. Women may be more susceptible to the effects of infection during pregnancy because of immunological changes leading to suppression of the immune system (most commonly cell-mediated immunity), probably as a result of increases in pregnancy-associated sex steroids, such as oestradiol or progesterone. Among the most common causes of diarrhea during pregnancy are several food- or water-borne pathogens (bacteria, protozoa, or viruses), including salmonella species, *Helicobacter pylori*, *Shigella*, *Escherichia coli*, and cryptosporidium. Hepatitis A is also a food- or water-borne pathogen of concern, particularly in countries where sanitation is poor. In pregnant women, severe vomiting and diarrhea may negatively affect the availability of important nutrients to the growing fetus. For example, impairment of the supply of folate (or the synthetic form, folic acid) during a critical stage of development could increase the risk of associated neural tube defects, such as spina bifida.

Although rare, infection with *Listeria* or *Toxoplasma* during pregnancy is of particular concern because even in a mild form these infections can prove fatal. Listeriosis caused by the consumption of food containing the bacterium *Listeria monocytogenes* leads to flu-like symptoms, such as fever, muscle aches, and sometimes nausea or diarrhea. If the infection spreads to the nervous system, it may also cause headaches, stiff neck, confusion, loss of balance, or convulsions. The bacterium has been found in a variety of raw foods, including unpasteurized (raw) milk, uncooked meats, and vegetables, and in processed foods that become contaminated after processing, such as soft cheeses and cold cuts of meat. According to the Centers for Disease Control and Prevention, pregnant women in the United States are approximately 20 times more likely than other healthy adults to get listeriosis and approximately one-third of listeriosis cases occur during pregnancy. The fetus and newborn are at greatest risk of this infection and its consequences can be severe, leading to miscarriage, stillbirth, and premature delivery or to meningitis in the newborn infant. When infection occurs during pregnancy, antibiotics given promptly to the pregnant woman can often prevent infection of the fetus or newborn, and infants developing the infection can also be treated in the same way.

Toxoplasma gondii is a parasite that can be transmitted to the fetus in utero through transplacental transmission, causing stillbirth, miscarriage, or mental retardation. The parasite has been found in raw, inadequately cooked or cured meat, cat feces, and unwashed raw fruit and vegetables. It has also occasionally been reported in unpasteurized goat milk. In the United Kingdom, toxoplasmosis occurs in approximately 2.5–5.5 in 1000 pregnant women (1750–2850 cases per year), generally causing flu-like symptoms, swollen lymph glands, or muscle aches and pains that last for a few days to several weeks. If a pregnant woman contracts the infection, there is an approximately 30–40% chance of fetal infection (congenital toxoplasmosis). Infants who became infected before birth may develop growth problems, vision and hearing loss, hydrocephalus, brain damage, epilepsy, and other problems. In Europe, congenital toxoplasmosis affects between 1 and 10 in 10 000 newborns, of whom 1 or 2% develop learning difficulties or die and 4–27% develop permanent loss of vision. Both the incidence of placental transmission and the severity of congenital disease depend on gestational age at which maternal seroconversion occurs. Although transmission rates from mother to fetus tend to be low early in pregnancy, fetal disease severity is highest when the fetus is infected early in gestation. Mothers can be tested to determine if they have developed an antibody to the infection. Fetal testing may include ultrasound and testing of amniotic fluid or cord blood. When

Table 1 General guidelines on good hygienic practices in the home

The risk of food poisoning can be minimized by adopting the following practices:

Cleanliness in the kitchen

Keeping all work surfaces scrupulously clean
Washing cooking utensils after coming into contact with raw meat, poultry, or eggs to prevent cross-contamination
Using separate chopping boards for foods that are to be cooked (e.g., raw meat)
Keeping kitchen cloths clean; rinsing crockery in hot water, leaving it to dry, and then wiping it clean with a tea towel
Using kitchen towels to mop up spills rather than a dishcloth
Ensuring waste bins are covered and away from food and keeping pets away from the kitchen

Hygienic food handling

Washing all equipment and work surfaces before and after touching raw food
Washing all foods to be eaten raw thoroughly
Cooking meat thoroughly to an internal temperature of at least 70 °C
Keeping raw and cooked foods separated during preparation and storage
Cooling cooked foods as quickly as possible if they are to be stored in a refrigerator or freezer
Covering foods and not leaving them standing around in the kitchen
Storing food at the correct temperature (<4 °C in the refrigerator or <−18 °C in the freezer)
Storing raw meat, well covered, at the bottom of the refrigerator
Storing eggs in a refrigerator, if possible
Never overloading the refrigerator because this can reduce the circulation of cool air
Keeping foods for as short a time as possible (especially meat and fish) and following storage instructions (i.e., not using beyond the 'use by' or 'best before' date)
Thawing frozen meat thoroughly before cooking
Avoiding reheating food more than once
Reheating foods thoroughly (if this is done in a microwave, the standing times recommended by the manufacturer should be observed to ensure that food attains an even temperature before it is eaten)

Personal hygiene

Washing hands thoroughly before preparing food, after visiting the toilet, and after emptying the trash bin
Never licking fingers or utensils and putting them back into food
Washing hands after blowing or touching the nose while handling food
Keeping nails clean and hair out of food
Wearing a clean apron
Not handling food during periods of illness (e.g., heavy cold, sickness, or diarrhea)
Covering all cuts, spots, and pimples, particularly on the hands, with a waterproof dressing and replacing it often
Washing hands thoroughly after handling cat feces or using rubber gloves

toxoplasmosis is diagnosed during pregnancy, antibiotic treatment can often help reduce the severity of symptoms in the newborn.

The risk of food poisoning can be minimized by ensuring adequate attention to good hygienic practice when preparing, cooking, and storing foods (**Table 1**). Pregnant women should therefore be advised of the need for a high regard for food hygiene and personal cleanliness during this vulnerable time. In addition, there are a few foods that may pose a particular (although small) risk, which should be avoided during pregnancy where possible (**Table 2**).

Alcohol

Excessive Alcohol Consumption during Pregnancy

Chronic alcohol abuse may result in a wide spectrum of secondary disturbances of the absorption and utilisation of many nutrients, including glucose, amino acids, fat, sodium, and some vitamins (especially thiamin, vitamin B_{12}, and folate). The inhibition of folate absorption by alcohol is of particular concern because of the risk of neural tube defects associated with an inadequate supply of this vitamin to the fetus before conception and during the first trimester of pregnancy. Alcohol may also directly impair the placental transfer of nutrients essential for growth

Table 2 Foods to avoid during pregnancy to minimize the risk of food-borne infections

Unpasteurized milk and milk products
Undercooked meats, poultry, eggs, fish (e.g., smoked salmon or trout, sushi), and shellfish (e.g., oysters)
Cook-chill meals and ready-to-eat poultry, unless they have been reheated until very hot
Raw egg or uncooked eggs or foods made from them, such as homemade mayonnaise, soft-whip ice cream, cake mix, mousses, and hollandaise sauce (eggs should be cooked until both the white and yolk are solid)
Unwashed fruit and vegetables
All types of pâté (including vegetable)
Soft, mould-ripened, or blue-veined varieties of cheese (e.g., Brie, Camembert, and Stilton) (foods containing these cheeses that have been properly cooked will be safe to eat)

Table 3 Symptoms of fetal alcohol syndrome (FAS)[a]

Prenatal and postnatal growth retardation
Intrauterine growth retardation, including smaller than normal head circumference, continued growth below the 10th centile, and failure to thrive
Central nervous system involvement
Neurological abnormalities, developmental delay, intellectual impairment, brain malformation, and hearing and visual disabilities
Physical anomalies
Characteristic facial deformity, including short upturned nose, receding forehead, and chin, smaller than normal eye appertures, absent philtrum, and asymmetrical ears

[a]The diagnosis of FAS requires signs in all three of the categories.

(e.g., amino acids), which at critical phases of fetal organogenesis could compound any direct fetotoxic effects of ethanol or acetaldehyde.

Both alcohol and its primary metabolite, acetaldehyde, are teratogenic. Excessive alcohol consumption (>80 g of ethanol or 10 units per day) during pregnancy can result in a child being born with a specific combination of physical and mental disabilities known as fetal alcohol syndrome (FAS). Such fetuses usually survive until birth but are growth retarded and display a characteristic range of clinical features, principally craniofacial abnormalities and neurological damage (**Table 3**).

It is estimated that 1 in 1000 infants worldwide are affected with the full syndrome, whereas 3 in 1000 may exhibit only some features. The damage varies depending on the stage of development at which high doses of alcohol are encountered. The fetus is most vulnerable to organ damage from the time the umbilical cord begins to function (5 weeks) to the completion of organ development (11 weeks). Inhibition of growth and neurobehavioral development occurs in the second and third trimester. Although the facial features of FAS become more subtle with age, growth deficits and central nervous system impairment may be permanent.

FAS is only seen in infants born to women who are excessive drinkers, but it is not an inevitable result of heavy drinking in pregnancy, and even children born to mothers who are active alcoholics may not show it. This differing susceptibility of fetuses to the syndrome is thought to reflect the interplay of genetic factors, social deprivation, nutritional deficiencies, and tobacco and other drug abuse, along with alcohol consumption.

Binge Drinking and Social Alcohol Consumption during Pregnancy

Binge drinking, generally defined as the consumption of alcohol equivalent to five or more standard drinks per occasion, may be particularly harmful because it exposes the fetus to high blood alcohol concentrations over relatively short periods of time and may be associated with repeated withdrawal episodes. Animal studies have demonstrated binge-like exposure to alcohol to be as teratogeneic as long-term exposure throughout gestation, even if the overall alcohol amount consumed by binge drinking is less than intake during more continuous drinking patterns. However, the findings of human studies have been inconsistent, possibly because of the problems of recording binge drinking during pregnancy.

The question of whether moderate or occasional alcohol consumption is safe during pregnancy has been widely debated. Currently, there is little evidence that modest drinking (<10 units per week) has any harmful effects. Although there is general agreement that women should not drink alcohol excessively during pregnancy, a consensus opinion of a safe limit to drink has not been established at any stage of pregnancy, and advice differs between countries. Many studies are confounded by factors such as cigarette smoking, social class, drug abuse, very high levels of caffeine intake, and different cross-country categorization of 'light,' 'moderate,' and 'heavy' drinking. A review of the evidence by the Royal College of Obstetricians and Gynaecologists in the United Kingdom concluded that alcohol consumption of more than 3 drinks per week during the first trimester increases the risk of spontaneous abortion, and consumption of more than 15 units per week can have a small negative effect on birth weight. However, it found no conclusive evidence of adverse effects on fetal growth or IQ at levels of intake less than 15 units per week.

Despite the lack of evidence of detrimental effects on any outcome at low/moderate maternal alcohol consumption, many professional bodies err on the side of caution. The Royal College of Obstetricians and Gynaecologists has suggested that alcohol intake during pregnancy should not exceed 1 unit/day (**Table 4**). The UK Department of Health advises that women who are pregnant or planning a pregnancy should not consume more than 1 or 2 units once or twice a week, and they should avoid

Table 4 Definition of a unit of alcohol

1 unit of alcohol approximately equals 8 g of absolute alcohol, which is equivalent to
1/2 pint of ordinary-strength beer, lager, or cider
1/4 pint of strong beer or lager
1 small glass of wine
1 single measure of spirits
1 small glass of sherry

intoxication. Advice in North America (United States and Canada) is that women should not consume alcohol at all during pregnancy, and there are warnings on products and advertisements. Anecdotally, many pregnant women develop a spontaneous aversion to the taste and/or smell of alcoholic beverages and so may limit their intakes anyway.

Vitamin A

During the period of early development, the supply of preformed vitamin A (retinol) must be carefully managed to ensure that the developing fetus is exposed to neither too little nor too much of the nutrient because either condition can have teratogenic consequences. Adequate vitamin A is required for normal embryonic development, and an insufficient supply during pregnancy can result in malformations in the offspring as well as increased mortality and morbidity during early childhood from infectious diseases, such as diarrhea, measles, and respiratory infections.

Excess vitamin A intake has also been associated with teratogenicity in animals and may represent a risk in humans, particularly within the first trimester of pregnancy. Characteristic features include severe motor deficit malformations of the heart, thymus, face, jaw, ears, palate, and brain. Although adverse effects from dietary sources are very rare, events have occurred with the ingestion of high-dose supplements. Although epidemiological data suggest there is no danger when consumption is less than 3000 RE per day, a threshold for any teratologic effects remains to be established.

In Western countries, where vitamin A deficiency is rare, women who are or might become pregnant are advised against taking vitamin A supplements (including cod liver oil), except on the advice of a doctor or antenatal clinic, and not to consume liver or liver products. In developing countries, vitamin A supplementation programs have resulted in decreased pregnancy-related mortality and lower rates of childhood morbidity and mortality, with benefits clearly outweighing any potential risks. The initiation of such programs in any population should be carefully examined in each case according to the risk–benefit ratio, with the final decision taking into account the estimated vitamin A status of the woman, the availability of vitamin A-rich foods, and whether supplementation can be supervised.

Fish and Pregnancy

Fish is a good source of protein, vitamins, and minerals. In particular, oil-rich fish (e.g., mackerel, salmon, kippers, herrings, trout, sardines, and fresh tuna) contain the long-chain n-3 fatty acids eciosapentenoic acid (EPA) and docosahexenoic acid (DHA), which may confer many health benefits to the developing fetus. For example, DHA is required for nerve and retinal development, and eating oily fish has been found to have a slight beneficial effect on birth weight and length of gestation. However, fish consumption has been positively associated with intakes of certain contaminants, namely mercury, dioxins, and polychlorinated biphenyls (PCBs), and concern has been expressed about the consequences of prenatal exposure to these toxic chemicals on the risk of brain and nervous system abnormalities.

Mercury

Mercury is a metal that is present in the environment from natural and man-made sources (e.g., coal-burning or other industrial pollution). It is converted primarily by microorganisms to a more toxic form, methylmercury, which is bioaccumulated in the aquatic food chain, reaching it highest levels in large, longer living predatory fish. Among humans, the sole source of exposure to methylmercury is the consumption of fish and sea mammals.

Methylmercury is neurotoxic and accumulates in the brain and central nervous system. It inhibits the division and migration of neuronal cells and disrupts the cytoarchitecture of the developing brain. Although a mother may show no signs of neurotoxicity, the developing fetus may be damaged following exposure to methylmercury. The concentration of methylmercury in fetal brain has been shown to be 5–7 times higher than that in maternal blood, and it has been estimated that the fetus is 5–10 times more sensitive to methylmercury exposure than an adult, although the reason for this is unknown.

Disasters in Minamata, Japan, in the 1950s and in Iraq in 1971–1972 demonstrated that acute prenatal exposure may result in severe mental retardation, cerebral palsy, blindness, and deafness. However, whether exposure to lower chronic doses, which may occur if pregnant women consume large amounts of fish, can also lead to adverse neurodevelopmental consequences is less certain. Large, long-term prospective epidemiological studies of high fish-eating populations have not found a consistent pattern of association between exposure and neuropsychological outcomes. Although subtle neuropsychological changes were reported in a study of children in the Faroe Islands study, where exposure was mainly from whale consumption, a similar study in the Seychelles found no adverse effects from fish consumption alone.

The Joint FAO/WHO Committee on Food Additives revised its safety guideline for weekly intake of

methylmercury, known as the provisional tolerable weekly intake (PTWI), to 1.6 μg/kg body weight per week. The UK government's independent expert Committee on Toxicity of Chemicals in Food, Consumer Products, and the Environment (COT) has applied a lower PTWI limit of 0.7 μg/kg body weight per week to women who are pregnant or those intending to become pregnant and to mothers who are breast feeding.

Any public health recommendations to pregnant women regarding fish consumption must recognize the important role that it plays as part of a healthy, balanced diet. Most fish contain trace amounts of methylmercury, but high concentrations of the metal have only been found in large, predatory fish, such as shark, marlin, and swordfish (**Table 5**). If a pregnant or breast feeding mother were to consume one portion of these predatory fish, she would exceed the lower PTWI set by COT and the EPA by 400%. Therefore, as a precaution, pregnant women, breast feeding mothers, and those who intend to become pregnant within the next 12 months are advised to avoid consumption of these types of fish (in the United States, this also includes king mackerel and tilefish). Some samples of tuna have also been found to have higher levels than other species. In the United Kingdom, pregnant women (and those who may become pregnant) are advised to restrict their weekly intake to two 140-g portions of fresh tuna or four 140-g portions of canned tuna.

Dioxins and Polychlorinated Biphenyls

Fish can also contain other organic pollutants, such as PCBs and dioxins. Whereas mercury accumulates in the muscles of larger predatory fish, PCBs and dioxins are found in the fatty tissues of fish. Most human exposure to PCBs and dioxins comes from dietary sources because they accumulate in the lipid fractions of meat, fish, milk and milk products, eggs, grains, and oils.

Table 5 Concentrations of methylmercury in surveyed fish in the United Kingdom

Fish	*Methylmercury (mg/kg)*
Shark	1.52
Swordfish	1.35
Marlin	1.09
Fresh tuna	0.40
Canned tuna	0.19
Herring	0.09
Pink shrimps	0.09
Cod	0.07
Plaice	0.06
Mackerel	0.05
Haddock	0.04
Scallops	0.01

PCBs and dioxins have been linked with increased rates of some cancers in studies of individuals exposed to high amounts through either vocational exposure or accidental environmental contamination. Prenatal exposure to large amounts of these pollutants (e.g., through contaminated fish) has been associated with neurobehavioral alterations in newborn children. Some studies have also suggested that exposure to smaller quantities of PCBs and dioxins in utero may lead to more subtle cognitive and motor developmental delays, although a favorable home environment appears to counteract any effect. However, the difficulty of separating the effects of PCBs and dioxins from potentially confounding factors (e.g., exposure to other contaminants, breast-feeding, smoking, and maternal education) makes it difficult to reach firm conclusions. Further research is also needed to ascertain whether any cognitive changes are temporary or persist into later life.

The potency of dioxins is expressed as toxic equivalents (TEQs), which have been internationally accepted. In the United Kingdom, the tolerable daily intake (TDI) recommended by COT is 2 pg TEQ/kg body weight, which is in line with recommendations of other international and European expert committees. In common with the United States and the European Union, approximately one-third of the UK population may exceed the TDI in their daily diet. The TEQ, therefore, provides a target to reduce dioxins and PCBs in the environment internationally. Since the 1960s, following the prohibition of many dioxins and PCBs by governments, concentrations have been declining in breast milk, which is commonly used to determine exposure. For example, between 1982 and 1997, consumption of dioxins and PCBs in the United Kingdom decreased by 75%, and due to strict controls concerning production, use, and disposal of PCBs and dioxins, it is anticipated that intakes will decrease further.

Caffeine

Caffeine is a methylated xanthine that acts as a mild central nervous system stimulant. It is found in a number of foods and beverages (**Table 6**), the main sources being coffee, tea, cocoa, chocolate, and soft drinks, as well as in prescription and nonprescription medicines, such as diet pills, headache treatments, and cold and flu medicines. Tea and cocoa also contain significant quantities of theophylline and theobromine, which are caffeine derivatives that have not been as widely researched.

Once caffeine and its derivatives are consumed, they are absorbed into the blood and body tissues and can cross the placenta to the fetus. The body

Table 6 Caffeine content of beverages and foods

Beverage or food	*Content (mg)*
1 cup (190 ml) of instant coffee	~75
1 cup (190 ml) of brewed coffee (filter or percolated)	~100–115
1 cup (190 ml) of decaffeinated coffee (brewed or instant)	~4
1 cup (190 ml) of tea	~50
1 cup (200 ml, using manufacturers' instructions) of drinking chocolate	1.1–8.2
250 ml serving of energy drinks (containing either caffeine or guarana)	28–87
330 ml serving of cola (regular and diet)	11–70
50-g bar of chocolate	5.5–35.5

metabolizes caffeine more slowly during pregnancy, especially in the last few months; the half-life of caffeine increases from approximately 5 to 18 h during the second and third trimesters. Blood caffeine concentrations are therefore raised during pregnancy with no change in intake. In contrast, smoking is known to increase caffeine metabolism appreciably.

Although very high doses of caffeine are teratogenic in animals, no link between consumption during pregnancy and birth defects has been demonstrated in humans. However, high maternal caffeine intakes (>500 mg/day) have been associated with increased fetal heart rate and newborn cardiac arrhythmias. Although there is no reliable evidence linking caffeine intake during pregnancy with sudden infant death syndrome or preterm birth, a number of studies have reported significantly increased risks of spontaneous abortion and low birth weight with caffeine intakes greater than 300 mg/day and some have also demonstrated increased risks at lower intakes (150 mg and higher). However, whether these associations are causal remains to be established because the effects of confounding factors such as smoking and alcohol consumption on these outcomes could not be determined. Concern has also been expressed about the reliability of using self-reported caffeine intake in many of these studies.

The lack of consistency between studies, particularly in relation to the dose at which an effect is reported, makes it very difficult to identify a threshold level of caffeine intake that presents an increased risk during pregnancy. The available data suggest that a moderate level of caffeine consumption is safe, although there may be concerns at high levels of intake. The Food Standards Agency in the United Kingdom advises pregnant women to moderate their caffeine intake to no more than 300 mg/day. This is equivalent to approximately three or four cups of coffee (**Table 7**). This figure is endorsed by the European Union Scientific Committee on Foodstuffs, which states that "up to 300 mg/day appear to be safe." In practice, many pregnant women reduce their coffee intake as a result of a spontaneous aversion to the taste and smell, particularly in early pregnancy.

Table 7 Approximate equivalent of 300 mg of caffeine

4 average cups or 3 average-size mugs of instant coffee
3 average cups of brewed coffee
6 average cups of tea
8 cans of regular cola drinks
4 cans of 'energy' drinks
400 g (8 standard 50-g bars) of milk chocolate

Avoiding Foods to Prevent Allergy

Food allergy has been estimated to affect approximately 1 or 2% of infants and young children in Western Europe and is assumed to be increasing in line with other forms of atopic disease, although evidence to support this is limited. Some food allergies (e.g., peanut allergies) can persist into adulthood and in severe cases can be life threatening. Most confirmed food allergies are associated with a relatively limited range of foods, including cow milk, eggs, tree nuts, peanuts, soybeans, wheat, fish, and shellfish.

The development of food allergy depends on several factors, including genetic factors and early exposure to allergenic proteins in the diet, food protein uptake and handling, and the development of tolerance. However, it remains uncertain whether sensitization occurs in utero and, if so, whether this occurrence is restricted to specific stages of gestation. There is little evidence to support any benefit of avoiding specific foods during pregnancy to reduce the risk of allergic disease in a genetically susceptible child. Indeed, such a strategy may be counterproductive because it has been suggested that exposure to foreign proteins that cross the placenta is important to establish a normal immune response that enables the infant to develop normal tolerance to the many foreign proteins in the environment. Because restrictive diets may limit the supply of essential nutrients, these should only be practiced under medical supervision. Inappropriate and unnecessary exclusion of foods could prevent both mother and infant from obtaining the nutrients they need, resulting in significantly reduced weight gain and a tendency toward lower birth weights. Nevertheless, because of the severity of reactions experienced from peanut allergy, COT advises against the consumption of peanuts and foods containing peanuts during pregnancy where there is a strong family history of atopic disease.

Food Additives and Herbal Supplements

Pregnant women often express concern about food additives. However, all additives have to be approved as safe for almost all but a small proportion of the population who may experience rare reactions to them before they can be used in foods. The presence of an 'E' number demonstrates that it has passed safety tests and been approved for use by the European Community. In the United Kingdom, COT sets an Acceptable Daily Intake for each additive, which is the amount that can be eaten daily with no risk to health. This may limit the amount of an additive used or restrict its use to certain food products. Even when an additive has been approved, new research is constantly reviewed and approval for any additive will be withdrawn if doubt is raised about its safety.

A number of herbal supplements and preparations may be used during pregnancy, most commonly to relieve gastrointestinal symptoms. Although the use of many herbal remedies is safe during pregnancy, this cannot be assumed simply because a product is described as 'natural.' Many plants, trees, fungi, and algae can be poisonous to humans, and many pharmaceuticals have been developed or derived from these sources because of the powerful compounds they contain. Very few randomized, clinical trials have examined the safety and efficacy of alternative therapies during pregnancy, and women should be warned to use any medicine, including herbal remedies, with care during pregnancy and with the advice of a doctor or pharmacist.

Summary

In addition to the consumption of a healthy, balanced diet, there are food safety precautions that need to be followed to ensure a safe pregnancy. A summary of the evidence and current advice described in this article is given in **Table 8**.

Table 8 Advice regarding dietary habits and food safe during pregnancy

- Pregnant women should pay careful attention to food and personal hygiene so as not to expose themselves to any risk of food poisoning, which is not only very unpleasant but also potentially very dangerous to the unborn child in some cases (e.g., with listeriosis and toxoplasmosis).
- Foods that have been linked with the bacteria *Listeria monocytogenese* should be avoided. These include pâtés and mold-ripened, soft cheeses (e.g., brie and camembert). Pre-prepared foods should be heated until they are piping hot, and fruit and vegetables should be washed well, especially if they are to be eaten raw.
- To reduce the risk of toxoplasmosis, pregnant women should avoid eating raw or uncooked meat, unpasteurized goat's milk or goat's cheese, or unwashed fruit and vegetables. After handling raw meat, chopping boards, utensils, and hands should be washed thoroughly. When gardening or emptying cat litter trays, rubber gloves should always be worn.
- Undercooked foods (e.g., meat, poultry, and eggs), foods containing raw egg (e.g., mayonnaise and soft-whip ice cream), and raw fish (e.g., sushi and smoked salmon) should also be avoided.
- Drinking alcohol heavily throughout pregnancy (>80 g or 10 units per day) is linked with fetal alcohol syndrome. Modest drinking (<10 units per week) does not appear to have harmful effects, but most professional bodies err on the side of caution and recommend that pregnant women abstain from drinking alcohol or limit their consumption to no more than 1 or 2 units per day.
- Supplements containing high doses of preformed vitamin A and foods containing large amounts of this vitamin (liver and liver products) are best avoided in countries where intake from a well-balanced diet should be sufficient. In areas of endemic vitamin A deficiency, supplementation can reduce pregnancy-related mortality and reduce rates of childhood morbidity and mortality. However, most experts agree that preformed vitamin A supplements in doses of more than 3000 RE should not be taken by women who may become pregnant. β-Carotene is safe for pregnant women.
- Although it is not clear if intakes of mercury and other contaminants, such as PCBs and dioxins, at levels that can be obtained from eating fish can influence children's neurological development, government organizations in a number of countries recommend that pregnant women avoid species of fish that have been found to contain high levels of these substances, including shark, marlin, swordfish, tilefish, and king mackerel. Some countries have also recommended limiting tuna intake (e.g., in the United Kingdom pregnant women are encouraged to consume no more than two tuna steaks or four medium-size cans of tuna per week).
- Consumption of caffeinated beverages (e.g., coffee, tea, and colas) has been associated with miscarriage and low birth weight, although many studies are confounded by high alcohol intakes, smoking, and drug and other substance abuse. In the United Kingdom, the Food Standards Agency recommends that pregnant women limit their caffeine intake to 300 mg/day.
- Avoiding specific foods during pregnancy is unlikely to reduce the risk of allergic disease in a susceptible child. However, where there is a strong family history of atopic disease, the avoidance of foods that may cause severe reactions (e.g., peanuts) may be advised. In the United Kingdom, the Food Standards Agency recommends that where a mother, father, or sibling suffers from allergic disease, peanuts or foods containing peanuts should be avoided during pregnancy and breast feeding and that infants should not given these foods until the age of 3 years.
- Additives permitted for use in foods undergo stringent safety tests over a long period of time before being approved and are safe for consumption during pregnancy by all but a small proportion of women who experience rare reactions to specific additives.
- Many pregnant women who would not consider taking over-the-counter medications often view herbal products as a safe and natural alternative. However, very few randomized, clinical trials have examined the safety and efficacy of alternative therapies during pregnancy. Pregnant women should be advised to seek advice from a doctor or pharmacist before taking any medication, including herbal supplements.

See also: **Alcohol**: Disease Risk and Beneficial Effects; Effects of Consumption on Diet and Nutritional Status. **Caffeine**. **Fish**. **Food Allergies**: Etiology; Diagnosis and Management. **Food Safety**: Bacterial Contamination; Other Contaminants; Heavy Metals. **Pregnancy**: Nutrient Requirements; Prevention of Neural Tube Defects. **Vitamin A**: Biochemistry and Physiological Role.

Further Reading

American College of Obstetricians and Gynaecologists Committee (1998) Committee opinion. Vitamin A supplementation during pregnancy. Number 196, January 1998 (replaces No. 157, September 1995). Committee on Obstetric Practice. American College of Obstetricians and Gynecologists. *International Journal of Gynaecology and Obstetrics* **51**: 286–287.

Azais-Braesco V and Pascal G (2000) Vitamin A in pregnancy: Requirements and safety limits. *American Journal of Clinical Nutrition* **71**: 1325S–1333S.

Clarkson TW (2003) The toxicology of mercury—Current exposures and clinical manifestations. *New England Journal of Medicine* **349**: 1731–1737.

Committee on the Toxicity of Chemicals in Food Consumer Products and the Environment (2000) *Adverse Reactions to Food and Food Ingredients*. London: Food Standards Agency.

Committee on the Toxicity of Chemicals in Food Consumer Products and the Environment (2001) *Statement on the Reproductive Effects of Caffeine. Statement 2001/06*. London: COT.

Committee on the Toxicity of Chemicals in Food Consumer Products and the Environment (2001) *Statement on the Tolerable Daily Intake for Dioxins and Dioxin-Like Polychlorinated Biphenyls*. London: COT.

Committee on the Toxicity of Chemicals in Food Consumer Products and the Environment (2002) *Statement on a Survey of Mercury in Fish and Shellfish*. Available at www.food.gov.uk/multimedia/pdfs/cotmercurystatement.pdf.

Expert Group on Vitamins and Minerals (2000) *Review of Vitamin A (EVM/00/02/P)*. Available at http://archive.food.gov.uk/committees/evm/papers.htm.

Food Standards Agency (2003) *Mercury in Imported Fish and Shellfish, UK Farmed Fish and Their Products (40/03)*. Available at www.food.gov.uk/science/surveillance/fsis-2003/fsis402003.

International Vitamin A Consultative Group (1998) *Safe Doses of Vitamin A during Pregnancy and Lactation. IVACG Statement*. Washington, DC: International Vitamin A Consultative Group.

Joint FAO/WHO Expert Committee on Food Additives (2003) *Sixty First Meeting. Summary and Conclusions*. Rome: FAO/WHO. Available at www.fao.org/es/ESN/jecfa/whatisnew_en.stm.

Royal College of Obstetricians and Gynaecologists (1999) *Alcohol Consumption in Pregnancy*. Available at www.rcog.org.uk.

Scientific Advisory Committee on Nutrition (2004) *Advice to FSA: On the Benefits of Oily Fish and Fish Oil Consumption from SACN*. Available at www.sacn.gov.uk/pdfs/fics-sacn-advice-fish.pdf/

Scientific Committee on Food (1999) *Opinion on Caffeine, Taurine and D-Glucurono-γ-lactone as Constituents of So-Called 'Energy' Drinks (Expressed on 21 January 1999)*. Available at www.europa.eu.int/comm/food/fs/sc/scf/out22_en.html.

Smith JL (1999) Foodborne infections during pregnancy. *Journal of Food Protein* **62**: 818–829.

World Health Organisation (1998) *Safe Vitamin A Dosage during Pregnancy and Lactation. Recommendations and Report of a Consultation*, Micronutrient Series WHO.NUT/98.4. Geneva: World Health Organisation.

Dietary Guidelines and Safe Supplement Use

L H Allen, J M Graham and J E Sabel,
University of California at Davis, Davis, CA, USA

During pregnancy there is an increased demand for nutrients for maternal metabolism and growth of fetal and maternal tissues. Nutritional goals during pregnancy include consumption of a high-quality, micronutrient-dense diet and the avoidance of, or a reduction in, intake of substances that are potentially harmful to the fetus or pregnant woman. The majority of additional nutrients required for pregnancy can be obtained by appropriate selection of foods. However, many women consume vitamin–mineral supplements, medications, and other forms of supplements during pregnancy. It is important to understand the risks, as well as the benefits, of these nondietary substances.

Dietary Guidelines for Pregnancy

In general, pregnant women should follow the dietary advice provided by reputable national or professional organizations for all members of the public. For example, the Dietary Guidelines of the US Department of Agriculture include advice to eat a variety of foods; maintain or improve your weight; and choose a diet low in fat, saturated fat, and cholesterol and moderate in sugars, salt, and sodium. However, advice to control or maintain body weight, or to consume alcohol in moderation if it is consumed, is inappropriate in pregnancy. Specific guidelines for pregnant women are available and provide more specific information on recommended nutrient intakes, weight gain, the need for vitamin mineral supplements, activity, and the use of alcohol and other substances. Examples include a position paper by the American Dietetic Association and guidelines from the March of Dimes.

Most of the increased requirements for nutrients during pregnancy can be met primarily by an adequate dietary pattern, with a few important changes that are listed in **Table 1**. Milk products should be increased to 4 C (cups) (1 l) of milk, which will increase energy, protein, calcium, and vitamin D intake. For women who are lactose intolerant, dairy products that contain digested lactose should be used. Cheese and yogurt have less lactose compared to milk and can be substituted. Whole grain breads and cereals, leafy green and yellow

Table 1 Daily food pattern for pregnancy

Food	*Servings/day*	*Serving size*
Milk, yogurt, and cheese	2 or 3	1 cup milk or yogurt 1.5 oz. natural cheese 2 oz. processed cheese
Meat, poultry, fish, eggs, dry beans, and nuts	2 (6 oz.)	2–3 oz. cooked lean meat, poultry, or fish 1 oz. = 1 egg 0.5 cup cooked dried beans 0.5 cup tofu 0.3 cup nuts 2 T. peanut butter
Vegetables (includes dark green and yellow vegetables)	4	1 cup raw leafy vegetables 0.5 cup other vegetables, raw/cooked 0.75 cup of vegetable juice
Fruits, fresh or canned	3	1 medium apple, banana 0.5 cup chopped fruit 0.75 cup fruit juice
Breads and cereals	9	1 slice bread 1 cup ready-to-eat cereal 0.5 cup cooked cereal, rice, or pasta

From Kaiser LL and Allen L (2002) Position of the American Dietetic Association: Nutrition and lifestyle for a healthy pregnancy outcome. *Journal of the American Dietetic Association* **102**: 1479–1490.

vegetables, and fresh fruits should be consumed daily to provide additional vitamins, minerals, and fiber. Approximately 10 glasses (2.3 l) of fluids per day (as total beverages, including milk and juices) should be consumed to prevent intestinal stasis, which can occur from reduced activity and the pressure of the enlarging uterus. There is no evidence that suggests the need for additional electrolytes (sodium, chloride, and potassium) during pregnancy. However, important considerations that may influence electrolyte balance include extreme loss through sweat and use of diuretics.

Energy

Energy requirements increase during pregnancy, depending on prepregnancy weight, amount and composition of weight gain, stage of pregnancy, and activity level of the mother. The recommended dietary allowance for energy is increased by 300 kcal (1.25 MJ) per day for adults and 500 kcal per day for young adolescents (<14 years) during the second and third trimesters. A total of 2500–2700 kcal/day should meet the energy needs of most pregnant women. This increase should be sufficient as long as body reserves are not depleted at the onset of pregnancy.

Observed energy intakes tend to be lower than recommended, but it should be recognized that as long as the amount and rate of weight gain are within the desirable range, there is a wide range of acceptable energy intakes. Energy expended during physical activity tends to be the largest variable in estimating energy expenditure. Because individuals vary considerably in their level and intensity of physical activity, the best advice is for pregnant women to eat enough to satisfy their physiologic appetite and support an appropriate rate of weight gain. Dieting during pregnancy is not recommended because low energy intakes can result in ketosis, which may pose a risk for the developing fetus, as well as inadequate consumption of other nutrients.

Protein

Additional protein is necessary to support the synthesis of maternal and fetal tissue, with greatest demands during the second and third trimesters when requirements are 20–28 g per day more than nonpregnant requirements. Women in industrialized countries generally consume more than the recommended intake of protein during pregnancy. Studies of protein deficiency during pregnancy are limited in number and often confounded by low energy intakes. However, the few studies that provided either extra energy or energy plus protein to undernourished women suggest that limited energy intakes are more unfavorable for pregnancy outcome than limited protein intakes. High-protein supplements are not recommended during pregnancy because they have been associated with lower birth weight and increased mortality and prematurity. However, when supplements provide protein-to-energy ratios comparable to those found in usual diets, higher rates of prematurity are not observed. Vegetarian women can substitute meat, poultry, fish, and eggs by additional servings of beans, tofu, nuts, and soymilk (preferably fortified with calcium

and vitamin D). Vitamin B_{12} will also need to be consumed as supplements to achieve the recommended intake of 2.6 μg/day.

Obesity in Pregnancy

Obesity is one of the most common nutritional problems complicating pregnancy in the Western world. Complications resulting from obesity include macrosomia in the infant (even when pregnancy weight gain is inadequate) and increased incidence of diabetes mellitus, hypertension, preeclampsia, and multiple gestation. More recently, it has become apparent that obese women are more likely than average weight women to produce an infant with spina bifida and several types of birth defects, and simply being overweight increases the risk of birth anomalies (although to a lesser extent than obesity). Maternal obesity is an important risk factor for failure to diagnose fetal abnormalities, particularly neural tube and cardiac defects, because ultrasonographic visualization of fetal anatomy is impaired. Obese women are also more likely to have delayed labor and repeated cesarean sections.

Dietary recommendations for obese women should consider optimal maternal weight gain. Obese or overweight women do not need to gain as much weight during pregnancy as normal weight or thin women, and in fact their weight gain is usually lower without intentional dietary restriction. In part this is explained by their larger increase in basal metabolic rate during pregnancy compared to women with normal or low fatness. However, weight gains of 6.8–11.4 kg are recommended to at least account for the weight of the fetus and maternal support tissues. Because obese women may be resistant to the idea of gaining further weight, special attention may be required to explain that pregnancy is not a time for weight loss. Some practical guidelines for obese pregnant women are listed in **Table 2**.

Table 2 Practical guidelines for overweight or obese mothers

1. Weight loss is not advised during pregnancy.
2. For women with a BMI of 26.1–29.0, a weight gain of 7–11.5 kg and a normal dietary intake for pregnancy should be recommended.
3. For women with a BMI of >29.0, the mother should be advised to gain at least 7 kg during pregnancy.
4. Nutritional counseling should focus on lowering intake of energy-dense foods that are low in other nutrients.
5. Obese women should be made aware of the increased risk for glucose intolerance. Screening for gestational diabetes should occur at the first prenatal visit, with repeated testing at 28 weeks of gestation if negative.
6. Frequent blood pressure monitoring with a cuff of the appropriate size is essential. If the patient has chronic hypertension, appropriate medication and a reduced sodium diet are indicated.
7. Weight loss should be encouraged after delivery.

Modified from Wolfe HM and Gross TL (1994) Obesity in pregnancy. *Clinical Obstetrics and Gynecology* **37**: 596–604.

Alcohol and Caffeine

There is abundant evidence that associates heavy alcohol consumption with teratogenicity, although the mechanisms by which this occurs are not completely understood. Heavy drinking during pregnancy increases the risk of prenatal and postnatal growth retardation, mental retardation, major birth defects, and learning disabilities such as those seen in fetal alcohol syndrome. The impacts of either binge drinking or moderate alcohol consumption, defined as no more than one drink per day, have not been satisfactorily evaluated and therefore the recommendation is that women abstain from consuming any alcohol during pregnancy.

Caffeine readily crosses the placenta to the fetus, where it may affect fetal heart rate and breathing. A meta-analysis of 12 studies found an increase in spontaneous abortions, risk of miscarriages, and lower birth weight in heavy coffee drinkers, although possible effects of smoking and alcohol could not be separated. No studies in humans have shown a link between caffeine consumption and birth defects, although massive doses of caffeine given to mice are teratogenic. Reports of an association between prenatal caffeine consumption and motor development in humans have been conflicting. Because it has not been proven without a doubt that caffeine does not cause birth defects and other problems in humans, the recommendation is that all pregnant and lactating women should consume no more than 300 mg caffeine per day. The caffeine content is approximately 85 mg/150 ml cup of percolated coffee, 60 mg/15 ml of instant coffee, 40 mg/oz. of espresso, 30 mg/150 ml tea, and 36 mg/12 oz. (350 ml) cola beverages.

Artificial Sweeteners

High doses of saccharin are weakly carcinogenic in rats so it is recommended that saccharin be consumed in no more than moderate amounts during pregnancy. Safety concerns about the use of aspartame are limited to pregnant women with phenylketonuria; individuals with this condition should only consume this sweetener in amounts compatible with their need to strictly monitor their intake of phenylalanine. Safety of acesulfame-K during pregnancy has been confirmed in animal studies.

Nutrient Supplements

Many pregnant women have special circumstances that prevent them from consuming an adequate diet. These include lactose intolerance, substance abuse, being an adolescent or a strict vegetarian, and having multiple fetuses. For these women, a multivitamin–mineral supplement may have positive benefits, including a reduced risk of some developmental defects, improvements in immune function, and a reduction in the onset and/or progression of some cancers. Anemia is common during pregnancy and justifies treatment with iron supplements. However, some micronutrients commonly taken as supplements are potentially harmful to the developing fetus or, less often, to the pregnant woman, especially when consumed in relatively large amounts.

Vitamin A

Although poor maternal vitamin A status is associated with preterm birth, intrauterine growth retardation, low birth weight, and increased maternal infection in developing regions of the world, vitamin A deficiency is quite rare in industrialized countries due the higher intake of the preformed vitamin (retinol). An increase of 50 μg/day for the entire pregnancy period is added to the EAR for nonpregnant adolescent girls and women, based on accumulation of vitamin A in the liver of the fetus. Excessive preformed vitamin A (>3000 μg RAE) should be avoided shortly before or during pregnancy, especially in the early months, because of its potential teratogenicity, so the upper limit for vitamin A intake in pregnancy is set at 3000 μg RAE/day for all women of childbearing age. An alternative dose schedule is up to 8500 RAE weekly during pregnancy. Fetal vitamin A toxicity and birth defects have also occurred from ingestion of isotretinoin and etrentinate, drugs used for treatment of severe cystic acne. High intakes of βcarotene (a precursor of vitamin A) do not have the same teratogenic effects.

Vitamin D

Because of its importance in increasing calcium retention, recommended intakes of vitamin D are doubled during pregnancy. Vitamin D deficiency during pregnancy causes disorders of calcium metabolism, including neonatal hypocalcemia and tetany, hypoplasia of infants' tooth enamel, and maternal osteomalacia. Because the prevalence of vitamin D deficiency during pregnancy is high during the winter months at northern latitudes in regions such as Europe, the United States, and Canada, and Japan, vitamin D supplements may be necessary for women who live in these regions or who have little exposure to sunlight. A national survey in the United States conducted in the 1990s revealed that approximately 40% of African American women in the southeastern region had low blood levels of 25-hydroxyvitamin D, and that not drinking vitamin D-fortified milk was a risk factor for deficiency. In the absence of vitamin D fortification or supplementation, infants in Paris, for example, have higher plasma levels of parathyroid hormone and other indications of vitamin D deficiency if they are born soon after the winter months. Vitamin D supplements reversed the indications of vitamin D deficiency. High maternal intakes of vitamin D are toxic and were implicated as the cause of a syndrome that included mental and physical growth retardation and hypercalcemia in British infants between 1953 and 1957. Excessive amounts of vitamin D taken during gestation have also caused aortic stenosis and abnormal skull development in infants. The upper limit for vitamin D in pregnancy is 50 μg per day, the same as for nonpregnant women.

Folic Acid and Vitamin B_{12}

There is a substantial increase in folate requirements during pregnancy, from 400 μg Dietary Folate Equivalents in the nonpregnant state to 600 μg per day, because of increased erythropoiesis and fetal–placental growth. Increased folate intakes throughout childbearing age are recommended to prevent neural tube defects such as spina bifida and anencephaly, the most common birth defects, and to lower the risk of abruptio placenta. To be effective for preventing neural tube defects in women at risk for producing an infant with this condition, increased folate intakes are needed preconception and early in pregnancy. The neural tube closes by 28 days of gestation, which is before many women realize that they are pregnant. It is for this reason that increased folate intakes are recommended throughout the childbearing years. In the United States and Canada, fortification of flour with folic acid in recent years has greatly increased folate intakes and improved status in the population; prior to fortification, typical intakes of folate were only about half of the recommended amount. No adverse effects were reported in recent studies in which pregnant women consumed up to 4 mg of folic acid per day during pregnancy.

The RDA for vitamin B_{12} increases slightly during pregnancy to 2.6 μg/day. Vitamin B_{12} supplements are definitely required by pregnant women who are strict vegetarians; the vitamin is found only in animal products and the usefulness of the form of the vitamin found in algae and bacteria is not clear. An

adequate intake of the vitamin during pregnancy is at least as important as the woman's vitamin B_{12} status at conception because the recently absorbed vitamin is more readily transported to the fetus than is the vitamin in maternal liver stores. Homocysteinemia is emerging as a common risk factor for several abnormal pregnancy outcomes, especially for preeclampsia, birth defects, and low birth weight, although there has been little research on whether maternal vitamin B_{12} deficiency causes these problems. Infants born to women with low vitamin B_{12} intakes are at high risk of growth failure and neurobehavioral problems that emerge when the infant is a few months old and may be permanent. Although supplements containing the recommended dietary intake of the vitamin are probably adequate for pregnancy, no adverse effects of consuming higher amounts have been reported.

Vitamin C

Low plasma vitamin C concentrations have been associated with preeclampsia and premature rupture of the membranes. There has been some concern about fetal vitamin C dependency induced by excessive maternal vitamin C intakes, but this is based on only one anecdotal report. Requirements for the vitamin in pregnancy increase to 80 mg/day for adolescents and 85 mg/day for adult women, an amount estimated to provide sufficient quantities for the fetus. Those who are heavy smokers (>20 cigarettes/day) may need twice the RDA for this vitamin.

Vitamin K

Usual diets provide adequate amounts of vitamin K for pregnant women. Newborn infants are routinely given a supplement of vitamin K by intramuscular injection because exclusively breast-fed infants are at risk of developing fatal intracranial hemorrhage secondary to vitamin K deficiency. This practice is quite safe.

Iron

Demands for iron are increased by approximately 700–800 mg during pregnancy and most of this is needed during the last two trimesters. Because the risk of becoming anemic is greater during pregnancy and there is an increased risk of a compromised pregnancy outcome for anemic women (including lower birth weight and less neonatal iron stores), most recommendations in the United States advise all pregnant women with a well-balanced diet take a supplement of 30 mg of ferrous iron daily starting at their first prenatal visit. If iron deficiency anemia is detected by routine testing, the recommendation is that 60–120 mg of ferrous iron be given in divided doses throughout the day. The World Health Organization recommends 60 mg/day throughout pregnancy (plus 400 μg folic acid) because of the higher prevalence of iron deficiency anemia in most low-income countries throughout the world.

Gastrointestinal side effects, mainly heartburn, nausea, upper abdominal discomfort, diarrhea, and constipation, increase with high iron doses and contribute to poor compliance with taking daily iron supplements. Supplements of 15 mg of zinc and 2 mg of copper daily are also recommended for pregnant women taking >30 mg iron per day because iron can interfere with the absorption of other minerals if given as a supplement without food. It is therefore important not to exceed recommended intakes of iron during pregnancy. Research suggests that it may be as effective to consume the recommended intakes once per week as it is to take them daily because daily iron supplements gradually block the absorption of subsequent doses.

Zinc

Typically, zinc intakes are below recommended amounts for pregnancy even in industrialized countries. In populations in which zinc deficiency is common, the prevalence of malformations and low birth weight is higher, although the causal role of zinc deficiency has not been proven. Zinc supplementation is recommended for pregnant women who ordinarily consume an inadequate diet, smoke, are substance abusers, or are carrying multiple fetuses. However, copper absorption may begin to be impaired at zinc intakes of approximately 18.5 mg/day, and a daily intake of 50 mg zinc impairs both iron and copper absorption. These negative effects are believed to be stronger if the minerals are taken without food.

Sodium

Due to hormonal changes during pregnancy, sodium metabolism is altered. At one time, dietary restriction of sodium was a common treatment for maternal edema, although it is ineffective. The newborn infants of women who had restricted their sodium intake drastically during pregnancy were observed to have hyponatremia. In animals, sodium restriction during pregnancy leads to water intoxication along with renal and adrenal tissue degradation of the pregnant animal. Therefore, sodium restriction during pregnancy is not advisable.

Iodine

Typically, the iodine intakes of pregnant women in the industrialized world easily meet recommended intakes, often as the result of consuming iodized salt. Maternal iodine deficiency and suboptimal iodine intake have been associated with cretinism, mental development impairments in utero, and infant mortality. Iodine deficiency before or during early pregnancy has the most severe effects, and in regions of endemic iodine deficiency cretinism should be prevented by treating maternal iodine deficiency before or during the first 3 months of pregnancy. However, hypothyroidism in the mother and fetus can be corrected by iodine administration in the third trimester. It appears to be safe to administer massive amounts (500 mg iodine) to pregnant women, orally or intramuscularly. There have been no reports of adverse effects of excessive iodine administration during pregnancy, and thus the upper limits are the same as for nonpregnant women (1100 μg/day for adults).

Teratogens

The World Health Organization estimates that 15% of all clinically recognizable pregnancies end in abortion. Of these, 50–60% are due to chromosomal abnormalities. In addition, 3–6% of all offspring are malformed. The causes of these malformations can be divided into three categories: unknown, genetic, and environmental. Environmental causes only account for 10% of all congenital malformations and can be further divided into maternal conditions, infectious agents, mechanical problems (deformations), and chemicals (including prescription drugs and high-dose ionizing radiation). Chemical environmental causes include consumption during pregnancy of the teratogenic agents discussed later. These account for less than 1% of all congenital malformations but are important in that the exposures to these chemicals may be preventable.

Several anticancer drugs cause problems for fetal development. Aminopterin can induce abortion within its therapeutic range, and it causes microcephaly, hydrocephaly, cleft palate, meningomyelocele, intrauterine growth retardation, abnormal cranial ossification, and mental retardation. Cyclophosphamide interacts with DNA and can result in cell death. Its use during pregnancy can result in growth retardation, ectrodactyly, syndactyly, and cardiovascular anomalies.

Some antibiotics cause abnormal fetal development if taken by the pregnant woman. Streptomycin can cause hearing problems, although the risk of this is quite low. Tetracycline may produce staining of the teeth and bones if taken late in the first trimester or during the last two trimesters.

Anticonvulsants can also cause adverse pregnancy outcomes. Carbamazepine produces minor craniofacial defects, fingernail hypoplasia, and developmental delays. Trimethadione causes 'fetal trimethadione syndrome,' characterized by V-shaped eyebrows, low-set ears, a high-arched palate, irregular teeth, central nervous system anomalies, and severe developmental delays. Valproic acid causes spina bifida and facial dysmorphology in the fetus of 1% of pregnant users.

Other potentially teratogenic drugs include androgens, which result in masculinization of the embryo and stimulate growth and differentiation of sex steroid receptor-containing tissues. Angiotensin-converting enzyme inhibitors are antihypertensive agents that have detrimental effects during the second and third trimesters, including fetal death, oligohydramnios, pulmonary hypoplasia, neonatal anuria, intrauterine growth retardation, and skull hypoplasia. The pregnant woman who uses cocaine risks preterm delivery, fetal loss, intrauterine growth retardation, microcephaly, neurobehavioral abnormalities, vascular disruptive phenomena, cerebral infarctions, and certain types of visceral and urinary tract malformations. Coumadin, a vitamin K analog, is an anticoagulant and in the first trimester can produce malformations, including nasal hypoplasia, stippling of secondary epiphysis, intrauterine growth retardation, and anomalies of the eyes, hands, neck, and central nervous system. Lithium carbonate, an antidepressant, has teratogenic effects in animals but these have not been confirmed in humans.

Contaminants

Most heavy metals, such as lead and mercury, are embryotoxic. High maternal serum lead concentrations increase the risk of abortion and adversely affect the central nervous system of the developing fetus, leading to a low IQ and abnormal behavior of the infant. PCBs are environmental contaminants that remain in the body up to 4 years after exposure. The fetus of a pregnant woman exposed to PCBs is at increased risk of fetal growth retardation, abnormal skull calcifications, deformed nails, and pigmentation of gums, nails, and the groin. Organic mercury compounds tend to accumulate in fat tissue and cause cell death due to the inhibition of cellular enzymes. These compounds cause cerebral palsy, microencephaly, mental retardation, blindness, and cerebellar hypoplasia in the infant.

Special Conditions

Nausea and Vomiting

Morning sickness or nausea is common in the early months of pregnancy. It is rarely a condition to cause alarm, except when there is excessive vomiting. In this situation, an acute protein and energy deficit and loss of minerals, vitamins, and electrolytes may result. Treatment of this condition is by consuming small frequent meals and a low-fat, high-carbohydrate diet. Prolonged, persistent vomiting (hyperemesis gravidarum) occurs in approximately 2% of pregnant women. Hospitalization is usually required, with intravenous fluid and electrolyte replacement to prevent dehydration.

Heartburn

Heartburn is a common complaint during the latter part of pregnancy due to the pressure of the enlarged uterus on the stomach in combination with the relaxed esophageal sphincter. This can usually be relieved by limiting the amount of food consumed at one sitting and avoiding lying in a reclining position after eating.

Constipation and Hemorrhoids

Pregnant women often develop constipation, most frequently during the latter stages of pregnancy. It is caused by reduced gut motility, physical inactivity, and the pressure exerted on the bowel by the enlarged uterus. The weight of the fetus and the downward pressure on the veins can lead to hemorrhoid formation. These conditions can be treated with increased consumption of high-fiber foods and dried fruits and higher fluid intake. Bulk-forming laxatives can also be used; however, there is a risk of alterations in electrolyte absorption with chronic use of laxatives.

Edema

Mild edema (fluid accumulation) is often present in the hands, feet, and legs in the third trimester. It is caused by the pressure of the enlarging uterus on the veins returning fluid from the legs. This fluid is often mobilized in the evening when the woman is lying down. This is a normal condition and does not require any special dietary or other treatment.

Diabetes in Pregnancy

For women with diabetes, nutritional counseling should include adequate dietary intake, frequent glucose monitoring, insulin management to meet the growth needs of the fetus, maintaining optimal blood glucose levels, and preventing ketosis and depletion of the mother's nutrient stores. The demands of pregnancy may impose a need for insulin in pregnant women whose condition was controlled through diet alone in the nonpregnant state. Because of hormonal changes during the first and second half of pregnancy, changes to the diet and the insulin dosage may be necessary.

Gestational diabetes occurs only during pregnancy and usually resolves after pregnancy. It occurs in 5–10% of pregnancies and most commonly arises after 20 weeks of gestation. Gestational diabetes can be treated largely through nutritional care and moderate exercise to achieve weight control. Nutritional recommendations are to limit protein intake to 15% of total calories, consume 55% of total calories as carbohydrate, and limit fat intake to 30% or less of total calories. Cholesterol intake should be 300 mg/day or less, simple carbohydrate intake should be limited, and sodium intake should not exceed 1000 mg/1000 kcal. Insulin is rarely needed, although blood glucose levels should be monitored daily.

Hypertension in Pregnancy

Pregnancy-induced hypertension is a syndrome characterized by hypertension, proteinuria, and edema. This condition usually develops in the third trimester and occurs in approximately 7 or 8% of pregnant women. It occurs more often in women who are young, pregnant for the first time, or are of low socioeconomic status. The exact cause of this condition is unknown, but most researchers agree that it is associated with a decreased uterine blood flow leading to reduced fetal nourishment. Previous treatments for this condition included sodium restriction and diuretics; however, neither of these has been successful in altering blood pressure, weight gain, or proteinuria in this condition.

Multiple Births

Women pregnant with twins or multiple fetuses should gain more weight than those with singleton births, approximately 15–20 kg. Nutrient supplementation should include at least zinc and vitamin B_6 in addition to the iron supplements recommended for all pregnant women.

See also: **Alcohol**: Absorption, Metabolism and Physiological Effects. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Caffeine**. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Early Origins of Disease**: Fetal. **Folic Acid**. **Food Safety**: Other Contaminants; Heavy Metals. **Hypertension**: Etiology. **Iodine**: Physiology, Dietary Sources and

Requirements; Deficiency Disorders. **Iron**. **Obesity**: Complications. **Pregnancy**: Role of Placenta in Nutrient Transfer; Nutrient Requirements; Energy Requirements and Metabolic Adaptations; Weight Gain; Safe Diet for Pregnancy; Prevention of Neural Tube Defects; Pre-eclampsia and Diet. **Sodium**: Physiology. **Vegetarian Diets**. **Vitamin A**: Deficiency and Interventions. **Vitamin D**: Rickets and Osteomalacia. **Vitamin K**.

Further Reading

Allen LH (1994) Nutritional supplementation for the pregnant woman. *Clinical Obstetrics and Gynecology* **37**(3): 587–595.

Allen LH (2001) Pregnancy and lactation. In: Bowman BA and Russell RM (eds.) *Present Knowledge of Nutrition*, 8th edn. Washington, DC: ILSI Press.

American Diabetes Association (1991) Position statement: Gestational diabetes mellitus. *Diabetes Care* **14**: 5–6.

Institute of Medicine (1987) *Committee on Nutrition of the Mother and Preschool Child. Laboratory Indices of Nutritional Status during Pregnancy*. Washington, DC: National Academy of Sciences.

Institute of Medicine (1990) *Nutrition during Pregnancy. National Research Council*. Washington, DC: National Academy Press.

Institute of Medicine (1997) *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride*. Washington, DC: National Academy Press.

Institute of Medicine (1998) *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin, and Choline*. Washington, DC: National Academy Press.

Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc*. Washington, DC: National Academy Press.

Institute of Medicine (2002) *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, and Amino Acids (Macronutrients)*. Washington, DC: National Academy Press.

Institute of Medicine (2004) *Dietary Reference Intakes for Water, Potassium, Sodium, Chloride, and Sulfate*. Washington, DC: National Academy Press.

Kaiser LL and Allen L (2002) Position of the American Dietetic Association: Nutrition and lifestyle for a healthy pregnancy outcome. *Journal of the American Dietetic Association* **102**: 1479–1490.

King JC, Bronstein MN, Fitch WL *et al.* (1987) Nutrient utilization during pregnancy. *World Reviews of Nutrition and Diet* **52**: 71–142.

Lewis DD and Woods SE (1994) Fetal alcohol syndrome. *American Family Physician* **50**: 1025–1032.

March of Dimes (2002) *Nutrition Today Matters Tomorrow: A Report from the March of Dimes Task Force on Nutrition and Optimal Human Development*. White Plains, NY: March of Dimes.

Neuhouser MLS (1996) Nutrition during pregnancy and lactation. In: Mahan LK and Escott-Stump S (eds.) *Krause's Food, Nutrition & Diet*, 9th edn. Philadelphia: WB Saunders.

Rosso P (1990) *Nutrition and Metabolism in Pregnancy: Mother and Fetus* New York: Oxford University Press.

Wolfe HM and Gross TL (1994) Obesity in pregnancy. *Clinical Obstetrics and Gynecology* **37**: 596–604.

Prevention of Neural Tube Defects

P N Kirke, The Health Research Board, Dublin, Ireland

J M Scott, Trinity College, Dublin, Ireland

Neural tube defects (NTDs) are major congenital malformations of the central nervous system resulting in fetal and perinatal death and severe handicap in the majority of survivors. The finding that folic acid can prevent most NTDs ranks as one of the most important medical research discoveries in recent times. In this article, the epidemiology of NTDs is reviewed, focusing primarily on the role of folic acid and, to a lesser extent, vitamin B_{12} in the etiology and prevention of these malformations. The causes of the approximately 30% of NTDs that are estimated not to be related to folate are also briefly considered. The mechanisms underlying the link between folate, vitamin B_{12}, and NTD etiology are examined, and the rapidly expanding research literature on genetic risk factors is reviewed. The main issues in using folic acid to prevent NTDs are discussed: ways to increase folate/folic acid intakes, supplementation, fortification, and safety. The role of other nutrients in NTD prevention is considered. Recommendations on using folic acid to prevent NTDs have been issued by various national health authorities and the main points in these recommendations are presented.

Epidemiology

Failure of the embryonal neural tube to close normally between 24 and 28 days after conception gives rise to a group of severe congenital malformations known as NTDs that includes spina bifida, anencephalus (approximately 50 and 40% of cases, respectively), encephalocoele, and iniencephaly. These anomalies are believed to be caused by an interaction of genetic predisposition and environmental factors, and many different factors have been pursued. Evidence of the importance of nutrition has accumulated since the 1960s, and the key role of folate/folic acid in the pathogenesis of these malformations was demonstrated conclusively in 1991.

Genetic and Environmental Factors

Evidence of a genetic component in the etiology of NTDs includes familial recurrence patterns, ethnic variation, and sex variation (more common in

females). More direct evidence of the role of genetic factors is the discovery that the gene encoding for the thermolabile variant of the 5,10-methylenetetrahydrofolate reductase enzyme is more common in individuals with spina bifida than in controls. The most striking environmental, or nongenetic, factors are the protective effect of folic acid and the marked variations in prevalence over time and between areas. The prevalence rates of NTDs at birth have been falling in most countries, particularly in regions that traditionally had high rates. It is assumed, but not scientifically proven, that better nutrition is a main factor determining this trend. Variations with season, social class (more common in disadvantaged groups), and, to a lesser extent, maternal age and reproductive history provide further evidence of the role of environmental factors. Several of these factors may be explained in whole or in part on nutritional grounds.

Folate/Folic Acid

There is a vast literature on the role of folate/folic acid in the etiology and prevention of NTDs. Evidence that folic acid can prevent NTDs comes from two main types of studies: observational studies of dietary folate intake and of supplementation with folic acid preparations and intervention studies. The strongest evidence on the efficacy of folic acid comes from randomized controlled trials, notably the Medical Research Council (UK) trial on NTD recurrence and the Hungarian trial on NTD occurrence (i.e., first-time NTDs). These and other intervention studies are summarized in **Table 1**. Following earlier research, the Medical Research Council trial published in 1991 conclusively established the efficacy of folic acid in preventing NTD recurrence. This trial used a research design to investigate the effects of both folic acid and a combination of other vitamins. The recurrence rate in the groups that received folic acid (1.0%) was significantly lower than that in the groups that did not take folic acid (3.5%), giving a 71% protective effect. Thus, 29% of NTDs were not prevented by folic acid, at least not at the very high pharmacological dose of 4 mg daily used in the trial. The multivitamin combination without folic acid had no protective effect. The main observational studies that have examined the effect of periconceptional use of vitamin supplements containing folic acid on NTD pregnancies are illustrated in **Table 2**. All of these studies but one found a marked protective effect of supplementation against NTD occurrence. In most of the studies of NTD occurrence, the daily dose of folic acid was between 0.4 and 0.8 mg. Studies of dietary folate intake also show a protective effect of high intakes during the periconceptional period. The consistent finding of a protective

Table 1 Intervention studies of periconceptional folic acid supplementation and NTD risk

Study	*Design*	*Daily dose folic acid (mg)*	*Outcome: No. of NTDs*	*Relative risk*	*Comments*
UK Medical Research Council Trial (1991)	Randomized controlled trial, international	4.0	6/593 suppl. 21/602 not suppl.[a]	0.29	Significant[b]
Laurence *et al.* (1981)	Randomized controlled trial, Wales	4.0	2/60 suppl. 4/51 not suppl.	0.42	Not significant Small numbers
Kirke *et al.* (1992)	Randomized controlled trial, Ireland	0.36	0/172 suppl. 1/89 not suppl.	0.00	Not significant Small numbers
Czeizel and Dudas (1992)	Randomized controlled trial, Hungary	0.8	0/2104 suppl. 6/2052 not suppl.	0.00	Significant[b]
Indian Council of Medical Research Trial (2000)	Randomized controlled trial, India	4.0	4/137 suppl. 10/142 not suppl.	0.41	Not significant Small numbers
Smithells *et al.* (1983)	Nonrandomized controlled trial, UK	0.36	3/454 suppl. 24/519 not suppl.	0.14	Significant[b]
Vergel *et al.* (1990)	Nonrandomized controlled trial, Cuba	5.0	0/81 suppl. 4/114 not suppl.	0.00	Not Significant Small numbers
Berry *et al.* (1999)	Nonrandomized controlled trial, China	0.4	Northern region 13/13 012 suppl. 16/3 318 not suppl. Southern region 34/58 638 suppl. 28/28 265 not suppl.	0.21 0.59	Significant[b] Significant[b]

[a]Six NTD pregnancies in 593 women supplemented with folic acid and 21 NTD pregnancies in 602 women not supplemented with folic acid.
[b]Statistically significant difference in NTD rate between supplemented and nonsupplemented groups.

Table 2 Main observational studies of the effect of periconceptional use of folic acid supplements on NTD risk[a]

Study	*Odds ratio*
Mulinare *et al.* (1988)	0.41
Mills *et al.* (1989)	0.94
Milunsky *et al.* (1989)	0.29
Werler *et al.* (1993)	0.60
Shaw *et al.* (1995)	0.65

[a]The difference between folate-supplemented and unsupplemented groups was statistically significant in all studies except Mills *et al.* (1989).

effect of dietary folate and folic acid supplementation in virtually all these different types of studies is very striking. Further evidence implicating folate comes from studies linking maternal folate status to pregnancies affected by NTDs. The main studies that have been published on serum/plasma folate and red cell folate (RCF) are summarized in **Tables 3** and **4**. The differences between affected and unaffected pregnancies are more pronounced in the first trimester of pregnancy. Maternal use of folic acid antagonists during early pregnancy increases the risk not only of NTDs but also of other congenital defects.

Vitamin B_{12}

The role of vitamin B_{12} in NTDs is of particular interest because of the close metabolic relationship between this nutrient and folate. The results of some studies of maternal levels of serum vitamin B_{12} in NTD pregnancies are shown in **Table 5**. As for folate, lower levels of vitamin B_{12} are generally seen in affected pregnancies, especially in the first trimester. Studies based on amniotic fluid have

Table 3 Serum folic acid (SFA) and central nervous system defects

Study		*No. of pregnancies*	*Mean SFA ($\mu g\ l^{-1}$)*	*Difference*	*Statistical significance*
Blood taken at antenatal booking					
Hall *et al.* (1977)	Affected	11	6.3	−0.3	No
	Unaffected[a]	>1000	6.6		
Molloy *et al.* (1985)	Affected	32	3.4[b]	0.0	No
	Unaffected	384	3.4[b]		
Kirke *et al.* (1993)	Affected	81	3.5[b]	−1.1	Yes
	Unaffected	247	4.6[b]		
Blood taken in first trimester					
Smithells *et al.* (1976)	Affected	5	4.9	−1.4	No
	Unaffected	953	6.3		
Mills *et al.* (1992)	Affected	89	4.1	−0.2	No
	Unaffected	172	4.3		
Wald *et al.* (1996)	Affected	16	4.3[b]	−1.4	No
	Unaffected	36	5.7[b]		
All women (antenatal booking and first trimester)				−0.6 95% CI (−1.0, −0.2)	Yes ($p = 0.005$)
Blood taken in second trimester					
Economides *et al.* (1992)	Affected	8	9.8[b]	2.4	Yes
	Unaffected	24	7.4[b]	95% CI (−0.04, 4.84)	($p = 0.054$)
Blood taken after delivery					
Emery *et al.* (1969)	Affected	19	4.9	0.3	No
	Unaffected	37	4.6		
Yates *et al.* (1987)	Affected	20	2.8	−0.5	No
	Unaffected	20	3.3		
Bower and Stanley (1989)	Affected	61	5.6	−0.1	No
	Unaffected	140	5.7		
Wild *et al.* (1993)	Affected	29	6.2[b]	0.7	No
	Unaffected	29	5.5[b]		
All women (after delivery)				−0.03 95% CI (−0.5, 0.4)	No ($p = 0.090$)

[a]Unaffected women were those without a neural tube defect pregnancy either before or during the particular study, except for Wald *et al.* (1996), in which women had at least one neural tube defect pregnancy before the study.
[b]Median value.
From Wald NJ, Hackshaw AK, Stone R and Sourial NA (1996) Blood folic acid and vitamin B_{12} in relation to neural tube defects. *British Journal of Obstetrics and Gynaecology* **103**: 319–324, Blackwell Scientific.

Table 4 Red cell folate (RCF) and central nervous system defects

Study		*No. of pregnancies*	*Mean RCF ($\mu g\ l^{-1}$)*	*Difference*	*Statistical significance*
Blood taken at antenatal booking					
Kirke *et al.* (1993)	Affected	81	269[b]	−69	Yes
	Unaffected[a]	247	338[b]		
Blood taken in first trimester					
Smithells *et al.* (1976)	Affected	6	141	−87	Yes
	Unaffected	959	228		
Wald *et al.* (1996)	Affected	14	156[b]	−6	No
	Unaffected	26	162[b]		
All women (antenatal booking and first trimester)				−77	Yes
				95% CI (−94, −60)	($p < 0.001$)
Blood taken in second trimester					
Laurence *et al.* (1981)	Affected	4	238	−43	No
	Unaffected	47	281		
Economides *et al.* (1992)	Affected	8	435[b]	35	No
	Unaffected	24	400[b]		
All women (second trimester)				5	No
				95% CI (−76, 86)	($p = 0.90$)
Blood taken after delivery					
Yates *et al.* (1987)	Affected	20	178	−90	Yes
	Unaffected	20	268		
Bower and Stanley (1989)	Affected	61	301	−7	No
	Unaffected	140	308		
Wild *et al.* (1993)	Affected	29	247[b]	24	No
	Unaffected	29	223[b]		
All women (after delivery)				−6	No
				95% CI (−33, 21)	($p = 0.66$)

[a]Unaffected women were those without a neural tube defect pregnancy either before or during the particular study, except for Wald *et al.* and Laurence *et al.*, in which women had at least one neural tube defect pregnancy before the study.
[b]Median value.
From Wald NJ, Hackshaw AK, Stone R and Sourial NA (1996) Blood folic acid and vitamin B_{12} in relation to neural tube defects. *British Journal of Obstetrics and Gynaecology* **103**: 319–324.

consistently found lower vitamin B_{12} levels in affected pregnancies. It is possible that the low levels of vitamin B_{12} coincide with low levels of folate, although the findings of a case–control study in Dublin suggest that they are independent risk factors and the distribution of the two ingredients in food is dissimilar. In another smaller study, lower levels of vitamin B_{12} in affected pregnancies were not independent of folate levels. On biochemical grounds, there is so much interaction between the pathways involving both nutrients that it is possible that deficiency of either could affect a common event in the closure of the neural tube. The role of vitamin B_{12} in NTDs is discussed further later.

Other Nutritional Factors

Vitamin C, vitamin A, and zinc have also been linked to NTDs. Lower maternal levels of white cell vitamin C were reported in affected compared to unaffected pregnancies in one small study. Large doses of natural or synthetic vitamin A consumed by the mother during pregnancy have been associated with congenital anomalies in her offspring. In a large US study of maternal vitamin A intake before and during early pregnancy, a total daily intake greater than 15 000 IU was associated with an increased risk of birth defects, especially of structures arising from the cranial neural crest (craniofacial, central nervous system, thymic, and heart defects), but the risk of NTDs was not raised. However, these findings have been challenged. Children born to women who have vitamin A supplements at levels found in current multivitamin preparations have not been shown to be at increased risk of birth defects. Although several studies have linked zinc deficiency or abnormalities in zinc metabolism to NTDs, the results have not been consistent. The role of zinc in NTD aetiology requires further clarification. Research on the association between riboflavin and folate and homocysteine levels in people homozygous for the

Table 5 Maternal serum vitamin B_{12} (SB12) and central nervous system defects

Study		*No. of pregnancies*	*Median SB12 ($ng\,l^{-1}$)*	*Difference*	*Statistical significance*
Blood taken at antenatal booking					
Kirke *et al.* (1993)	Affected	81	243	−53	Yes
	Unaffected[a]	247	296		
Blood taken in the first trimester					
Schorah *et al.* (1980)	Affected	6	288	−129	Yes
	Unaffected	48	417		
Molloy *et al.* (1985)	Affected	28	297	20	No
	Unaffected	363	277		
Mills *et al.* (1992)	Affected	89	483[b]	−37	No
	Unaffected	178	520[b]		
Wald *et al.* (1996)	Affected	18	230	−10	No
	Unaffected	75	240		
All women (first trimester and antenatal booking)				−38	Yes
				95% CI (56, −20)	($p < 0.001$)
Blood taken in second trimester					
Economides *et al.* (1992)	Affected	8	205	−25	No
	Unaffected	32	230	95% CI (−58, 8)	($p = 0.12$)
Blood taken after delivery					
Yates *et al.* (1987)	Affected	20	300[b]	−20	No
	Unaffected	20	320		
Wild *et al.* (1993)	Affected	29	449	−40	No
	Unaffected	29	489		
All women (after delivery)				−34	No
				95% CI (−83, 15)	($p = 0.17$)

[a]Unaffected women were those without neural tube defect pregnancy either before or during the particular study, except for Wald *et al.* (1996), in which women had a least one neural tube defect pregnancy before the study.
[b]Median value.
From Wald NJ, Hackshaw AK, Stone R and Sourial NA (1996) Blood folic acid and vitamin B_{12} in relation to neural tube defects. *British Journal of Obstetrics and Gynaecology* **103**: 319–24, Blackwell Scientific.

5,10-methylenetetrahydrofolate reductase C677T genetic polymorphism suggests a possible role for riboflavin in the aetiology and prevention of NTDs.

Research in the United States has shown that women who are obese (defined as prepregnancy body weight of more than 80 kg or body mass index greater than $29\,kg\,m^{-2}$) are more likely to have infants with NTDs and some other congenital malformations than women of average prepregnancy weight. In one study it was found that this association was independent of folate intake. Although the underlying mechanism is unclear, these findings suggest that it may involve something other than folate. Studies in the United States found that dieting behaviors involving restricted food intake during the first trimester of pregnancy and diarrheal illnesses during the periconceptional period were associated with increased NTD risk.

Other Causes of NTDs

It is estimated from the results of the Medical Research Council trial that approximately 30% of NTDs are not folate-related. The causes of this group of NTDs are unknown but are likely to include genetic and environmental factors. In this context, recent reports on obesity and NTDs are most interesting. Further research on this subject should result in a better understanding of the complex aetiology of NTDs. Nutritional factors other than folate may be involved—for example, vitamins B and C and zinc, as noted previously, and other nutrients.

Mechanisms

The possible mechanisms underlying the involvement of folate/folic acid in the etiology and prevention of NTDs are examined in this section.

Functions of Folate and Vitamin B_{12} and NTD Etiology

Folate acts as the intermediary in the transfer of methyl groups for two important processes in metabolism, namely the methylation reactions and the synthesis of the nucleic acids DNA and RNA

(**Figure 1**). The folate cofactor, N^5-methyltetrahydrofolate, acts via the vitamin B_{12}-dependent enzyme, methionine synthase, to remethylate homocysteine to produce methionine, which is converted to *S*-adenosylmethionine (SAM) via *S*-adenosylmethionine synthase. SAM is the universal methylator necessary for the synthesis of essential proteins, lipids such as myelin, and DNA. The folate cofactor also acts via methionine synthase to synthesize tetrahydrofolate, which, unlike N^5-methyltetrahydrofolate, can be polyglutamated and thereafter used to produce the nucleic acids DNA and RNA. Simple deficiency or metabolic impairment in the biochemical functions of either folate or vitamin B_{12} could, by interrupting DNA biosynthesis or methylation reactions, interfere with cell growth and function and tissue development during a period of very rapid cell proliferation of the fetal neural crest, thereby preventing normal closure of the neural tube.

Folate/Folic Acid and NTDs: Mechanisms

Does folic acid prevent NTDs by correcting simple dietary deficiency, by overcoming a problem in gastrointestinal absorption, or by overcoming some type of metabolic block? Recent research has helped to clarify the role of folate/folic acid in the etiology and prevention of these malformations. Blood samples were collected from women at their first antenatal clinic in the Dublin maternity hospitals and 81 women in this cohort subsequently had infants affected by NTDs. Folate and B_{12} status were compared in these 81 cases and in a control sample of 247 unaffected pregnancies by measuring plasma and RCF and plasma vitamin B_{12}. Although folate levels were significantly lower in the cases than in the controls, more than 91 and 86% of the cases had normal plasma and RCF levels, respectively. Thus, the vast majority of women who had an

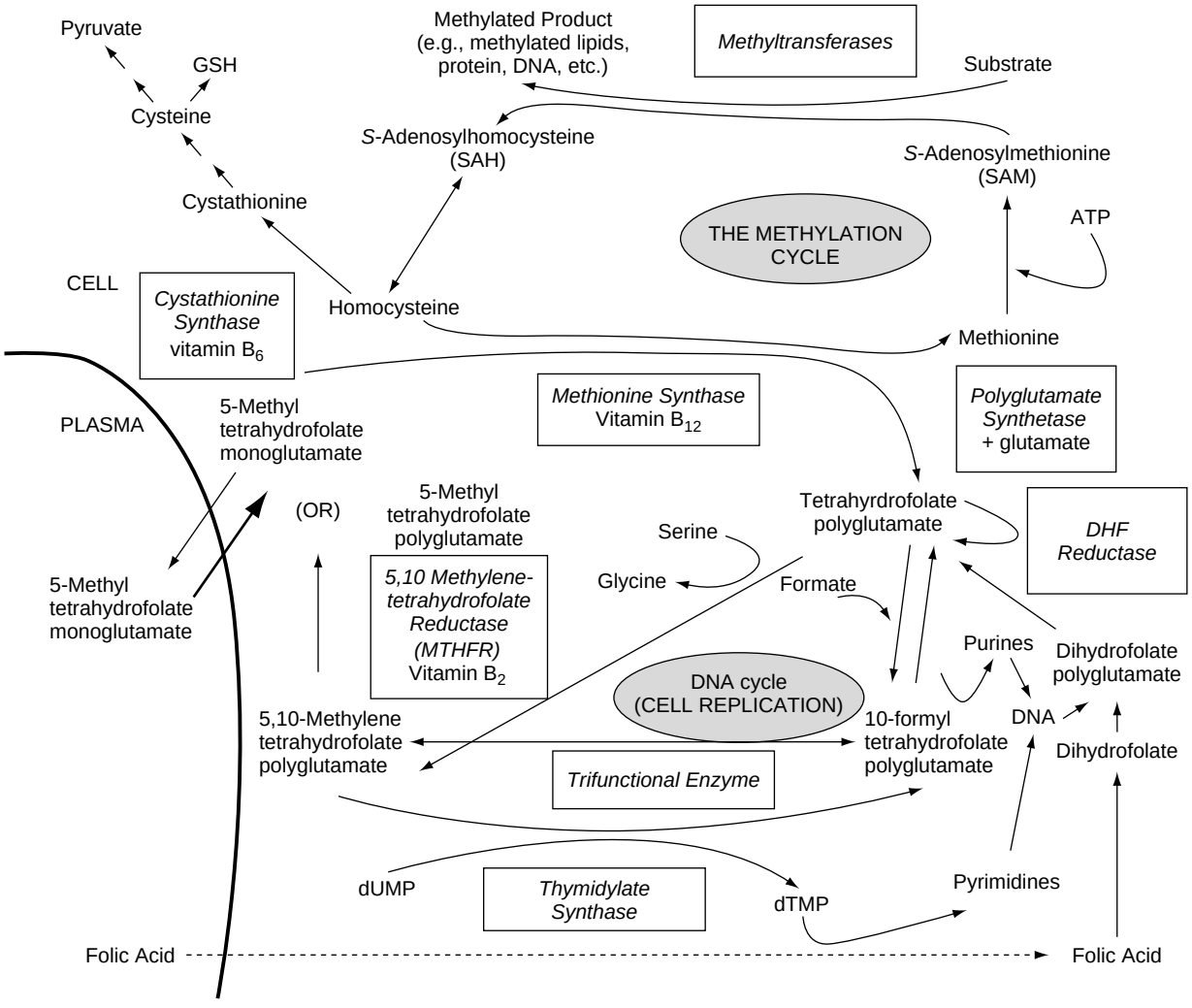


Figure 1 Intracellular pathways of folate and homocysteine metabolism and their relation to vitamin B_{12} function.

NTD birth were not folate deficient, as defined by conventional levels.

It has been suggested that women who have had children with NTDs may have a defect in gastrointestinal absorption of folate or folic acid, but there is no strong evidence to support this hypothesis. In a study designed to overcome the methodological problems of earlier investigations, folic acid absorption was similar in a group of nonpregnant women with a history of an NTD pregnancy and in control women with a normal pregnancy history. These findings suggest that the absorption of folic acid routinely consumed in supplements and fortified food products is not impaired in women with a history of an NTD pregnancy. However, autoantibodies against folate receptors have been reported in women who have had a pregnancy complicated by an NTD. These autoantibodies bind to the folate receptors and can block the cellular uptake of folate.

A woman's risk of having an NTD baby has been shown to be closely related to her early pregnancy levels of plasma folate and RCF, the relationship being stronger for RCF (**Table 6**). There is a strong dose–response effect. Those with RCF levels less than $150\,\mu g\,l^{-1}$ have more than eight times the risk of those with levels of more than $400\,\mu g\,l^{-1}$. Although the most marked absolute reductions in risk occur by elevating the lower RCF levels, risk continues to decrease as RCF levels increase well beyond what would be considered normal levels, with little further protection apparently being gained at levels higher than $400\,\mu g\,l^{-1}$. Most of the NTDs were born to women whose RCFs would have been considered to be in the normal range (i.e., $>150\,\mu g\,l^{-1}$). Thus, views on what constitutes desirable levels of RCF need to be reconsidered.

The lack of evidence of a simple dietary deficiency or of malabsorption and the marked dose–response relationship between maternal RCF level and risk of NTD point to a metabolic explanation for the aetiology of these conditions. Since it is estimated that folic acid can prevent up to 71% of NTDs, defects in folate-related enzymes or processes have been candidates for study. There are 16 folate-dependent enzymes in the internal metabolism of mammalian cells. The finding in the Dublin study of significantly higher plasma homocysteine levels in case mothers than in controls suggested that one or more enzymes involved in homocysteine metabolism may be abnormal. The main folate-related enzymes involved in homocysteine metabolism are illustrated in **Figure 1**. Homocysteine levels in the amniotic fluid of women carrying a fetus with an NTD have been reported as being higher compared with those of normal pregnancies. Evidence of deranged homocysteine metabolism also comes from metabolic studies conducted in Holland. In a study in which women who had given birth to an NTD baby were given a methionine-loading test, methionine intolerance and very high peak levels of homocysteine were found in a subgroup of the NTD women. Cystathionine synthase levels in skin fibroblasts taken from the methionine-intolerant women were normal. Plasma folate and vitamin B_{12} were found to be independent risk factors for NTDs in the Dublin study. Although the results of this study pointed to an abnormality in the methionine synthase enzyme, there is no strong evidence linking genetic variants of the enzyme to NTDs. However, it is possible that vitamin B_{12} status may influence NTD risk in ways other than directly affecting the activity of this enzyme.

Table 6 Distribution of cases and controls and risk of NTDs by red cell folate level

Red cell folate ($\mu g\,l^{-1}$)	*No. of cases (%)*	*No. of controls (%)*	*Risk of NTD per 1000 births*	*95% confidence interval*
0–149	11 (13.1)	10 (3.8)	6.6	3.3–11.7
150–199	13 (15.5)	24 (9.0)	3.2	1.7–5.5
200–299	29 (34.5)	75 (28.2)	2.3	1.6–3.3
300–399	29 (23.8)	77 (29.0)	1.6	1.0–2.4
≥400	11 (13.1)	80 (30.0)	0.8	0.4–1.5
Total	84 (100.0)	266 (100.0)	1.9	1.5–2.3

From Daly LE, Kirke PN, Molloy A, Weir DG and Scott JM (1995) Folate levels and neural tube defects—Implications for prevention. *Journal of the American Medical Association* **274**: 1968–1702. Copyright © 1995, American Medical Association.

Genetic Risk Factors

The main focus of research on NTDs during the past decade has been the investigation of genetic risk factors with particular emphasis on the genes encoding the enzymes in the folate/homocysteine metabolic pathways. A common variant (C677T) in the gene for one of the folate-related enzymes, 5,10-methylenetetrahydrofolate reductase (MTHFR) (**Figure 1**), was identified in 1993 and has been shown to be associated with reduced enzyme function and lower blood folate and higher homocysteine levels. The most frequently studied association between a genetic polymorphism and a congenital malformation has been the relationship between NTD risk and this variant. In initial reports from Holland and Ireland published in 1995, homozygosity for the C677T allele was associated with an increased risk of having spina bifida or having an affected child. From the numerous studies on the link between this polymorphism and NTD that have been conducted in

many countries, it is clear that homozygosity for the variant in the child or mother is a risk in some populations but not others. A review showed that homozygosity for the variant in the child or mother doubles the risk of NTD. A large study of baby–mother pairs showed that the embryo's MTHFR genotype was more important than that of the mother in conferring risk. Studies have also shown evidence of a strong gene–nutrient interaction in that low maternal blood folate levels in early pregnancy and no periconceptional folate supplementation increase the risk associated with the variant allele. Although how the polymorphism causes NTDs is unclear, it may do so through its association with lower folate or higher homocysteine levels or by some other metabolic mechanism. It is estimated that homozygosity for the MTHFR C677T variant is likely to account for not more than approximately 13% of NTDs. Because it is considered that approximately 71% of NTDs can be prevented by folic acid, other mechanisms, possibly including variants in genes coding for other folate-dependent enzymes, problems with folate absorption, or even dietary deficiency of folate, may play a role.

In one study it was shown that the elevated homocysteine levels associated with homozygosity for the MTHFR C677T variant were seen only in those with low riboflavin status; the homocysteine levels did not differ by genotype in those with medium or high riboflavin status. The fact that the activity of the MTHFR C677T variant is influenced by the prevailing riboflavin status may help to explain why the variant is a risk factor for NTDs in some countries but not others. These findings, if confirmed, may be important in view of research reports that substantial proportions of populations have suboptimal riboflavin status.

Genes encoding other enzymes in the folate/homocysteine pathways have been studied. Polymorphisms in some of the genes that may be expected to be important because of their position in the pathways have been shown not to be important (e.g., methionine synthase and cystathionine beta synthase). Other polymorphisms in these pathways have been reported to increase the risk of NTD (i.e., NTD risk was estimated to be significantly higher in either cases or mothers compared to controls)—for example, the MTHFR 1298A → C variant, the reduced folate carrier 80A → G variant, the methionine synthase 919D → G variant, the methionine synthase reductase 66A → G variant, and the R653 Q variant in the trifunctional enzyme methylenetetrahydrofolate dehydrogenase/methenyltetrahydrofolate cyclohydrolase/formyltetrahydrofolate synthetase. For each of these five polymorphisms, however, the increased NTD risk is based on just one study, and other studies have reported negative results. Interactions between some genes in these pathways have been reported as increasing NTD risk—for example, the MTHFR 677C → T and MTHFR 1298A → C variants (two studies); the MTHFR 1298A → C and reduced folate carrier 80A → G variants (one study); the MTHFR 1298A → C, the MTHFR 677C → T, and the reduced folate carrier 80A → G variants (one study); and the methionine synthase 2756A → G and methionine synthase reductase 66A → G variants (one study). Again, other studies have not confirmed these findings. There have also been reports of increased NTD risk associated with polymorphisms in the genes encoding the thymidylate synthase and glutamate carboxypeptidase enzymes. These results need to be replicated in larger studies and in different populations to provide a clearer picture of whether these polymorphisms truly increase NTD risk.

The rationale for studying variants in genes involved directly or indirectly in folate metabolism is clear. However, the closure of the neural tube is a complex process involving the orchestration of many genes. It seems possible that polymorphisms in genes that are far removed from folate metabolism may also cause NTDs that are responsive to folic acid. The MTHFR gene variant is the first specific genetic risk factor to be linked to NTDs. This is the strongest evidence for the involvement of a metabolic derangement in the aetiology of these conditions. This breakthrough gives added impetus to the search for other genetically determined risk factors. Although there is no strong evidence implicating other genetic polymorphisms in the aetiology of NTDs, it is likely that such evidence will soon emerge.

Prevention

In the context of the prevention of NTDs, it is necessary to distinguish between primary and secondary prevention. Primary prevention concerns measures that prevent the development of NTD in the embryo. Secondary prevention refers to screening and termination of affected pregnancies. Primary prevention became a reality following the demonstration of the efficacy of folic acid in preventing NTDs and represented a major public health breakthrough.

A substantial body of research shows that taking extra folate/folic acid before conception and during the early months of pregnancy prevents approximately 50–75% of NTDs and is effective in preventing both occurrent and recurrent NTDs. For all women who may become pregnant, it is recommended that they take an extra 0.4 mg of folic acid per day for the primary prevention of NTDs. This is

in addition to the usual dietary folate intake, which is estimated to be, on average, approximately 0.2 mg per day in the United Kingdom. The most effective ways of using this knowledge to reduce the number of NTD births are discussed next.

Ways of Increasing Folate/Folic Acid Intake

As already noted, a woman can increase her folate intake in three ways: eating more folate-rich foods, eating foods fortified with folic acid, and taking folic acid as a medicinal or food supplement. Although all three methods are known to increase folate status, taking folic acid either as supplements or in fortified foods has been shown to be much more effective in achieving this goal than eating folate-rich foods. It is very difficult to achieve a total daily intake of 0.6 mg folate/folic acid from folate-rich (unfortified) foods alone. Furthermore, folate in unfortified food is not as bioavailable as folic acid in fortified food or in supplements. The only practical way of obtaining an extra 0.4 mg folic acid daily, as recommended, is by consuming fortified foods or folic acid supplements, and this should be made clear by health professionals. However, an improved general diet, especially consuming more vegetables and fruit, should be advocated preconceptionally and during pregnancy because it results in an increased intake of other vitamins and nutrients that are important for normal fetal development.

Supplementation

Since 1992, women of reproductive age have been advised to take an extra 0.4 mg of folic acid daily before pregnancy and during the first 12 weeks of pregnancy. Compliance with the recommendation has been examined in numerous studies conducted throughout the world. Knowledge of the appropriate use of folic acid for NTD prevention in women of childbearing age and in health workers increased markedly throughout the 1990s. Because folic acid can only work if it is taken before closure of the neural tube, the best indicator of periconceptional supplementation is the proportion of pregnant women who take a folic acid supplement before the pregnancy begins, and this proportion increased during the 1990s. In seven studies published from 1999 to 2003 and based on representative study samples in North America and Europe, the proportion of women reported as taking folic acid before pregnancy ranged from 33 to 49%, with a median of 36%. Supplementation is less common in unplanned pregnancies; in young, socially or educationally disadvantaged, and single mothers; and in those with no knowledge about the protective effect of folic acid. The most important predictor of nonsupplementation is unplanned pregnancy. Because unplanned pregnancy is very common (e.g., approximately half of all pregnancies are reported as being unplanned in the United States and Ireland), this factor constitutes the greatest logistical obstacle to planning optimal protection against NTD by periconceptional supplementation. The low supplementation rates reflect the relative lack of effectiveness of promotional campaigns as currently formulated. Public health programs promoting folic acid must be sustained and must pay particular attention to those at greatest risk of not supplementing.

Fortification

Supplementation is probably the most efficient method for ensuring individual protection against an NTD pregnancy but not for a general public health strategy. The disappointing results of supplementation programs led experts to consider fortification of foodstuffs as another public health strategy. There are two approaches to food fortification—voluntary and mandatory. In the former, it is left to individual manufacturers to add folic acid to specific products, whereas in mandatory fortification the relevant authority, with government approval and legislation, requires that a specified dietary staple or staples be fortified to a specified agreed level. The objective of a food fortification policy to prevent NTDs is to increase folate intakes for the target childbearing population as near as possible to the recommended intakes while maintaining safe levels of intake for the entire population. The United States first introduced mandatory food fortification. The Food and Drug Administration (FDA) authorized the addition of folic acid to enriched grain products in 1996 and made compliance mandatory by January 1998. The FDA decided on a level of fortification of 140 μg of folic acid per 100 g flour, and this was estimated to increase average daily intakes of folic acid by 100 μg in women of reproductive age. Studies of fortified foods in the United States have found considerably higher folate levels for many products than those required by the regulations, and the actual average daily increase is estimated to be 150–200 μg. The Canadian government introduced a similar fortification plan in 1998. Studies in both countries have shown that the markers of body folate status (serum and red cell folate and serum homocysteine) have increased dramatically in the population postfortification.

The effect on NTD rates has been striking, especially in Canada. Comparing NTD rates before and after fortification showed decreases of 55% in Nova Scotia and 49% in Ontario postfortification. A study in the United States reported a decrease of 19% in the NTD rate after fortification, and the smaller decrease was considered to be mainly due to the fact that pregnancy terminations for NTDs were not included in the US data and were included in the Canadian studies. These studies provide strong evidence of the effectiveness of mandatory food fortification of folic acid in preventing NTDs and point the way forward for other countries interested in solving this problem.

Fortification of flour was introduced in Chile in 2000 at a level of 220 μg folic acid per 100 g flour. Approximately 38 countries currently either fortify flour (including the United States, Canada, Chile, Argentina, and Israel) or have agreed to do so. In the United Kingdom, the government nutritional advisory committee (the Committee on Medical Aspects of Food and Nutrition Policy) recommended fortification at the level of 240 μg of folic acid per 100 grams of flour, but the Board of the Food Standards Agency decided to defer its implementation. No European Union country has decided to fortify flour to date.

Dose

The appropriate dose of folic acid in relation to mandatory food fortification and supplementation programs continues to be debated. The data from the Dublin study, which showed a marked relationship between early pregnancy maternal RCF levels and NTD risk, were used to examine the effectiveness of a food fortification intervention to increase maternal folate levels to prevent NTDs. The analysis showed that if, as a result of food fortification, all women in a population doubled their RCF level, the prevalence of folate-responsive NTDs would be reduced by 66%. If this increase were 150%, which could be achievable with sufficient fortification, the level of protection would be 73% of folate-responsive NTDs (equivalent to 53% of all NTDs). The finding that a woman's risk of having an NTD pregnancy is related to her early pregnancy levels of RCF in a continuous dose–response relationship suggested that folic acid intake is also related to risk in a continuous dose–response-type relationship, and this was demonstrated in a randomised trial that studied the effect of three different doses of folic acid on RCF levels and reduction in NTD risk. The results of this study in women of reproductive age showed that an extra 0.1, 0.2, or 0.4 mg daily during a 6-month period would be expected to reduce NTD rates by 22, 41, and 47%, respectively. According to another dose–response model that examined the effect of increases in a wide range of daily folic acid intakes on NTD prevention, a 5.0 mg daily dose of folic acid was estimated to decrease NTD risk by 85% in women with a presupplementation serum folate level of 5 ng per liter. These authors argue that the current recommended daily dose of 0.4 mg for folic acid supplements is too low and should be increased to 5.0 mg. They also argue, somewhat controversially, that no known or suspected adverse effects of the 5.0 mg dose have been recorded.

The 4.0 mg daily supplement recommended internationally by National Departments of Health for the prevention of NTD recurrence is based mainly on the unequivocal evidence of the efficacy of this dose in the UK Medical Research Council Trial. In another large nonrandomized intervention study, a daily dose of 0.36 mg folic acid seemed to offer similar protection because the recurrence rate in the treated groups was similar to that in the Medical Research Council trial. So while the recommendation of 4.0 mg to prevent recurrence is quite correctly based on the best scientific evidence, it is likely that a much smaller dose would be equally effective.

Duration of Supplementation

The minimum duration of supplementation necessary for prevention is not known. Although most national health authorities recommend that women take extra folic acid for at least 4 weeks before conception and until week 12 of pregnancy, supplementation for a shorter duration before closure of the neural tube may also be effective. Until more data are available, the official guidelines should be followed. Given the estimate that approximately half of all pregnancies are unplanned, however, it is important that a woman who has not been taking extra folic acid and who suspects that she may be pregnant immediately starts taking a folic acid supplement. This point requires greater emphasis.

Safety

The main concern about taking folic acid at levels greater than 0.4 mg per day is the possibility that the diagnosis of pernicious anemia, which is caused by vitamin B_{12} malabsorption and is more common in the elderly, would be missed since folic acid at high levels prevents the development of the anemia and thus its diagnosis. In this situation, nerve damage

progresses and becomes irreversible. To ensure that 95% of all women get 0.4 mg of folic acid per day through fortification of staple foods would mean that, depending on the diet and differences in eating habits, more than half of the population would get approximately 0.7 mg and 5% would get more than 1.0 mg per day. A compromise is to select a lower target figure for universal fortification that would aim to prevent most folate-responsive NTDs and not put the elderly at risk. This is what the FDA has done. At the 140 μg per 100 grams flour fortification level in the United States, it is estimated that 15–25% of children aged 1–8 years and 0.5% of men and women >70 years would have daily folic acid intakes higher than the tolerable upper intake level. When account is taken of the fact that actual fortification levels in the United States have been estimated to be as much as twice the planned level, the proportions of children and elderly with intakes higher than the upper level may be greater than projected. At the proposed fortification level of 240 μg per 100 grams flour in the United Kingdom, it is estimated that 10% of males would have folic acid intakes >1.0 mg per day. Folic acid in the dose range 0.5–1.0 mg may be absorbed in its original unmetabolized form and may be found in this form in the bloodstream. Circulating folic acid would then be taken up by body cells by a vitamin B_{12}-independent mechanism (**Figure 1**) and would have the potential to switch on the megaloblastic bone marrow in a vitamin B_{12}-deficient person, thereby preventing the development of the anemia and masking the B_{12} deficiency. Further research is needed, therefore, to determine the amount of folic acid in fortified food that is safe from this potential hazard. The incidence of B_{12} deficiency in the elderly has been reported to be as high as 15%, but estimates vary considerably. Some reassurance is provided by a U.S. study that suggests that fortification with folic acid has not been associated with an increase in masking of vitamin B_{12} deficiency.

Adding vitamin B_{12} as well as folic acid to fortified food has been suggested as a solution, but this is problematic. The vast majority of cases of vitamin B_{12} deficiency are due to the autoimmune disease pernicious anemia. In this condition, the absence of intrinsic factor prevents the absorption of physiological amounts of vitamin B_{12}. Thus, including vitamin B_{12} at levels of the dietary reference value (DRV) or less is unlikely to benefit such people because it is not absorbed in sufficient amounts. It has been suggested that if a large enough dose of vitamin B_{12} is added to the diet, then a sufficient amount will be absorbed by passive diffusion to prevent vitamin B_{12} deficiency. However, the amounts required to do this are between 200 and 400 times the DRV for vitamin B_{12} and most experts would be concerned about adding such a vast excess of an albeit apparently safe nutrient to the food chain.

Vitamin B_{12} deficiency is rare in children, but concern has been raised about possible unknown negative health effects of long-term exposure of children to levels of folic acid that are several times the DRV.

It is established that anticonvulsant drugs impair folate status and there is concern that folic acid supplements may reduce the efficacy of these drugs. However, folic acid supplementation of 4 mg daily is recommended for women taking anticonvulsant medication, and there is no evidence of negative effects of supplementation on the control of epilepsy.

A number of studies have suggested the possibility that periconceptional use of vitamin supplements containing folic acid may be associated with an increase in multiple births. A systematic review of the three randomized trials of periconceptional supplementation with folic acid or multivitamins or both (see **Table 1**), updated with new information from the Medical Research Council trial in women who took folic acid, found a consistent increase in the twinning rate. The pooled relative risk was 1.40 (95% confidence interval (CI), 0.93–2.11), but the increase did not reach statistical significance. Increased rates of multiple births were reported in mothers who took multivitamin supplements in two other studies. In a large prospective study of young women in China, there was no increase in the rate of multiple births in women who had taken periconceptional folic acid supplements (0.62%) compared to nonsupplementers (0.67%) (rate ratio, 0.92; 95% CI, 0.83–1.01). The other studies raised the question of whether folic acid or some other component of the multivitamin supplements was responsible for the increased multiple birth rate, and the Chinese study data suggest that folic acid is not associated with this effect. The association between periconceptional use of folic acid and multiple pregnancy has been shown to be confounded by use of in vitro fertilization (IVF). Pregnancies following IVF are strongly associated with both multiple pregnancies and periconceptional use of folic acid. A number of studies have reported no association between use of folic acid supplements and multiple pregnancy when adjustment is made for use of IVF. Research in the United States suggests that fortification has not resulted in an increase in multiple pregnancy rates.

There are reports of increased miscarriage rates in women who took vitamin supplements containing folic acid before and during early pregnancy. A review of the randomized trials of periconceptional supplementation of folic acid or multivitamins found a statistically nonsignificant 12% increase in the miscarriage rate among those who took folic acid alone or as part of a multivitamin supplement. It has been suggested that folic acid may extend the viability of fetuses that would otherwise miscarry at earlier stages of pregnancy and be unrecognized as such. In the largest study that has addressed this question, the rates of miscarriage were similar in Chinese women who had (1981 of 21 935 or 9.0%) and had not (174 of 1871 or 9.3%) taken supplements containing only 400 μg folic acid before and during early pregnancy. This study is the most scientifically rigorous examination of the hypothesis that periconceptional folic acid may increase the miscarriage rate, and the findings indicate that this is not so.

Other Nutrients

As noted previously, nutritional factors may be involved in the aetiology of some of the nonfolate-related NTDs. Although there is increasing evidence for the involvement of vitamin B_{12} and less evidence for vitamin C, riboflavin, and zinc, the available evidence is not strong enough to support dietary supplementation with these or other nutrients.

Because large doses of vitamin A are known to be teratogenic, women at risk of pregnancy and those in the early months of pregnancy should avoid liver products which can contain high quantities of vitamin A. In order to keep daily vitamin A intake below 10 000 IU, women in these groups should not take a multivitamin tablet that contains a dose of more than 5000 IU.

Recommendations

The Department of Health in the United Kingdom and the Department of Health and Human Services in the United States were the first to issue recommendations on the use of folic acid to prevent NTDs. The recommendations relate to the prevention of occurrent (first-time) and recurrent NTDs, and other national health authorities have adopted similar recommendations. The main points in these recommendations are given next.

Prevention of NTD Recurrence

To prevent NTD recurrence in the offspring of women or men who have spina bifida or encephalocoele, or a history of a previous child with NTD,

Such women and men should be counseled about the increased risk in subsequent pregnancies and about the protective effect of supplementation with folic acid.

Women with a previously affected pregnancy should, unless contraindicated, be advised to take 4.0 mg of folic acid daily from at least 4 weeks before conception until the end of the third month of pregnancy. In countries in which a 5.0 mg rather than a 4.0 mg preparation is available, the former can be used but the lower 4.0 mg dose should be used as soon as this preparation becomes available.

The 4.0 mg dose should be taken only under the supervision of a doctor because giving high doses of folic acid can complicate the diagnosis of vitamin B_{12} deficiency and epileptic women on anticonvulsant therapy require individual counseling before starting folic acid.

Prevention of NTD Occurrence

For the prevention of occurrence of NTDs, the US Public Health Service recommends that all women capable of becoming pregnant consume 0.4 mg of folic acid per day and that total folate consumption should not be more than 1.0 mg per day to avoid the possible risks of high intakes. The UK Expert Advisory Group recommends that women should take an extra 0.4 mg of folic acid daily from when they begin trying to conceive until week 12 of pregnancy. If a woman who has not been taking this additional amount of folic acid suspects that she may have just started a pregnancy, she should begin taking extra folic acid immediately and continue until week 12 of pregnancy. The US Public Health Service and the UK Expert Advisory Group have outlined three possible ways of achieving an extra intake of folate/folic acid: eating more folate-rich foods, eating foods fortified with folic acid, and taking folic acid as a medicinal or food supplement. It is recommended that women should use whatever source or combination of sources they prefer to ensure that they obtain the necessary extra folic acid. The effectiveness of these approaches in achieving the recommended increased population intake of folate/folic acid was considered under Prevention. As already noted, it is very difficult to achieve a total daily intake of 0.6 mg folate/folic acid from only foods naturally rich in folate. The only practical ways of obtaining the recommended extra 0.4 mg folic acid daily is by consuming folic acid supplements or fortified foods, and this should be emphasized when advising women. In countries in which there is mandatory food fortification, it is important that women

are advised to continue taking supplements because fortification is designed to deliver considerably less than the recommended extra 0.4 mg daily intake of folic acid. When taking supplements, the folic acid dose should be obtained from pills containing only folic acid rather than from multivitamin preparations because of the risk of taking harmful levels of vitamins A and D in early pregnancy.

See also: **Bioavailability**. **Cobalamins**. **Folic Acid**. **Food Fortification**: Developed Countries; Developing Countries. **Fruits and Vegetables**. **Homocysteine**. **Nutrient–Gene Interactions**: Health Implications. **Obesity**: Definition, Etiology and Assessment; Complications. **Older People**: Nutrition-Related Problems. **Socio-economic Status**. **Supplementation**: Role of Micronutrient Supplementation.

Further Reading

Berry RJ, Li Z, Erickson JD *et al.* for the China–U.S. Collaborative Project for Neural Tube Defect Prevention (1999) Prevention of neural tube defects with folic acid in China. *New England Journal of Medicine* **341**: 1485–1490.

Botto LD and Yang Q (2000) 5,10-Methylenetetrahydrofolate reductase gene variants and congenital anomalies: A HuGe review. *American Journal of Epidemiology* **151**: 862–877.

Botto LD, Moore CA, Khoury MJ, and Erickson JD (1999) Neural-tube defects. *New England Journal of Medicine* **341**: 1509–1519.

Centers for Disease Control and Prevention (1991) Use of folic acid for prevention of spina bifida and other neural tube defects—1983–1991. *Morbidity and Mortality Weekly Report* **40**: 513–516.

Centers for Disease Control and Prevention (1993) Recommendations for the use of folic acid to reduce the number of cases of spina bifida and other neural tube defects. *Morbidity and Mortality Weekly Report* **41**(RR-14): 1–7.

Committee on Medical Aspects of Food and Nutrition Policy (2000) *Folic Acid and the Prevention of Disease*. London: Department of Health.

Daly LE, Kirke PN, Molloy A, Weir DG, and Scott JM (1995) Folate levels and neural tube defects—Implications and prevention. *Journal of the American Medical Association* **274**: 1698–1702.

Daly S, Mills JL, Molloy AM *et al.* (1997) Minimum effective dose of folic acid for food fortification to prevent neural tube defects. *Lancet* **350**: 1666–1669.

Elwood JM, Little J, and Elwood JH (1992) *Epidemiology and Control of Neural Tube Defects*. Oxford: Oxford University Press.

EUROCAT Working Group (2003) *EUROCAT Special Report: Prevention of Neural Tube Defects by Periconceptional Folic Acid Supplementation in Europe*. Belfast: University of Ulster. Available at www.eurocat.ulster.ac.uk/pubdata/folic%20acid.html.

Expert Advisory Group (1992) *Folic Acid and the Prevention of Neural Tube Defects*. London: Department of Health.

Scott JM, Kirke PN, and Weir DG (1995) Folate and neural tube defects. In: Bailey L (ed.) *Folate in Health and Disease*, pp. 329–360. New York: Marcel Dekker.

Shields DC, Kirke PN, Mills JL *et al.* (1999) The "thermolabile" variant of ethylenetetrahydrofolate reductase and neural tube defects: An evaluation of genetic risk and the relative importance of the genotypes of the embryo and the mother. *American Journal of Human Genetics* **64**: 1045–1055.

Wald NJ, Hackshaw AK, Stone R, and Sourial NA (1996) Blood folic acid and vitamin B_{12} in relation to neural tube defects. *British Journal of Obstetrics and Gynaecology* **103**: 319–324.

Pre-eclampsia and Diet

E Abalos, Centro Rosarino de Estudios Perinatales, Rosario, Argentina
J Villar, World Health Organization, Geneva, Switzerland

Introduction

Hypertensive disorders during pregnancy are one of the main causes of maternal death worldwide, most of these deaths being attributed to eclampsia. Eclampsia is the occurrence of fits in a pre-eclamptic woman that cannot be attributed to other causes (such as epilepsy, etc.). Hypertensive disorders occur in 6–8% of all pregnancies contributing significantly to stillbirths and neonatal morbidity and mortality. Babies are also at increased risk of intrauterine growth restriction, low birth weight, and preterm delivery. Pregnant women with hypertension, either newly diagnosed or pre-existing, are prone to the development of potentially lethal complications, notably abruptio placentae, disseminated intravascular coagulation, cerebral hemorrhage, pulmonary edema, hepatic failure, and acute renal failure. The etiology of hypertensive disorders related to pregnancy, particularly pre-eclampsia, remains unknown.

The most important consideration in the classification of the disease is differentiating hypertensive disorders that antedate pregnancy from those that are pregnancy specific, of which the more ominous are pre-eclampsia and eclampsia. Pre-eclampsia is a pregnancy-specific syndrome of reduced organ perfusion secondary to vasospasm and activation of the coagulation cascade. Although our understanding of this syndrome has increased, the criteria used to identify the disorder remain a subject of confusion and controversy. In chronic hypertension, elevated blood pressure is the cardinal pathophysiologic feature, whereas in pre-eclampsia, increased blood pressure is important primarily as a sign of the underlying disorder. As might be expected, the impact of the two conditions on mother and fetus is different, as is their management.

Classification

There is controversy about the definition of hypertensive disorders during pregnancy, and several classifications have been suggested. Recently, the USA National High Blood Pressure Education Program Working Group on High Blood Pressure in Pregnancy updated the 1990 report, and classified the hypertensive disorders during pregnancy as: (a) chronic hypertension defined as hypertension observable before pregnancy, or diagnosed before the 20th week of gestation; (b) pre-eclampsia, which is a pregnancy-specific syndrome occurring usually after 20 weeks' gestation, determined by hypertension with proteinuria; (c) pre-eclampsia superimposed on chronic hypertension; and (d) pregnancy-induced hypertension or gestational hypertension, which is transient hypertension detected for the first time after mid-pregnancy if pre-eclampsia is not present at the time of delivery and blood pressure returns to normal by 12 weeks post-partum (a retrospective diagnosis). The system suggested by the International Society for the Study of Hypertension in Pregnancy (ISSHP) defines hypertension as a diastolic blood pressure of 90 mmHg or above on two consecutive occasions at least 4 hours apart, or a single diastolic blood pressure of 110 mmHg or more. The definition of pre-eclampsia has the same criteria for high blood pressure, but with the addition of significant proteinuria, usually at least 300 mg per 24 h or 1+ on dipsticks.

Pathophysiology of Pre-eclampsia

Pre-eclampsia is a syndrome with both fetal and maternal manifestations. The maternal disease is characterized by vasospasm, activation of the coagulation system, and perturbations in many humoral and autacoid systems related to volume and blood pressure control. The pathologic changes in this disorder are primarily ischemic in nature and affect placenta, kidney, liver, and brain. Of importance, and distinguishing pre-eclampsia from chronic or gestational hypertension, is that pre-eclampsia is more than hypertension; it is a systemic syndrome, and several of its 'nonhypertensive' complications can be life-threatening even when blood pressure elevations are quite mild.

The cause of pre-eclampsia is not known. Many consider the placenta as the pathogenic focus for all manifestations of pre-eclampsia because the delivery of both the baby and the placenta is the only definitive cure of this disease. There is no disease without the placenta. Thus, research has focused on the changes in the maternal blood vessels that supply blood to the placenta. Failure of the spiral arteries to remodel is postulated as the morphologic basis for decreased placental perfusion in pre-eclampsia, which may ultimately lead to early placental hypoxia. Oxidative stress and inflammatory-like responses may also be important in the pathophysiology of pre-eclampsia.

Research on how alterations in the immune response at the maternal interface might lead to pre-eclampsia addresses the link between placenta and maternal disease. A nonclassical human leucocyte antigen (HLA), HLA G, is expressed in normal placental tissue and may play a role in modulating the maternal immune response to the immunologically foreign placenta. Placental tissue from pre-eclamptic pregnancies may express less or different HLA G proteins, resulting in a breakdown of maternal tolerance to the placenta. Additional evidence for alterations in immunity in pathogenesis includes the higher frequency of nulliparous gestations with subsequent normal pregnancies, a decreased prevalence after heterologous blood transfusions, a long period of cohabitation before successful conception, and observed pathologic changes in the placental vasculature in pre-eclampsia that resemble allograft rejection. Finally, there are increased levels of inflammatory cytokines in the placenta and maternal circulation, as well as evidence of increased 'natural killer' cells and neutrophil activation in pre-eclampsia.

The mechanisms underlying vasoconstriction and altered vascular reactivity in pre-eclampsia remain obscure. Research has focused on changes in the ratio of vasodilative and vasoconstrictive prostanoids, since prostacyclin may be suppressed and thromboxane may be raised. More recently, investigators have postulated that the vasoconstrictive potential of pressor substances (e.g., angiotensin II and endothelin) is magnified in pre-eclampsia as a consequence of a decreased activity of nitric oxide (NO) synthesis and decreased production of NO-dependent or NO–independent endothelium relaxing factor (EDRF). Also under investigation is the role of endothelial cells (the site of prostanoid, endothelin, and EDRF production), which in pre-eclampsia may be dysfunctional, due perhaps to inflammatory cytokines (e.g., tumor necrosis factor alpha) and increased oxidative stress. Other systems postulated to play a role in pre-eclamptic hypertension are the sympathetic nervous system, calciotrophic hormones, insulin, and magnesium metabolism.

Finally, some nutritional deficiencies have been postulated as playing a role in the pathogenesis of pre-eclampsia. Their possible role in the hypertensive disorders of pregnancy are discussed below.

The Possible Role of Nutrition in the Pathophysiology of Pre-eclampsia

Epidemiological observations have long suggested a role for nutritional deficiencies (i.e., calcium, proteins, vitamins, etc.) in pre-eclampsia. However, intervention evaluations have failed to confirm such promising observations. We will describe here the evidence from randomized controlled trials that supports the relationship between different nutrients and pre-eclampsia.

Calcium

There is considerable evidence linking calcium intake and hypertension during pregnancy from observational and experimental studies. However, there is still no satisfactory explanation for the mechanisms involved in the calcium-mediated effect on blood pressure reduction. It has been postulated that parathyroid hormone could be involved in this relationship. Demonstrated alterations in extracellular calcium homeostasis in pre-eclampsia include hypocalciuria and decreased serum levels of calcitriol. Increased parathyroid hormone (PTH) and decreased plasma ionized calcium concentration have not been consistently observed. Also, consistent abnormalities of intracellular calcium metabolism have been described in pre-eclamptic women, such as increased intracellular free calcium concentration in platelets and lymphocytes. Increases in intracellular free calcium concentration in circulating cells are hypothesized to result from fluctuation in hormones or vasoactive substances that cause similar alteration in vascular smooth muscle. Pregnancy is a state of high calcium requirements as a result of fetal demands while maternal adaptive mechanisms are partially inhibited. These phenomena lead to the hyper-parathyroid state of pregnancy. An increase of parathyroid hormone serum levels would involve an increase of free intracellular calcium. Then, the concentration of intracellular free calcium in vascular smooth muscle cells determines the degree of tension, and is the trigger for muscular contraction. So the vasoconstrictive effect, with a rise in blood pressure, results from an increase in vascular smooth muscle tension.

Antioxidant Agents

An additional role for nutrition in the genesis of pre-eclampsia could be nutritional factors that strengthen oxidative stress, leading to pre-eclampsia. A nutritional factor could be the deficiency of antioxidant intake, specifically vitamin C and E. Vitamin C is central for the neutralization of both water-soluble and lipid-soluble free radicals; as a water-soluble molecule its ability to neutralize free radicals in the aqueous compartment is clear. Also, ascorbate is not made in humans and must come from diet. Vitamin E, a potent antioxidant, has been suggested to play a role in preventing pre-eclampsia.

Other Nutrients

Nutritional factors other than antioxidants can also contribute to oxidative stress. Hyper-homocysteinemia can occur as a result of dietary deficiencies. Hyper-homocysteinemia as a risk factor for pre-eclampsia is said to be altered, at least in part, by the genesis of oxidative stress. Vitamin B_6 and B_{12} and folic acid are involved at different steps in the metabolic pathway for removing or recycling homocysteine to methionine. Dietary deficiencies of any of these micronutrients can increase circulating homocysteine. Pre-eclampsia is characterized by increased triglycerides that favor the formation of small, dense low-density lipoproteins (LDLs). This lipoprotein variant has increased access to the subendothelial space where it is sequestered from blood-borne antioxidants. The relevant role of triglycerides in the genesis of pre-eclampsia is indicated by the fact that they are increased long before clinically evident disease. Similarly, free fatty acids are increased in pre-eclampsia and this increment can be observed months before the diagnosis. Recent studies indicate that this effect may be secondary to altered copper binding by albumin to which large amounts of free fatty acids are bound. Unbound copper is a potent stimulator of free radical formation. Ordinarily this effect of copper is prevented by protein binding (quantitatively, primarily to albumin). However, with fatty acid binding, albumin binds copper differently. In this configuration, copper bound to albumin maintains its ability to participate in redox reactions. Thus, it appears that increased free fatty acids can also contribute to oxidative stress.

All of these nutritional alterations may be amenable to dietary modification raising the possibility of nutritional prophylaxis.

Nutritional Interventions and Hypertensive Disorders of Pregnancy

Prevention

The ability to prevent hypertensive disorders of pregnancy is limited by lack of knowledge of its underlying etiology. Prevention is focused on identifying women at higher risk of developing pregnancy-induced hypertension or pre-eclampsia during

pregnancy, followed by close clinical and laboratory monitoring to recognize the clinical symptoms of the disease in its early stages. These women and their pregnancies can then be selected for more intensive monitoring or delivery. Although these measures do not prevent the disease, they may be helpful for preventing some adverse maternal and fetal sequelae.

As part of many other nonpharmacological interventions, some dietary interventions have been proposed to prevent the development of pregnancy-induced hypertension and pre-eclampsia.

Nutritional advice in pregnancy The relevant literature was reviewed in order to assess the effects of advising pregnant women to increase their energy and protein intakes on the outcome of pregnancy, and maternal and fetal/infant morbidity and mortality. Nutritional advice was assessed on a Cochrane systematic review and appears to be effective in increasing pregnant women's energy and protein intake, but the implications for fetal, infant, or maternal health cannot be judged from the available evidence. Pre-eclampsia prevention was assessed only in one small trial involving 136 women with no beneficial effects.

Protein/energy supplementation The effect of balanced protein/energy supplements for pregnant women on gestational weight gain and pregnancy outcomes was also evaluated. Pre-eclampsia prevention was assessed in three trials involving 516 women, with no significant beneficial effects. However, these trials had methodological flaws, so the results should be interpreted cautiously. In another pre-specified subgroup, only one trial involving 782 women evaluated pre-eclampsia prevention when isocaloric balanced protein/energy supplements were given to underweight pregnant women, showing no effect.

Energy/protein restriction for obese pregnant women Excessive weight gain during pregnancy has long been recognized as a risk factor for edema and impending pre-eclampsia. Epidemiological studies suggested that high maternal weight was positively associated with the risk of pre-eclampsia.

Energy/protein restriction for high weight-for-height or weight gain during pregnancy was another subgroup assessed in this systematic review. Pre-eclampsia was evaluated in two trials (284 women), which showed no reduction in the risk of occurrence. Similarly, there was no influence on pregnancy-induced hypertension (3 trials, 384 women). The limited evidence available suggests that protein/energy restriction of pregnant women who are overweight or exhibit high weight gain is unlikely to be beneficial and may be harmful to the developing fetus. Although weight reduction may be helpful in reducing or preventing high blood pressure in nonpregnant women, there is no effect on preventing pre-eclampsia, even in obese women. Clinicians frequently ask pregnant women to restrict their food intake in an attempt to prevent pre-eclampsia, despite the absence of evidence that such advice is beneficial.

Salt restriction Even in the early phase of pregnancy, marked hemodynamic changes occur including a fall in vascular resistance and blood pressure and a rise in cardiac output. To compensate for the increased intravascular capacity the kidney retains more sodium and water. Apparently, the set point of sodium homeostasis shifts to a higher level at the expense of an expansion of extracellular volume. In nonpregnant individuals, a strong positive association of sodium intake with blood pressure has been established, but the relationship between sodium intake and blood pressure in human pregnancy remains obscure to date. For decades a low-salt diet has often been recommended as treatment for edema, in the hope that restricting salt intake would treat, and also prevent, pre-eclampsia. Recently, this practice has been questioned, and even a high sodium intake has been proposed for pre-eclampsia treatment and prevention.

The concerns about the effect of a low-sodium diet during pregnancy on maternal nutritional status led researchers to investigate if such changes could alter other nutrient intake. It was shown that the reduction in sodium intake also caused a significant reduction in the intake of energy, protein, carbohydrates, fat, calcium, zinc, magnesium, iron, and cholesterol. Even though the majority of clinicians no longer advise women to alter their salt intake during pregnancy, this is still current practice in many countries worldwide.

A recently published Cochrane systematic review evaluates the effect of the advice about low dietary salt intake during pregnancy. The review includes two trials with data reported for 603 women. Both trials compared nutritional advice to restrict dietary salt with advice to continue a normal diet. Women with established pre-eclampsia were not enrolled, so this review provides no information about the effects of advice to restrict salt intake for treatment of pre-eclampsia. No effect was found in preventing pre-eclampsia or pregnancy-induced hypertension (1 trial, 242 women). Women's preferences were not reported, but the authors presumed that a

low-salt diet was not very palatable and was therefore difficult to follow.

Calcium supplementation A role for altered calcium metabolism in the pathogenesis of pre-eclampsia is suggested by epidemiological evidence linking low dietary levels of calcium with increased incidence of the disease. In agreement with these observations, several modifications in calcium metabolism have been observed in pre-eclamptic women and in calcium supplemented mothers.

A Cochrane systematic review of calcium supplementation during pregnancy has been published. Authors prespecified comparison groups taking into account the women's risk of hypertensive disorders of pregnancy (low versus increased), and the women's baseline dietary calcium intake (low: $<900\,mg\,day^{-1}$ versus adequate: $\geq 900\,mg\,day^{-1}$).

High blood pressure with or without proteinuria was evaluated in 9 trials involving 6604 women. Overall, there was less high blood pressure with calcium supplementation (relative risk (RR) 0.81; 95% confidence interval (CI) 0.74–0.89), but there was a variation in the magnitude of the effect across the subgroups. The effect was considerably greater in women at high risk of developing hypertension (3 trials, 297 women: RR 0.35; 95% CI 0.21–0.57) than in those at low risk (6 trials, 6307 women: RR 0.84; 95% CI 0.76–0.92). Taking into account the women's calcium intake, the effect was also greater in those with low baseline dietary calcium (5 trials, 1582 women: RR 0.49; 95% CI 0.38–0.62) than in those with adequate calcium intake (4 trials, 5022 women, RR 0.90; 95% CI 0.81–0.99).

There was a reduction in the risk of pre-eclampsia when evaluated from 10 trials involving 6864 women (RR 0.70; 95% CI 0.58–0.83). When predefined subgroups were considered, there was a significant reduction in women with low baseline dietary calcium intake (6 trials, 1842 women: RR 0.32; 95% CI 0.21–0.49), but not in those with adequate calcium intake (4 trials, 5022 women: RR 0.86; 95% CI 0.71–1.05). Pre-eclampsia was considerably reduced in women at high risk of hypertension (4 trials, 557 women: RR 0.22; 95% CI 0.11–0.43), and less consistently in those at low risk of hypertension (6 trials, 6307 women: RR 0.79; 95% CI 0.65–0.94).

The results from the largest trial conducted by the National Institutes of Health (NIH), which studied low-risk women with adequate baseline calcium diet, and in whom all women in both groups received low-dose calcium supplementation as part of their routine antenatal care, showed no significant effect on hypertension and that pre-eclampsia. Based on this, authorities from developed countries where adequate dietary calcium intake is common, discourage the use of routine calcium supplementation during pregnancy. Evidence from this review support the view that calcium supplementation might benefit women at high risk of gestational hypertension and women with low dietary calcium intake are at risk of developing pre-eclampsia, and current guidelines suggest supplementing calcium intake in these groups.

This recommendation is currently being evaluated in a large (8300 women), double-blind randomized controlled trial by the World Health Organization (WHO), conducted in seven locations around the world where calcium intake is low ($<600\,mg\,day^{-1}$) in which pregnant women received an extra $1.5\,g\,day^{-1}$ of calcium carbonate or a placebo from the 20th week of gestation. Results should be available in 2005.

Iron and folate supplementation Numerous trials involving various populations of pregnant women with normal hemoglobin levels have evaluated the effects of iron and/or folate supplementation on several outcomes, some of them including hypertensive disorders of pregnancy. A Cochrane systematic review of 2 trials involving 87 women with normal hemoglobin levels in which iron and folic acid were compared with no treatment showed no effect on the occurrence of gestational hypertension. Pre-eclampsia was not evaluated. In another Cochrane review of two trials involving 696 pregnant women already receiving iron, where some women were allocated to receive folic acid and others received no treatment/placebo, again there was no effect on the prevention of gestational hypertension.

Although evidence shows that iron and folate supplementation is not effective in preventing hypertensive disorders during pregnancy, they should be prescribed for other established beneficial effects on pregnancy such as prevention of anemia.

Magnesium supplementation Magnesium is one of the essential minerals needed by humans in relatively large amounts. Magnesium works with many enzymes regulating body temperature and synthesizing proteins as well as maintaining electrical potentials in nerves and muscle membranes. Magnesium occurs widely in many foods; dairy products, breads and cereals, vegetables, and meats are all good sources. It is therefore not surprising that frank clinical magnesium deficiency has never been reported to occur in healthy individuals who eat standard diets. However, dietary intake studies during

pregnancy consistently demonstrate that many women, especially those from disadvantaged backgrounds, have intakes of magnesium below recommended levels. Observational studies based on medical records reported that magnesium supplementation during pregnancy was associated with a reduced risk of fetal growth retardation and pre-eclampsia and that magnesium intake was associated with increased birth weight. Stimulated by these encouraging epidemiological studies, randomized clinical trials have been undertaken to evaluate the potential benefits of magnesium supplementation during pregnancy on pregnancy and neonatal outcomes.

A Cochrane systematic review of these randomized controlled trials was carried out in order to assess the effects of magnesium supplementation during normal or high-risk pregnancies on maternal, neonatal, and pediatric outcomes. Results from two trials (474 women) showed no apparent effect of magnesium treatment on prevention of pre-eclampsia. However, these results may have been confounded by the fact that in the largest trial all women (both magnesium supplemented and placebo groups) received a multivitamin and mineral preparation containing low doses of magnesium. Several of the trials also have poor methodological quality, especially related to concealment of allocation, which could give biased results. These authors conclude that dietary magnesium supplementation of pregnant women cannot be recommended for routine clinical practice because of the poor methodological quality of the current evidence.

Fish oil supplementation Studies of non-pregnant subjects suggest that fish oil, rich in long-chain n-3 fatty acids, has a moderate effect on blood pressure in normotensive as well as hypertensive individuals. A meta-analysis of controlled clinical trials of the effect of fish oil on blood pressure has demonstrated a significant reduction in systolic and diastolic blood pressure in untreated hypertensive non-pregnant individuals, but found no significant effect on normotensives. Fish oil has been shown to modify prostaglandin metabolism, and its effect on blood pressure has often been assumed to be due to such interference. Epidemiological studies suggested that marine diets could have a preventive effect on early delivery and hypertensive disorders of pregnancy.

Fish oil supplementation during pregnancy was evaluated in 1995 in a systematic review of 2 trials (5135 women), showing no effect on pregnancy-induced hypertension (2 trials, 5135 women) RR: 0.98, 95% CI 0.91 to 1.04. There was a statistically significant but modest reduction in the rate of pre-eclampsia (RR: 0.81, 95% CI 0.69 to 0.93). However, this reduction is strongly influenced by a single large trial conducted in 1942. Four other trials of fish oil supplementation involving more than 2000 women have been published recently none of which demonstrates any differences in the incidence of hypertension and pre-eclampsia between groups. Based on current evidence, fish oil supplementation is not recommended during pregnancy for the prevention of pre-eclampsia.

Zinc supplementation Zinc is proposed as playing an important role in many biological functions, including protein synthesis and nucleic acid metabolism. There is controversy in the literature in demonstrating the relationship between low serum zinc levels and abnormalities of pregnancy outcomes such as pregnancy-induced hypertension, prolonged labor, post-partum hemorrhage, preterm or post-term pregnancies, small-for-gestational age babies, or poor perinatal outcomes.

The role of routine zinc supplementation during pregnancy on outcomes for both mother and newborn was assessed in a Cochrane systematic review. Routine zinc supplementation in pregnancy had no detectable effect on gestational hypertension (four trials, 1962 women). However, there appears to be inconsistency among trials regarding the effects from other pregnancy outcomes. This may be related to variable population characteristics of women recruited in the various trials, as some included normal pregnant women with no systemic illness, other studies specifically selected women at high risk of low-zinc status, and in one study, participants were selected on the basis of proven low plasma zinc levels. There is at present no evidence of overall benefit from routine as opposed to selective zinc supplementation in pregnancy in pregnancy-induced hypertension or pre-eclampsia.

Vitamin (A, E, and C) supplementation An oxidant/antioxidant imbalance has been suggested among the possible pathogenic factors involved in pre-eclampsia. As vitamin E is one of the most important antioxidants, its levels and their relation with circulating levels of lipid peroxides in pre-eclamptic women has been intensively studied in recent years. As with other antioxidants, several studies found decreased vitamin E levels in serum from women with gestational hypertension and pre-eclampsia compared with controls. However, these findings could not be demonstrated in other studies. Increased ascorbate radical formation and ascorbate depletion were also found in plasma from women

with pre-eclampsia. Recently, a randomized controlled trial involving 283 women at very high risk of developing pre-eclampsia was conducted. Women were randomly assigned to receive vitamin C ($1000\,mg\,day^{-1}$) and E ($400\,IU\,day^{-1}$) or placebo at 16–22 weeks of gestation. The authors found a significant reduction in the risk of developing pre-eclampsia in the vitamin-supplemented group compared to controls (RR: 0.46; 95% CI 0.24–0.91). The authors concluded that supplementation with vitamins C and E may be beneficial for preventing pre-eclampsia in women at increased risk of the disease. However, these findings come from a single trial of 283 women and need to be further assessed in different settings and populations, as well as in low-risk women. The preventative potential of vitamins C and E is currently being evaluated in three large multicentre double-blind randomized trials in North America, in several institutions in the UK and in a new WHO multicentre trial in India, Peru, and Vietnam. Results are expected during 2006.

The role of vitamin A in pregnancy-induced hypertension and pre-eclampsia is another subject of controversy. It was proposed as a chain-breaking antioxidant in the free radical cascade. Some studies found significantly reduced serum vitamin A levels in pre-eclamptic and eclamptic women when compared to levels in healthy women in the third trimester. No trials have been published to date to assess the effect of vitamin A supplementation on pregnancy-induced hypertension or pre-eclampsia. A double-blind cluster randomized trial of low-dose supplementation with vitamin A or beta-carotene carried out in Nepal in 44 646 married women showed a 40% reduction in maternal mortality related to pregnancy in vitamin A supplemented women. However, differences in cause of deaths, including pre-eclampsia and eclampsia, could not be reliably distinguished between supplemented and placebo groups. Use of vitamin A supplements for the prophylaxis and management of pregnancy-induced hypertension and pre-eclampsia needs to be evaluated further before it can be recommended.

Treatment

The objectives of treatment for established pre-eclampsia or pregnancy-induced hypertension are to prevent eclampsia as well as other severe maternal complications. Close maternal evaluation is aimed at observing progression of the condition, both to prevent maternal complications and to determine whether fetal well-being can be assessed. As this disorder is often completely reversible and usually begins to abate with delivery, an imbalance between the mother's condition and the risk for fetus survival without significant neonatal complications *in utero* or in the nursery must be continuously evaluated. Even though the only definitive treatment of pre-eclampsia is delivery, some nonpharmacological approaches were proposed as part of an overall strategy of management of the disease to achieve these goals.

Unfortunately, there is no information from randomized controlled trials related to dietary approaches to the management of the disease in its mild to moderate stage, at which point conservative management is generally decided (**Table 1**).

Pre-existing (Chronic) Hypertension

Mild and uncomplicated chronic hypertension during pregnancy has a better prognosis than pre-eclampsia. However, there is an increased risk of superimposed pre-eclampsia and possible complications if pre-existing renal disease or systemic illness is present. The primary aim of therapy, if necessary, is to prevent cerebrovascular complications and to avoid progression to superimposed pre-eclampsia with its worse prognosis. Nonpharmacological management of this condition during pregnancy remains controversial.

In a published review of management of mild to chronic hypertension during pregnancy, no trials were found that compared nonpharmacological interventions with either pharmacological agents or no intervention in pregnant women. This comprehensive search identified 50 randomized controlled trials, but they involved either normotensive women or women with a history of pre-eclampsia. For the management of established chronic hypertension during pregnancy, no relevant evidence could be located to assess the effects of nonpharmacological interventions, such as limiting activity, diet modifications, or stress reduction.

Weight reduction during pregnancy, even in obese women, in general is not recommended to improve pregnancy outcomes. As weight reduction may be helpful in reducing blood pressure in nonpregnant individuals, for obese hypertensive women planning a pregnancy, weight reduction before conception is advisable. Even though obesity may be a risk factor for superimposed pre-eclampsia, there is no evidence that limiting weight gain during pregnancy reduces its occurrence.

Pregnant women with hypertension have a lower plasma volume than normotensive women, and some studies suggest that the severity of hypertension correlates with the degree of plasma volume

reduction. For this reason, sodium restriction is generally not recommended during pregnancy for the reduction of blood pressure. In addition, an increase in plasma volume reduction is a risk factor for intrauterine growth restriction. If, however, a pregnant woman with chronic hypertension is known to have salt-sensitive hypertension and has been treated successfully with low salt diet before pregnancy, it is reasonable to continue some sodium restriction for blood pressure control during pregnancy, but not for preventing superimposed pre-eclampsia.

High alcohol intake is related to hypertension in nonpregnant subjects, but is not associated with an increased risk for gestational hypertension, pre-eclampsia, or eclampsia. There is no conclusive evidence of adverse effects on pregnancy outcomes, including fetal growth, at levels of consumption below 120 g of alcohol per week. However, there are suggestions that excessive consumption of alcohol can cause or aggravate maternal hypertension.

There is no reliable information from well-designed randomized controlled trials assessing the best dietary approach for the management of pre-existing hypertension during pregnancy. Recommendations come from expert's consensus and authorities' statements. It seems that mild-to-moderate pre-existing (chronic, essential) hypertension without any risk factor should be managed in the same way as in the nonpregnant state. However, additional concerns are effects on fetal well being (mainly intrauterine growth restriction) and worsening of hypertension, particularly as a result of superimposed pre-eclampsia.

Conclusions

In short, based on the available data from systematic reviews (see **Table 1**) we can conclude that there is some evidence that calcium supplementation in populations with low calcium intake and/or at risk of developing pregnancy-induced hypertension could be beneficial. Data on antioxidants (particularly vitamins E and C) are promising, but there is a need for adequately designed randomized controlled trials of sufficient size to confirm these findings before widespread recommendation. Although pregnant women living in developing countries could be exposed to several other nutrient deficiencies, a lack of evidence precludes recommending other nutrient supplementation as part of their routine antenatal care in order

Table 1 Effectiveness of nutritional interventions in hypertension during pregnancy and pre-eclampsia

Intervention	*Hypertension during pregnancy*		*Pre-eclampsia*	
	Practice	*Research*	*Practice*	*Research*
Nutritional advice	No evidence	–	No effect; RR = 0.89 (0.42–1.88)	–
Balanced protein (<25%)/energy	No evidence	–	No effect; RR = 1.20 (0.77–1.89)	–
Isocaloric balanced protein (<25% of total energy)	No evidence	–	No effect; RR = 1.00 (0.57–1.75)	–
Energy/protein restriction for high PI or high weight gain	No effect; RR = 0.97 (0.75–1.26)	–	No effect; RR = 1.13 (0.59–2.18)	–
Salt restriction	No effect; RR = 0.97 (0.49–1.94)	–	No effect; RR = 1.11 (0.46–2.66)	–
Calcium	Possibly beneficial for women at high risk (RR = 0.35 (0.21–0.57)) and with low baseline intake (RR = 0.49 (0.38–0.62))	RCT in progress	Possibly beneficial for women at high risk (RR = 0.22 (0.11–0.43)) and with low baseline intake (RR = 0.32 (0.21–0.49))	RCT in progress
Iron and folate	No effect; RR = 1.15 (0.41–3.18)	–	No evidence	–
Folate	No effect; RR = 1.26 (0.90–1.76)	–	No evidence	–
Magnesium	No evidence	Needed	No effect; RR = 0.87 (0.57–1.32)	Needed
Fish oil	No effect; RR = 0.98 (0.91–1.04)	–	Possibly beneficial (data from low-quality studies); RR = 0.81 (0.69–0.93) Systematic review in process	–
Zinc	No effect; RR = 0.87 (0.65–1.15)	Needed	No evidence	Needed
Antioxidants	No evidence	RCT in progress	Possibly beneficial for vitamins C and E (data from one RCT); RR = 0.46 (0.24–0.91)	RCT in progress

to prevent the occurrence of pregnancy-induced hypertension or pre-eclampsia.

See also: **Ascorbic Acid**: Physiology, Dietary Sources and Requirements; Deficiency States. **Calcium**. **Copper**. **Cytokines**. **Fatty Acids**: Omega-3 Polyunsaturated. **Folic Acid**. **Hypertension**: Etiology; Dietary Factors. **Iron**. **Lipoproteins**. **Magnesium**. **Obesity**: Definition, Etiology and Assessment. **Pregnancy**: Dietary Guidelines and Safe Supplement Use. **Sodium**: Physiology; Salt Intake and Health. **Supplementation**: Dietary Supplements. **Vitamin A**: Deficiency and Interventions. **Vitamin E**: Metabolism and Requirements; Physiology and Health Effects. **Zinc**: Deficiency in Developing Countries, Intervention Studies.

Further Reading

Atallah AN, Hofmeyr GJ, and Duley L (2004) Calcium supplementation during pregnancy for preventing hypertensive disorders and related problems (Cochrane Review). In: *The Cochrane Library*, Issue 1. Chichester: John Wiley & Sons Ltd.

Belizán JM, Villar J, and Repke J (1988) The relationship between calcium intake and pregnancy-induced hypertension: up-to-date evidence. *American Journal of Obstetrics and Gynecology* 1988; **158**: 898–902.

Duley L and Henderson-Smart D (2004) Reduced salt intake compared to normal dietary salt, or high intake, in pregnancy (Cochrane Review). In: *The Cochrane Library*, Issue 1. Chichester: John Wiley & Sons Ltd.

Ferrer RL, Sibai BM, Murlow CD, Chiquette E, Stevens KR, and Cornell J (2000) Management of mild chronic hypertension during pregnancy: a review. *Obstetric Gynecology* **96**: 849–860.

Kramer MS and Kakuma M (2004) Energy and protein intake in pregnancy (Cochrane Review). In: *The Cochrane Library*, Issue 1. Chichester: John Wiley & Sons Ltd.

Kulier R, de Onis M, Gülmezoglu AM, and Villar J (1998) Nutritional interventions for the prevention of maternal morbidity. *International Journal of Gynecology and Obstetrics* 1998; **63**: 231–246.

Mahomed K (2000) Iron and folate supplementation in pregnancy (Cochrane Review). In: *The Cochrane Library*, Issue 4. Oxford: Update Software.

Mahomed K (2004) Folate supplementation in pregnancy (Cochrane Review). In: *The Cochrane Library*, Issue 1. Chichester: John Wiley & Sons Ltd.

Mahomed K (2004) Zinc supplementation in pregnancy (Cochrane Review). In: *The Cochrane Library*, Issue 1. Chichester: John Wiley & Sons Ltd.

Makrides M and Crowther CA (2004) Magnesium supplementation in pregnancy (Cochrane Review). In: *The Cochrane Library*, Issue 1. Chichester: John Wiley & Sons Ltd.

Moutquin JM, Garner PR, Burrows RF, Rey E, Helewa ME, Lange IR, and Rabkin SW (1997) Report of the Canadian Hypertension Society Consensus Conference: 2. Nonpharmacologic management and prevention of hypertensive disorders in pregnancy. *Canadian Medical Association Journal* **157**: 907–919.

National High Blood Pressure Education Program Working Group (1990) Report on high blood pressure in pregnancy. *American Journal of Obstetrics and Gynecology* **163**: 1689–1712.

Olsen SF, Secher NJ, Tabor A, Weber T, Walker JJ, and Gluud C (2000) Randomised clinical trials of fish oil supplementation in high risk pregnancies. *British Journal of Obstetrics and Gynaecology* **107**: 382–395.

National High Blood Pressure Education Program Working Group (2000) Report of the National High Blood Pressure Education Program Working Group on high blood pressure in pregnancy. *American Journal of Obstetrics and Gynecology* **183**: S1–S22.

Roberts JM and Cooper DW (2001) Pathogenesis and genetics of pre-eclampsia. *Lancet* **357**: 53–56.

Villar J and Belizan JM (2000) Same nutrient, different hypotheses: disparities in trials of calcium supplementation during pregnancy. *American Journal of Clinical Nutrition* **71**(Supplement): 1375S–1379S.

PREMENSTRUAL SYNDROME

M C de Souza, Universidad de Mogi das Cruzes, São Paulo, Brazil

A F Walker, The University of Reading, Reading, UK

Premenstrual syndrome (PMS), the common disorder which affects woman of reproductive age, is characterized by changes in mood, behavior, or mental and physical functioning in the luteal phase of the menstrual cycle, and yet has no defined cause for its etiology. The different hypotheses proposed in the literature indicate several mechanisms and hence a range of possible therapeutic strategies. The involvement of nutritional factors in PMS has allowed a dietary approach for the management and treatment of women experiencing the condition. This article focuses on the definition, classification, and prevalence of PMS, emphasizing the hypotheses of etiology, including dietary factors, in the relief of PMS symptomatology.

Definition, Classification, and Prevalence

Premenstrual syndrome, or premenstrual dysphoric disorder (PMDD), is an association of distressing

physical, psychological, and/or behavioral symptoms which occur in the luteal phase (second half) of the menstrual cycle of sufficient severity to interfere with the normal activities and personal relationships of many women. Although the late luteal phase is the most common time for symptoms of PMS to be experienced, occasionally symptoms may occur as early as ovulation. To be classified as PMS, symptoms must be relieved by the onset of or during menstruation. Indeed, a symptom-free week after menstruation is necessary for differential diagnosis from other gynecological or psychiatric disorders.

Despite the plethora of research studies on the subject, there exists no commonly agreed definition of PMS among gynecologists and researchers. The lack of consensus is probably due to the large number of symptoms described. PMS was previously associated only with nervous tension and termed premenstrual tension (PMT). However, this term is no longer used as it only describes a limited range of the numerous symptoms experienced by many women. The most commonly mentioned symptoms of PMS are now grouped into psychological and somatic categories (**Table 1**).

Table 1 Summary of the most common premenstrual symptoms

Psychological	*Somatic*
Aggression	Abdominal bloating
Agitation	Acne/spots
Anorexia	Breast swelling
Anxiety	Breast tenderness
Craving for sweets	Change in bowel habit
Argumentative	Clumsiness
Confusion	Constipation
Crying/easy upset	Cramps
Decreased alertness	Diminished activity
Decreased libido	Diminished efficiency
Depression	Diminished performance
Emotional liability	Dizziness
Forgetfulness	Heart pounding
Hopelessness	Hot/cold flushes
Hunger	Headache
Impulsive behavior	General pain/aches
Increased appetite	Infections (e.g., cold)
Lethargy	Passing water frequently
Insomnia	Migraine
Irritability	Nausea/sickness
Lack of inspiration	Oedema
Loss of attention	Poor concentration
Loss of concentration	Swelling of extremities
Loss of confidence	Weight gain
Loss of self-esteem	
Pessimism	
Loss of self-control	
Nervous tension	
Mood swings	
Sadness	
Violent feelings	
Social isolation	
Suicidal tendency	
Tiredness	

Classification

The variable symptoms of PMS have been classified into four main categories by the American clinician and researcher Abraham in an attempt to facilitate research and elucidate the etiology of PMS and its links with lifestyle, including diet (**Table 2**). Abraham contended that each PMS category may exist alone or in combination with other categories. For example, PMS-D normally manifests itself in association with PMS-A, which is usually exhibited first.

Although other classification systems for PMS symptoms have been suggested, there is little evidence to suggest any greater merit of them over Abraham's system. In 1992 an attempt was made to systematize criteria and procedures for diagnosing PMS as LLPDD (late luteal phase dysphoric disorder) in the *Diagnostic and Statistical Manual of Mental Disorders*. Despite this, there remains much confusion surrounding the classification of PMS symptomatology, which has led to difficulties of data interpretation and diagnosis.

In the absence of quantifiable signs of PMS, most studies have relied upon the self-reporting of symptoms. The methods have largely fallen into two categories: the use of a retrospective Menstrual Health Questionnaire (MHQ) or use of a prospective Menstrual Diary (MD). While the MHQ may be helpful in screening prospective volunteers for studies of PMS, the requirements for a useful instrument for assessing PMS must include the use of prospective recording and the

Table 2 Abraham's classification of premenstrual symptoms

PMS category[a]	*Symptoms*
PMS-A (Anxiety)	Anxiety, irritability, mood swings, nervous tension
PMS-H (Hydration)	Weight gain, abdominal bloating and tenderness, breast tenderness, swelling of the extremities
PMS-C (Craving)	Premenstrual increased appetite, craving for sweets, fatigue, palpitations, headache
PMS-D (Depression)	Depression, withdrawal, lethargy, forgetfulness, confusion, insomnia, difficulty verbalizing

[a]PMS (premenstrual syndrome) is used in place of Abraham's PMT (premenstrual tension).

quantification of symptoms which identify and exclude psychiatric occurrences.

Prevalence

PMS can first appear at any stage in the reproductive life of a woman. However, the most prevalent age of onset is usually from 28 to 34 years and symptoms may be first noticed following pregnancy or oral contraceptive use. The reported prevalence of PMS differs greatly from 21 to 90% of specific female populations studied. Some of this variability relates to the difficulty of a precise definition of PMS, its classification, and the reporting method used. In addition, other factors, such as age, parity, race, culture, psychopathology, menstrual characteristics, occupation, social activities, family life, lifestyle, and stress, may play a part. Indeed, reaction to stress has been advocated by a number of gynecologists as an underlying cause of PMS.

The Etiology of PMS

Various hypotheses have been implicated in the etiology of PMS, including interaction between ovarian steroid hormones, endogenous opioid peptides, central neurotransmitters, eicosanoids, and peripheral autonomic and endocrine secretions. Despite all the investigations relating to these diverse fields, there is no firmly established biological pattern for PMS and its pathophysiology still remains obscure.

Gonadal Hormone Imbalance

As early as the 1930s, it was suggested that PMS was caused as a result of excessive levels of the female sex hormones in the blood. Later, deficient blood levels of progesterone were blamed, so that by the 1960s PMS treatment was dominated by progesterone administration. Despite this, the findings from several studies showed no clear link between low blood progesterone levels and the severity of the condition.

More recently, attention has been given to the significance of the oestrogen-to-progesterone ratio as a factor triggering PMS. In particular, it has been postulated that women with PMS-A (anxiety) have increased plasma concentrations of oestrogen relative to progesterone in the luteal phase compared with normal women. One possible mechanism is that low plasma blood levels of oestrogen and progesterone early in the luteal phase in PMS sufferers lead to an increased secretion of gonadotrophic hormones from the pituitary which, in turn, leads to a rise in oestrogen in the late luteal phase, owing to the stimulation of ovarian follicles. These follicles rapidly regress under the influence of the luteal secretion of progesterone. Excess oestrogen in the luteal phase may be the cause of fluid retention, breast tenderness, changes in carbohydrate metabolism, and mood swings associated with PMS. Although the ovarian hormone imbalance hypothesis remains unproven, administration of progesterone in the latter half of the cycle continues to be the first-line treatment for PMS by general practitioners in the UK.

The existence of subgroups of PMS of varying etiology may account for clinical observations that some women with severe premenstrual depression report, paradoxically, a worsening of symptoms during progesterone treatment. Indeed, such women have been reported to respond to oestrogen treatment. Their condition may be linked to progesterone excess, which has been suggested to occur in some women who experience PMS symptoms at midcycle. High luteal progesterone levels may lead to depletion of oestrogen receptors of the hypothalamus, which is consequently less sensitive to oestrogen, requiring a higher midcycle oestrogen surge for normal pituitary response. This is followed by an abrupt and pronounced oestrogen drop at midcycle which may even result in symptoms of hot flushes.

Angiotensin

It has been demonstrated that symptoms such as irritability, withdrawal, depression, hopelessness, tension, lack of initiative, and weight gain are linked to changes in urinary potassium/sodium ratio, which in PMS women is higher 3–4 days before the onset of menses and lower for the rest of the cycle compared with that in normal women. The reason for this may lie in the enhanced secretion of angiotensin promoted by higher progesterone levels during the luteal phase of the menstrual cycle.

Angiotensin is a hormone with vasoconstricting properties which acts on the kidneys by constricting efferent arterioles. This reduces blood flow from the kidneys and enhances excretion of the electrolytes sodium and potassium. An electrolyte imbalance occurring in PMS may alter neurotransmitter activity, in particular monoaminergic (noradrenalin and serotonin) and cholinergic neurotransmitters, affecting behavior and mood regulation. Angiotensin is also known to act directly on the adrenal cortex to promote aldosterone release (see below) and may be involved in stimulating antidiuretic hormone (ADH) secretion by the pituitary. Both of these hormones are directly involved in the maintenance of electrolyte balance.

Aldosterone

Aldosterone, the steroid hormone secreted by the adrenal cortex, promotes retention of sodium and excretion of potassium by the kidneys. An elevated secretion of aldosterone in the luteal phase would tend to lead to sodium retention and, as a result, promote fluid retention. Whether the high aldosterone levels found in PMS are a consequence of direct adrenal cortex stimulation or are promoted by increased adrenocorticotrophic hormone (ACTH) secretion from the pituitary is unknown, although circulating levels of gonadal hormones are known to influence electrolyte homeostasis via the angiotensin–aldosterone system, as stated above.

Prolactin

Fluctuation in prolactin levels has also been suggested as a cause for PMS. Prolactin is a pituitary hormone which regulates mammary gland development in women and is necessary for successful lactation. Latent hyperprolactinemia is thought to predispose nonlactating women to premenstrual breast pain. Prolactin levels rise acutely due to stress and higher levels promote sodium, potassium, and water retention. Noradrenalin has been implicated in promoting prolactin release, while the presence of dopamine reduces it. Both stress and high oestrogen levels promote noradrenalin secretion, while dopamine may be reduced in those consuming a diet low in certain nutrients (see below).

Thyroid Stimulating Hormone

In one study, a high percentage of women with PMS who were treated with thyrotrophin releasing hormone (TRH) were found to have low thyroid function. As thyroid hormone supplementation has been used in the past as antidepressant therapy, some recent studies have attempted to link thyroid dysfunction to PMS. Indeed, there are suggestions that PMS may be an early symptom of a progressive thyroid disorder and the severity of PMS increases as thyroid dysfunction progresses. Nevertheless, the use of thyroid hormone for the treatment of PMS remains controversial.

Opioids, Peptides, and Endorphins

Opioid and neuropeptides receptors are found in nerve synapses of the brain and the gastrointestinal tract, and their suppression by morphine is known to alter the perception of pain. Endorphins are substances also found in the brain and in the pituitary gland that have opiate-like activity. Indeed, the endogenous opioid of the pituitary, β-endorphin, has been described as the body's own analgesic, absence of which leads to symptoms sometimes described by PMS sufferers: cramping, craving for carbohydrates, insomnia, irritability, and nausea. This observation has led to the hypothesis that β-endorphin deficiency may be the cause of PMS. If this is the case, then it is only likely to hold true for a minority subgroup of PMS because although lacrimation, diarrhea, and pupillary dilation are common in β-endorphin deficiency, they are not common in PMS. On the other hand, PMS symptoms such as depression, breast swelling or tenderness, and weight gain are not common features of β-endorphin withdrawal.

Neurotransmitter Imbalance

Depression is a commonly reported symptom of PMS. Monoamines, such as noradrenalin and serotonin, are known chemical mediators of mood. Monoamine oxidases (MAOs) and catechol-*O*-methyl transferase are enzymes that metabolize monoamines and thus decrease the amounts available for neural transmission. Their activity can be affected by the greater fluctuation in progesterone and oestrogen levels found in women with PMS compared with normal women.

Oestrogen suppresses MAO type A activity and increases MAO type B activity. While MAO type A enzymes are involved in the breakdown of adrenalin, noradrenalin, serotonin, and dopamine, MAO type B enzymes deactivate only dopamine. Thus suppression of type A and increase in type B enzyme activity results in excess serotonin, adrenaline, and noradrenalin and a relative deficiency of dopamine, a situation which may trigger anxiety and depression.

PMS and Dietary Factors

Various reports, many of a preliminary nature, or based on clinical experience, suggest that women suffering from PMS consume more sugar, refined carbohydrate, and dairy products and less fiber, B complex vitamins, iron, zinc, and magnesium than normal women.

Carbohydrate

Carbohydrate-rich meals have been shown to improve mood in women with premenstrual depression. The reason for this has been indicated in animal studies: The availability of tryptophan to the brain increases following such meals. As tryptophan is a substrate for the synthetic pathway to serotonin, levels of this neurotransmitter rise, while levels of other neurotransmitters are maintained. Hence

premenstrual craving has been suggested as a compensatory response to deal with a relative lack of serotonin during this phase of the cycle. Indeed, PMS symptoms of craving, particularly for sweet foods, may be a useful indicator of the cyclical changes that occur in brain neurotransmitters in women.

Increased energy intake due to craving has been shown to range from 380 to 2000 kJ per day (90–500 kcal per day) during the luteal phase compared with the follicular phase in women with PMS. However, there is no obvious pattern of macronutrient intake. Some women consume more carbohydrate, some more fat, and some more protein. As might be expected, the increase in energy intake is accompanied by an increase in certain nutrients, such as magnesium, vitamin D, potassium, phosphorus, and riboflavin. Enhanced intake of several of these nutrients has been linked to alleviation of PMS symptoms (see below).

Magnesium

Magnesium deficiency has been proposed as a causative factor in PMS. Magnesium has a sedative effect on neuromuscular excitability and is involved as an enzyme cofactor in many reactions in the body. In the metabolism of essential fatty acids it also acts as a cofactor, working together with vitamin B_6, zinc, niacin, and vitamin C. At the cell membrane, magnesium acts as a regulator of both membrane rigidity and ion exchange, helping to maintain electrolyte balance. In addition, it moderates the action of calcium in stimulating cell functions such as hormone secretion.

Modern Western diets high in refined cereals lack magnesium. Many dietary surveys, including those sponsored by governments throughout the Western world, have shown that the mean intake of magnesium for women is below recommended dietary standards, with subgroups having exceptionally low intakes. Decreased intake or absorption or increased renal excretion may lead to a reduced intracellular magnesium. Indeed, perhaps the most consistent physiological abnormality yet found for PMS subjects has been the reduced magnesium level in red blood cells compared with controls.

While severe magnesium deficiency is characterized by a progressive muscle weakness, failure to thrive, neuromuscular dysfunction, and tachycardia, symptoms of marginal deficiency are more subtle. Nevertheless, there is accumulating evidence to suggest that magnesium supplementation of the diet can alleviate a variety of conditions in which there is an element of muscular overcontraction, such as hypertension and tension headaches. The stress of modern living plays a part by enhancing magnesium excretion in the urine even in otherwise normal subjects. This sets up a vicious circle, as magnesium deficiency itself increases susceptibility to stress by increasing the secretion of ACTH-mediated adrenal androgen, which is a central nervous system depressant.

Several mechanisms proposed for the development of PMS symptoms have been claimed to be promoted by magnesium deficiency. Low magnesium status may also be responsible not only for exacerbating gonadal hormone imbalance in women, but may promote an increase in the aldosterone-to-oestrogen ratio. Enhanced aldosterone levels promote potassium and magnesium excretion and sodium retention, thus inducing fluid retention as found in PMS-H. In addition, deficient levels of magnesium decrease blood glucose control in two ways: by decreasing the ability of the liver to metabolize glucose and by increasing insulin secretion in response to glucose. Hence, changes in appetite and craving, both common PMS symptoms, may be closely linked to magnesium deficiency through loosening of blood glucose control. A low blood glucose supply to the brain may cause craving as a signal for increased energy intake. Even the decreased brain dopamine levels postulated to be responsible for anxiety and irritability of PMS (see above) may be exacerbated by magnesium deficiency.

In support of some of these hypotheses, several scientific reports have demonstrated a role for magnesium supplementation in relieving symptoms of PMS. In particular, Italian workers have shown that in 32 women, who were given 360 mg magnesium or placebo per day from the 15th day of the menstrual cycle to the onset of menstrual flow, magnesium supplementation was an effective treatment for low mood in PMS.

Vitamin B_6

Vitamin B_6, in the form of pyridoxal phosphate (PLP), is a cofactor in a large number of important enzymic reactions throughout the body. Therefore it is a cofactor in serotonin and dopamine production in the hypothalamus. High oestrogen levels may lead to a relative deficiency in vitamin B_6 by altering tissue distribution and by inducing hepatic enzymes which increase the rate of vitamin B_6 breakdown. An oestrogen-induced deficiency in vitamin B_6 may reduce the synthesis of both serotonin and dopamine, an action that may alter the delicate balance between the two, normally maintained by adequate synthesis and breakdown. Adequate intakes of vitamin B_6 are also thought to be necessary for the

maintenance of normal intracellular magnesium levels, as this vitamin plays a fundamental role in the active transport of magnesium through the cell membrane. Thus, a synergetic role for magnesium and vitamin B_6 has been suggested, although this remains to be tested.

High doses of vitamin B_6 have been found to be effective in treating most of the most common symptoms of PMS in several double-blind, placebo-controlled trials. For this reason administration of dietary supplements is a popular therapy for PMS used by many medical practitioners. However, as large doses have been associated with dependency and sensory neuropathy, doses higher than 50 mg per day should be avoided.

Essential Fatty Acids

Gamma-linolenic acid (GLA) is thought to be the major active constituent of evening primrose oil (EPO), which is self-administered by many women for the relief of PMS symptoms. GLA is a fatty acid belonging to the n-6 essential fatty acid family. It is formed in the body from linoleic acid (from seed oils such as sunflower). In the body linoleic acid is elongated and further desaturated in a several-step process leading to arachidonic acid. GLA is one of the intermediates in this pathway, which, in response to a stimulus, can act as a substrate for a series of enzyme reactions giving rise to series 1 eicosanoids (biologically active substances, including prostaglandins), which have a broad range of activities in the body.

Under similar circumstances arachidonic acid present in the cell membrane gives rise to series 2 eicosanoids, which tend to be proinflammatory unless moderated by the presence of series 1 and 3 eicosanoids (series 3 are from n-3 fatty acids, which are high in fish oils). In people consuming a Western diet, it is common to find that when cell membranes are stimulated (e.g., stressed), production of series 2 eicosanoids is dominant. This is because arachidonic acid can be provided in the preformed state in the diet in meat, and body status of GLA and ω-3 fats can be low. Although low status of the latter may derive from poor diet choice, low GLA status may result because of the slow action of the enzyme δ-6-desaturase, which is involved in the first desaturation step in the metabolism of linoleic acid. Its action is further slowed by viral infection, age, alcohol, stress, and lack of magnesium or zinc in the diet.

The suggestion that PMS may be caused by eicosanoid imbalance is based on the assumption that failure of the normal conversion of linoleic acid to GLA results in low levels of prostaglandin E_1 (PGE_1) eicosanoids in relation to the other eicosanoids, and this sensitizes tissues so that they respond abnormally to normal levels of oestrogen and progesterone. In support of this hypothesis, studies from Japan, Finland, and the UK have shown that women suffering from PMS have lower blood levels of GLA and DGLA (dihomo-γ-linolenic acid, a compound related to GLA), although linoleic acid levels are higher. Additional support for a therapeutic role of GLA in PMS comes from six double-blind, placebo-controlled studies which have shown a significant improvement in symptoms from a daily supplement of GLA in the form of EPO, although one further study carried out in Australia showed no benefit.

Cyclical mastalgia, or breast pain, in the premenstrual phase may or may not be accompanied by other symptoms of PMS. In any case, similar abnormalities in n-6 fatty acid profile of cell membranes to that described above for PMS have been found in this condition and good response to GLA supplementation has been reported. Indeed, EPO is commonly prescribed by breast surgeons as the first-line treatment for the condition. In one study nearly half of hospital outpatients with this condition showed a benefit of EPO treatment without side effects, although doses of at least 4 g per day are required.

Other Nutrients

Apart from those already discussed, a role for deficiency of other nutrients in the etiology of PMS has been suggested, although few well-designed studies have been reported. Studies in vitamin E-deficient animals suggest that vitamin E supplementation may enhance the production of eicosanoids of series 1 and reduce the release of arachidonic acid from phospholipids. Hence, this combined action would reduce the inflammatory tendency implicated in some forms of PMS. Two double-blind, placebo-controlled studies in the 1980s indicated that supplementation with vitamin E may alleviate PMS symptoms, but at least one other study reported no effect. In one of the positive studies, daily supplements of 300 IU significantly alleviated PMS symptoms of anxiety after 2 months, while 600 IU per day was required for the same duration to reduce PMS symptoms of craving and depression. These levels of vitamin E are far greater than can be obtained through diet.

Zinc deficiency may be involved in the etiology of PMS. This suggestion stems from observations of low luteal-phase zinc levels in women suffering from PMS. Several mechanisms for involvement of zinc deficiency in PMS have been proposed. Zinc is involved in the regulation of pituitary hormone secretion, influencing,

in particular, prolactin and luteinizing hormone activity, which may affect predisposition to PMS. Zinc is a modulator of endogenous opiate-receptor binding in the central nervous system, a system also implicated in the condition. Zinc also takes part in the synthesis of PGE_1 by its involvement in the release of DGLA and hence may influence eicosanoid balance (see above). Nevertheless, no placebo-controlled study has been carried out to show the effects of supplementary zinc as therapy for PMS.

Women with PMS have been reported to have higher intakes of calcium than normal women owing to excessive intake of dairy products. Foods in this group are characterized by having very high calcium-to-magnesium ratios and PMS sufferers have been reported to have diets with higher ratios than normal women. As a high intake of calcium is known to reduce magnesium absorption, a high calcium intake has been proposed to result in a chronic magnesium deficiency and PMS. It has also been postulated that excessive calcium intake may cause the behavioral changes of PMS by calcium interference with glucose breakdown as a source of energy to the brain. However, a controversial placebo-controlled study on 33 women showed that daily supplementation of 1 g of calcium for 3 months significantly reduced PMS symptoms of depression and fluid retention.

Caffeine

Several surveys of unselected women have shown that those who consume large amounts of beverages containing caffeine are more likely to suffer from PMS. Although constant consumption of low doses of caffeine may exacerbate the stress reaction and tendency to PMS, paradoxically, acute, high-dose consumption has been used to treat migraine headaches, although it was not reported whether these headaches were present premenstrually.

Botanicals

There has been renewed interest in recent years in the therapeutic applications of herbal medicine for a wide range of conditions. The active phytochemicals of the majority of commonly used herbs and their physiological effects are well reported. It is only recently, and mostly in Germany, that clinical studies of efficacy in treatment have been undertaken. There is no doubt from the clinical experience of practitioners that phytotherapy has much to offer for treatment of hormone imbalance syndromes in women, including PMS, but more research-based evidence is required.

An important herb used by phytotherapists to treat PMS is the chaste tree (*Vitex agnus-castus* L.). Extracts of the berries have been shown to reduce the abnormally high prolactin secretion of PMS via the ability of certain of its phytochemicals to mimic the action of dopamine by binding to dopamine receptors in the pituitary. Other herbs traditionally used in phytotherapy for PMS contain phyto-oestrogens. These molecules may have oestrogen-like action, either due to the steroidal nature of their active constituents (false unicorn root, *Chamaelirium luteum* A. Gray) or to the spatial similarity of active groups in their constituents, which allow them to bind to oestrogen receptors. Among the latter group are isoflavonoids and lignans, which appear to have 'adaptogenic' properties: They are weakly oestrogenic at low circulating oestrogen concentrations and antioestrogenic at high oestrogen concentrations. Isoflavonoids are present in soya bean and its products and in medicinal herbs such as black cohosh (*Cimicifuga racemosa* Nutt.); these show a beneficial effect in reducing symptoms of PMS and the menopause. Lignans are present in high concentration in seed coats, including wheat, and are especially high in linseed (*Linum ussitatissimum* L.). The presence of lignans may explain why women who eat high quantities of whole grains, fruit, and vegetables are less likely to suffer from PMS.

See also: **Appetite**: Psychobiological and Behavioral Aspects. **Behavior**. **Brain and Nervous System**. **Carbohydrates**: Regulation of Metabolism. **Cofactors**: Inorganic; Organic. **Hunger**. **Magnesium**. **Phytochemicals**: Epidemiological Factors. **Supplementation**: Dietary Supplements. **Vitamin B_6**.

Further Reading

Abraham GE (1982) Magnesium deficiency in premenstrual tension. *Magnesium Bulletin* **1**: 68–73.

Abraham GE and Rumley RE (1987) Role of nutrition in managing the premenstrual tension syndromes. *Journal of Reproductive Medicine* **32**: 405–422.

Backstrom T and Hammarback S (1986) Endocrinological aspects of the premenstrual syndrome. *Progress of Clinical and Biology Research* **225**: 421–428.

Brush MG, Watson SJ, Horrobin DF, and Manku MS (1984) Abnormal essential fatty acid levels in plasma of women with premenstrual syndrome. *American Journal of Obstetrics and Gynecology* **150**: 363–366.

Chuong CJ and Dawson EB (1992) Critical evaluation of nutritional factors in the pathophysiology and treatment of premenstrual syndrome. *Clinical Obstetrics and Gynecology* **35**: 679–692.

Dalton K (ed.) (1984) *Premenstrual Syndrome and Progesterone Therapy*, 2nd edn. London: William Heinemann Medical Books Ltd/Year Book Medical Publishers Inc.

Facchinetti F, Borella P, Sances G, Fioroni I, and Nappi RE (1991) Oral magnesium successfully relieves premenstrual mood changes. *Journal of the American College of Obstetrics and Gynecology* **78**: 177–181.

Gallant SJ, Popiel DA, Hoffman DM, Chakraborty PK, and Hamilton JA (1992) Using daily rating to confirm premenstrual syndrome/late phase dysphoric disorder. *Psychosomatic Medicine* **54**: 149–166.

Janowsky DS, Berens SC, and Davis J (1973) Correlation between mood, weight, and electrolytes during the menstrual cycle: A renin–angiotensin–aldosterone hypothesis of premenstrual tension. *Psychosomatic Medicine* **35**: 143–154.

London RS, Murphy L, Kitlowski KE, and Reynolds MA (1987) Efficacy of alpha-tocopherol in the treatment of the premenstrual syndrome. *Journal of Reproductive Medicine* **32**: 400–404.

Piesse JW (1984) Nutrition factors in the premenstrual syndrome. *International Clinical Nutrition Review* **4**: 54–81.

Reid RL and Yen SSC (1981) Premenstrual syndrome. *American Journal of Obstetrics and Gynecology* **1**: 85–104.

Shangold GA (1993) The premenstrual syndrome: Theories of etiology with relevance to the therapeutic use of GnRH agonists. *Seminars in Reproduction Endocrinology* **11**: 172–186.

Wurtman JJ, Brzezinski A, Wurtman RJ, and Laferrere B (1989) Effect of nutrient intake on premenstrual depression. *American Journal of Obstetrics and Gynecology* **161**: 1228–1234.

PROSTAGLANDINS AND LEUKOTRIENES

G E Caughey, M J James and L G Cleland, Royal Adelaide Hospital, Adelaide, SA, Australia

Introduction

Prostaglandins (PGs) and leukotrienes (LTs) belong to a large, heterogeneous group of lipid mediators, collectively named eicosanoids, that exhibit a diverse array of physiological activities (**Figure 1**). Eicosanoids are synthesized by oxygenation and remodeling of their precursor 20-carbon polyunsaturated fatty acids (PUFAs), namely arachidonic acid (AA; 20:4 *n*-6). Whilst pivotally involved in many homeostatic processes, eicosanoids are also implicated in the pathophysiology of many chronic disorders (**Figure 1**). Since the discovery in the mid 1930s of prostaglandins as a component of human semen that potently induced uterine contractility, the field of eicosanoid biology has expanded to include the PGs, LTs, thromboxanes (TXs), hydroxyeicosatetraenoic acids (HETEs), lipoxins (LXs, including epi-lipoxins), isoprostanes, and the cyclopentaeone PGs (**Figure 2**). Except the latter two classes, which are generated by nonenzymatic oxidation, synthesis of these mediators is tightly regulated by a number of enzymes. Eicosanoids generally act as paracrine or autocrine agents, in that they exert their biological effects locally, either on the cell from which they were synthesized or on neighboring cells. This chapter will focus primarily on the synthesis and physiological roles of the PGs and LTs and the regulation of their synthesis by dietary fatty acids.

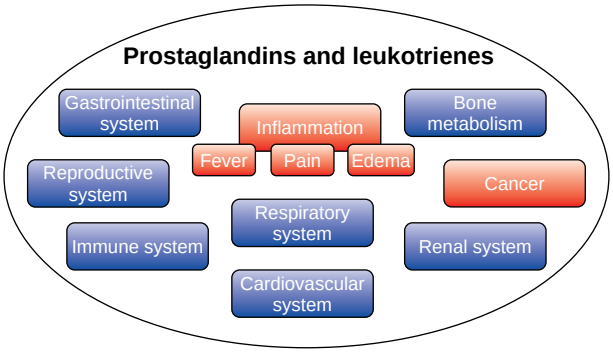


Figure 1 The diverse activities of prostaglandins and leukotrienes are reflected by their involvement in both normal homeostasis (blue) and pathophysiology (red).

Synthesis

Following an appropriate physiological or pathological stimulus, AA is released from cell membrane phospholipids by one of the many forms of phospholipase A_2 (PLA_2), which is generally regarded as the rate-limiting step in eicosanoid synthesis (**Figure 2**). There are two major biosynthetic pathways: the cyclooxygenase (COX/PGH synthase) pathway, which synthesizes the PGs and TXs; and the 5-lipoxygenase (5-LO) pathway, which synthesizes the LTs, HETEs, and lipoxins. The predominant cellular origins of PG and LT synthesis, their receptors, and their major physiological activities are summarized in **Table 1**.

Major Biosynthetic Pathways

Cyclooxygenase (COX)

COX catalyzes two enzymatic activities; namely, the conversion of AA to the hydroperoxy endoperoxide PGG_2, followed by its subsequent reduction to the labile product PGH_2. PGH_2 is the common substrate for a number of different cell-specific synthases, which convert PGH_2 to the individual PGs or TX, including PGE_2, PGI_2 (prostacyclin), PGD_2, $PGF_{2\alpha}$, and TXA_2 (**Figure 3**). Two isoforms of COX named

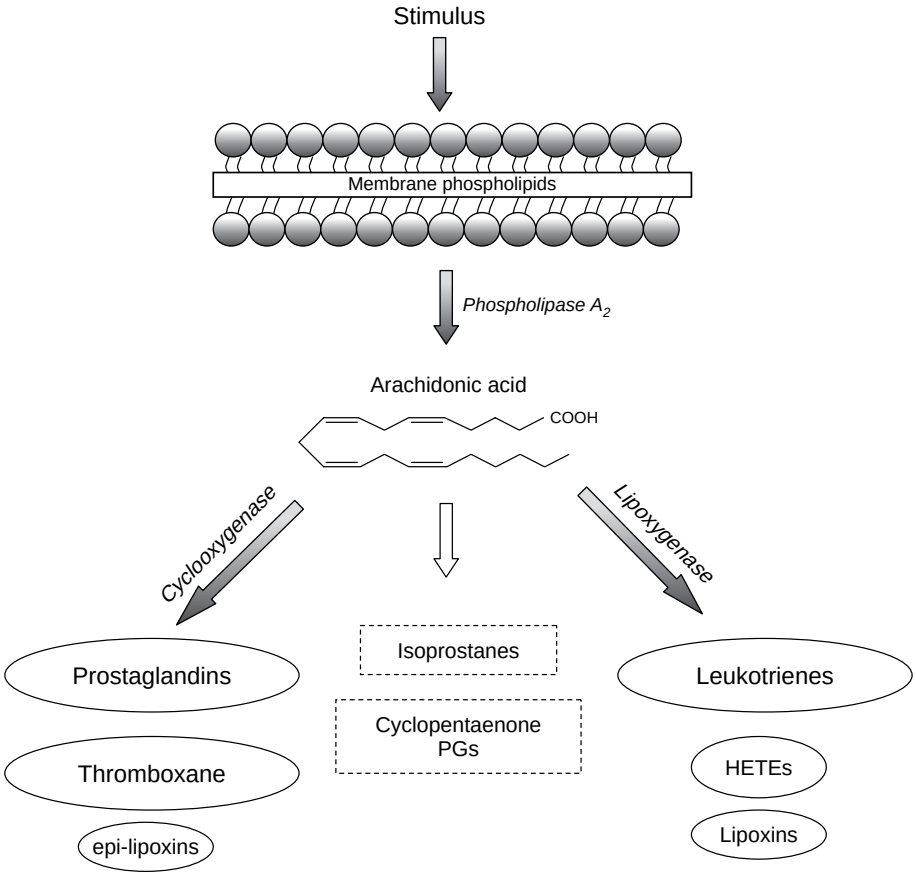


Figure 2 Metabolism of arachidonic acid to either the prostaglandins and thromboxane via the cyclooxygenase pathway or leukotrienes (and HETEs, lipoxins) via the lipoxygenase pathway. The isoprostanes and cyclopentaenone PGs are generated by nonenzymatic oxidation (dashed boxes). The epi-lipoxins are formed via interactions with cyclooxygenase and aspirin.

COX-1 and COX-2 were identified in the early 1990s and this led to renewed interest in the field of PG biology. Both isoforms catalyze the same reactions but are produced by different genes and although they share only about 61% sequence identity, the 3-dimensional crystal structures are virtually identical. After the discovery of the two isoforms, it quickly became apparent that their roles in many physiological processes were distinctive and that their expression and tissue profiles were differentially regulated. In general terms, COX-1 is constitutively expressed in most tissues and cell types and is responsible for the synthesis of the PGs required for the maintenance of normal physiology in the noninflamed state. Whereas COX-2 is generally undetectable in most tissues, its expression can be rapidly induced by a variety of inflammatory stimuli, such as bacterial LPS, cytokines, and growth factors. It is this isoform that synthesizes most PGs in inflammation and carcinogenesis. However, this division of the biological roles of COX-1 (physiological PGs) and COX-2 (inflammatory PGs) is an oversimplification of the biological reality. More recent studies have shown COX-1 can be induced or upregulated under certain conditions and that COX-2 is constitutively expressed in the brain and kidney. Thus, both COX-1 and COX-2 are involved in physiological as well as pathological responses.

The COX pathway is of major clinical importance because it is the major pharmacological target of nonsteroidal anti-inflammatory drugs (NSAIDs) and aspirin. Inhibition of PG synthesis is considered the primary mechanism responsible for both the therapeutic (anti-inflammatory, analgesic) and the toxic effects of NSAIDs. The clinically significant side effects of NSAIDs include renal impairment, dyspepsia, and upper gastrointestinal bleeding, the latter being particularly associated with inhibition of COX-1. By comparison, the anti-inflammatory and analgesic effects are associated with COX-2 inhibition. These observations provided the rationale for fast-track development of selective COX-2 inhibitors, which were promoted under the premise that they would have similar anti-inflammatory efficacy to conventional NSAIDs but would have significantly

Table 1 Predominant cellular origins and physiological activities of prostaglandins and leukotrienes

Eicosanoid/receptor		*Major cell origins*	*Physiological activities*
PGE_2	*EP_1–EP_4*	Most cell types	Potent vasodilator Stimulates bone and cartilage resorption Increases microvascular permeability Mediator of febrile responses Hyperalgesic
PGI_2	*IP*	Endothelial cells	Potent vasodilator Inhibits platelet aggregation
TXA_2	*TP_α, TP_β*	Platelets Monocytes	Potent vasoconstrictor and inducer of platelet aggregation
PGD_2	*DP_1, DP_2*	Mast cells	Vasodilator Inhibits platelet aggregation
$PGF_{2\alpha}$	*FP*	Monocytes Macrophages Uterine cells Epithelial cells	Potent vasoconstrictor and bronchoconstrictor Myometrial and smooth muscle cell contraction
LTB_4	*BLT_1, BLT_2*	Neutrophils Monocytes Macrophages Eosinophils Mast cells	Potent neutrophil chemotactic and chemokinetic agent Induces leucocyte adhesion Induces release of reactive oxygen species and hydrolytic enzymes by neutrophils
LTC_4 **LTD_4** **LTE_4**	*$CysLT_1$, $CysLT_2$*	Eosinophils Mast cells Macrophages Smooth muscle cells	Potent bronchoconstrictor Promotes vasoconstriction

fewer gastrointestinal (GI) side effects. Highly selective COX-2 inhibitors have been relatively successful with regard to their reduced GI toxicity. However, based on the role of COX-2 derived PGs in normal physiology, there exists the potential for other side effects such as increased cardiovascular events for at-risk patients and aggravated renal impairment in patients with reduced renal function.

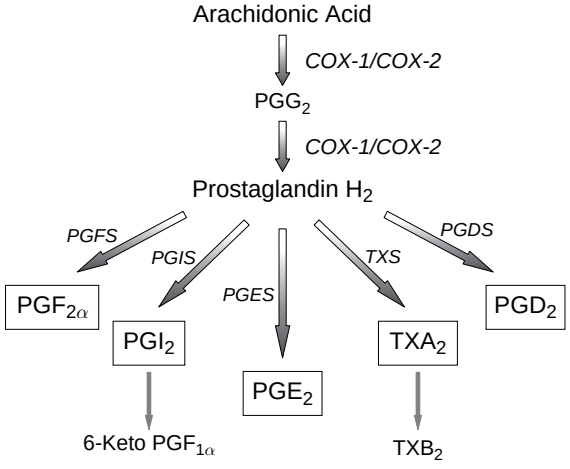


Figure 3 Metabolism of arachidonic acid (AA) to prostaglandin (PG) $PGF_{2\alpha}$, PGI_2 (prostacyclin), PGE_2, TXA_2, and PGD_2 by the COX pathway. PGI_2 and TXA_2 have very short half-lives (30 s) and are converted to the stable but inactive 6-keto $PGF_{1\alpha}$ and TXB_2, respectively.

Lipoxygenase

An alternative pathway available for the metabolism of AA is the 5-lipoxygenase (5-LO) pathway, which gives rise to the LTs which contain a conjugated triene structure. The 5-LO enzyme catalyzes the addition of oxygen at the 5th carbon of AA to produce 5-hydroperoxyeicosatetraenoic acid (5-HPETE) as well as the subsequent conversion of 5-HPETE to LTA_4. 5-LO is dependent on ATP and Ca^{2+} for activation, following which it translocates from the cytosol to the cell membrane in association with a transmembrane protein termed FLAP (5-LO activating protein). This translocation step facilitates substrate presentation, since the majority of AA is found in the cell membrane. LTA_4 may undergo one of two enzymatic reactions depending on the cell type. In the first, glutathione S-transferase (LTC synthase) catalyzes the addition of glutathione to the 6 position of LTA_4 to produce the first of three cysteinyl LTs (CysLTs), LTC_4. LTC_4 is then exported to the extracellular space through a specific transmembrane transporter. In the extracellular space subsequent peptide cleavage yields LTD_4 and then LTE_4. This represents metabolism from one active mediator to another and not a catabolic inactivation process. Collectively, these three CysLTs (historically known as the slow-reacting substance of anaphylaxis) contribute to the bronchoconstricting activity generated during anaphylaxis and they

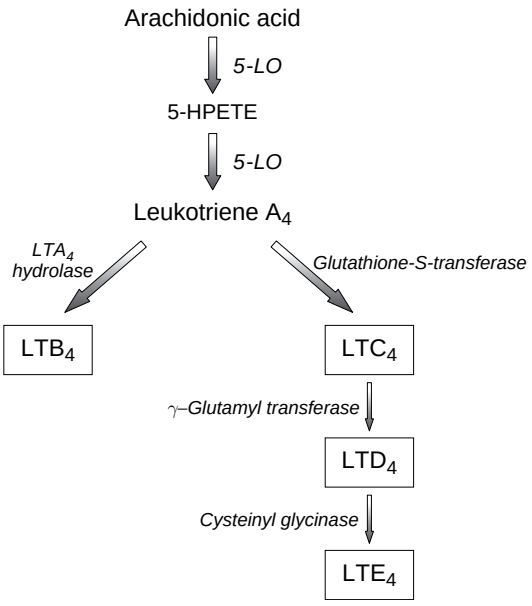


Figure 4 Metabolism of arachidonic acid (AA) to leukotriene B_4 (LTB_4) and the cysteinyl LTs, LTC_4, LTD_4, and LTE_4 by the 5-lipoxygenase (5-LO) pathway.

play a key role in asthma and allergic reactions. In the second reaction, LTA_4 hydrolase converts LTA_4 to the dihydroxy fatty acid LTB_4. Once formed, LTB_4 is actively exported from the cells where it acts as a potent chemoattractant and triggers adherence and aggregation of leucocytes. In addition, LTB_4 modulates immune responses and host defense against infections. Release of LTB_4 may contribute to the pathology of many inflammatory disorders, including asthma, arthritis, and inflammatory bowel disease.

PG and LT Receptors

The biological actions of both PGs and LTs are mediated through G-protein-coupled cell surface receptors, which are coupled to specific signal transduction pathways. Eight subtypes of PG receptors encoded by separate genes are characterized and include the PGE receptors (EP1, EP2, EP3, EP4), the TX receptor (TP), the PGI receptor (IP), the PGF receptor (FP), and the PGD receptor (DP). The tissue distributions of these receptors is linked to specific functional roles and can be grouped into three categories based on their signal transduction and activities:

1. Relaxant receptors (EP2, EP4, IP, DP): mediate an increase in cAMP and smooth muscle relaxation.
2. Contractile receptors (EP1, TP, FP): mediate an increase in intracellular calcium and smooth muscle cell contraction.
3. Inhibitory receptors (EP3): mediate inhibition of cAMP and inhibition of smooth muscle cell relaxation.

Two receptors for LTB_4, termed BLT1 and BLT2, have been identified at the molecular level and these differ in their affinity and specificity for LTB_4. BLT1 is a high-affinity receptor specific for LTB_4 and BLT2 is a low-affinity receptor, to which other eicosanoids can also bind. The major activities of LTB_4 appear to be mediated via BLT1. The precise role of BLT2 remains to be identified.

Two subtypes of the receptor for the CysLTs termed $CysLT_1$ and $CysLT_2$, were postulated based on pharmacological studies. This classification has recently been confirmed by molecular identification. To date, several high-affinity $CysLT_1$ receptor antagonists have been developed, which have been shown to be clinically efficacious in chronic asthma. By contrast, no high-affinity selective $CysLT_2$ antagonists have yet been described.

The development of specific agonists and antagonists for each of the PG and LT receptors, in addition to receptor knockout animal models, have aided in the characterization and understanding of the roles of these mediators in both normal physiology and disease states.

Physiological Activities of Prostaglandins and Leukotrienes

Bone Metabolism

Bone remodeling, the continuous process of bone resorption by osteoclasts and bone formation by osteoblasts, is mediated by a number of factors, one of which is PGE_2. However, the role of PGs in bone metabolism is somewhat contradictory, in that PGE_2 can have both anabolic and catabolic effects. For example, PGs can stimulate the differentiation *in vitro* of precursors of both osteoclasts (responsible for bone resorption) and osteoblasts (bone growth). Both bone resorption and formation by PGE_2 is mediated by the EP4 receptor. It has been suggested that the opposing actions of PGE_2 may serve to maintain a coordinated regulation of bone resorption and formation and may be dose related with stimulation of bone formation at low concentrations and inhibition at high concentrations.

Cancer

The involvement of PGs, in particular those arising from COX-2, in the causation and prevention of cancer has been identified recently. Much of the

evidence has come from epidemiological studies that indicate chronic use of aspirin or other NSAIDs can significantly decrease the risk of developing certain cancers (e.g., colorectal cancer). Additionally, aspirin and NSAIDs can lower the mortality rates and induce tumor regression from colorectal cancer and other forms of cancer. Angiogenesis, the development of new blood vessels, which is an essential step in tumor growth, is associated with the upregulation of COX-2 and presumably PG synthesis. COX-2 is upregulated in a variety of premalignant and malignant states and there is evidence that selective COX-2 inhibitors can inhibit the early development of malignant tumor growth, cause premalignant tumors to regress, and lead to death of established cancer cells.

Cardiovascular

The PGs and TX have a central role in the regulation of platelet aggregation and vascular tone and as such are particularly important regulators of the cardiovascular system. TXA_2 is a potent vasoconstrictor and inducer of platelet aggregation, whilst PGI_2 dilates blood vessels and prevents platelet aggregation. Under normal conditions, a dynamic balance based on the opposing actions between TXA_2 produced by platelets and PGI_2 produced by vascular endothelial cells maintains cardiovascular homeostasis and prevents thrombotic events (**Figure 5**). Altered metabolism of these mediators has been reported in association with atherosclerosis, in which there is a shift in the TXA_2/PGI_2 balance due to both increased TXA_2 and decreased PGI_2 synthesis. A proatherogenic state with increased platelet adhesion and aggregation at sites of endothelial injury is a result. This may lead to subsequent thrombus formation and vessel occlusion.

Platelets only express COX-1 and the cardioprotective effect of low-dose aspirin is attributed to its ability to inhibit irreversibly platelet COX-1 and hence TXA_2 synthesis. In the vasculature, it appears that endothelial cell COX-2, possibly upregulated by shear blood flow, is primarily responsible for the synthesis of PGI_2. Traditional NSAIDs inhibit both COX-1 and COX-2, thereby inhibiting platelet-derived TXA_2 and endothelium derived PGI_2. The balance may therefore be maintained but in a less stable state. In contrast, the selective COX-2 inhibitors reduce PGI_2 synthesis but importantly have no effect on platelet COX-1 activity, potentially altering the balance of TXA_2/PGI_2 to a prothrombotic state that may explain the potential thrombotic side effects associated with use of these drugs. Low-dose aspirin inhibits platelet COX-1 which cannot be regenerated by these anuclear cells thereby shifting the balance between TXA_2 (principally platelet COX-1 derived) and PGI_2 toward a less coagulable state.

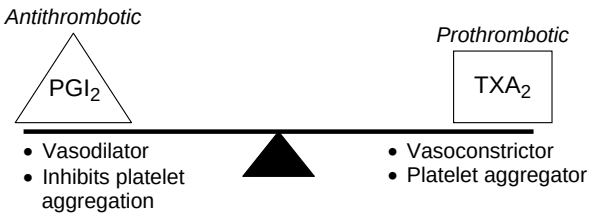


Figure 5 A dynamic balance between TXA_2 production by platelets and PGI_2 by vascular endothelial cells maintains cardiovascular homeostasis.

LTB_4 can also be implicated in cardiovascular events by virtue of its potent chemotactic effects and ability to induce leucocyte adhesion to vascular endothelial cells at sites of injury.

Gastrointestinal System

Of the COX isotypes, only COX-1 is constitutively expressed throughout the gastrointestinal (GI) system, where the main PGs produced are PGE_2 and PGI_2. Both have important cytoprotective effects on the GI mucosa, including reducing gastric acid secretion from stomach parietal cells, increasing mucosal blood flow, and stimulating the release of protective mucus. As stated previously, the upper GI toxicity commonly associated with classical NSAIDs is thought to arise from the nonselective inhibition of COX-1 activity in the stomach. Clinical trials with highly selective COX-2 inhibitors have demonstrated clinically meaningful reductions in the incidence of serious upper GI events by comparison to conventional NSAIDs.

COX-2 is expressed in peptic ulcers and the inhibition of COX-2 has been associated with delayed ulcer healing.

Immune System

Within the immune system, PGE_2 regulates a wide range of functions, particularly in the cell populations central to the cell-mediated immune response, namely T cells and macrophages. In these immune-modulating cells, the actions of PGE_2 are generally immunosuppressive. For example, PGE_2 inhibits antigen-induced T-cell proliferation and activation, cytokine production, cytokine receptor expression and macrophage proliferation, and class II major histocompatibility complex expression. Additionally, PGE_2 can

regulate the overall characteristic of an immune response by its ability to promote a Th2-type response, which is characterized by immunoglobulin (Ig) class switching to IgG1 and IgE and increased production of IL-4, IL-5, and IL-10. The ability of PGE_2 to inhibit many of the responses initiated by T-cell activation supports a central role for PGE_2 within the immune response. Moreover, PGE_2 has inhibitory and protective functions in autoimmune disease. Administration of PGE_2 or its analogs can ameliorate the manifestations of autoimmunity and reduce immune-mediated organ injury, and can also delay or prevent allograft rejection.

Inflammation

Inflammation is a complex of sequential and partly recursive cellular and biochemical changes in tissues in response to injury or infection. It is a normal homeostatic process that protects the host against the effects of everyday and incident trauma and invasive microorganisms. However, when this process becomes dysregulated unwanted inflammation and tissue destruction arises. Acute inflammation is characterized by hyperemia, pain, edema, and leucocyte infiltration. PGs are involved in these processes, as further illustrated by the analgesic effects and reduction in inflammatory swelling of NSAIDs. PGE_2 has been regarded as the principal PG mediator of pain and edema, but both PGI_2 and PGD_2 can exert similar effects. These PGs exert their hyperalgesic effects by increasing sensitivity of pain receptors to peripheral inflammation. PGE_2 dilates vessels and with LTB_4 leads to tissue swelling with both edema and leucocyte infiltration.

While levels of PGs and LTs are generally low in uninflamed tissues, their synthesis is increased substantially during an inflammatory response. As immune cells infiltrate the tissues, further increases in levels of PGs and LTs are observed. Induction of PLA_2 and COX-2 by inflammatory stimuli accounts for the high levels of PGs found at sites of inflammation. Cellular infiltration mediated by LTB_4 is contributed to by the chemotactic effects on leucocytes and altered adhesion molecule expression on endothelial cells.

Fever

Fever, a common symptom of many diseases, is elicited by exogenous pyrogens (such as bacterial LPS) or an inflammatory insult, which results in the production of cytokines such as IL-1β that act as endogenous pyrogens. These cytokines then stimulate the neural pathways that increase body temperature. PGE_2 is an important contributor to the febrile response and NSAIDs are used to treat pyresis. In COX-2-deficient mice, the febrile response to LPS is ameliorated, suggesting an important role for COX-2-derived PGs in fever production. Furthermore, mice deficient in the EP3 receptor fail to respond to either endogenous or exogenous pyrogens.

Respiratory System

The bronchoconstrictor activity of the CysLTs underlies their pathogenic role in asthma. Other relevant biological actions of CysLTs include: (1) increased microvascular permeability, which leads to airway edema; and (2) mucus hypersecretion. Both the 5-LO inhibitors and CysLT receptor antagonists have been used to treat asthma.

PGs, in particular PGD_2, are involved in many processes within the lung, including regulation of pulmonary vascular tone, maintenance of lung surfactant, regulation of capillary and alveolar permeability and control of bronchial mucous secretion.

Renal

Maintenance of normal kidney functions is dependent on PGE_2, which regulates vascular tone, blood flow, sodium, and water homeostasis, and renin secretion. PGE_2 can reduce sodium and water reabsorption and mediate the release of renin, which in turn can act to regulate blood pressure control. Under conditions of increased sodium reabsorption, PGE_2 can act as a counter-regulatory factor. PGI_2 is involved in potassium secretion by stimulating the renin–angiotensin system. Both isoforms are constitutively expressed in the kidney with quite selective and distinct localization. For example, COX-2 is highly expressed in the macula densa, which plays an important role in the coordinated regulation of glomerular filtration, proximal tubule function, and renin production, processes that are responsible for sodium and water homeostasis. In those with poor renal function, reversible renal failure has been associated with NSAIDs and highly selective CXO-2 inhibitors. This presumably reflects a crucial compensatory role for PGs in the compromised failing kidney.

Reproduction

PGs play important regulatory roles in the reproductive processes of ovulation, implantation, and parturition. Just prior to ovulation, there is an increase in PGE_2 synthesis by the preovulatory follicle in response to an increase in leutenizing hormone (LH). Induction of COX-2 (by LH) is necessary for this increase in PGE_2 synthesis and for the successful rupture of the follicle. After fertilization, PGs (PGE_2 and PGI_2) play a role in the successful implantation

of the embryo and, again, PGE_2 production appears to be COX-2 dependent. The importance of COX-2 in the reproductive process is further emphasized by studies in which COX-2 knockout mice (but not COX-1 knockout mice) have impaired fertility based on multiple reproductive failures, at the level of both ovulation and implantation, which can be restored by exogenous administration of PGE_2. Both PGE_2 and $PGF_{2\alpha}$ have potent uterotonic activities and are involved in uterine contraction during the initiation of labor and parturition. Administration of $PGF_{2\alpha}$ and PG analogs have been used extensively to induce labor since the 1960s. In addition, at the time of parturition, there is an increase in the uterine expression of EP1, EP3, and FP receptors, which act to potentiate smooth muscle cell contraction. In the neonate, COX-2-derived PGE_2 is required for closure of the ductus arteriosis, through an action by EP4 receptors on smooth muscle cells.

Regulation of Prostaglandin and Leukotriene Synthesis by Dietary Fatty Acids

The diverse physiologic and pathologic functions mediated by eicosanoids highlight the importance of their fatty acid precursors in the diet. Unlike cellular proteins that are genetically predetermined, the PUFA composition of cell membranes is dynamic and is pivotally dependent on dietary intake. The typical Western diet is high in the *n*-6 family of PUFA (up to 25-fold more *n*-6 fats than *n*-3 fats are consumed). This predominance of *n*-6 fat is due to the abundance in the diet of the 'parent' 18-carbon PUFA linoleic acid (LA; 18:2 *n*-6), which is present in high concentrations in corn, soy, safflower, and sunflower oils. Once ingested, LA can be converted to AA by a series of elongation and desaturation enzymes (**Figure 6**). Hence AA is the predominant PUFA of membrane phospholipids and substrate for eicosanoid biosynthesis in the Western context.

The enzymes involved in the metabolism of the 20-carbon PUFAs to PGs and LTs can use either *n*-9 (eicosatrienoic acid; EtrA 20:3), *n*-6 (arachidonic acid; AA 20:4), or *n*-3 (eicosapentaenoic acid; EPA 20:5) PUFAs as the substrate (**Figure 7**). When *n*-3 PUFAs are included in the diet, EPA, the *n*-3 homolog of AA competes with AA for incorporation into the cellular phospholipids. An increase in the concentration of EPA in cell membranes displaces AA, which will result in reduced substrate for the synthesis of the *n*-6 eicosanoids. EPA can also compete with AA as the substrate for either COX or 5-LO enzymes. This results in inhibition of the synthesis of *n*-6-derived PGs and LTs and the formation of the *n*-3 PGs and LTs. The *n*-3-derived PGs and LTs are similar in structure but can be considerably different in their biological activity. On balance, the *n*-3 PGs and LTs are less thrombotic and less inflammatory than the homologous *n*-6-derived mediators (**Figure 6**).

Although the *n*-9 fatty acid oleic acid (OA; 18:1 *n*-9) is consumed in substantial amounts in the diet, the elongase and desaturase enzymes that catalyze the conversion of OA to ETrA (20:3 *n*-9)

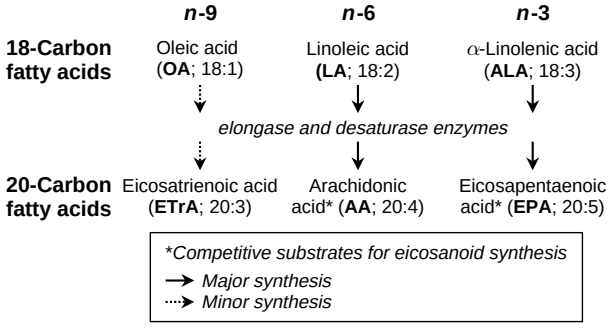


Figure 6 Dietary fatty acids and their metabolism after ingestion via the desaturase/elongase pathways.

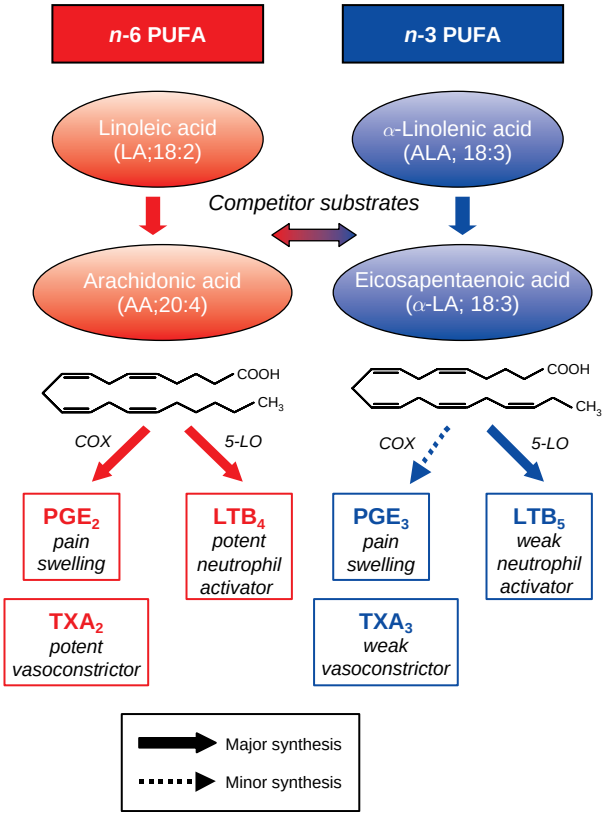


Figure 7 A comparison of the physiological activities between the *n*-6- and *n*-3-derived PGs and LTs.

preferentially metabolize the *n*-3 and *n*-6 20-carbon PUFAs, α-LA and LA, respectively. Metabolism of OA to ETrA is only quantitatively significant in essential fatty acid deficiency, which is very rare due to the abundance of essential fatty acids available in the diet and the small amounts required to avoid deficiency. Furthermore, ETrA can be metabolized by 5-LO but not COX because it lacks the *n*-6 bond necessary for PG and TX formation.

While Western diets are rich in *n*-6 and relatively poor in *n*-3 fats, there are populations in which *n*-6 fats are less dominant and more *n*-3 fats are consumed in total and relative terms (e.g., Greenland Eskimo, Japanese, and Mediterranean diets). In the extreme case of the Eskimos eating their aboriginal diet, which is based almost entirely on marine foods, there is a striking reduction in thrombotic vascular events and inflammatory diseases. The cardiovascular benefit has also been associated with traditional Japanese and Mediterranean diets. These benefits may be, in part, ascribed to a more favorable balance of *n*-6 and *n*-3 derived eicosanoids, although a myocardial membrane stabilizing effect of *n*-3 fats, independent of PG and LT synthesis, is also important.

See also: **Bone**. **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Cytokines**. **Fatty Acids**: Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated. **Immunity**: Physiological Aspects. **Pregnancy**: Nutrient Requirements. **Stomach**: Structure and Function.

Further Reading

Bergström S, Danielsson H, Klenberg D, and Samuelsson B (1964) The enzymatic conversion of essential fatty acids into prostaglandins. *Journal of Biological Chemistry* **239**: 4006–4009.

Boers M (2001) NSAIDs and selective COX-2 inhibitors: competition between gastroprotection and cardioprotection. *Lancet* **357**: 1222–1223.

Dyerberg J and Bang HO (1979) Haemostatic function and platelet polyunsaturated fatty acids in Eskimos. *Lancet* **2**: 433–435.

Drazen JM, Israel E, and O'Bryne PM (1999) Treatment of asthma with drugs modifying the leukotriene pathway. *The New England Journal of Medicine* **340**: 197–206.

Fitzgerald GA and Patrono C (2001) The coxibs, selective inhibitors of cyclooxygenase-2. *The New England Journal of Medicine* **345**: 433–442.

Funk CD (2001) Prostaglandins and leukotrienes: Advances in eicosanoid biology. *Science* **294**: 1871–1875.

Harris RC and Breyer MD (2001) Physiological regulation of COX-2 in the kidney. *American Journal of Physiology* **281**: F1–11.

James MJ, Gibson RA, and Cleland LG (2000) Dietary polyunsaturated fatty acids and inflammatory mediator production. *American Journal of Clinical Nutrition* **71**: 343s–348s.

Hui Y and Funk CD (2002) Cysteinyl leukotriene receptors. *Biochemical Pharmacology* **64**: 1549–1557.

McAdam BF, Catella-Lawson F, Mardini IA, Kapoor S, Lawson JA, and Fitzgerald GA (1999) Systemic biosynthesis of prostacyclin by cyclooxygenase (COX)-2: The human pharmacology of a selective inhibitor of COX-2. *Proceedings of the National Academy of Science USA:* **96**: 272–277.

Narumiya S and FitzGerald GA (2001) Genetic and pharmacological analysis of prostanoid receptor function. *The Journal of Clinical Investigation* **108**: 25–30.

Prescott SM (2000) Is cyclooxygenase-2 the alpha and the omega in cancer? *The Journal of Clinical Investigation* **105**: 1511–1513.

Samuelsson B, Dahlen SE, Lindgren JA, Rouzer CA, and Serhan CN (1987) Leukotrienes and lipoxins: structures, biosynthesis and biological effects. *Science* **237**: 1171–1176.

Samuelsson B and Funk CD (1989) Enzymes involved in the biosynthesis of leukotriene B_4. *Journal of Biological Chemistry* **264**: 19469–19472.

Serhan CN and Levy B (2003) Success of prostaglandin E_2 in structure-function is a challenge for structure-based therapeutics. *Proceedings of the National Academy of Science USA* **100**: 8609–8611.

Smith WL and Lanenbach R (2001) Why there are two cyclooxygenase isozymes. *The Journal of Clinical Investigation* **107**: 1491–1495.

Smith WL, DeWitt DL, and Garavito RM (2000) Cyclooxygenase: structural, cellular and molecular biology. *Annual Review of Biochemistry* **69**: 145–182.

PROTEIN

Contents

Synthesis and Turnover

D J Millward, University of Surrey, Guildford, UK

Whole Body Protein Homeostasis

The regulation of the protein mass of the body requires mechanisms that control the protein content within cells, organs, and tissues and that coordinate this control during growth and body weight maintenance. Because all intracellular proteins exhibit turnover, regulation of the cellular protein content involves control of both protein synthesis and proteolysis. For some proteins, such control is understood in considerable detail. Less is known about the coordinated control of intracellular protein turnover to maintain an appropriate cellular composition, and even less is known about whole body coordination. There are two aspects of control: acute control during feeding and fasting and chronic control during growth and long-term maintenance.

The diurnal cycle of feeding and fasting characteristic of human nutrition results in gains and losses of body protein in overall nitrogen balance. Furthermore, such diurnal cycling occurs with an amplitude that increases with increasing dietary protein intakes. These oscillations in body protein content involve changes in whole body and tissue protein synthesis, proteolysis, and amino acid oxidation, and much effort has been invested in identifying the control mechanisms, particularly those that influence the efficiency of postprandial protein utilization.

Long-term homeostasis is a less well understood phenomenon. For the slow-growing long-lived human, most of the life span involves a constant body weight and the remarkable phenomenon of this long-term constancy of body protein at a characteristic mass is a particularly challenging problem. Regulatory mechanisms exist that allow restoration of body weight and especially protein content to its target size after an insult that induces wasting (i.e., catch-up growth) and that prevents the continuation of growth after the target size has been reached. However, such mechanisms are poorly understood.

One approach to the problem has involved the concept of a protein-stat mechanism, the central feature of which is an interaction between linear growth of bone, protein deposition in skeletal muscle, and dietary protein intake, with the growth of most other organs secondary to this interaction (**Figure 1**). Within this context, whole body protein content is controlled through an amino-static appetite mechanism, acting primarily to maintain skeletal muscle mass at a level set by the linear dimensions of the organism. Bone lengthening occurs at rates determined by genetic programming and an appropriate hormonal anabolic drive, exerted by dietary protein. Bone lengthening controls, by passive stretching, net protein deposition in skeletal muscle mainly through the regulation of new connective tissue synthesis, which controls muscle volume. Some level of muscle activity is also required for maximal muscle size. Provision of amino acids to allow muscle to accumulate myofibrillar protein and increase to its phenotypic size is regulated through appetite stimulation, which monitors net amino acid flow into muscle. This is most obvious in catch-up growth. After muscle wasting with loss of myofiber protein there is potential for expansion within the preexisting connective tissue framework. Muscle growth ceases in the absence of passive stretch when bone length growth ceases. The growth of most other organs is secondary to this main interaction, determined primarily by the level of protein intake and the consequent metabolic work and functional demand for the organ, and is not specifically limited in size.

Protein Turnover

Protein turnover occurs because of the presence within cells of proteolytic systems that degrade

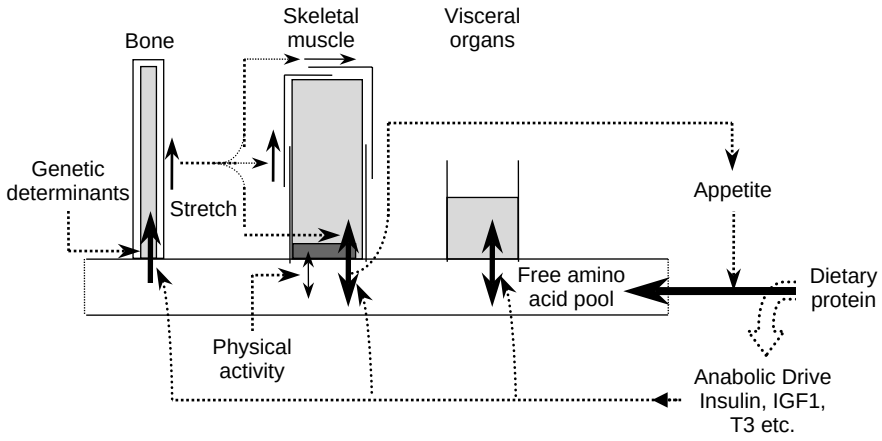


Figure 1 Protein-stat mechanism for coordinated control of body protein growth and maintenance. The protein content of skeletal muscle is controlled by long bone growth mediated through a passive stretch mechanism and anabolic signals in response to dietary protein intake, with the growth of most other organs secondary to this interaction—that is, growth rate for these organs is a function of metabolic and functional demand in response to food intake.

proteins for a variety of reasons ranging from the removal of proteins with an incorrect primary amino acid sequence ("error" proteins) to the provision of free amino acids during nutrient deprivation. However, the half-lives of individual proteins vary over at least three orders of magnitude within the same cells, identifying the process as specific. The nature and control of proteolysis are poorly understood. The physicochemical structure, especially hydrophobicity and ionic charge, influences susceptibility to proteolysis. Also, an amino acid sequence (PEST: proline, glutamate, serine, and threonine) influences susceptibility to proteolysis and occurs in several rapidly degraded proteins. The molecular basis for this remains largely unknown. For heterogeneous turnover of proteins within structures such as mitochondria, myofilaments, and multienzyme complexes to occur while functional integrity is maintained, complex coregulation is required, and there is evidence for at least three different systems involved in the case of skeletal muscle.

The lysosomal-autophagic system is present in all cells and involves acid proteinases (cathepsins) within a distinct vacuolar structure capable of engulfing and degrading complete organelles, ribosomes, as well as individual intracellular proteins and proteins entering cells via endocytosis. Lysosomal proteolysis is complete, and most is known about hepatic macroautophagy in which hepatic protein mass appears to be regulated by a receptor-mediated amino acid-dependent inhibitory process.

The ubiquitin–proteasome system is widely distributed among tissues, with a relatively broad protein specificity, catalyzing the hydrolysis of protein to peptides averaging approximately 8 amino acids long and exhibiting an ATP dependency. It involves two components. One is a recognition system involving ATP-dependent formation of a covalent link between the protein and a short polymer of ubiquitin, which is responsible for targeting the protein substrates toward proteolysis. This phase involves three separate reactions: ATP-dependent activation of ubiquitin, conjugation of ubiquitin to cellular proteins, and proofreading of the conjugates to either regenerate the target protein by removal of ubiquitin or commit the target protein to proteolysis by further ubiquitinylation. Proteolysis is mediated by the giant multifunctional protease, the proteasome. This comprises a core particle made of duplicate sets of at least 14 different proteins assembled in groups to form rings that are in turn stacked to form a donut-like structure within which the ubiquitin-conjugated proteins are unfolded by another ATP-dependent process and proteolytically cleaved to form peptides. A regulatory particle both delivers the ubiquitin-conjugated proteins to the core particle and removes ubiquitin from the peptides released after proteolysis. Proteolysis of the peptides is achieved by other systems, including the lysosomal system since degradation of ubiquitinylated proteins can also be achieved by the lysosome, or other poorly described proteinases and peptidases including the giant protease tripeptidyl peptidase II (TPP II) and various aminopeptidases. The relative importance of the two main systems capable of complete proteolysis, the lysosome and the proteosome, in various tissues remains uncertain. Most work on the regulation and activation of the ubiquitin–proteasome system has involved the accelerated

proteolysis in skeletal muscle atrophy, antigen processing, and removal of aberrant proteins rather than basal and nutritionally sensitive proteolysis. The third proteolystic system involves calpain and calpastatin, a calcium-activated proteolytic pathway that can initiate but not complete proteolysis. This comprises a highly conserved family of nonlysosomal calcium-dependent cysteine proteases composed of two ubiquitous isoforms (calpain I and II), several tissue-specific isoforms, and a 28-kDa regulatory subunit (calpain 4). *In vivo* calpain activity is tightly regulated by its endogenous and highly specific inhibitor, calpastatin. There is little evidence to suggest a role in nutritionally sensitive basal proteolysis.

Models and Tracer Methods for the Study of Protein Turnover

Studies of protein turnover have utilized isotope tracer techniques, radioactive tracers (^{14}C and ^{35}S) in animals, and stable isotopes (^{13}C, ^{15}N, and ^{2}H) in humans. Most studies utilize simplified models; the simplest and most widely used (**Figure 2**) is based on the measurement of the amino acid flux through the plasma amino acid pool. An example is the primed continuous intravenous infusion of ^{13}C-1-labeled leucine. During the infusion, the tracer isotopic enrichment is diluted by unlabeled amino acid from proteolysis (D) and the diet (I). At isotopic equilibrium constant labeling of the tracee is achieved, the magnitude of which (tracer/tracee ratio), in relation to the infusion rate (i), indicates the flux (Q), which is the total entry or exit rate of leucine through the pool. With the free leucine pool relatively small and turning over rapidly, isotopic equilibrium can be reached in 2–4 h if a priming dose is given as a bolus injection at the start of the infusion.

At isotopic and metabolic equilibrium, rates of entry and exit from the free leucine pool are equal for both labeled and unlabeled leucine so that Q is the rate of appearance or irreversible loss. Appearance is partitioned into dietary intake that is known (I) and entry from proteolysis of body protein (D) (i.e., no *de novo* synthesis of leucine occurs). Loss is partitioned into protein synthesis (S) and oxidative catabolism (O) to CO_2 and urinary N. The rate of leucine oxidation (O) is calculated from measurement of the production of labeled $^{13}CO_2$ in the breath and the labeling of the leucine or its keto acid in the plasma. This allows the components of protein turnover (D and S) to be calculated. Using the leucine content of tissue proteins, rates of leucine appearance and loss can be converted into rates of whole body protein synthesis and proteolysis.

^{13}C-1 leucine is especially useful since it is enables D to be calculated; it has a small pool enabling equilibrium to be achieved in a short period; decarboxylation is the first irreversible step in its catabolism, releasing $^{13}CO_2$ quantitatively; and its transamination product, α-ketoisocaproate, appears in the plasma and can serve as a measure of the labeling of the intracellular pool.

The latter advantage of leucine is especially important since determining isotopic enrichment of the precursor amino acid pool for protein synthesis is the most serious problem in these studies. Thus, amino acid pools are compartmentalized in the

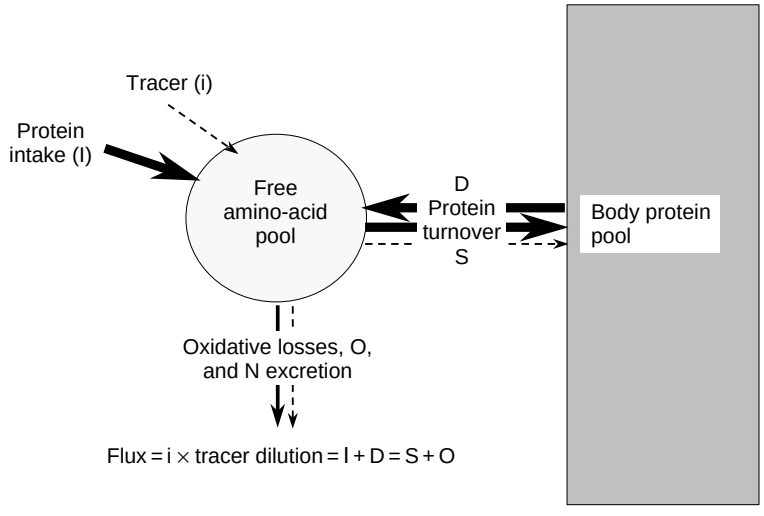


Figure 2 Single pool model for study of amino acid flux and protein turnover by tracer-labeled amino acid infusion. Movement of tracer is shown as dotted lines.

body, and the isotopic enrichment of tRNA-bound amino acids is lower than the measurable extracellular amino acid pool. Because it is difficult to measure labeling of amino acyl tRNA, a variety of indirect approaches have been used in an attempt to circumvent the problem. Equilibrium labeling of apolipoprotein B-100 has also been used to measure isotopic enrichment of hepatic amino acids. Thus, because apo B-100 turns over with a half-life of less than 1 h, during an infusion of several hours the protein labeling will reach a plateau representative of the hepatic precursor pool, and this has indicated a complex relationship between plasma and precursor enrichments for phenylalanine and leucine.

Flux rate measurements define whole body rates of protein synthesis and proteolysis, and each measurement is subject to error associated with the precursor assumption. Alternative methods have attempted to measure protein synthesis and proteolysis separately. In animal studies, a "flooding large dose" measurement of protein synthesis has been developed that enables all free pools to become equally labeled. Protein synthesis rates are calculated from measurement of isotope uptake into protein and free amino acid labeling during short periods after the dose. The method has been adapted for human use with a stable isotope, but there is evidence that the large quantities of the single amino acid stimulate protein synthesis.

Quantification of rates of proteolysis is especially problematic. In animal studies, proteolysis can be estimated from rates of synthesis and growth of tissue protein, and with careful design proteolysis rates can be measured over relatively short periods (e.g., 6 h following the administration of an endotoxin). Urinary excretion rates of 3-methyl histidine, a post-translationally modified amino acid not metabolized in the organism and excreted quantitatively in the urine, were proposed as a measure of myofibrillar protein degradation. Although the substantial contribution of small, rapidly turning over pools in microfilaments invalidates this approach for whole body studies, its release from incubated or perfused muscle can be used to determine myofibrillar protein degradation.

Simultaneous determination of protein synthesis and degradation can be made, in principle at least, from organ tracer balance studies (i.e., measurements of concentrations and isotopic enrichments of tracer amino acids across tissues such as the leg or forearm combined with measurements of 3-methyl histidine release). Such studies have identified a selective inhibitory effect of insulin on nonmyofibrillar protein degradation and stimulatory influences of amino acids on muscle protein synthesis.

All of these methods allow study of turnover of individual amino acids in protein and measurement of their nonprotein metabolic fate (e.g., oxidation). A different approach is to use ^{15}N glycine to study overall amino nitrogen turnover. Because of nitrogen exchange between amino acids by transamination, this label acts as a tracer for total free amino nitrogen rather than for any individual amino acid. The whole body nitrogen flux is estimated from the relative proportion of administered tracer excreted in the end product. This is then resolved into protein synthesis and proteolysis from measurements of N intake and excretion. The application of the method can be made simple by giving the ^{15}N label orally as a single dose. Although simple in concept, this approach is metabolically complicated with two urinary end products of nitrogen metabolism, urea and ammonia, each deriving from different pathways and each giving different flux values.

The choice of method must depend on the questions asked and circumstances of the subjects under study. ^{13}C carbon labeling is more suited to short-term (e.g., 3 or 4 h to 24 h) clinical measurements for which frequent blood and breath sampling is possible. Thus, the efficiency and mechanisms of postprandial protein utilization during meal feeding can be measured by means of ^{13}C leucine balance and turnover measurements. ^{15}N methods are more suitable for free-living subjects and patients, when urine sampling is possible but regular blood and breath sampling is inconvenient. The most famous example is the use of this method in an unassisted Antarctic crossing. Both methods involve many assumptions, but in practice the two approaches have been shown to give similar results.

Applications

Extent and Physiological Implications of Protein Turnover

In the human adult, approximately 300 g of protein turnover occurs each day (4 g/kg/day)—that is, three or four times the daily dietary intake. Rates vary between tissues with rapid turnover in visceral tissues and those with slow turnover in muscle. Liver and intestine account for approximately 8% of the lean body mass (LBM) and up to 50% of whole body protein turnover, with skeletal muscle, at 55% of the LBM, accounting for only approximately 25% of total protein turnover. Thus, whole body protein turnover varies with body composition and this largely explains developmental

changes. In the infant, turnover rates are much higher, ranging between 10 and 20 g protein turnover per kilogram per day, consistent with the higher proportion of metabolically active tissue and lower muscle mass. However, animal studies indicate a developmental decline in protein turnover in skeletal muscle that may be an additional component of the marked decline in protein turnover with age. In the elderly there is little evidence of any change other than that associated with the decline in LBM.

Protein turnover constitutes an appreciable fraction of the maintenance energy expenditure. On the basis of 5 mol ATP/GTP per mole of protein turnover (4 mol per peptide bond with 1 mol for amino acid transport, RNA turnover, and proteolysis), an energy cost of 22 kcal/mol ATP, and a molecular weight of 110 per mole of peptide bond, this is equivalent to approximately 1 kcal/g protein turnover. Thus, in the normal adult, protein turnover at 300 g/day accounts for approximately 20% of the basal metabolic rate. Therefore, changes in the protein turnover and metabolic rate would be expected to occur in parallel to some extent, and this is observed. Thus, the decline with age in both protein turnover and metabolic rate from birth to adulthood involves a factor of three or four in each case.

Regarding protein turnover and protein requirements, there is no a priori reason for any interrelationship, and there is little evidence of any. Thus, turnover does not consume amino acids, and amino acid catabolism and oxidation is not linked to turnover. Maintenance protein requirements decrease relatively little with age (<20%) compared with the 3- or 4-fold decrease in turnover.

Regulatory Mechanisms of Protein Turnover Control

The physiological importance of protein turnover is undoubtedly the regulatory flexibility it allows. With opportunities for control of both synthesis and proteolysis, the number of potential control sites is increased. In addition, because of the continuing turnover in the steady state, changes in the amount of protein can be achieved with low energy costs through inhibition of proteolysis to allow growth or through inhibition of synthesis to allow mobilization.

At the molecular level, regulation of protein synthesis is necessarily complex at both transcriptional and translational levels. Advances in molecular biology have revealed many examples of transcriptional control to the extent that changes in specific mRNA concentrations have become a surrogate measure of changes in rates of synthesis for specific proteins. Notable nutritional examples include control of hepatic export protein synthesis. Thus, the downregulation of albumin synthesis in response to either protein deficiency or the proinflammatory cytokine-mediated acute phase response is mediated largely at the level of transcriptional control of mRNA levels, with reductions in mRNA for albumin and other hepatic export proteins and increases in mRNA for acute phase proteins.

The concentration of ribosomes in tissues determines the capacity for protein synthesis and in this way controls overall tissue protein turnover rate and the changes associated with postnatal development. Cellular ribosome concentrations can change both acutely (e.g., during the diurnal cycle of feeding and fasting) and chronically in response to protein and energy intakes, increased functional demand, and hormones such as insulin, thyroid, growth hormone, and the glucocorticoids. Furthermore, these influences are tissue specific, with glucocorticoids, for example, increasing hepatic ribosome concentrations (as part of the hepatic acute phase response) and decreasing ribosome concentrations in muscle. In contrast, thyroid hormones increase ribosomes (and proteolytic enzymes) in both muscle and liver in association with a generalized increase in protein turnover.

Acute regulation of translation is exerted mainly through initiation, with reversible phosphorylations known to regulate at least four separate steps of the initiation cycle enabling very rapid changes in protein synthesis. Peptide hormones (insulin and insulin-like growth factor-1 (IGF-1)), glucocorticoids, and amino acids have all been implicated in such regulation, although the specific targets of control remain uncertain. Furthermore, there are major differences between the mechanisms observed in the young, rapidly growing animal and the adult animal. Thus, in skeletal muscle in the young rat, an insulin-mediated stimulation occurs. In adult human, muscle insulin is relatively ineffective, with amino acid levels the main stimulatory influence. Indeed, because insulin inhibits proteolysis and lowers amino acid levels, insulin alone appears to inhibit protein synthesis in human muscle.

Regarding the nutritional regulation of proteolysis, most is known about lysosomal proteolysis, especially hepatic autophagy, with both amino acids and insulin having inhibitory roles. Leucine, alanine, and insulin interact to regulate this pathway, with a leucine-sensitive receptor-mediated inhibitory pathway identified in liver. In the case of the ubiquitin–proteasome system, its activation in

skeletal muscle during fasting and following glucocorticoid treatment supports a role in the physiological regulation of protein turnover. On the other hand, both lysosomal and calcium-activated proteolysis are activated under the same conditions. Similarly, in response to protein deficiency when protein turnover rates generally decrease in tissues, in part through the decline in thyroid hormone levels, the activities of all three systems decrease. One control mechanism involves changes in cell volume. Thus, swelling acts like a proliferative anabolic signal, inhibiting proteolysis, whereas cell shrinkage is catabolic, stimulating proteolysis. These effects have been shown in liver, and there is evidence for such a mechanism in skeletal muscle.

Postprandial Protein Utilization

Overall nitrogen homeostasis within the LBM is maintained within a diurnal cycle of postprandial protein gain and postabsorptive loss. The amplitude of these diurnal changes increases as habitual protein intakes increase, with implications for the qualitative nature of the metabolic demands for amino acids and hence dietary protein. The key questions are how both acute and chronic protein intakes mediate such responses and, most important, what influences the efficiency of postprandial protein utilization and consequent protein requirements. Nutritional regulation of dietary protein utilization, protein turnover, and amino acid oxidation involve both hormonal responses to food intakes and direct substrate influences. Whereas interactions between insulin, thyroid hormones, and IGF-1 mediate the anabolic drive of dietary protein on muscle and bone growth in the growing animal, the control mechanisms involved in the transient gains and losses of protein during diurnal cycling differ since neither thyroid hormones nor IGF-1 levels vary from meal to meal or in relation to habitual protein intakes. On the basis of several studies on either insulin or amino acids alone or variations in meal protein levels, it appears that insulin and amino acids act as main acute regulators.

The mechanisms involved are best understood in the context of the interrelationships between the free and protein-bound amino acid pools. Many indispensable amino acids are potentially toxic and are maintained at very low concentrations in tissues, so rapid and regulated postprandial disposal is important. After a protein meal there are two pathways for amino acid disposal. The first comprises the various high-capacity, finely regulated oxidative pathways activated by a protein meal. In most cases, rates of amino acid oxidation are influenced by their tissue concentrations (generally similar to the K_m of the rate-limiting enzymes), together with substrate activation and covalent enzyme modification. Examples are phenylalanine hydroxylase and branched-chain α-keto acid dehydrogenase, which are both regulated by substrate binding and reversible phosphorylation and dephosphorylation. The second pathway is net protein deposition. This can be achieved by stimulation of protein synthesis or inhibition of proteolysis so that a regulatory link between postprandial hyper-amino acidemia and protein synthesis and proteolysis can be expected and does indeed exist.

For protein synthesis, amino acids cannot exert simple kinetic concentration-related influences since the low K_m of amino acyl tRNAs synthesis means that they are usually fully charged. Nevertheless, there is ample evidence for regulatory stimulation by amino acids. This remains poorly understood but in some cells is known to involve signaling events linked to the mammalian target of rapamycin, mTOR, which in turn regulates S6 (an initiation factor) as well as eEF2 (an elongation factor). However, stimulation of protein synthesis through increased amino acid levels may also stimulate amino acid oxidation pathways, as discussed previously. Although this allows effective removal of amino acids, in the context of an efficient protein utilization this would not be a preferred mechanism.

For proteolysis, amino acids exert an inhibitory influence as described previously, and this inhibition will reduce endogenous amino acid supply. This will prevent undue increases in amino acid levels and will therefore minimize amino acid oxidation and maximize dietary protein utilisation. Furthermore, since inhibition of proteolysis and lowering of intracellular amino acid levels can be achieved by receptor-mediated mechanisms involving insulin as well as specific amino acids (e.g., leucine), this allows the postprandial increases in plasma amino acids to mediate substantial amino acid transport into cells, resulting in protein deposition without any increase in intracellular amino acid levels and with minimal increases in amino acid oxidation. Thus, as a strategy for mediating postprandial protein utilisation, inhibition of proteolysis is predicted to be more efficient.

^{13}C leucine studies have provided clear experimental support for such a mechanism. The meal protein-dependent responses of protein synthesis, proteolysis, and leucine oxidation are shown in **Figure 3** based on measurements in adult subjects fed isoenergetic meals of increasing protein intake

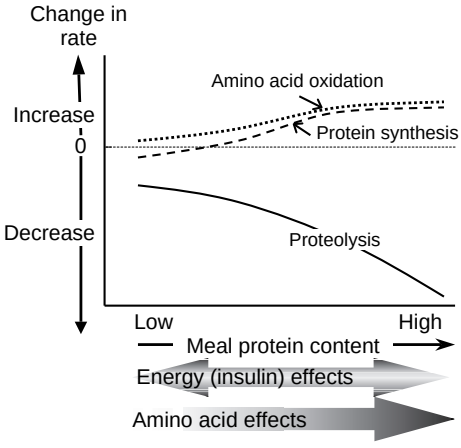


Figure 3 Feeding-induced responses of leucine kinetics shown as protein synthesis, proteolysis, and leucine oxidation. Patterns of responses reflect actual changes observed in ^{13}C-1 leucine infusion studies of feeding responses to frequent small meals of increasing protein intake, equivalent to daily intakes between 0.3 and 2.0 g kg^{-1} day^{-1}. (Adapted from Pacy PJ, Price GM, Halliday D *et al.* (1994) Nitrogen homeostasis in man: 2. The diurnal responses of protein synthesis, degradation and amino acid oxidation to diets with increasing protein intakes. *Clinical Science* **86**:103–118; and Gibson NR, Fereday A, Cox M *et al.* (1996) Influences of dietary energy and protein on leucine kinetics during feeding in healthy adults. *American Journal of Physiology* **33**: 282–291.)

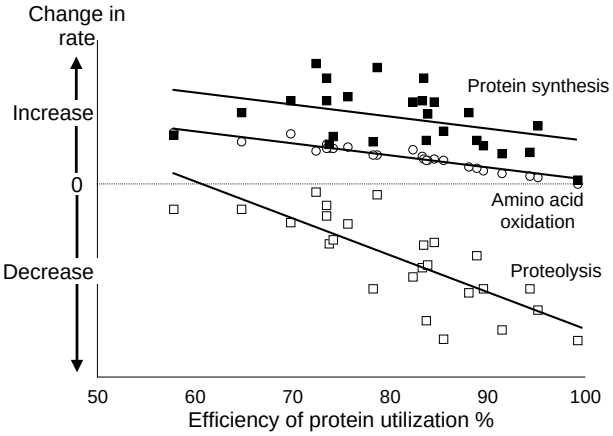


Figure 4 Relationships between the efficiency of protein utilization (Δbalance/Δintake) and responses of leucine turnover and oxidation to protein feeding observed in 24 normal adults fed frequent small meals containing protein intakes similar to the habitual intakes of the subjects. (Adapted from Fereday A, Gibson NR, Cox M *et al.* (1998) Variation in amino acid mediated, insulin activated inhibition of proteolysis determines the efficiency of protein utilization. *Clinical Science* **95**: 725–733.)

from 0.36 to 2.07 g protein/kg/day. Inhibition of proteolysis occurs at all levels of protein intake, but it increases with intake. However, the direction and magnitude of the response of protein synthesis reflect the level of dietary protein intake, with slight inhibition or no change at low intakes and stimulation at high intakes. Such studies clearly establish the importance of proteolysis as a regulator of tissue protein balance in the postabsorbtive and postprandial state. These and other ^{13}C leucine studies of postprandial protein utilization have allowed the separate influences of dietary energy and amino acids protein to be identified as shown in **Figure 3**. The response to energy alone involves insulin-mediated changes allowing leucine balance to becomes less negative through inhibition of proteolysis with minimal changes in protein synthesis or amino acid oxidation. In fact, because this tends to lower amino acid levels, there is a decrease in protein synthesis. With increasing amino acid supply as dietary protein intake increases, there is further inhibition of proteolysis by amino acids with increases in protein synthesis and, to some extent, amino acid oxidation, allowing net protein deposition as tissue protein.

The increase in amino acid oxidation with protein feeding is an unwanted response that reduces the efficiency of protein utilization. Although utilization of proteins such as milk is higher than that of wheat, as would be expected because of the lysine limitation of wheat gluten utilization, there is variability between individuals in the efficiency of postprandial protein utilization, ranging from 50 to 100% with milk protein. **Figure 4** shows the results of studies that have examined this variation. Efficient utilization involves maximal inhibition of proteolysis by protein feeding with minimal increases in free amino acid concentrations and consequent amino acid oxidation and stimulation of protein synthesis, indicating that the efficiency of protein utilization in individuals is determined by the sensitivity of the insulin-mediated inhibition of proteolysis to amino acid supply.

Current understanding suggests a mechanism indicated in **Figure 5**, in which the major target of insulin is inhibition of proteolysis, with amino acids acting to both enhance the inhibition of proteolysis and stimulate synthesis and oxidation. With tissue amino acid levels controlled by both diet and endogenous supply from proteolysis, inhibition of proteolysis will minimize any increase in amino acid levels, minimize oxidation, and maximize protein utilization. Since protein synthesis and amino acid oxidation appear to be stimulated in parallel, the optimum strategy for maximum efficiency of postprandial protein utilization appears to involve maximal inhibition of proteolysis and minimal postprandial increases in tissue amino acid levels.

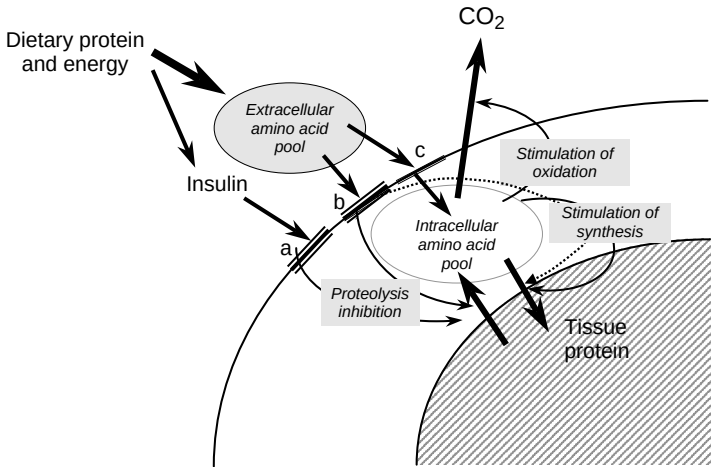


Figure 5 Scheme for the action of insulin and amino acid supply on postprandial protein utilization. Insulin and extracellular amino acids exert inhibitory influences on proteolysis and protein synthesis through receptor-mediated mechanisms (a and b), whereas amino acid uptake (c) and proteolysis regulate intracellular amino acid levels, amino acid oxidation, and protein synthesis in parallel. Maximal inhibition of proteolysis and maintenance of low intracellular amino acid levels is the optimal response. (Modified from Millward DJ, Fereday A, Gibson NR *et al.* (1996) Postprandial protein metabolism. *Baillier's Clinical Endocrinology and Metabolism* **10**: 533–549.)

In summary, postprandial protein utilization appears to be mediated by an insulin-mediated, protein-conserving influence of dietary energy that inhibits proteolysis, lowers amino acid levels, and reduces oxidation, with dietary amino acids augmenting the inhibition of proteolysis. The response of protein synthesis is primarily determined by the resultant intracellular amino acid levels that reflect the balance between the decreasing endogenous supply following insulin-mediated inhibition of proteolysis and the increasing exogenous supply as dietary protein intake increases, stimulating protein synthesis and increasing oxidation when amino acid dietary supply exceeds the capacity for its net deposition.

See also: **Amino Acids**: Metabolism. **Protein**: Requirements and Role in Diet; Digestion and Bioavailability; Quality and Sources.

Further Reading

Fereday A, Gibson NR, Cox M *et al.* (1998) Variation in amino acid mediated, insulin activated inhibition of proteolysis determines the efficiency of protein utilization. *Clinical Science* **95**: 725–733.

Gibson NR, Fereday A, Cox M *et al.* (1996) Influences of dietary energy and protein on leucine kinetics during feeding in healthy adults. *American Journal of Physiology* **33**: 282–291.

Haussinger D and Lang F (1992) Cell volume and hormone action. *Trends in Pharmacological Science* **13**: 371–373.

Jagoe RT and Goldberg AL (2001) What do we really know about the ubiquitin–proteasome pathway in muscle atrophy? *Current Opinion in Clinical Nutrition and Metabolic Care* **4**: 183–190.

Millward DJ (1995) Insulin and the regulation of amino acid catabolism and protein turnover. In Cynober L (ed.) *Amino Acid Metabolism in Health and Disease*, pp. 127–136. Boca Raton, FL: CRC Press.

Millward DJ (1995) A protein-stat mechanism for regulation of growth and maintenance of the lean-body mass. *Nutrition Research Reviews* **8**: 93–120.

Millward DJ, Fereday A, Gibson NR *et al.* (1996) Postprandial protein metabolism. *Baillier's Clinical Endocrinology and Metabolism* **10**: 533–549.

Millward DJ, Fereday A, Gibson NR *et al.* (2000) Human adult protein and amino acid requirements: [^{13}C-1] leucine balance evaluation of the efficiency of utilization and apparent requirements for wheat protein and lysine compared with milk protein in healthy adults. *American Journal of Clinical Nutrition* **72**: 112–121.

Millward DJ and Rivers JPW (1989) The need for indispensable amino acids: The concept of the anabolic drive. *Diabetes Metabolism Reviews* **5**: 191–212.

Miotto G, Venerando R, Marin O *et al.* (1994) Inhibition of macroautophagy and proteolysis in the isolated rat hepatocyte by a non-transportable derivative of the multiple antigen peptide Leu8-Lys4-Lys2-Lys-betaAla. *Journal of Biological Chemistry* **269**: 25348–25353.

Pacy PJ, Price GM, Halliday D *et al.* (1994) Nitrogen homeostasis in man: 2. The diurnal responses of protein synthesis, degradation and amino acid oxidation to diets with increasing protein intakes. *Clinical Science* **86**: 103–118.

Waterlow JC, Garlick PJ, and Millward DJ (1978) *Protein Turnover in Mammalian Tissues and in the Whole Body* Amsterdam: Elsevier/North-Holland Biomedical Press.

Requirements and Role in Diet

D J Millward, University of Surrey, Guildford, UK

Defining minimum amino acid and protein requirements is inherently difficult. Humans are exposed to a wide range of protein intakes, which enable full expression of their genotypical lean body mass throughout the range, and identifying the lower limits of this range has proved intractable. Without unequivocal symptoms of deficiency, the adequacy of an intake can only be assessed in terms of nitrogen or amino acid balance, which is unsatisfactory for several reasons. In particular, adaptation causes major difficulties in designing balance studies and interpreting results. Furthermore, balance methods are inherently imprecise and logistically extremely difficult. It is therefore not surprising that there is much debate about both the nature and the extent of protein requirements.

Terminology

Protein requirements are best discussed in terms of metabolic demand, dietary requirement, and dietary allowances. Metabolic demand concerns amino acids and is determined by the nature and extent of those metabolic pathways (e.g., net protein synthesis) that consume amino acids and that vary with the phenotype and the developmental and physiological state of the individual. The dietary requirement is the amount of protein and/or its constituent amino acids that must be supplied in the diet in order to satisfy the metabolic demand. The requirement will usually be greater than the metabolic demand. Thus, factors associated with digestion and absorption may limit digestibility (i.e., dietary nitrogen lost in the feces) and biological value (i.e., the availability of the absorbed amino acid pattern in relation to cellular needs, which influences urinary nitrogen excretion). Dietary allowances are a range of intakes derived from estimates of individual requirements taking into account variability between individuals. They are designed to meet the dietary requirements of the population. In the United Kingdom, these allowances are described in terms of Dietary Reference Values (DRVs) and in the United States as Dietary Reference Intakes (DRIs).

Metabolic Demands for Amino Acids

Current evidence supports the representation of the metabolic demands as in **Figure 1**. The metabolic demand for amino acids is to maintain tissue protein at appropriate levels and to provide for all amino acid-derived metabolites and any additional needs during growth, rehabilitation, pregnancy, and lactation. Tissue proteins are diverse, including structural or fibrous insoluble types and soluble globular

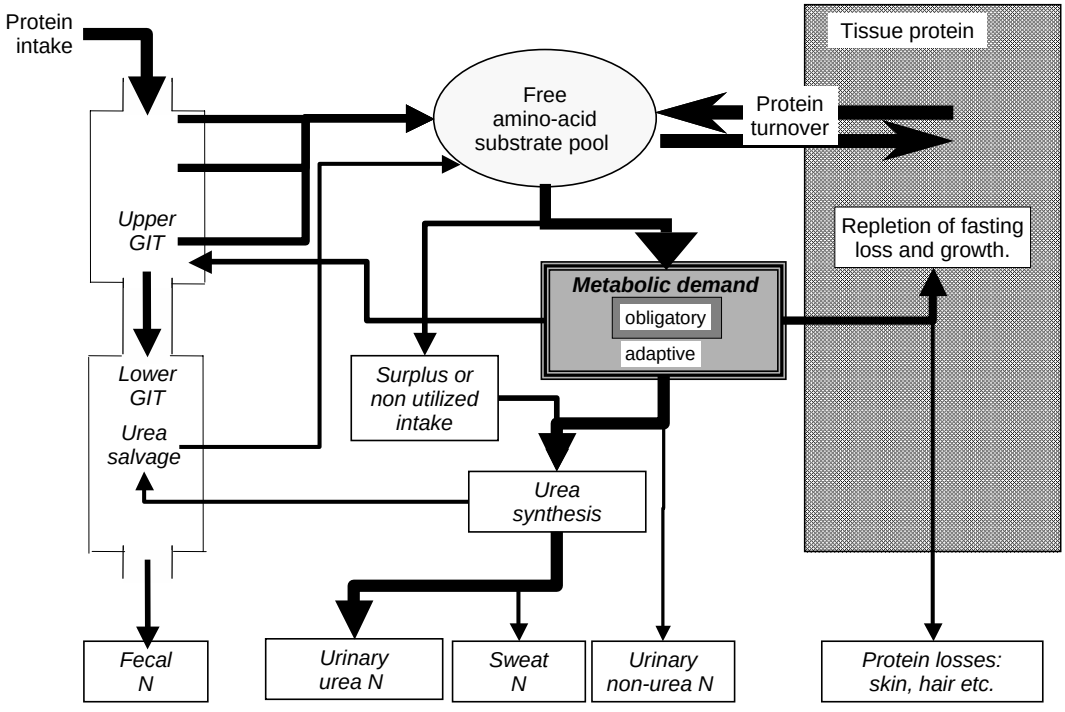


Figure 1 Schematic representation of the metabolic demands for amino acids.

species, with characteristic properties and functions that are determined by their amino acid sequence. All proteins are in a dynamic state of constant turnover (i.e., breakdown to constituent amino acids and resynthesis), although for the structural proteins this is slow or minimal. Nonprotein products include nucleic acids and a diverse range of smaller molecules, such as creatine, taurine, glutathione, hormones (e.g., catecholamines and thyroxine), neurotransmitters (serotonin and dopamine), and nitric oxide, a key regulator of blood flow and other physiological processes.

The metabolic demand is supplied from the free amino acid pool, the size of which, for most amino acids, is regulated within narrow limits. Regulation involves supply from three sources: dietary proteins after digestion and absorption from the upper gastrointestinal tract (GIT), tissue protein after proteolysis during protein turnover, and de novo formation, which may include amino acids and ammonia, deriving from urea salvage after hydrolysis and bacterial metabolism in the large bowel. Removal of free amino acids occurs by reactions in which they act as substrates, and these reactions are shown as three pathways, one of which is the metabolic demand. This pathway involves a number of irreversible pathways, including net protein synthesis and other irreversible metabolic transformations of individual amino acids. The second and quantitatively largest pathway is the removal for protein synthesis during protein turnover. At nitrogen equilibrium, because turnover involves the reversible removal of amino acids, with replacement through proteolysis, it does not exert a net metabolic demand (other than for those amino acids irreversibly modified during or subsequent to protein synthesis). The third pathway is the irreversible removal of amino acids by oxidation and nitrogen excretion provoked, for example, by the transient increases in some or all free amino acids after a protein meal. This would represent an inefficient utilization.

The metabolic demand for amino acids appears to involve obligatory and adaptive components. The obligatory component for subjects at equilibrium (i.e., maintenance) comprises conversion of some individual amino acids into important metabolites that are further transformed into nitrogenous end products, mainly urea and other compounds in urine, feces, or sweat, as well as net synthesis of proteins lost from the body as skin, hair, and any other secretions. These diverse biological demands for amino acids for maintenance represent an essential but probably quite small intrinsic metabolic demand for protein. The magnitude of this maintenance component is assumed empirically to be equal to the obligatory nitrogen loss (ONL)—that is, the sum of all nitrogen losses from the body observed in subjects fed a protein-free but otherwise nutritionally adequate diet after 7–14 days, by which time nitrogen losses have declined to a stable and reproducible low level with the subjects losing body protein at a constant daily rate. In normal adults, the obligatory urinary, fecal, and subcutaneous and other losses are approximately 29, 13, and 5 mg N/kg, respectively (i.e., 47 mg/kg/day), in total equivalent to 0.29 g protein/kg/day tissue protein mobilized to meet such demands. The ONL is a function of body weight and, as far as is known, varies little with age. After adaptation to a protein-free diet, net tissue proteolysis is assumed to provide for the nonprotein components of the obligatory demand at a rate determined by the metabolic consumption of the rate-limiting amino acid (the amino acid with the highest ratio of molar proportion in the metabolic demand to molar proportion in protein). Because the obligatory metabolic demand is for a mixture of amino acids with a profile that is unlikely to match that of tissue protein, the actual nitrogen content of the metabolic demand is likely to be less than that indicated by the ONL (i.e., less than an equivalent of 0.29 g protein/kg/day). This is because all amino acids mobilised to provide for the metabolic demand must be oxidized and will contribution to the nitrogen excretion, whereas only some of them will serve useful functions. The evidence for this is the lowering of the ONL in response to feeding selective amino acids, such as threonine, tryptophan, and methionine. In addition to these metabolic demands for maintenance, any net protein synthesis associated with growth, pregnancy, and lactation also constitutes an obligatory metabolic demand.

The adaptive component of the metabolic demand represents amino acid oxidation at a rate varying with the habitual protein intake that occurs as a result of the increasing activities of the pathways of oxidation of amino acids that regulate free amino acid pool sizes. Although this aspect of amino acid metabolism is least understood, it is likely that it is a consequence of the fact that humans grow slowly or maintain constant weight on diets that contain protein considerably in excess of minimum needs. Thus, in order to be able to rapidly dispose of excess protein and maintain the very low tissue concentrations of those amino acids, such as the branched-chain, aromatic, and sulfur amino acids, that may be toxic at higher concentrations, pathways of oxidative amino acid catabolism adapt (increase their V_{max}), enabling them to operate at the appropriate rate set by habitual protein intakes. Importantly, the

adapted rate of amino acid oxidation, characteristic of habitual intake, changes only slowly in response to either a change in dietary protein intake level or feeding and fasting. This has two main consequences. First, when intake falls below habitual intake mobilization of tissue protein occurs with a negative nitrogen balance for as long as it takes to adapt to the lower level of intake. This was previously identified as the labile protein reserves. It can be assumed that for intakes greater than the minimum requirement, full adaptation to the new level will include not only a change in the adaptive metabolic demand to match intake but also repletion of most tissue nitrogen lost during the adaptive transition. Second, because the adaptive rate of amino acid oxidation continues to some extent into the postabsorptive state, there are varying postabsorptive losses of tissue protein and nitrogen excretion with varying habitual intake—that is, a diurnal cycle of postabsorptive losses and postprandial gains with an amplitude that increases with the increasing habitual level of protein intake as shown in **Figure 2**. As such, the adaptive metabolic demand includes a component of net protein synthesis that repletes postabsorptive losses. The magnitude of this varies in a complex way with eating pattern and with the amount and quality (amino acid score) of the habitual protein intake.

Although amino acid oxidation and urea synthesis is assumed to be irreversible, this is not entirely true because of urea salvage. Thus, the rate of urea synthesis is usually in excess of the rate of urea excretion because some urea enters the large bowel and is hydrolysed by bacteria. Most of this nitrogen is utilized by bacteria, and since little is lost as fecal nitrogen, it is eventually returned to the systemic pool as ammonia and amino acids after bacterial death and proteolysis, including indispensable amino acids. Although the extent and nature of this salvaged urea nitrogen are poorly understood, it may provide nutritionally important amounts of amino acids.

The main practical implications of the previously discussed model are that true minimum metabolic demands and consequent protein requirements will occur when the adaptive metabolic demand has fallen to the lowest possible level, and it is not known with any certainty how long such adaptation would take. However, studies that have examined balance responses to changes in protein intakes suggest it is likely to be longer than the periods employed in short-term balance studies. This implies that short-term balances from which our estimates of the minimal protein requirement (MPR) derive may overestimate the value and also that some of the variability in protein requirements between studies may reflect variable completeness of adaptation to the test diets. Another implication of the adaptive metabolic demands model is that intakes and requirements are correlated, which has implications for the definition of risk of deficiency and safe intakes.

Protein Requirements

Plant versus Animal Sources

The nutritional requirement for protein will be the minimum intake that satisfies metabolic demands and that maintains appropriate body composition and growth rates, after taking into account any inefficiency of digestibility and metabolic

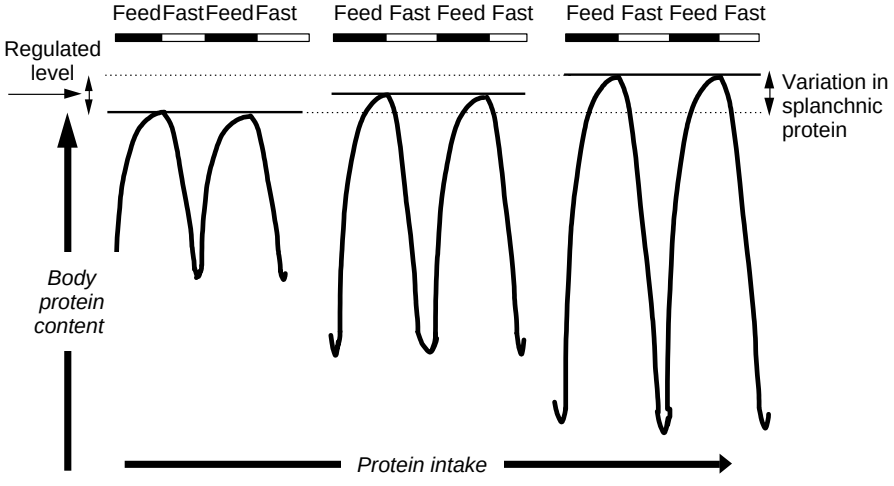


Figure 2 Balance regulation throughout the diurnal cycle with increasing habitual protein intakes.

consumption. With continuous and extensive amino acid interconversion, the pattern of dietary amino acids need not match that of the composition of tissue proteins or the maintenance metabolic demand exactly because some amino acids (aspartic acid, asparagine, glutamic acid, alanine, and serine) are dispensable and can be replaced by sufficient total amino acid nitrogen supplied from other amino acids or sources of nonessential nitrogen. However, there will be a minimum dietary requirement for those amino acid that are not interconverted, classified as indispensable amino acids (histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine), and for those that are formed only slowly from other amino acids and become indispensable under specific physiological or pathological conditions (conditionally indispensable; e.g., cysteine, tyrosine, taurine, glycine, arginine, glutamine, and proline). Traditionally, dietary proteins have been classified by their nutritional value (quality) measured in terms of their ability to provide for tissue growth in rapidly growing rats. In this case, marked differences are observed between most animal protein and plant protein sources, with relative nutritional value reflecting mainly relative amounts of specific indispensable amino acids. The similarity between overall tissue protein amino acid composition and that of most animal dietary protein sources and the contrast with plant protein sources resulted in clear distinctions between their quality, although when combined it is clear that plant proteins can provide the appropriate balance of essential amino acids.

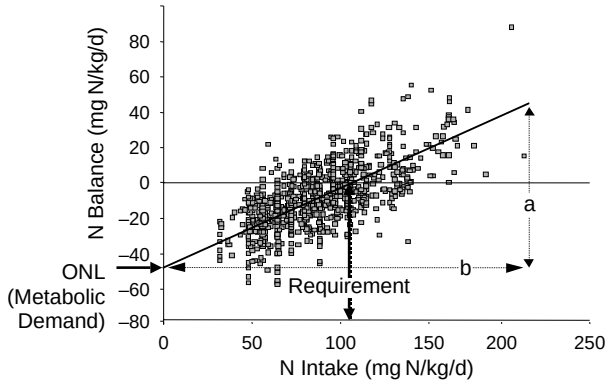
However, in human nutrition with growth occurring very slowly after the first few months of life, the nutritional demand for indispensable amino acids for tissue growth is much less and may be minimal. Little metabolic demand for amino acids is generated by protein turnover because of amino acid recycling. Some net protein synthesis is associated with skin and hair growth and with gastric secretions (e.g., threonine-rich mucus glycoproteins) that pass into the colon to be utilized for bacterial metabolism. The metabolic demand for maintenance of normal function and composition is a poorly understood pattern of amino acids utilized in the various metabolic pathways other than protein synthesis, but this pattern is almost certainly different from that required for growth (i.e., mainly the amino acid pattern of tissue protein) and may contain a much lower overall amount of indispensable amino acids. As such, a distinction between plant and animal dietary protein sources is much more difficult to demonstrate and is probably less relevant in human nutrition. Currently, there is considerable controversy regarding the magnitude of the requirements for indispensable amino acids in the human diet and there are different views about the relative importance of dietary protein quality in human nutrition. Some national bodies have stressed that in most mixed, nutritionally balanced diets, sufficient indispensable amino acids will be provided regardless of the relative amounts of plant or animal protein sources (e.g., UK Department of Health). However, others (e.g., US FNB/IOM) have argued that the requirements for indispensable amino acids in the human diet are higher than previously believed and that protein quality is important.

Nitrogen Balance

Nitrogen balance studies were initiated in the mid-nineteenth century by Carl Voit, and such studies have been central to the definition of protein requirements. The aim of nitrogen balance studies is simple—to define the relationship between intake and all losses (urinary, fecal, and surface—mainly sweat, skin, hair, breath ammonia, nail clippings, etc.) so that the intake that allows equilibrium and provides for all losses can be identified. Thus, when the intake equals the requirement,

$$\text{Balance} = \text{intake} - \text{losses} = 0$$

As indicated previously, the lowest level of losses observed, the ONL, is approximately 46 mg/kg/day, equivalent to a daily loss of 0.29 g protein/kg/day. When such subjects are re-fed with protein, losses of body protein decrease as the dietary protein provides for some of the metabolic demand. However, nitrogen losses increase with intake so that the required intake for balance is more than the ONL. The main objective of nitrogen balance studies has been to define how much extra protein above the ONL must be fed to achieve equilibrium. The literature on human nitrogen balance studies has been assessed in a meta-analysis, as shown in **Figure 3**. A linear regression of balance against intake will allow prediction of the ONL as the zero intake intercept. The slope of the balance curve ($a/b = e$) will indicate the efficiency of utilization, and the maintenance requirement (i.e., the amount that must be fed to balance all losses and produce equilibrium) is ONL/e. Thus, the currently accepted maintenance requirement (0.66 g/kg/day) derives from an analysis of the data shown in **Figure 2** (the median requirement calculated from individual regressions on each individual studied at more than three levels of protein intake).



Balance = e × intake-ONL
(e = efficiency of utilisation = Δbalance(a)/Δintake(b))

Balance = 0.47.intake – 48.1 mg N/kg/d
Requirement = intake for zero balance = ONL/e
= 105 mg N/kg/d = 0.66 g protein/kg/d

Figure 3 Meta-analysis of nitrogen balance studies.

Inherent Difficulties with Nitrogen Balance Studies

This apparently simple but laborious approach, which is currently the main method for investigating protein and amino acid requirements, is in fact beset with a large number of quite serious problems, as listed in **Table 1**.

The lack of precision results in balance being a small value compared with the much larger values of nitrogen intake and nitrogen excretion, resulting in considerable error. The various systematic errors mean that balance is usually overestimated, often with unrealistic positive balances (protein gains) at high intakes. The nonlinearity of the balance curve as losses increase to match intakes when body protein reaches the maximum level means that there is no simple term to define the overall shape of the balance curve allowing prediction of the requirement (as the zero balance intake). In practice, prediction of a zero balance–intake intercept is made from a few balance points by linear regression, and this will result in requirement values that will vary according to where the intake values lie on the balance curve; that is, studies conducted using low intakes will underestimate requirements, whereas studies conducted with supramaintenance intakes will overestimate requirements. The logic of this is that (i) reliable balance studies are those that are conducted with intakes very close to the actual requirement, and (ii) studies with intakes based on preconceived requirement values will tend to confirm such preconceptions.

Table 1 Potential problems relating to nitrogen balance methodology

Imprecise
Systematic errors: intake overestimated, loss underestimated
due to problem of accounting for all losses, for example,
Skin surface and secretions
Loss of N_2 gas
Expired ammonia
Endogenous NO production gives urinary nitrate, faecal ammonia, and nitrite
Changing size of the body urea pool
Nonlinearity of the balance curve
Design
Dietary energy intake and physical activity influences balance
Accounting for adaptation

Energy–Protein Interactions

Body protein equilibrium can be influenced by intakes of energy, and ensuring that energy intakes are sufficient is difficult. Excess energy intake leads to weight and some lean tissue gain, whereas with too little intake, dietary and/or body protein is oxidised as an energy source. This means that the protein requirement is a function of the state of energy balance and the actual influence is quite marked. According to one analysis of nitrogen balance (NB) on intakes of energy (EI; kcal/kg) and N (NI; mg N/kg), $NB = 0.171NI + 1.006EI - 69.13$. This means that the intake for N equilibrium (the requirement) will vary from 1.4 to 0.32 g/kg/day according to whether energy intakes are equal to the resting metabolic rate (RMR) or equal to twice the RMR. In fact, two–thirds of the overall variability reported in the meta–analysis shown in **Figure 2** (SD = 31.9 mg N/kg) could be accounted for by an error of only approximately 0.2 of basal metabolic rate (BMR) in estimating the true energy needs of a subject.

Since nitrogen balance varies as a function of energy intake, it may be argued that protein requirements can only be defined in terms of a specified energy intake level, but what is the appropriate energy intake? Should populations with low protein staples consume more energy to achieve body protein equilibrium? Will this predispose to obesity? To what extent does variation in energy intakes at energy balance (i.e., with increasing levels of physical activity) influence nitrogen balance? These are difficult and currently unanswered questions.

Adaptation

With the metabolic demand for amino acids including both fixed and variable demands, the relationship between intakes and balance will be a function of time and the rate of adaptation. This is undoubtedly why the determination of human protein requirement by nitrogen balance has proved to be so difficult. Thus, when protein intake changes, the

metabolic adjustments involved with matching amino acid oxidation and urea excretion rates to the new intakes take considerable time to adapt to the new level of intake. The actual time taken for complete adaptation is poorly understood and controversial. In practice, most balance studies are 'short term,' with dietary periods of 2 weeks at each intake studied and with diet periods randomized to minimize metabolic carryover of prior diets. Two weeks is comparable to the time taken to stabilize excretion in subjects fed a protein-free diet while establishing the magnitude of the obligatory nitrogen losses. It may be that adjustment to a protein-free diet, an extreme metabolic change, occurs more rapidly than the adjustment from one intake to another, with evidence of changes over several months to an intake similar to the ONL, and more than 1 month is required to adjust to a lower but adequate intake (1 g/kg/day) after 2 months of a high-protein diet of 3 g/kg/day.

Dietary Protein Allowances and Implications of Adaptation

The Estimated Average Requirement (EAR) defines the notional mean requirement for the population group. The Recommended Nutrient Intake (RNI) is defined according to the range of interindividual variability and is two standard deviations above the EAR. The RNI (or Recommended Dietary Allowance) is thus an intake that will meet the requirement of most of the population assuming normal distribution of requirements and is therefore a 'safe allowance.' The Lower Nutrient Reference Intake, which is two standard deviations below the EAR, defines the lowest intake that will meet the requirement of some of the population. It follows from these definitions that in deriving dietary allowances from nitrogen balance studies, the variability in the reported values is very important since this is used to set the RNI. The currently agreed value is based on an EAR of 0.66 g/kg and a SD = 12% (i.e., 0.82 g/kg). Such calculations try to assess true between-subject variation rather than measurement errors. The range of individual values from the reported nitrogen balance meta-analysis is shown in **Figure 4**. Analysis of individual risk of deficiency (intake < requirement) assumes that the requirement is not correlated with the intake so that for an individual with an intake equal to the mean requirement value, the risk of deficiency is 50%, declining to less than 2.5% at the higher intake equal to the RNI.

The serious implication of lack of complete adaptation in short-term multilevel balance studies is that

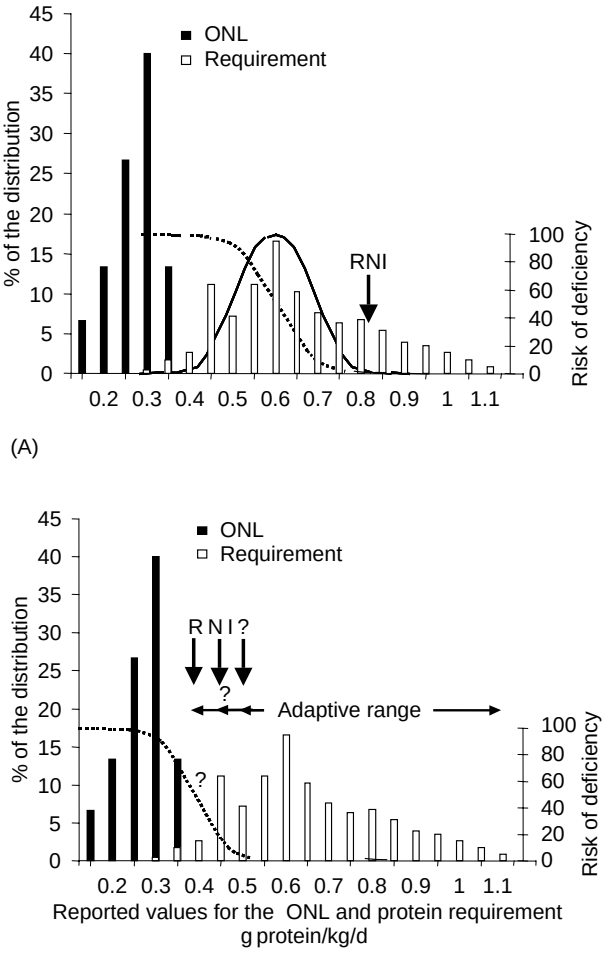


Figure 4 Distribution of reported values for the protein requirements and obligatory nitrogen loss (ONL) and calculation of risk of deficiency for an individual for current protein requirements model (A) and the adaptive metabolic demands model of protein requirements (B). The bars represent the distribution of reported mean values for the obligatory N loss (solid bars; $n = 15$ studies on 273 subjects) and individual values of intakes for N equilibrium expressed as protein equivalents (open bars; $n = 224$ subjects from 32 studies, after a 5% trim of outliers) from a meta-analysis of N balance data reported by Rand *et al.* (A) A normal distribution of requirements is shown (solid line), with the Recommended Nutrient Intake (RNI) and risk of deficiency for an individual (broken line) calculated assuming no correlation between protein intake and requirement. (B) RNI and suggested risk of deficiency (broken line), assuming most of the variation in reported requirement values reflects incomplete adaptation to the test diets with the true minimum requirement at the lower end of the observed range. Risk of deficiency in fully adapted individuals will not become significant until intakes fall below the upper range of the true minimum requirement (value currently unknown).

because of the very wide range of protein intakes in the human diet, mainly through variable meat intake, the apparent requirement indicated in a study may still reflect the prior habitual diet. That is, the apparent metabolic demands are higher than minimum

levels because of an adaptive component of amino acid oxidation set to balance previous intakes. This may explain the very wide range of reported apparent requirements analyzed in the nitrogen balance meta-analysis from approximately 0.4 to more than 1.1 g protein/kg/day. If adaptation does account for the variability, than a quite different analytical model would be implied (**Figure 4B**). Thus, the RNI would be much lower, with risk of deficiency for fully adapted individuals not increasing until intakes decrease to very low levels—close to values equivalent to the ONL. Such adaptive models pose difficult questions for public health nutrition.

Protein Requirements for Growth and Special Needs

For infants, children, and pregnant and lactating women, protein requirements are derived by a semi-factorial analysis of the components of the metabolic demands shown in **Figure 1**, with an assumed efficiency of utilization, all adjusted for individual variation to give the RNI. The main components as reported in the recent US Dietary Reference Intake (DRI) report are shown in **Figure 5**. The maintenance value is derived from nitrogen balance studies on children at 0.69 and 0.66 g protein/kg/day for ages 0.75–13 years and older than 13 years, respectively. The dietary requirements for growth derive from measured rates of protein accretion adjusted for an efficiency of utilization of 58%. To account for interindividual variability, the RNI includes the addition of 2 SDs for maintenance and dietary growth needs, calculated from a critical value (CV) that is the weighted mean of the CVs for maintenance (12%) and growth (43%).

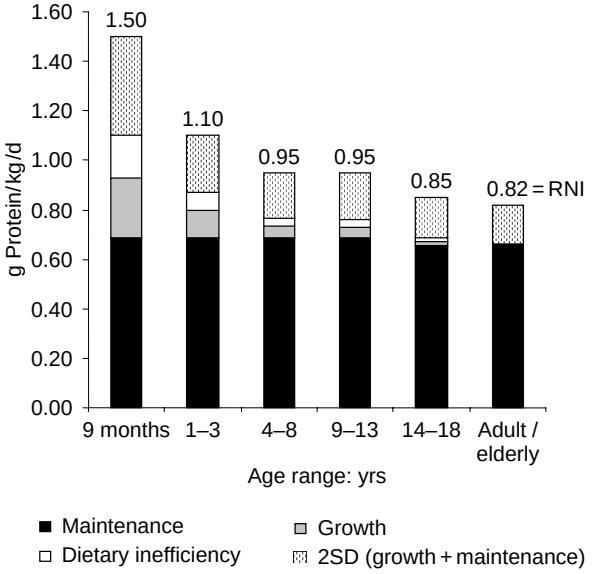


Figure 5 Factorial estimates of protein requirements throughout the life cycle. Overall values are averages of the separate values for boys and girls.

Pregnancy Requirements

These allow for protein retention in the products of conception and in the maternal tissues associated with the birth of an 'ideal' 3.3-kg infant. It is assumed that protein gain occurs in the maternal tissues in the early part of pregnancy and in the fetus mainly in the latter stages so that the metabolic demand is uniform throughout pregnancy. Thus, in the United Kingdom a single daily additional amount of 6 g protein throughout pregnancy is recommended. Lactation requirements of 11 and 8 g/day derive from estimates of the protein content of breast milk of healthy mothers (milk nitrogen $\times$ 6.25) assuming that daily breast milk protein content is constant for the first 6 months and declines thereafter. For the elderly, requirements are assumed to be the same as for younger adults since there is no evidence that they are higher than those of younger adults.

Areas of Uncertainty

Requirements of Infants

Definitions of protein requirements have historically been problematic and controversial, and current values are no exception. It has been suggested that values for the protein requirement of infants and children proposed in the 1985 FAO/WHO/UNU report were overestimates, and this problem was not entirely resolved in the US DRI report. The argument derives from a comparison of the requirement values with the protein intake of the breast-fed infant. Thus, the average requirement defined by FAO/WHO/UNU for the 3-month-old infant is the same as the average protein intake of the breast-fed infant. Since infants of healthy, well-nourished mothers consuming habitual amounts of breast milk are assumed to be optimally nourished, average intakes of breast milk protein are assumed to represent the safe level of the requirement, which is higher than the average requirement. Resolution of the problem requires use of a lower value for maintenance and a higher value for the assumed efficiency of dietary protein utilization for growth, giving values for the EAR and RNI at 3 months of 1.06 and 1.37 g protein/kg compared with a mean protein intake of the breast-fed infant of 1.44 g/kg. Such values would be lower than those implied by the factorial model used in the US DRI report. However, at this age of most rapid growth, the nitrogen in breast milk is utilized with unusual efficiency—an indication of the special properties and

qualities of breast milk that are poorly understood. Therefore, it may be proposed that formula-fed infants require more protein because of less efficient protein utilization. Indeed, some have questioned whether the breast-fed infant is the ideal model for protein requirements, with breast milk protein levels being a compromise between feeding the infant and minimizing losses of maternal protein stores. However, given the lower rates of morbidity of breast fed compared to formula-fed infants, it is difficult to sustain arguments that breast-fed infants are less than optimally nourished.

Optimal Protein Intakes and Implications of Adaptation for Nutrition Policy

In general, protein requirements serve two purposes. One is as a basis for prescription (i.e., advice on safe diets through recommending appropriate dietary intakes). Adaptation implies a low but difficult to define RNI. Indeed, since natural diets, providing sufficient energy and other nutrients, usually provide more than the minimal amount of protein, the magnitude of the minimal requirement becomes to some extent an issue of scientific curiosity only. Formulation of policy in relation to prescriptive matters will inevitably and correctly be most concerned with satisfying the upper range of demands for protein and, where there is uncertainty, include positive margins of error. In this case, it is arguably unwise to adopt an adaptive model and reduce the RNI, even if agreement could be reached on the likely lower limit of adaptation. Indeed, an adaptive model does not imply that protein is an unimportant nutrient for the maintenance of human health and well-being but that indicators other than balance (nitrogen, protein, or amino acid) need to be identified. Thus, the most relevant measure is an optimal requirement allowing balance and supporting both optimal body function and minimum risk of chronic disease. There is increasing experimental evidence for the potential benefit of protein intakes considerably higher than the current RNI for bone health in the elderly and epidemiological evidence for benefit with respect to hypertension and ischemic heart disease. However, such influences are unproven, with no plausible mechanism identified in the latter cases. In any case, there are no quantifiable indicators. This results in a dilemma for those attempting to frame prescriptive dietary guidelines. From this perspective, it is probably wise to retain current values as an operational expedient until it becomes possible to quantify the benefits (and any risks) of protein intakes within the adaptive range.

The other purpose of requirement recommendations is as a diagnostic indicator of risk, often within an epidemiological context in which population groups rather than individuals are considered. In this case, indicators used to estimate prevalence of disease states or deficit risk are carefully chosen so as to strike an acceptable balance between false positives and false negatives. The main implication of adaptation for estimating risk of deficiency as intakes become less than requirements is a dramatic reduction in the prevalence of risk for most populations compared with that assessed according to the traditional model, which does not account for adaptation. As in the prescriptive context, this low risk of deficiency applies only to that of being unable to maintain nitrogen balance after full adaptation with otherwise nutritionally adequate diets satisfying the energy demands. Whether such populations enjoy optimal protein-related health in terms of immune function, bone health, or any other function is a separate issue and needs to be addressed as such. From this perspective, it follows that maintenance of nitrogen balance can no longer be used as a surrogate of adequate protein-related health, and that current lack of quantifiable alternative indicators is no excuse for ignoring the issue of adaptation.

See also: **Amino Acids**: Chemistry and Classification; Metabolism; Specific Functions. **Breast Feeding**. **Infants**: Nutritional Requirements. **Osteoporosis**. **Pantothenic Acid**. **Pellagra**. **Pregnancy**: Nutrient Requirements. **Protein**: Synthesis and Turnover; Digestion and Bioavailability; Quality and Sources; Deficiency.

Further Reading

Department of Health (1991) *Dietary Reference Values for Food Energy and Nutrients for the United Kingdom*, Report on Health and Social Subjects No. 41. London: H.M. Stationery Office.

Dewey KG, Beaton G, Fjeld C, Lonnerdal B, and Reeds P (1996) Protein requirements of infants and children. *European Journal of Clinical Nutrition* **50**: 5119–5150.

FAO/WHO (1991) *Protein Quality Evaluation. Report of a Joint FAO/WHO Expert Consultation.* Rome: FAO.

FAO/WHO/UNU (1985) *Energy and Protein Requirements. 15. Report of a Joint FAO/WHO/UNU Expert Consultation*, Technical report series No. 724. Geneva: WHO.

Food and Nutrition Board and Institute of Medicine (2002) *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, and Amino Acids (Macronutrients).* Washington, DC: National Academy Press.

Institute of Medicine (2000) *Dietary Reference Intakes: Application in Dietary Assessment.* Washington, DC: National Academy Press.

Millward DJ (1998) Metabolic demands for amino acids and the human dietary requirement: Millward and Rivers (1988) revisited. *Journal of Nutrition* **42**: 367–393.

Millward DJ (1999) The nutritional value of plant based diets in relation to human amino acid and protein requirements. *Proceedings of the Nutrition Society* **58**: 249–260.

Millward DJ (1999) Optimal intakes of dietary protein. *Proceedings of the Nutrition Society* **58**: 403–413.

Millward DJ (2001) Workshop on "Protein and Amino Acid Requirements and Recommendations" methodological considerations. *Proceedings of the Nutrition Society* **60**: 1–4.

Millward DJ (2003) An adaptive metabolic demand model for protein and amino acid requirements. *British Journal of Nutrition* **90**: 249–260.

Millward DJ, Forrester T, Ah-Sing E *et al.* (2000) The transfer of ^{15}N from urea to lysine in the human infant. *British Journal of Nutrition* **83**: 505–512.

Millward DJ and Jackson AA (2004) Reference protein:energy ratios of diets in relation to current diets in developed and developing countries: Implications of proposed protein and amino acid requirements values. *Public Health Nutrition*, **7**: 387–405.

Pellett PL and Young VR (1992) The effects of different levels of energy intake on protein metabolism and of different levels of protein intake on energy metabolism: A statistical evaluation from the published literature. In: Scrimshaw NS and Schurch B (eds.) *Protein Energy Interactions*, pp. 81–136. Waterville Valley, NH: International Dietary Energy Consultative Group.

Rand WM, Pellett PL, and Young VR (2003) Meta-analysis of nitrogen balance studies for estimating protein requirements in healthy adults. *American Journal of Clinical Nutrition* **77**: 109–127.

Digestion and Bioavailability

Z A Bhutta, The Aga Khan University, Karachi, Pakistan

Proteins are the principal nitrogenous constituents of the protoplasm of all animal and plant tissue, and it is estimated that almost half of the dry weight of animal cells is composed of proteins. Proteins are crucial for the synthesis of body tissues and regulatory proteins, and it is also recognized that approximately 90% of all cellular proteins are present as enzymes.

The basic structural units of proteins are the amino acids, which are characterized by the presence of an amino NH_3 component and an acid or carboxyl group. Nitrogen thus comprises approximately 16% of all proteins by weight. Most naturally occurring amino acids are of the L configuration. These amino acids are in turn linked together by peptide bonds. Units of 2 or 3 amino acids are called dipeptides or tripeptides, respectively, whereas by convention any protein structure of less than 100 amino acid residues is called a polypeptide. The primary structure of a protein refers to the chains of amino acids constituting it, whereas the secondary structure is formed by the linkages between close amino acids by hydroxyl or sulfide bonds. More complex proteins have a tertiary structure due to the amino acids being held together by strong interatomic forces. The quarternary structure of a protein refers to the manner of association or binding between different units.

Dietary proteins are the major sources of protein intake and constitute on average approximately 10–20% of daily energy intake. In addition, they are the main sources for the essential amino acids, which cannot be synthesized by humans. Despite wide variations in dietary composition, the average daily protein intakes in different populations of the world range from 50 to 70 g, although it must be recognized that the intake may be much lower in deprived populations, in both qualitative and quantitative terms. Almost half of the total protein entering the gastrointestinal tract daily is derived from endogenous sources, mainly intestinal secretions and cellular desquamation. Salivary, gastric, biliary, pancreatic, and intestinal secretions contribute approximately 20–30 g per day, whereas desquamated villus epithelial cells contribute an additional 30 g, and a relatively smaller amount (2 g) is derived from plasma proteins leaking into the lumen.

An intricate and coordinated system of digestion ensures that under normal conditions, approximately 95% of ingested protein is digested and absorbed.

Digestion

The purpose of digestion is to hydrolyze proteins to small peptides and amino acids so that these can be absorbed. The daily protein load requiring digestion within the gastrointestinal tract includes both the exogenous protein derived from the food consumed and that from endogenous intestinal enzymes and cellular debris. The latter may constitute approximately 40% of the total gastrointestinal protein load, approximately 160–170 g. daily. The digestion of proteins in the gastrointestinal tract involves a coordinated series of events at different levels, with sequential digestion by proteolytic enzymes to a form that can be absorbed into the bloodstream. **Figure 1** is a sequential representation of the various sites of protein digestion and absorption in the gastrointestinal tract. The main gastric and pancreatic proteolytic enzymes and their physiological functions are summarized in **Table 1**.

Stomach Peptic Activity

The digestion of proteins begins in the stomach by the actions of pepsins, which are secreted as the

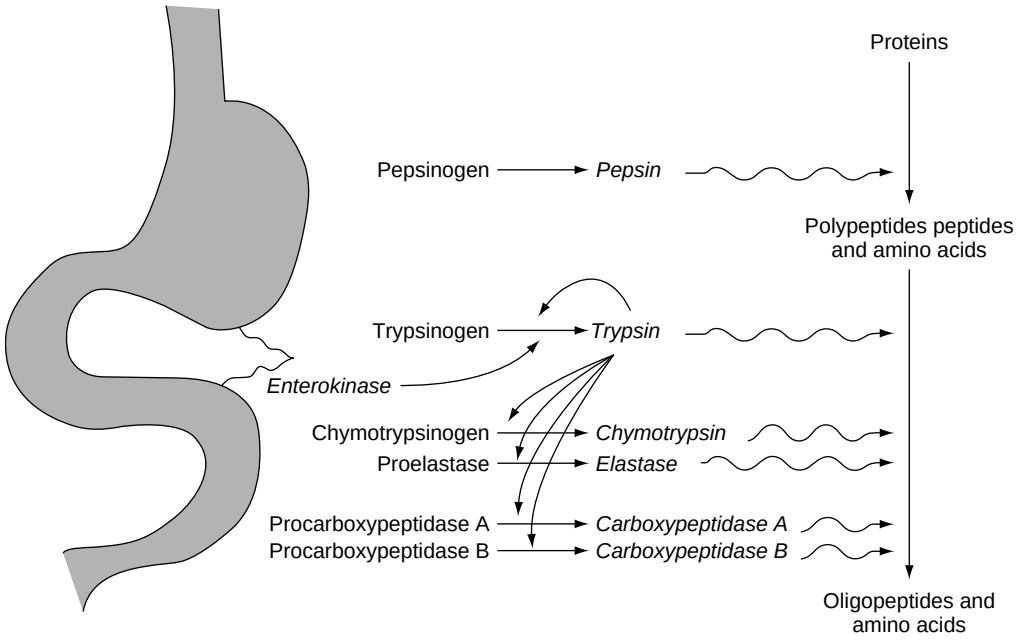


Figure 1 Cascade of protein hydrolysis in the gastrointestinal tract.

precursor form pepsinogen by the gastric mucosa main cells. The release of pepsinogen is stimulated by gastrin, histamine, and cholinergic stimulation and sinogens are converted to the active form pepsin by the loss of a small basic peptide. Pepsins remain active in the acid pH of the stomach and have a broad proteolytic specificity, splitting peptide bonds mostly involving phenylalanine, tyrosine, and

Table 1 Proteolytic enzyme activity in the gastrointestinal tract

Enzyme	*Precursor*	*Products*	*Catalyst*	*Substrate*	*Action*
Stomach					
Pepsins	Pepsinogens	Polypeptides of diverse sizes and some amino acids	Acid pH	Protein	Hydrolyse bonds between aromatic amino acids (e.g., phenylalanine or amino acid)
Pancreatic proteases					
Trypsin	Trypsinogen	Oligopeptides	Enterokinase Trypsin	Proteins Polypeptides	Cleaves internal bonds at lysine or arginine amino acids; cleaves other pancreatic proenzymes
Chymotrypsin	Chymotrypsinogen	Oligopeptides	Trypsin	Protein Polypeptides	Cleaves bonds of aromatic or neutral amino acids
Elastase	Proelastase	Oligopeptides	Trypsin	Elastin Other proteins	Cleaves bonds of aliphatic amono acids (e.g., alanine, glycine, and serine)
Carboxypeptidase A	Procarboxypeptidase A	Aromatic amino acids and peptides	Trypsin	Polypeptides at the free C-terminal end of the chain	Cleaves aromatic amino acids from C-terminal end of protein and peptides
Carboxypeptidase B	Procarboxypeptidase B	Arginine, lysine, and peptides	Trypsin	Polypeptides at the free C-terminal end of the chain	Cleaves arginine or lysine from C-terminal end of protein and peptides

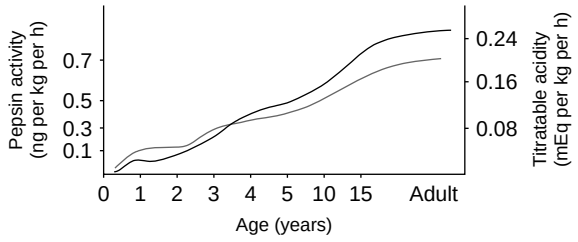


Figure 2 Postnatal development of gastric acid secretion and titratable acidity. Modified from Koldovsky (1987). Digestion and absorption of carbohydrates, proteins and fat in infants and children. In Walker WA and Watkins JB (eds), *Nutrition in Pediatrics*, Boston, Little Brown. Reproduced with permission from Little Brown & Company.

leucine. The level of peptic activity and acid production is lower in premature infants and increases in relation to gestational age; pepsin activity increases approximately twofold between infancy and adulthood (**Figure 2**). Immunohistochemistry indicates two distinct forms of pepsinogen: Pepsinogen I is only found in acid-secreting regions of the stomach, whereas pepsinogen II is also found in the mucous cells of the oxynctic and pyloric regions of the stomach as well as in the duodenal Brunner's glands. Although these two forms of pepsinogen have slightly different pH optima, their substrate specificity is very similar and both are rapidly inactivated by the alkaline pH beyond the pylorus.

A gelatinase liquefying gelatin is also found in the stomach. There is controversy regarding the presence of rennin (a peptidyl peptide hydrolase) in the stomach of young infants; however, the mild clotting activity in human infants is fairly rapid.

The completeness of gastric protein digestion is dependent on several factors, including the rate of gastric emptying, the pH of intragastric contents, and the type of protein ingested. Given the significant buffering capacity of food, it is unlikely that gastric proteolysis plays a major role in protein digestion. This is also verified by the fact that neither patients with achlorhydria nor those recovering from major gastric surgery appear to have a major problem with protein digestion.

Pancreatic Proteases

The pancreatic proteases are secreted as proenzymes and are activated in the lumen. The enteropeptidase (also called enterokinase) released from the brush border membrane removes a hexapeptide from the N-terminal end of trypsinogen, converting it to the active form trypsin. Trypsin, in turn, activates the other protease proenzymes and also autocatalytically promotes further activation of trypsinogen. The pancreatic proteases include the endopeptidases trypsin, chymotrypsin, and elastase, primarily splitting peptide bonds located within the protein molecules resulting in the production of short-chain polypeptides. These are further hydrolyzed by the exopeptidases carboxypeptidase A and B, acting on aromatic/aliphatic C terminals or basic C terminals, respectively, to remove single amino acids. The end product of this coordinated intraluminal digestion by these endopeptidases and exopeptidases is a mixture of neutral and basic amino acids (30%) with peptide chains varying in length from two to six amino acids (70%). The presence of excess amino acids in the lumen can further limit peptide hydrolysis (product inhibition).

The activity of enterokinase is noticeable after 26 weeks of gestation and its activity at term is approximately 10% of that of adults. Although pancreatic trypsin levels are substantial in both preterm and term infants, the secretory response to secretin and pancreozymin stimulation is somewhat blunted at birth compared with that at 2 years of age. However, such comparatively lower levels of protease activity in newborn infants do not appear to limit protein digestion significantly.

Brush Border Membrane and Cytoplasmic Peptidases

An important step in the final hydrolysis of peptides is their proteolysis to amino acids, either at the level of the intestinal brush border or within the cytoplasm of the intestinal mucosa. An important physiological observation is that protein absorption can occur both as amino acids and as peptides; indeed, absorption as peptides is considered a more efficient way of amino acid absorption compared with that of single amino acids (**Figure 3**). Even when a di- or tripeptide is subject to rapid hydrolysis by brush border peptidases, 30–50% of it is directly absorbed unconverted. The recognition that peptides are the main physiological routes of entry of amino acids into the enterocytes is a point of fundamental importance in the formulation of special protein hydrolysates and enteral feeds.

A range of peptidases are present at the level of the brush border membrane or cytoplasm with the capability of hydrolyzing oligopeptides of up to eight amino acid residues (**Table 2**). These oligopeptidases are synthesized in the rough endoplasmic reticulum of enterocytes and, after transfer through the Golgi apparatus, are transported to the brush border and extruded by exocytosis. There is little posttranslational processing of these peptidases, and they are attached to the brush border membrane by short anchoring pieces. The brush border peptidases differ in several ways from the cytoplasmic

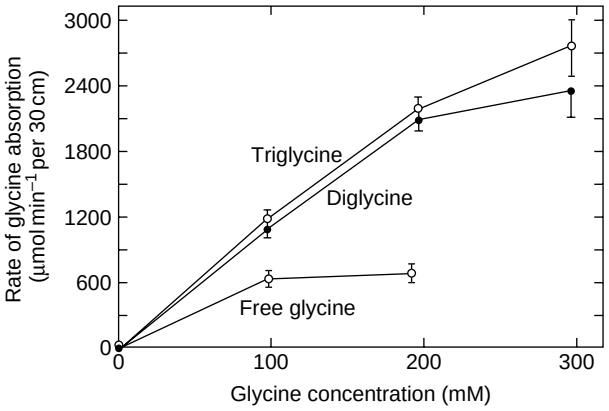


Figure 3 Rates of glycine absorption (mean ± SEM) from perfusion solutions containing equivalent amounts of glycine in free or peptide form. (From Adibi SA, Morse EL, Masilamani SS and Amiu P (1975) Evidence from two different modes of tripeptide appearance in human intestine: Uptake by peptide carrier systems and hydrolysis by peptide hydrolases. *Journal of Clinical Investigation* **56**: 1355–1363. Reproduced with permission from the American Society of Clinical Investigation.)

peptidases; the bulk of the hydrolysis of tetrapeptides and longer peptides occurs at the brush order, whereas the converse is true for dipeptidase activity, which is primarily within the cytoplasm. Most oligopeptidases are aminopeptidases, acting at the N-terminal amino acid. The brush border proteolysis rate is most rapid for tripeptides and least rapid for dipeptides, whereas the rates of hydrolysis of tetrapeptides and pentapeptides are somewhat intermediate. The brush border peptidases are capable of hydrolyzing all peptide bonds except those with praline at the C terminal.

In general, the cytoplasmic peptidases are more heat labile than brush border peptidases. Of the cytoplasmic peptidases, the most abundant is a dipeptidase that cleaves neutral dipeptides, whereas the aminotripeptidase has a high specificity toward tripeptides with N-terminal amino acids or those containing praline terminally.

Very little is known about the developmental aspects of brush border and cytoplasmic proteases. However, the activity of many of these proteases is discernible by 10–16 weeks of gestation and progressively increases during development. In contrast, γ-glutamyl transpeptidase activity decreases with increasing gestational age, but the significance of this transition is unknown.

Colonic Digestion

Although colonic digestion and fermentation is an important mechanism for energy production in plant-eating animals, its role in human nutrition is of minor importance. Colonic fermentation may lead to the production of short-chain fatty acids from undigested starch, nonstarch polysaccharides, or proteins reaching the colon, providing approximately 5–10% of daily energy requirements from this source. The contributory role of colonic protein digestion may become important for people with reduced small intestinal function such as short bowel syndrome.

Table 2 Peptidases present at the brush border membrane and cytoplasm of villous epithelial cells

Peptidase	*Action*	*Products*
Brush border membrane peptidase		
Aminooligopeptidases (at least two types)	Cleave amino acids from C terminal of 3–8 amino acid peptides	Amino acids dipeptides
Aminopeptidase A	Cleaves dipeptides with acidic amino acids at N terminal	Amino acids
Aminopeptidase I	Cleaves dipeptides containing methionine	Amino acids
Aminopeptidase III	Cleaves glycine-containing dipeptides	Amino acids
Dipeptidyl aminopeptidase IV	Cleaves praline-containing peptides with free C terminal	Peptides and amino acids
Carboxypeptidase P	Cleaves praline-containing peptides with free C terminal	Peptides and amino acids
Angiotensin I converting enzyme (ACE) γ-glutamyl transpeptidase	Cleaves γ-glutamyl bonds and transfers	γ-Glut amino and/or peptide
Endopeptidases (two, including PABA peptidase)		
Folate conjugase	Cleaves pteroyl polyglutamates	Monoglutamate
Cytoplasmic peptidases		
Endopeptidases (several, including Gly–Leu dipeptidase)	Cleaves most dipeptides	Amino acids
Aminotripeptidase	Cleaves tripeptides	Amino acids
Proline depeptidase	Cleaves X–Pro bonds in praline-containing depeptides	Proline and amino acids

PABA, *para*-aminobenzoic acid.

Absorption

As already indicated, although the final end product of protein digestion is amino acids, small peptides are the dominant form of entry of amino acids into enterocytes, where they are further hydrolyzed into amino acids and absorbed into the bloodstream (**Figure 4**). Thus, the vast majority of products of protein digestion that reach the bloodstream are single amino acids. Amino acid transport systems develop *in utero* by the end of the first trimester, whereas peptide transport systems can be demonstrated by the beginning of the second trimester.

It is recognized that the intestinal permeability of the preterm and newborn infant may be high, allowing the entry of small amounts of undigested proteins. The maternal antibodies from colostrum can enter the newborn's bloodstream relatively unaltered by a process of endocytosis and subsequent exocytosis. Although the intestinal permeability decreases with age, adults can still absorb larger proteins in abnormal circumstances. However, the predominant form of absorption and presentation of large foreign proteins is through the specialized microfold or M cells overlying the lymphoid Peyer's patches. This mode of absorption of intact proteins or polypeptides, however, is nutritionally insignificant.

Peptide Absorption

Di- and tripeptides can cross the brush border membrane by a main peptide transport system with broad specificity. This carrier protein can transport dibasic as well as diacid peptides and peptides consisting of up to three amino acid residues. However, there is some stereospecificity for this transporter because the longer the length of the amino acid side chain on the peptides, the easier the absorption. The transporter system also has greater affinity for dipeptides than tripeptides, and the acidic and basic amino acid residues in dipeptides lower the affinity for the transport system compared with neutral amino acids. In general, the absorption of L-isomers of amino acids in dipeptides is preferred over the D forms. The peptide transport system is coupled to the proton pump system rather than the sodium gradient. The oligopeptide transporter (Pept-1) in the brush border membrane is the major mechanism for protein absorption in the human intestine and is primarily responsible for the transport of di- and tri-peptides. Several factors may determine the levels of Pept-1, such as insulin, which may stimulate membrane insertion of the oligopeptide transporter from a preformed cytoplasmic pool, and cholera toxin, which decreases the activity of Pept-1 through an increase in the intracellular concentration of cyclic AMP.

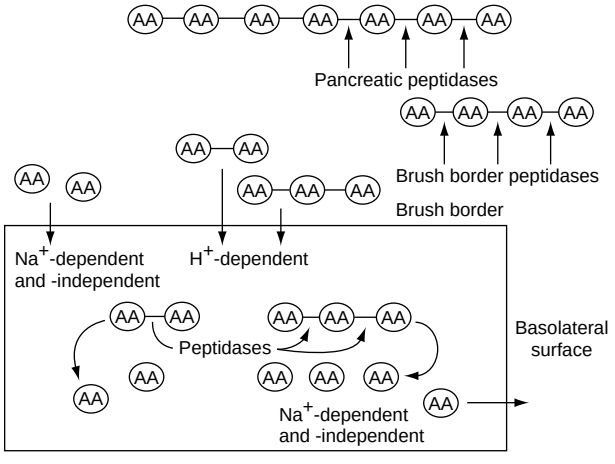


Figure 4 Small intestinal protein digestion and absorption. (Adapted from Shulman RJ (1996) Intraluminal digestion and absorption in the small intestine. In: Gluckman PD and Hayman MA (eds.) *Pediatrics and Perinatology: The Scientific Basis*, 2nd edn. London: Arnold. Reproduced with permission from Arnold (UK).)

Once in the absorbing cell, the di- and tripeptides are further hydrolyzed to the constituent amino acids by the cytoplasmic peptidases before absorption. The only small peptides that are known to enter the portal blood directly are those from gelatin that contain proline and hydroxyproline, and those from certain meats containing carnosine and anserine. However, their relative proportion in comparison to amino acids is inconsequential.

Amino Acid Absorption

Although some diffusion of amino acids does occur, they are mostly absorbed by active transport. Unlike peptides, which are absorbed equally well in both proximal and distal small intestine, amino acids are absorbed more rapidly in the duodenum and jejunum. Also in contrast to the parsimonious peptide transport system, there are multiple transport mechanisms for various amino acids at both the luminal end and the basolateral membrane of the enterocyte (**Table 3**). At the luminal end, the transporters are mostly located at the villous enterocytes. The villous enterocytes utilize approximately 10% of the absorbed amino acids for their own protein production, whereas the crypt cells derive their amino acid supply from the portal circulation. Of the various amino acids, glutamine appears to have a major role in the nutrition and regeneration of enterocytes, and it is now recognized that in the human intestine the predominant mechanism for assimilation of glutamine dipeptides is absorption as intact dipeptide rather than hydrolysis.

Table 3 Major amino acid transport systems in the intestinal epithelial cells

Transport system	*Substrates*	*Sodium gradient-dependent*
Brush border membrane		
B	Dipolar α amino acids	+
$B^{0,+}$	Dipolar α amino acids Basic amino acids Cystine	+
$B^{0,+}$	Dipolar α amino acids Basic amino acids Cystine	−
Y^+	Basic amino acids (e.g., lysine) Cysteine	−
IMINO	Imino acids (e.g., proline) β-Alanine	+
X^-_{GA}	Acidic amino acids (e.g., glutamate, aspartate)	+
β	β-Amino acids (e.g., alanine)	+
Basolateral membrane		
L	Broad selectivity	−
A	Broad selectivity	−
ASC	Neutral amino acids (e.g., alanine, serine) Cysteine	+
N	Clutamine, histidine, asparagines	+

Modified from Shulman RJ (1996) Intralumunal digestion and absorption in the small intestine. In: Gluckman PD and Heyman MA (eds.) *Pediatrics and Perinatology: The Scientific Basis*, 2nd edn. London: Arnold.

There are at least five different sodium-dependent transport systems for amino acid uptake. The sodium-dependent transport is facilitated by energy derived from Na^+/K^+-exchanging ATPase at the basolateral membrane. Most energy-dependent transporters are coupled either to cotransport of Na^+ or Cl^- or to the countertransport of K^+. An additional system of sodium-independent facilitated diffusion also exists and is predominantly geared toward basic and dipolar α amino acids. These passive transporters are either facilitated transporters or channels.

Digestibility

The digestibility of a protein is a measure of the amount of protein available from it for absorption after digestion; this is usually obtained from estimates of dietary nitrogen and fecal and urinary nitrogen. Digestibility is different from the other measures of protein quality, such as the amino acid or chemical scores and biological value, which respectively represent the essential and nonessential amino acid composition of the protein and the proportion of available nitrogen retained for growth or maintenance. Thus, a protein-based diet of high amino acid or chemical score may be poorly digested and of limited nutritional value. The digestibility of a protein is also dependent on the physical shape of the protein and the relative ease with which peptide bonds can be hydrolysed. Fibrous proteins with long polypeptide chains, such as collagen, keratin, and elastin, are relatively insoluble. In contrast, globular proteins, which are coiled and tightly packed, are comparatively soluble and thereby more digestible. Such proteins are insulin enzymes, hemoglobin, and albumin.

The apparent protein digestibility is a measure of the amount of protein intake (%) available for absorption and is usually calculated by estimation of fecal nitrogen and corresponding dietary intake:

$$\text{Apparent digestibility} = \frac{(\text{dietary nitrogen} - \text{fecal nitrogen})}{\text{dietary nitrogen}} \times 100$$

However, since not all fecal nitrogen is of dietary origin and some is derived from obligatory endogenous intestinal losses, a more appropriate measure is that of 'true protein digestibility.' This is derived as

$$\text{True protein digestibility} = \frac{[\text{dietary nitrogen} - (\text{fecal nitrogen} - \text{obligatory fecal nitrogen})]}{\text{dietary nitrogen}} \times 100$$

The obligatory fecal intestinal protein losses have been variably estimated to range from 20 mg/kg/day in young infant and preschool children to approximately 12 mg/kg/day in adults. These losses may result in some difficulty in interpreting digestibility findings. The estimated value of true digestibility of food and feed proteins is dependent on the excretion of metabolic fecal nitrogen (MFN). Results of many studies show that a high-fiber content of the diet increases MFN excretion and lowers the true digestibility of the diet protein. The exact estimation of MFN is only possible with isotopic methods. Experimental studies indicate that for human subjects the fecal digestibility values are significantly higher than the ileal values for Arg, Asp, Gly, Phe, Pro, Ser, Thr, and Trp, with the exception of faecal digestibility of Met, which is significantly lower than the ileal value.

Table 4 gives the true digestibility values for several common foods and diets. In general, milk and eggs have the highest true digestibility values of approximately 97%, followed by meats, fish, and

Table 4 Illustrative values of protein digestibility in humans

Protein sources	*True digestibility (%)*	*Digestibility relative to reference protein (%)*
Eggs	97	100
Milk and cheese	95	100
Meat and fish	94	100
Maize	85	89
Oatmeal	86	90
Whole wheat	86	91
Refined wheat	96	101
Polished rice	88	93
Soy flour	86	91
Soybean isolate	94	99
Millet	79	83
Peanut butter	95	100
Beans	78	82
Chinese mixed diet	94	99
Brazilian mixed diet	78	82
Guatemalan mixed diet	79	92
Indian rice and milk diet	87	92
Mixed American diet	96	101

Modified from Torun B (1985). Proteins: chemistry, metabolism and nutrition requirements. In Brunser O, Carraza F, Gracey M, Nichols B, Senterre J (eds). *Clinical Nutrition of the young Child.* New York, Raven Press, pp 99–119. Reproduced with permission from Raven Press.

poultry. Plants and legumes have comparatively lower protein digestibility values, ranging from 75 to 85%. Thus, in mixed diets, increasing the relative amounts of animal proteins compared with plant-based proteins results in increased protein digestibility of the diet. However, some fibrous animal proteins, such as keratin and collagen, are relatively indigestible. A useful approximation is to assume a protein digestibility of 75–80% for diets based on whole grain cereals and vegetables, 95% for diets based on refined cereals and animal proteins, and 85–90% for mixed diets. In general, the lower the true digestibility of a protein, the greater the amount required to achieve nitrogen equilibrium.

In addition to the differences in the nature of proteins highlighted previously, several other factors affect protein digestibility of a diet, including the presence of additional dietary factors such as trypsin inhibitors. The latter may be present in certain foods, such as navy beans and soybeans, and can be largely inactivated by heating, thus improving protein digestibility. Although moderate heating can promote digestibility by promoting breakdown of peptide cross-linkages and inactivation of protease inhibitors in natural food, strong heating, especially in the presence of a carbohydrate or oxidized lipids, may make the protein resistant to enzymatic hydrolysis. The Maillard or 'browning reaction' occurs after high, usually prolonged heating of a protein in the presence of a reducing sugar such as lactose or glucose, resulting in cross-linkages of the sugar with the free side chain of the lysine residues. This may make up to 30% of the lysine biologically unavailable. These changes are of particular importance in situations of marginally sufficient protein intake, in which cooking procedures may further aggravate protein malnutrition. The effect of heat treatment on the protein digestibility of a formulation was highlighted by a study using an elegant suckling rat model to investigate the digestibility different infant milk formulations. The data indicate that proteins from ultraheat-treated milk formulate were most rapidly digested (84%), resulting in an amino acid profile closest to that of breast-milk-fed pups, whereas the digestibility from powdered formulations (77–82%) and soy milk-based formulas was slower. The slowest digestion of protein was found in sterilized milk formula (72–74%), where the canned formulation was exposed to high temperatures for extended time period.

The Maillard reaction is highly influenced by the pH of foodstuffs or other agents. The reduction of pH that may be performed by increasing fermentation in the baking industry lessens the decomposition of lysine and tryptophan in proteins. Fermentation is widely used as a strategy to increase the digestibility of starch and improve the organoleptic properties of weaning foods and cereal-based preparations in developing countries. However, although the impact of fermentation on starch digestion is well established, the impact on protein digestibility is variable. Although some studies have suggested an impact on protein quality and digestibility of legumes and finger millet-based foods, other studies suggest that fermentation only modifies the gastric emptying rate and does not significantly affect the level of diet hydrolysis, the endogenous nitrogen stimulation, or the digestibility rate.

Despite several limitation, the digestion and absorption of ingested proteins is remarkably complete, with only a small fraction (3–5%) of ingested protein nitrogen escaping hydrolysis and excreted in the stools. In the context of infant nutrition, although breast milk is well digested, some proteins, such as secretory IgA, lactoferrin, and α_1-antitrypsin, escape digestion.

The protein digestibility-corrected amino acid score (PDCAAS) has been adopted by FAO/WHO as the preferred method for the measurement of the protein value in human nutrition. The PDCAAS is a combination of the chemical score of the limiting amino acid multiplied by true digestibility of the protein. The method is based on comparison of the

concentration of the first limiting essential amino acid in the test protein with the concentration of that amino acid in a reference (scoring) pattern. This scoring pattern is derived from the essential amino acid requirements of the preschool-aged child. The chemical score obtained in this way is corrected for true fecal digestibility of the test protein. PDCAAS values higher than 100% are not accepted as such but are truncated to 100%. Although the principle of the PDCAAS method has been widely accepted, critical questions have been raised in the scientific community about the validity of the preschool-aged child amino acid requirement values, the basis of correction for fecal instead of ileal digestibility, and the truncation of PDCAAS values to 100%. At the time of the adoption of the PDCAAS method, only a few studies had been performed on the amino acid requirements of the preschool-aged child, and there is still a need for validation of the scoring pattern. Also, the scoring pattern does not include conditionally indispensable amino acids.

These amino acids also contribute to the nutrition value of a protein. There is strong evidence that ileal, and not fecal, digestibility is the correct parameter for correction of the amino acid score. The use of fecal digestibility overestimates the nutritional value of a protein because amino acid nitrogen entering the colon is lost for protein synthesis in the body and is, at least in part, excreted in urine as ammonia. The truncation of PDCAAS values to 100% can be defended only for the limited number of situations in which the protein is to be used as the sole source of protein in the diet. For evaluation of the nutritional significance of proteins as part of mixed diets, the truncated value should not be used. In these cases, a more detailed evaluation of the contribution of the protein to the amino acid composition of the mixed diet is required. From such an evaluation, it appears that milk proteins are superior to plant proteins in cereal-based diets. Other studies have assessed the validity of the PDCAAS method in predicting the quality of protein products compared with the commonly used protein quality methods, protein efficiency ratio and net protein ratio. These data demonstrate that the PDCAAS method is inappropriate for predicting protein quality of protein sources that may contain naturally occurring growth-depressing factors or antinutritional factors formed during alkaline and/or heat processing.

See also: **Amino Acids**: Chemistry and Classification; Metabolism. **Protein**: Synthesis and Turnover; Requirements and Role in Diet; Quality and Sources; Deficiency.

Further Reading

Adibi SA (2003) Regulation of expression of the intestinal oligopeptide transporter (Pept-1) in health and disease. *Am J Physiol Gastrointest Liver Physiol* **285**: G779–88.

Brodin B, Nielsen CU, Steffansen B, and Frokjaer S (2002) Transport of peptidomimetic drugs by the intestinal Di/tri-peptide transporter, PepT1. *Pharmacol Toxicol* **90**: 285–96.

Castanga M, Shayakul C, Trotti D, Sacchi VF, Harvey WP, and Hediger MA (1997) Molecular characteristics of mammalian and insect amino acid transporters: implications for amino acid homeostasis. *Journal of Experimental Biology* **200**: 269–286.

Darragh AJ and Hodgkinson SM (2000) Quantifying the digestibility of dietary protein. *J Nutr* **130**: 1850S–6S.

Evenepoel P, Hiele M, Geypens B, Geboes KP, Rutgeerts P, and Ghoos Y (2000) 13C-egg white breath test: a non-invasive test of pancreatic trypsin activity in the small intestine. *Gut* **46**: 52–7.

Gaudichon C, Mahe C, Luengo C, Laurent C, Meaugeais P, Krempf M, and Tome D (1996) A '5N-leucine- dilution method to measure endogenous contribution to luminal nitrogen in the human upper jejunum. *European Journal of Clinical Nutrition* **50**: 261–268.

Hopfer U (1987) Membrane transport mechanisms for hexoses and amino acids in the small intestine. In: Johnson LR (ed.) *Physiology of the Gastrointestinal Tract*, 2nd edn., p. 1499. New York: Raven Press.

Lee VH (2000) Membrane transporters. *Eur J Pharm Sci* **11**(Suppl): S41–50.

Lonnerdal B (1994) Digestibility and absorption of protein in infants. In: Raiha NCR (ed.) *Protein Metabolism During In fancy*, pp. 53–65. New York: Raven Press.

Nordgarrd I and Mortensen PB (1995) Digestive processes in the human colon. *Nutrition* **11**: 37–45.

Proteins and amino acids. In: Robinson CH, Lawler MR, Chenoweth WL and Garwick AE (eds), Normal and Therapeutic Nutrition, 17th edn., pp. 44–63. New York: Macmillan.

Seidler U, Rossmann H, Jacob P, Bachmann O, Christiani S, Lamprecht G, and Gregor M (2000) Expression and function of Na^+HCO_3-cotransporters in the gastrointestinal tract. *Ann N Y Acad Sci* **915**: 1–14.

Shulman RJ (1996) Intraluminal digestion and absorption in the small intestine. In: Gluckman PD and Heyman MA (eds.) *Pediatrics and Perinatology: The Scientific Basis*, 2nd edn., pp. 634–637. London: Arnold.

Torun B (1985) Proteins: chemistry, metabolism, and nutritional requirements. In: Brunser O, Carraza F, Gracey M, Nichols B, and Senterre J (eds.) *Clinical Nutrition of the Young Child*, pp. 99–119. New York: Raven Press.

Quality and Sources

B Torun, Center for Research and Teaching in Latin America (CIDAL), Guatemala City, Guatemala

The amino acid composition of food proteins and the efficiency with which they are digested to allow amino acid absorption determine their capacity to

provide nitrogen and essential amino acids for human growth and functions. This capacity, known as protein quality, influences dietary requirements: The lower the quality, the higher the required dietary protein intake. The nutritive value of food proteins is also influenced by protein concentration and the bioavailability of its amino acids. The latter can be affected by some forms of food storage and processing.

This article examines the ways of assessing the protein quality of foods and diets and the quality inherent to various protein sources. The following operational terms are used:

- Protein (or nitrogen) digestibility: The proportion of dietary nitrogen that is absorbed. 'True' protein digestibility is calculated correcting for endogenous or obligatory fecal nitrogen losses (i.e., nitrogen in epithelial cells, gastrointestinal secretions, and intestinal flora) (**Table 1**).
- Nitrogen balance (NB): The average amount of nitrogen that is retained or lost from the body. It is calculated from measurements of dietary, urinary, and fecal nitrogen and estimates of integumental (sweat, skin, nails, and hair) nitrogen losses (**Table 1**).
- Essential amino acids (EAA) (also called 'indispensable amino acids'): Amino acids that the diet must provide because humans cannot synthesize them from other components at a rate commensurate with normal bodily needs.
- Amino acid scoring pattern: Amino acid composition of a hypothetical reference protein that contains all EAAs in the amounts necessary to satisfy requirements.
- Amino acid scoring procedure: Calculation of the proportion of each EAA in a protein or diet relative to the scoring pattern (**Table 1**). It can be expressed as percentage or as a fractional value.
- Limiting amino acids: EAAs in food proteins that are present in lower proportions than in the reference protein (i.e., with fractional value <1.00, relative to the reference protein) (**Table 1**).
- Amino acid score (or 'chemical score'): Value of the limiting amino acid with the lowest score in a protein (i.e., the 'most limiting amino acid'). A protein is assigned a percentage score of 100 (or a fractional score of 1.00) when none of its EAAs are limiting.

Table 1 Calculation of operational definitions[a]

Definition	*Calculation*
Apparent digestibility	$\frac{I_N - F_N}{I_N}$
True digestibility	$\frac{I_N - (F_N - F_E)}{I_N}$
Nitrogen balance	$I_N - U_N - F_N - \text{Integ}_N$
Amino acid score	$\frac{\text{mg of EAA in 1 g of food protein}}{\text{mg of EAA in 1 g of reference protein (or EAA scoring pattern)}}$
Limiting amino acid	EAA with a score <1.00 (or $<100\%$)

[a]Digestibility and amino acid scores can be expressed as fractional values (≤ 1.00) or multiplied by 100 and expressed as percentages.
EAA, essential amino acid; F_E, endogenous fecal nitrogen; F_N, total fecal nitrogen; I_N, nitrogen intake; Integ_N, integumental nitrogen; U_N, total urinary nitrogen.

Assessment of Protein Quality

Metabolic Studies

The most accurate assessment of protein quality of foods for humans is through clinical or metabolic studies that measure nitrogen balance. A fixed amount of protein is fed to a group of individuals until a steady state is reached. At that point, excreta are collected and analyzed for their nitrogen content, and integumental nitrogen losses are generally estimated at approximately $5\,\text{mg N kg}^{-1}$ to calculate NB as follows: $\text{NB} = I_N - U_N - F_N - \text{Integ}_N$ (See abbreviations in **Table 1**). Measurements are repeated with different amounts of food protein and the relationship between nitrogen intake and nitrogen balance is evaluated (**Figure 1**). The slope of the line before NB reaches a plateau and the amount of dietary protein needed to attain "0" nitrogen balance are indicators of protein quality: The steeper the slope and the lower the amount of dietary protein to achieve balance, the higher the quality of the protein being tested.

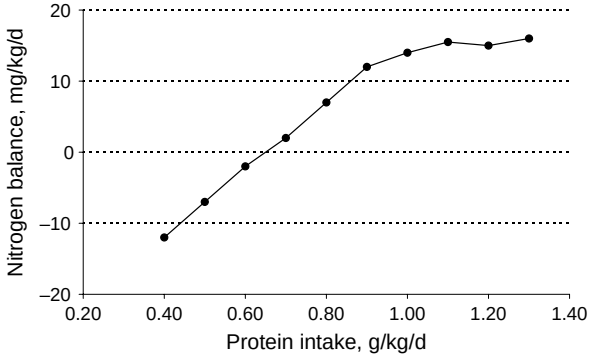


Figure 1 Plot of a nitrogen balance study. Nitrogen retention increases with protein intake until a plateau is reached.

Influence of Energy Intake on Nitrogen Balance

When food energy intake is insufficient to satisfy energy needs, amino acid oxidation increases in an effort by the human body to satisfy energy requirements. This raises urinary nitrogen excretion and reduces nitrogen balance. On the other hand, increased energy intake may reduce amino acid oxidation and urinary nitrogen excretion, thereby improving N balance until it reaches a plateau. This response, known as the protein-sparing effect of dietary energy, can be attenuated if the quantity or quality of food protein intake is inadequate. It has been postulated that the protein-sparing effect of dietary carbohydrates is mediated by increased insulin secretion, which inhibits proteolysis, hepatic gluconeogenesis, and renal ammoniagenesis. The protein-sparing effect of dietary fat may be due to a reduction of amino acid oxidation through an effect of free-fatty acid oxidation in the liver, whereby the increase in NADH/NAD inhibits branched-chained keto-acid dehydrogenase. For these reasons nitrogen balance must not be used to estimate protein quality when the amount of dietary energy is such that it produces weight loss or gain in an otherwise well-nourished individual.

Because of their high cost and experimental complexity, metabolic studies are done mainly to evaluate new, nonconventional protein sources and novel food processes that may affect protein quality. Other methods that can predict protein quality for humans rapidly and at low cost are used to evaluate diets and conventional foods routinely.

Assays in Laboratory Animals

Biological assays in laboratory animals have been used to assess food protein quality, based either on a protein's ability to support growth in young rats (protein efficiency ratio) or on nitrogen retention (net protein utilization). However, these assays underestimate the quality of some vegetable and animal proteins for humans. For example, the proteins of pulses and milk casein have a lower quality for rats than for humans because rats have a higher requirement of sulfur-containing amino acids. Thus, application of rat assay results to human nutrition can result in important quantitative errors. The discrepancy usually has economic rather than public health implications because rat assays generally err by underestimating protein quality for humans, but the value of certain animal proteins can be overestimated because of higher efficiency of utilization by the rat.

Amino Acid Score Adjusted for Digestibility

The concept of assessing protein quality on the basis of a protein's constituent amino acids was introduced in the late 1940s. It was later suggested that the calculations be corrected by the protein's digestibility. The validity of this approach and its correlation with results of metabolic and clinical studies were initially limited by lack of accurate procedures to measure tryptophan and sulfur amino acids, insufficient information on digestibility of proteins from various sources, and uncertainty about human amino acid requirements to prepare an adequate scoring pattern. Significant scientific and technological advancements now allow the use of an amino acid scoring procedure adjusted for digestibility as a good and practical predictor of protein quality for humans.

This method is recommended by expert committees of the World Health Organization (WHO), United Nations Food and Agriculture Organization (FAO), United Nations University (UNU), and the Codex Committee on Vegetable Proteins (CCVP), as well as by regulatory agencies of several countries, for routine evaluation of protein quality for humans. The elements required for its application are knowledge about the amino acid composition and digestibility of the food protein(s) under evaluation and a scoring pattern based on human amino acid requirements.

Amino Acid Analysis of Food Proteins

Modern methods that involve acid or alkaline hydrolysis of the protein followed by separation and quantification of the released amino acids by ion exchange, gas–liquid or high-performance liquid chromatography, and other chemical and microbiological methods for specific amino acids, such as lysine, methionine, cysteine, and tryptophan, provide data with a repeatability within laboratory of approximately 5% and a reproducibility between laboratories of approximately 10%. Although several national and international food composition tables include amino acid contents of foods, it is preferable to use analytical results from a reliable laboratory owing to technical shortcomings in the preparation of some tables and to the considerable variability between the reported values, especially for tryptophan, cysteine, and methionine.

Amino acid data are usually calculated as milligrams amino acid per gram of protein. If they are reported as milligrams amino acid per gram of nitrogen, they can be converted to the protein equivalents multiplying by specific protein factors that range from 5.7 (17.5% nitrogen) to 6.4 (15.6% nitrogen) for the major protein sources in the diet. The factor used for a mixture of protein sources is 6.25, corresponding to a nitrogen content of 16%.

Table 2 Calculating the lysine (lys) content of a rice, lentil, and chicken mixture

1. Protein sources in 100 g of the cooked mixture:

10 g dry polished rice
10 g dry lentil
20 g raw white chicken meat

2. Chemical composition:

	Rice	*Lentil*	*Chicken*
Protein (g per 100 g food)	7.0	23.7	19.2
Lysine (mg per 100 g food)	255	1739	1590

3. Lysine content of the mixture (mg per g protein):

Food	*mg lys per g protein*		*g component per 100 g mixture*
Rice	(255/7.0)	×	10
Lentil	(1739/23.7)	×	10
Chicken	(1590/19.2)	×	20

Weighted mean = (364 + 734 + 1656)/(10 + 10 + 20) = **60 mg lys per g protein**

Table 3 Amino acid scoring patterns for infants under 1 year and for older children and adults (mg amino acid per g protein)[a]

Amino acid	*Infant < 1 year*	*Older children and adults*	*Egg, cow's milk, and beef protein*
Histidine	26	(19)[b]	22–34
Isoleucine	46	28	47–54
Leucine	93	66	81–95
Lysine	66	58	70–89
Methionine + cysteine	42	25	33[c]–57
Phenylalanine + tyrosine	72	63	80–102
Threonine	43	34	44–47
Tryptophan	17	11	12[c]–17
Valine	55	35	50–66

[a]Composition of animal proteins shown for comparison.
[b]Essentiality of histidine not clearly determined after 1 year of age.
[c]Cow's milk proteins have less sulfur-containing amino acids and tryptophan than human milk.

To calculate the amino acid content of a combination of food proteins, as in a processed food based on several protein sources or in a mixed diet, a weighted mean of the published or analytical results of each component should be used, as illustrated in **Table 2**.

Amino Acid Scoring Pattern

For infants younger than 1 year, the scoring pattern should be based on the amino acid composition of breast milk, even if some EAAs in human milk exceed minimum requirements for infants of this age. For example, infants consuming cow's milk proteins, which have less sulfur-containing amino acids than human milk, show adequate growth and nitrogen balance. Thus, although the use of a scoring pattern based on human milk composition may somewhat underestimate the protein quality of some foods for infants, there is consensus to accept errors on the side of safety for this highly vulnerable age group.

International expert committees (WHO, FAO, UNU, and CCVP) have agreed that the scoring pattern proposed in the 1980s for preschool children—based on studies of amino acid requirements at the Institute of Nutrition of Central America and Panama and on recommendations of protein intake by FAO, WHO, and UNU—is robust and represents the best available estimate of EAA requirements for this age group. Published nitrogen balance studies on older children and adults have experimental flaws, and only a limited number of EAAs have been studied with amino acid oxidation techniques in adults. Since proteins with amounts of EAAs that satisfy the needs of young children will probably be adequate for older children and adults, the scoring pattern for preschool children is currently used for all after 1 year of age.

Table 3 shows the internationally accepted patterns for amino acid scoring applicable to infants and to persons after 1 year of age; the composition of high-quality animal foods is shown for comparison. The content of each EAA in a food protein is evaluated relative to the age-specific scoring pattern, to determine the protein's amino acid score and to identify the limiting amino acids as shown in **Table 1**. All EAAs present in proportions that exceed requirements are assigned a fractional score of 1.00 (or a percentage score of 100%), even if mathematical calculation gives a higher value. The EAA with the lowest value (i.e., the most limiting amino acid) determines the protein's amino acid score.

The only EAAs that are likely to limit the protein quality of mixed diets for humans are lysine, the sulfur-containing amino acids (methionine and cysteine), threonine, and tryptophan. Consequently, when information on all EAAs is not available, protein quality can be estimated on the basis of its score for these four amino acids.

Correction for Protein Digestibility

A protein may have a good amino acid composition relative to the scoring pattern, but if it is not fully digested and its constituent amino acids are not absorbed, its capacity to provide nitrogen and EAAs for human function will diminish. Not all food proteins are digested,

absorbed, and utilized to the same extent because of inherent differences in their source (e.g., inside vegetable cells with indigestible membranes), their physicochemical nature (e.g., protein configuration and amino acid binding), the presence of food constituents that modify digestion (e.g., dietary fiber, tannins, and other polyphenols), the presence of anti-physiological factors that interfere with protein breakdown (e.g., trypsin inhibitors and lectins), and processing conditions that alter the nature or release of amino acids (e.g., Maillard reaction and formation of polyamino acids and methylmercaptan). Consequently, amino acid scores as predictors of protein quality must be adjusted for protein digestibility and amino acid availability.

The standard for obtaining digestibility data is through metabolic studies in humans, in which the nitrogen excreted in the feces is subtracted from the amount ingested with the diet and expressed as a percentage of intake. This apparent digestibility value must be corrected for the amount of fecal nitrogen excreted when a person is consuming a protein-free diet to calculate "true" digestibility (**Table 1**). Ethical constraints and practical complexities do not permit the determination of obligatory fecal nitrogen losses on a protein-free diet in all age and physiological groups. It is recommended that existing published values for daily obligatory fecal losses in preschool children (approximately 20 mg $N\,kg^{-1}$) and adults (approximately 14 mg $N\,kg^{-1}$) be used to correct apparent protein digestibility values.

Protein digestibility values of specific foods and well-defined diets may be taken from reliable published data. **Table 4** shows some examples. When such data are not available for a mixed diet, a weighted average can be calculated from the true digestibilities of its constituent protein sources, as illustrated in **Table 5**. For new or novel products or processes, digestibility must be determined, preferably in humans. When cost and practicality do not permit performing metabolic studies in humans, standardized fecal balance methods in rats have been used. These methods have given true protein digestibility values of 93–100% for animal foods or food products (casein, beef salami, skim milk, tuna, and chicken sausage) and soya protein isolate; 86–92% for beef stew, chick peas, rolled oats, and whole-wheat cereal; and 70–85% for lentils and different types of beans. These value ranges are similar to those from human studies. Nevertheless, rat data must be used with caution for foods and diets that are known or suspected of being handled differently by the human and rat intestines. *In vitro* procedures have also been developed using combinations of trypsin, chymotrypsin, peptidase, and bacterial protease. Further research is needed to validate their use as predictors of protein digestibility in humans.

Table 4 True protein digestibility of selected foods and diets

	True protein digestibility (%)
Egg white	97
Whole egg, milk, beef, poultry, fish	95
Wheat, refined flour	95
Soya protein isolate	94
Polished rice	88
Soya flour	86
Wheat, whole	86
Maize products	85
Rice, whole	84
Beans	69
Mixed diets	
USA	96
China	94
Colombia, high-income	93
Phillippines, urban	88
Chile, middle class	82
Mexico, rural	80
Guatemala, rural	79
Brazil, rural	78
India, vegetarian	78

Table 5 Calculation of true digestibility of a mixed diet of rice, beans, wheat, and egg

Diet	*True protein digestibility (%)*	*Proportion of total protein (g per 100 g protein in whole diet)*
Polished rice	88	40
Black beans	69	35
Whole wheat	86	15
Whole egg	95	10
Estimated digestibility of whole diet	$(0.88 \times 40) + (0.69 \times 35) + (0.86 \times 15) + (0.95 \times 10) = \mathbf{82\%}$	

Calculations and Examples

The EAA composition and protein digestibility of the food or mixed diet being tested are determined. Then the percentage or fractional value of the most limiting EAA (noncorrected amino acid score) is multiplied by the percentage or fractional value of 'true' protein digestibility to obtain the corrected score, which is equivalent to protein quality. This value can be used as such or it can be expressed in relation to the corrected amino acid score of a reference protein or food, usually casein or an animal food (milk, egg, or beef).

Proteins that have no limiting amino acids are assigned an amino acid score of 100% (or 1.00) that must be only corrected for digestibility.

Table 6 Calculation of amino acid scores of single protein sources corrected for digestibility and in relation to the protein quality of cow's milk

Food	*Most limiting amino acid*	*Noncorrected amino acid score*	*True protein digestibility*	*Corrected amino acid score*	*Protein quality relative to milk*
Cow's milk	None	$>100 \rightarrow 100\%$	$\times 95\%$	$=$ **95%**	—
Polished rice	Lysine 36 mg per g protein	$(36/58) \times 100 = 62\%$	$\times 88\%$	$=$ **55%**	$(55/95) \times 100 =$ **58%**
Egg white	None	$>100 \rightarrow 100\%$	$\times 97\%$	$=$ **97%**	$(97/95) \times 100 =$ **102%**

Similarly, if the clinical or experimental assessment of 'true' protein digestibility gives a value greater that 100% (generally due to experimental variability), a digestibility correction factor of 100% (or 1.00) is applied to the amino acid score. **Table 6** shows examples of calculations for a single food as protein source. The same procedure can be used for food mixtures using a weighted average procedure based on the protein content, amino acid composition, and digestibility of the individual components. **Table 7** shows an example of those calculations. For simplicity, the example uses only the four EAAs that are most often limiting.

Protein Concentration

Protein concentration or density (i.e., the amount of protein per unit of food) is another factor of a food's protein quality. Protein-dense foods are especially important for young infants, whose small gastric capacity limits the amount they can eat, and for elderly people with poor appetite. Evaluation of a food's protein concentration must be done for ready-to-eat preparations because food processing and cooking can result in significant changes relative to raw foods. Meats, poultry, and fish usually have a higher concentration of protein after cooking or frying, whereas vegetable food preparations contain more water and less protein than the raw products (**Table 8**).

Protein/Energy Ratio

The percentage of protein energy in the diet (P/E ratio) has been used to describe whether a diet provides adequate amounts of protein. The reasoning is that energy requirements are the main driving force for food intake. Therefore, a diet is adequate if it satisfies the requirements for all nutrients when it is eaten in amounts that will satisfy energy needs.

P/E ratio is calculated by dividing the amount of metabolizable energy derived from dietary protein (grams of protein $\times$ 16.7 kJ or 4 kcal) by the total amount of metabolizable energy in the diet, multiplied by 100 to avoid using fractional values. However, the use of P/E ratio as an index of food's protein adequacy may be misleading because it only gives information about protein concentration and does not indicate the biological value or quality of the proteins. Its usefulness improves when amino acid score is taken into account to calculate what can be defined as a desirable P/E ratio, as in the examples discussed later.

The P/E ratio indicates the amount of protein that the diet provides relative to energy and does not imply a constant relationship between protein and energy requirements. For example, the lower limit of the desirable P/E ratio of a diet with an amino acid score of 85% is 6.2 for a young child whose daily requirements are 16 g protein and 5.1 MJ energy $((16\,g \times 16.7\,kJ/0.85)/5100\,kJ \times 100)$. For an adult male with daily requirements of 55 g protein and 12.8 MJ, the desirable P/E ratio is 8.4 $((55 \times 16.7/0.85)/12{,}800 \times 100)$.

Diets, especially those eaten by adults, often provide protein in amounts that surpass requirements, which elevates the P/E ratio. For example, almost all adult populations eat diets with P/E ratios between 10 and 15%. This is related to culture and food availability and does not reflect a biologically optimal ratio. Consistent with the calculations in the preceding paragraph, P/E ratios of 10 and 15 are adequate and it cannot be argued that one is nutritionally better than the other.

Improvement of Protein Quality

Amino Acid Profile

The amino acid profile of a food or diet can be improved by increasing the amount of constituent amino acids in its proteins, adding specific amino acids, or combining foods in proportions that result in a better amino acid pattern.

Genetic handling This has resulted in cereals with higher contents of the amino acids that limit their protein quality. For example, varieties of Opaque-2 corn have approximately 50% more lysine and 35%

Table 7 Calculation of protein quality of a mixed diet based on whole wheat, polished rice, and chicken breast

Raw ingredients	*Data from analysis or literature*							*Quantities calculated for the mixed diet*				
	Weight (g)	*Protein (g per 100 g)*	*Lys (mg per g protein)*	*SAA*	*Th*	*Trp*	*True digestibility (%)*	*Total protein (g)* $H = A \times B/100$	*Lys (mg)* $I = H \times C$	*SAA (mg)* $J = H \times D$	*Thr (mg)* $K = H \times E$	*Trp (mg)* $L = H \times F$
	A	*B*	*C*	*D*	*E*	*F*	*G*					
Whole wheat	300	11	28	37	29	11	86	33	924	1221	957	363
Polished rice	200	7	36	38	33	13	88	14	504	532	462	182
Chicken breast	150	19	83	38	40	12	95	28.5	2366	1083	1140	342
Totals								75.5	3794	2836	2559	887
M Weighted mean digestibility of the mixed diet (sum of ($G \times H$) for each food component divided by total protein, *H*)							**0.90**					
N mg amino acid per g protein (totals for *I, J, K*, or *L* divided by total *H*)									50	38	34	12
P Amino acid scoring pattern, mg amino acid per g protein									58	25	34	11
Q Score for each amino acid in the mixed diet *(N/P)*									0.86	1.52	1.00	1.09
R Amino acid score adjusted for digestibility (*Q* of the limiting amino acid multiplied by *M*)								$0.86 \times 0.90 =$ **0.77** (or **77%**)				

Lys, lysine; SAA, sulfur-containing amino acids: Thr, threonine; Trp, tryptophan.

Table 8 Protein concentration in selected raw and ready-to-eat foods (g protein per 100 g food)

Food	*Ready-to-eat*	*Raw*
Beef, lean	36.8 (cooked)	21.4
Fish	31.8 (fried)	20.0
Wheat flour	12.0 (white bread)	11.0
Egg, hen	11.3 (hard-boiled)	11.3
Lentils	7.1 (cooked)	23.7
Common beans	6.2 (cooked)	22.0
Maize	4.2 (tortilla)	9.4
Milk powder, cow	3.2 (12% in water)	26.1
Rice	2.5 (boiled)	7.2
Potato, no skin	1.1 (cooked)	1.8

more tryptophan than native corn, both of which are limiting amino acids in this cereal.

Fortification and enrichment The addition of synthetic amino acids eliminates or reduces the magnitude of limiting amino acids, for example, in lysine-enriched wheat flour.

Complementation The combination of a food that has one or more limiting EAAs with another food(s) that has a surplus of these amino acids results in an improved combined amino acid profile. A double complementation effect has been achieved in the formulation of vegetable mixtures based on protein sources in which one has a surplus of the EAA that is limiting in the other and vice versa (**Figure 2**).

Digestibility and Bioavailability

Various food processing procedures can improve protein digestibility by removing food constituents that reduce digestibility (such as dietary fiber), breaking down poorly digestible vegetable cell membranes, destroying or neutralizing anti-physiological factors, and increasing the food surface area that can come into contact with gastrointestinal enzymes. For example, soya protein isolate, polished rice, and refined wheat flour have higher protein digestibilities than soya flour, whole rice, and whole wheat, respectively (**Table 4**).

Food storage and processing in adverse circumstances can reduce protein quality by making some EAAs unavailable for use in the human body. These conditions should be avoided to preserve protein quality. Some examples are the storage of dried milk under mild to moderate heat and humidity, which renders lysine side chains unavailable after reacting with the reducing sugar, lactose (Maillard

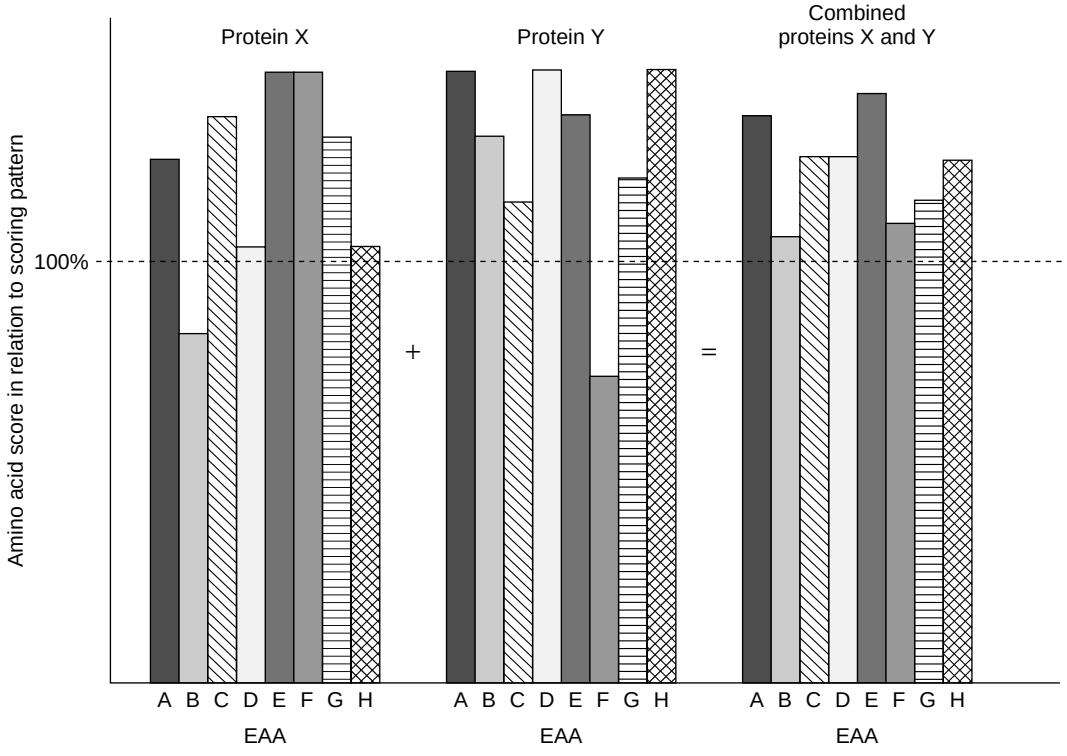


Figure 2 Amino acid complementation: When the two proteins are combined, protein X compensates the deficiency of amino acid F in protein Y, and protein Y compensates the deficiency of amino acid B in protein X. EAA, essential amino acids. From Torun B, Menchu MT and Elias LG (1994) Recommendaciones Dietéticas Diarias del INCAP. INCAP publication ME/057. Guatemala: Institute of Nutrition of Central America and Panama.

or 'browning' reaction); the severe treatment of protein with alkali, which causes lysine and cysteine residues to react and form lysinoalanine; and the treatment of proteins with oxidizing agents, which can result in a loss of methionine. Severe heating conditions in the presence of reducing sugars or oxidized lipids can make some food proteins resistant to digestion, thereby reducing the availability of all their amino acids.

Protein Concentration

Protein concentration can increase by genetic selection of protein sources, as in improved varieties of rice that have approximately 30% more protein than native rice, by the use of nitrogen-concentrating fertilizers that can raise the protein contents of several cereals or by industrial and home processing that reduce the water content of food preparations. Addition of concentrated protein sources, such as casein, soya protein isolates, soya flour, milk powder, or dehydrated egg, will also increase the protein concentration of foods and diets, as well as their amino acid score in some instances.

Protein Quality and Dietary Sources

Foods of animal origin, such as milk and milk products, eggs, meats, poultry, and fish, have excellent amino acid composition with a score of 100% and true protein digestibility of 95–98%. In addition, their protein concentrations often increase after cooking. Consequently, they are used as the reference for comparison of protein quality, provided that they are processed in ways that will not decrease amino acid bioavailabilty.

Almost all vegetable foods have one or more limiting amino acids. Soya beans and soya products are notable exceptions. In general, proteins in natural vegetable foods have digestibilities of 70–85%. Vegetable protein isolates, flours, and extruded products have higher digestibilities.

Among vegetables, pulses have the highest protein concentrations, ranging from 20–25% in most raw beans and peas to approximately 36% in soya beans. Pulses usually have limiting sulfur-containing amino acids.

Cereals and cereal products are the largest sources of protein in most areas of the world. Cereal grains and flours contain approximately 7–12% protein with a quality that is limited by their lysine content and, in many instances, also by threonine and/or tryptophan. Although deficient in lysine and threonine, rice has one of the best amino acid compositions among cereals, whereas sorghum and native maize (i.e., not genetically improved) are among the lowest.

Most nuts and edible seeds contain 8–18% protein. Many oil seeds have 12–20% protein, and the cake that remains after oil extrusion can have as much as 30–40% protein.

See also: **Amino Acids**: Chemistry and Classification; Metabolism; Specific Functions. **Cereal Grains**. **Energy**: Balance. **Fruits and Vegetables**. **Nuts and Seeds**. **Protein**: Synthesis and Turnover; Requirements and Role in Diet; Digestion and Bioavailability; Deficiency.

Further Reading

Banch-Knudsen KE, Wisker E, Daniel M, Feldheim W, and Eggum BO (1994) Digestibility of energy, protein, fat and non-starch polysaccharides in mixed diets: Comparative studies between man and the rat. *British Journal of Nutrition* **71**: 471–487.

Bodwell CE, Carpenter KJ, and McDonough FE (eds.) (1989) A collaborative study of methods of protein evaluation. *Plant Foods in Human Nutrition* **39**: 3–127.

FAO (2003) *Food Energy—Methods of Analysis and Conversion Factors*, FAO Food and Nutrition Paper No. 77. Rome: Food and Agriculture Organization of the United Nations.

FAO/WHO (1991) *Protein Quality Evaluation*, FAO Food and Nutrition Paper No. 51. Rome: Food and Agriculture Organization of the United Nations.

FAO/WHO/UNU (1985) *Energy and Protein Requirements*, WHO Technical Report Series No. 724, pp. 64–70, 117–127. Geneva: World Health Organization.

McLarney MJ, Pellett PL, and Young VR (1996) Pattern of amino acid requirements in humans: An interspecies comparison using published amino acid requirement recommendations. *Journal of Nutrition* **126**: 1871–1882.

Pellett PL and Young VR (eds.) (1980) Nutritional evaluation of protein foods. *Food and Nutrition Bulletin*, supplement 4.

Pineda O, Torun B, Viteri FE, and Arroyave G (1981) Protein quality in relation to estimates of essential amino acid requirements. In: Bodwell CE, Adkins JS, and Hopkins DT (eds.) *Protein Quality in Humans: Assessment and In Vitro Estimation*, pp. 29–42. Westport, CT: AVI.

Sarwar G and McDonough FE (1990) Evaluation of protein digestibility-corrected amino acid score method for assessing protein quality of foods. *Journal of the Association of Official Analytical Chemists* **73**: 347–356.

Torun B (1990) Current concepts on requirements of essential amino acids. In *New Era! Global Harmony through Nutrition. 14th International Congress of Nutrition: Symposium Lectures*, pp. 87–91. Seoul: Ewha Women's University.

Torun B, Durnin JVGA, Garza C, Jequier E, and Shetty PS (1992) Dietary protein/energy ratios for various ages and physiological states. In: Scrimshaw NS and Schurch B (eds.) *Protein-Energy Interactions*, pp. 379–384. Lausanne: International Dietary Energy Consultancy Group.

Torun B, Pineda O, Viteri FE, and Arroyave G (1981) Use of amino acid composition data to predict protein nutritive value for children with specific reference to new estimates of their essential amino acid requirements. In: Bodwell CE, Adkins JS, and Hopkins DT (eds.) *Protein Quality in Humans:*

Assessment and In Vitro Estimation, pp. 374–393. Westport, CT: AVI.

Young VR (ed.) (2004) Proceedings of the 3rd Amino Acid Assessment Workshop. *Journal of Nutrition* **134**(supplement): 1553S–1672S.

Young VR and Borgonha S (2000) Nitrogen and amino acid requirements: The Massachusetts Institute of Technology Amino Acid Requirement Pattern. *Journal of Nutrition* **130**: 1841S–1849S.

Young VR and Pellett PL (1994) Plant proteins in relation to human protein and amino acid nutrition. *American Journal of Clinical Nutrition* **59**(supplement): 1203S–1212S.

Deficiency

Z A Bhutta and H L Dewraj, The Aga Khan University, Karachi, Pakistan

Introduction

The term protein deficiency represents a state of deficit in body protein or one or more of the essential amino acids. Thus, the term protein deficiency can also be considered synonymous with negative nitrogen balance. The deficiency can result from a protein-deficient diet or other diseases and, in general, can also result from a global deficit of food. While protein-energy malnutrition is the more common form of protein deficiency, in general the features are comparable to those seen with kwashiorkor.

Dietary protein contributes all of the amino acids and fixed nitrogen necessary for the biosynthesis of tissue proteins and nonprotein nitrogenous compounds such as purines and pyrimidines. Dietary amino acids are required for the synthesis of new tissue constituents at all ages, particularly during growth. Amino acids consumed in excess of these needs are not stored but are degraded, the nitrogen being excreted and the carbon skeleton recycled. Each day more amino acids are degraded and resynthesized in the body than are ordinarily consumed in the diet.

All of the 20 fundamental amino acids must be present for protein synthesis to occur. The remarkable range of functions mediated by proteins is a function of the diversity and versatility of these 20 distinct building blocks of proteins. Nine of these amino acids (histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine) are not synthesized in the human body and are therefore 'essential' for well-being. In addition, arginine is essential in infancy, and in the preterm infant there is a transient need for dietary tyrosine and cysteine as well. The nutritional quality of dietary protein is influenced by its essential amino acid content. This implies that the protein intake contains sufficient 'nonessential' amino acids to minimize metabolic diversion of essential amino acids to cover nonspecific nitrogen requirements.

An adequate protein intake contains all of the essential amino acids in sufficient quantities to satisfy maintenance needs and to provide a surplus sufficient for the processes of normal growth and development. Serum concentrations of albumin and total protein serve as clinical indicators of the sufficiency of dietary protein intake in the absence of systemic disease. While the protein foods of animal origin (milk, meat, fish, and eggs) supply all of the essential amino acids; it must be recognized that some foods of vegetable origin supply most amino acids in adequate amounts.

Although the importance of hormonal regulation of protein metabolism is well recognized, there is increasing evidence that dietary protein may play a regulatory role by modulating the hormonal milieu leading to tissue accretion. Studies in recovering malnourished children and in normal children have shown a significant increase in circulating insulin-like growth factor I associated with higher protein intakes. Protein amino acid composition, however, appears to be of less relevance for these effects, which were observed with both animal and vegetable protein sources. Thus, it appears that dietary protein quantity is the major factor for this observed response. The amino acid composition of dietary proteins has a direct effect on growth by determining the supply of amino acids at the cellular level. Protein synthesis requires the presence of each component amino acid at the time of chain elongation. Thus, a dietary protein intake deficient in one or more essential amino acids will not be able to sustain protein synthesis. Many vegetable proteins have one or more limiting amino acids, e.g., with levels below those in high-quality reference proteins such as egg or milk.

It is important to maintain a balanced intake of amino acids in the diet, and to understand the relationships between different groups of amino acids and other nutrients such as vitamins. For example, when the most limiting amino acid in a diet generally poor in protein is increased, a deficiency of the next most limiting amino acid may be precipitated. Excessive intakes of certain amino acids, which may or may not be limiting, when added to a diet that is marginal in certain of the B vitamins, may result in an increased severity of the vitamin deficiency. In

other cases, an excess of an amino acid may reduce the utilization of another amino acid that is provided in normally adequate amounts to such an extent that a deficiency occurs.

Protein Turnover and Regulation

The term protein turnover reflects the balance of protein degradation and resynthesis. From a quantitative standpoint, by far the greatest influence on amino acid turnover and metabolism is this turnover cycle in which proteins are continuously degraded and resynthesized. Co-regulation of synthetic and degradative arms of the cycle is crucial to maintaining cellular viability, to regulation of growth and cellular protein mass, and to control of enzyme levels. At least 20% of basal energy expenditure is used in maintaining whole-body protein synthesis. Body protein mass and rates of protein gain or loss in a cell are entirely dependent on the balance of these mechanistically distinct processes, i.e., the relative rates, of protein synthesis and degradation. Although both processes are influenced by protein and energy nutritional status and by the same hormones (e.g., insulin, growth factors, growth hormone, and glucocorticoids), direction and magnitude of a response of either process are not easily predicted.

Nutritional status, especially amino acid intake, and the response of protein turnover to endocrinological changes interact in a complex way. As a result of these complexities it has proved difficult to identify a common response even when the same outcome variable (e.g., increased protein deposition) is achieved. For example, stimulation of proliferative growth involves a simultaneous increase in protein synthesis and decrease in protein degradation, while hypertrophic growth (e.g., of a muscle in response to increased workload) involves simultaneous increases in both protein synthesis and degradation. Similarly, increases in whole-body protein retention brought about by either increased intake of energy, limiting amino acid, or insulin infusion appear to involve primary changes in whole body protein degradation. On the other hand, separate evidence implies that changes associated with total protein intake or following growth hormone administration involve primarily protein synthesis. Furthermore, the magnitude of changes in whole-body protein turnover, even in response to a common nutritional manipulation, can depend on the prior nutritional status of the individual.

Developmental factors influence regulation of protein turnover as it relates to protein deposition. Protein synthesis appears to be of particular importance to nutritional regulation of growth of immature tissues during childhood, but the response of protein synthesis to protein intake becomes progressively smaller as subjects approach adulthood. In adults, protein degradation seems to be the critical factor regulating protein balance in the short term.

General Nutritional Factors Regulating Amino Acid Catabolism

Essential amino acid catabolism is primarily influenced by the following nutritional factors:

1. The degree to which total nitrogen intake approximates total nitrogen needs of the individual. This factor affects amino acid catabolism in general and is reflected in adaptations in urea synthesis.
2. The degree to which the pattern of amino acids in dietary protein matches amino acid needs. This is reflected directly in the efficiency with which a given dietary protein is utilized in productive processes (e.g., growth, lactation) and is the principal factor underlying differences in biological value of dietary proteins. This factor determines the regulation of the catabolism of individual indispensable amino acids independently of the total. This is the premise underlying recent nutrition interventions with specific high-quality protein intake in infected malnourished children, as a means of preventing amino acid diversion to acute phase protein synthesis.
3. The balance between essential and nonessential amino acids. Dietary indispensable amino acids represent 45% of total amino acid needs for protein deposition and 30% of total for maintenance and the rest consists of dispensable amino acids. Although nonessential amino acids do not have to be supplied in the diet, the organism still has a metabolic need for these nutrients, and if the diet fails to provide them, dispensable amino acids must be synthesized endogenously. An imbalance between dietary essential and nonessential amino acids intake will lead to catabolism of essential amino acids to supply nitrogen for nonessential amino acid synthesis.
4. The degree to which energy intake matches energy needs. Amino acid catabolism is also part of the body's energy supply in order to maintain ATP synthesis. Variations in nonprotein energy intake can have rapid and marked effects on overall amino acid catabolism.

Relationship between Protein Intake and Protein Need

Amino acid catabolism changes rapidly after protein intake. Even in the fed state, amino acid catabolism changes within hours in response to a change in overall level of dietary protein. An important factor in the immediate response to protein intake is the concentration of amino acids. The quantitative relationship between circulating amino acid concentrations and their rate of catabolism is not uniform, either between individuals or between diets. A persistently high or low intake of protein leads to an overall increase or decrease in rate of amino acid catabolism that is partially independent of circulating amino acid concentrations.

Both short- and long-term changes in protein intake alter the levels of insulin, glucagon, and glucocorticoids, all of which are capable of altering the function of amino acid catabolic enzymes. Glucagon, for example, both activates and induces a wide range of amino acid catabolic enzymes. The positive relationship between glucocorticoid level and hepatic amino acid catabolism has been known for many years. There is now additional evidence for direct regulation of catabolic enzyme synthesis by amino acids that is independent of hormonal effects. For example, the initial enzyme of the urea cycle, carbamoyl-phosphate synthase, is immediately responsive to changes in ammonia production via activity of glutaminase, which is in turn activated by ammonia.

Adaptation to Low Protein Intakes

Nitrogen Balance

Nitrogen equilibrium is a state in which, for given intake of nitrogen, an equivalent amount of nitrogen is lost from the body via urine, feces, skin, sweat, etc. In general, when protein intake is low, dietary protein is used more efficiently, urea nitrogen excretion is reduced, and amino acid synthesis pathways are stimulated. The liver plays an important role in the adaptive process since it is the only organ that can transform the nitrogen from amino acids into urea. The metabolic activity of the gastrointestinal tract is important in this adaptive process. Normally, one-third of the urea produced is passed into the bowel and can be hydrolyzed by the microflora. As a process of adaptation to reduced protein intake, the body retains a greater proportion of urea. Similarly, the intravascular circulating albumin mass is maintained by reduced breakdown and a shift of albumin from the extravascular to the intravascular compartment.

Factors Affecting Adaptation

Among the factors that can affect the adaptation to low protein intake are infections, diarrheal disease, and injuries. In infections, protein from muscle and skin is needed for the immune response and synthesis of acute phase proteins. This immunostimulation can lead to an overall negative nitrogen balance. **Table 1** indicates the relative contribution of various

Table 1 Estimated amino acid requirements for synthesis of some acute phase proteins (grams of amino acid per kilogram of protein)

Amino acid	*C-reactive protein*	*Fibrinogen*	*α-Acid glycoprotein*	*α-Antitrypsin*	*Serum amyloid A*	*Haptoglobulin*
Valine	77	48	46	59	18	84
Leucine	91	62	101	124	29	82
Isoleucine	54	32	48	49	29	47
Threonine	58	60	74	66	30	54
Tryptophan	42	35	30	11	45	32
Phenylalanine	105	46	64	83	103	30
Serine	84	91	31	49	47	40
Arginine	36	84	52	23	116	28
Alanine	31	29	36	43	106	54
Lysine	71	77	75	92	33	92
Histidine	16	27	17	37	35	38
Cysteine	13	15	18	6	0	24
Tyrosine	50	56	74	27	67	70
Methionine	16	32	11	28	16	22
Proline	44	48	34	41	34	44
Glycine	46	59	19	33	61	44

Data from Reeds PJ, Fjeld CR, Jahoor F. Do the differences between the amino acid composition of acute-phase and muscle proteins have a bearing on nitrogen loss in traumatic states? *J Nutr* 1994; **124**: 1754S–1764S

amino acids to the production of acute phase proteins.

In other disorders such as injuries or burns, there may be more severe direct losses of nitrogen and altered adaptation. Energy balance is critical for nitrogen balance because of its nitrogen-sparing effect. Thus, if protein deficiency is accompanied by energy deficiency, the adaptation to a low protein intake cannot be achieved completely.

The process of adaptation is clearly dependent on prior nutritional status and overall protein deficits or reserves. It is estimated that the body of a human adult (65 kg) contains 12 kg of protein, about 50% of which is found in muscles. The well-fed human adult can lose about 3 kg of protein without disturbances to his or her health. The amount of body protein depends on, among other things, the dietary protein and carbohydrate intake; if carbohydrate is lacking, the amino acids are utilized for gluconeogenesis.

Protein reserves are not comparable to special fat depots, and not all body proteins can serve as protein reserve. Reserves are primarily organs that contain labile body protein such as liver, plasma (with protein such as albumin and enzymes), and the gastrointestinal tract. Although the protein turnover rate in muscle is very slow, this tissue is a very important protein reserve owing to its large mass. In general, however, during protein deficiency the labile body proteins are metabolized first, sparing the reserves. However, when deficiency is long term, all organs are affected to various extents. **Table 2** indicates the rates of loss of protein from various organs and tissues in rats on a protein-deficient diet. **Table 3** shows some different amino acid responses to a range of stresses.

Table 2 Relative losses of protein in different organs and tissues from rats over 7 days

Organs or tissues	*Loss (percentage of primary content)*
Liver	40
Prostate gland	29
Seminal vesicle	29
Gastrointestinal tract	28
Kidney	20
Blood plasma	20
Heart	18
Muscle, skin, skeleton	8
Brain	5
Eyes	0
Testicle	0
Adrenal gland	0

From Kraut H (ed) (1981) Der Nahrungshedarf des Menschen. I-Stroffwechsel Ernabrung and Nahrungshedarf. Energiebedarf, Proteinbedrarf, pp. 140–53, Darmstadt, Steinkopff Verlag.

Table 3 Plasma amino acid response to different disease states and different intakes

	Starvation	*Protein-free diet*	*Infection*	*Malnutrition*
Valine	↑	↓	↓	↓
Leucine	↑	↓	↓	↓
Isoleucine	↑	↓	↓	↓
Phenylalanine			↑	↓
Alanine	↓	↑	↓	↓
Glycine	↑	↑	↓	↓

Causes of Protein and Amino Acid Deficiency

Causes

Although the main cause of protein deficiency is a protein-deficient diet, the disorder can commonly occur in a variety of pathologic states. In particular, the disorder can be seen in the general context of starvation (although the deficits may be both protein and energy) or in disorders where there are specific protein losses from the body as in nephrotic syndrome or after burns.

Secondary protein deficiencies can be ascribed to six causes:

1. Irregular food habits and starvation states; this may be seen in both developed and developing countries in a variety of pathological states.
2. Inability to digest and absorb the protein that is consumed; this occurs in patients with chronic gastrointestinal disorders such as celiac disease, persistent diarrhea, or protein-losing enteropathy.
3. A disturbed protein metabolism, which may exist in patients with cirrhosis of the liver, but also in patients with hormonal disorders or in some cases of diabetes.
4. A continuous loss of protein; this predominates in patients with disease such as chronic renal disease, bleeding, or exudative gastroenteropathy. High losses of albumin into the urine are indicators of the nephrotic syndrome.
5. Increased protein turnover, which is characteristic in cases of systemic infection or fever. In many instances this may be subclinical and associated with protein diversion due to immunostimulation.
6. Enhanced catabolism of protein, with increased nitrogen losses, seen in patients with severe injuries, especially burns, or in postoperative stress.

Principles of Treatment of Protein Deficiency

The dietary treatment of protein deficiency depends on the cause of the deficiency and must depend upon a sound understanding of the underlying disorder. In most instances, isolated protein deficiency due to deficient intake is rare and most deficits include both macronutrients and micronutrients. In this situation, isolated repletion of protein or amino acids alone is inadequate and may even cause harm. This is well illustrated by the great 'protein fiasco' of the past when attempts to provide high protein supplements to malnourished children were found to be both inadequate and deleterious. Similarly, high protein supplements in pregnancy have been shown to actually increase rates of adverse pregnancy outcomes. Thus, the mainstay of treatment in such states of global deficiency includes balanced energy-protein and micronutrient supplementation.

In other instances, protein supplementation is critical. For example, in children with nephrotic syndrome, the daily intake of protein should be increased to 3–4 g kg^{-1} day^{-1} so that hepatic synthesis of albumin can compensate in part for the urinary losses. In other acute circumstances, infusion of albumin can be used to acutely correct deficits and circulatory abnormalities. However, in states of metabolic adaptation, care should be used in increasing protein intakes. For example, in cases of cirrhosis, the protein intake should be restricted to 20 g day^{-1} to reduce the risk of precipitating hepatic encephalopathy.

See also: **Amino Acids**: Chemistry and Classification; Metabolism; Specific Functions. **Protein**: Synthesis and Turnover; Requirements and Role in Diet; Digestion and Bioavailability; Quality and Sources.

Further Reading

Dean RFA and Whitehead RG (1963) The metabolism of aromatic amino acids in kwashiorkor. *Lancet* **1**: 188–191.

Jackson AA and Grimble RF (1990) Malnutrition and amino acid metabolism. In: Suskind RM and Suskind LL (eds.) *The Malnourished Child*, pp. 73–94. New York: Raven Press.

Reeds PJ and Beckett PR (1996) Protein and amino acids. In: Ziegler EE and Fibr LJ Jr (eds.) *Present Knowledge in Nutrition,* 9th edn, pp. 67–86. Washington, DC: ILSI Press.

Reeds PJ, Fjeld CR, and Jahoor F (1994) Do the differences between the amino acid composition of acute-phase and muscle proteins have a bearing on nitrogen loss in traumatic states? *Journal of Nutrition* **124**: 1754S–1764S.

Waterlow JC, Golden M, and Picou D (1977) The measurement of rates of protein turnover, synthesis, and breakdown in man and the effects of nutritional status and surgical injury. *American Journal of Clinical Nutrition* **30**: 1333–1339.

Pulses *see* **Legumes**

Pyridoxine *see* **Vitamin B_6**

R

REFUGEES

R Bhatia and L N Richardson, United Nations World Food Programme, Rome, Italy

Since World War II, more than 100 million people have been forced to flee persecution or the violence of war to seek refuge in neighboring countries or in a different areas of their own countries. The optimism following the end of the Cold War was short-lived because an epidemic of civil conflicts erupted in several areas of the world. In 1993, 47 conflicts were active, of which 43 were internal. Armed conflicts have increasingly affected civilian populations, resulting in high mortality, widespread human rights abuses, forced migration, famine, and total collapse of governance in some countries.

The 1951 United Nations Convention defines refugees as "any person who owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group, or political opinion is outside the country of his nationality and is unable, or owing to fear is unwilling to avail himself of the protection of the country." In 1969, the Organization of African Unity expanded this definition to include persons fleeing from war, civil disturbance, and violence of any kind.

'Refugees' cross international borders, but 'internally displaced persons' (IDPs) do not. However, both groups have been forced to leave their homes and undergo physical and mental trauma as they settle in harsh and unhealthy environments, where they are often unable to take responsibility for their own welfare. The terms refugee and internally displaced persons have major implications for the people concerned, particularly regarding their rights to protection and assistance, which are embedded in international law. The United Nation's High Commissioner for Refugees (UNHCR) is mandated by the international community to protect and provide assistance to refugees. Due to state sovereignty, the internally displaced are not included within UNCHR's mandate. Only on an ad hoc basis, at the request of the secretary general of the United Nations and the nation concerned, does UNHCR provide assistance to IDPs.

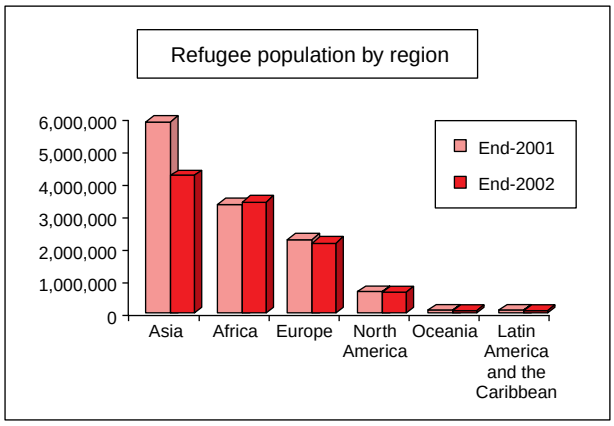
Trends

The escalation in crises and numbers affected since the early 1990s has had a dramatic impact on the nutrition and health of refugees. The number of people affected by natural disasters increased from 50 million in 1980 to 250 million in 2000. Similarly, approximately 30 million people were affected by conflict each year during the 1990s in more than 60 countries. The number of refugees has steadily increased from approximately 5 million in 1980 to a peak of more than 20 million in 1994, with a slow decline by 2003 to approximately 10.4 million. This is primarily due to the fact that more refugees are repatriating than are being forced to leave their countries, and new refugee flows have declined. The large numbers of repatriated refugees from Afghanistan, Angola, and Sierra Leone have contributed to the reduction in the number of refugees. In addition to the large numbers of refugees, in 2003 UNHCR assisted approximately 5.8 million of the estimated 20–25 million IDPs worldwide.

The largest numbers of refugees are in Asia, which is also the region from which more than half the world's refugees originate. Included in this region is Afghanistan, which in 2001 accounted for an estimated 3.8 million refugees, or one-third of the global refugee population. Africa is the second largest refugee region. Approximately 48% of the UNHCR persons of concern are female, 12% are children younger than the age of 5 years, and half of the population is between 18 and 59 years of age (**Figure 1**).

Nutritional Implications of Displacement

The public health and nutrition consequences of war and population displacement have been well documented during the past 25 years. The major determinants of high death rate among affected populations and the major priorities for intervention have also been identified. Clean water, sanitation,



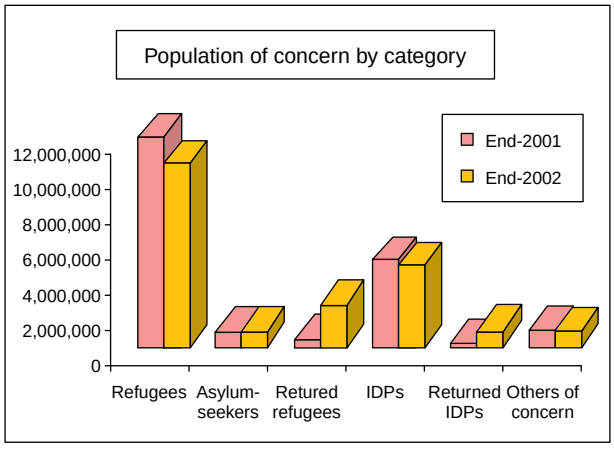


Figure 1 UNHCR Statistics, 2002.

shelter, and immunizations are essential components, but an adequate and diverse food supply remains the central factor in refugee nutrition. Today, it is acknowledged that acute malnutrition is a strong predictor of excess mortality among young children; even moderate malnutrition contributes to increased mortality in emergencies because a larger proportion of the affected population is usually moderately, rather than severely, malnourished; and; micronutrient deficiencies contribute significantly to diseases in emergencies.

Malnutrition results from a lack of food and/or prolonged inadequacies of food consumption, infection, or both. Malnutrition comprises a broad range of clinical conditions in children and adults that result from deficiencies in one or a number of nutrients. It has been defined as a state in which the physical function of an individual is impaired to the point at which he or she can no longer maintain adequate bodily performance processes, such as growth, pregnancy, lactation, physical work, and resisting and recovering from disease. The link between acute malnutrition and excess mortality has been documented for decades. The close correlation between these two factors was demonstrated during a Somali refugee operation in Ethiopia in 1988–89. During the period of peak incidence of mortality and prevalence of acute malnutrition, the food rations provided were less than 1400 kcal/person/day instead of the recommended 1900 kcal/person/day at the time (**Figure 2**).

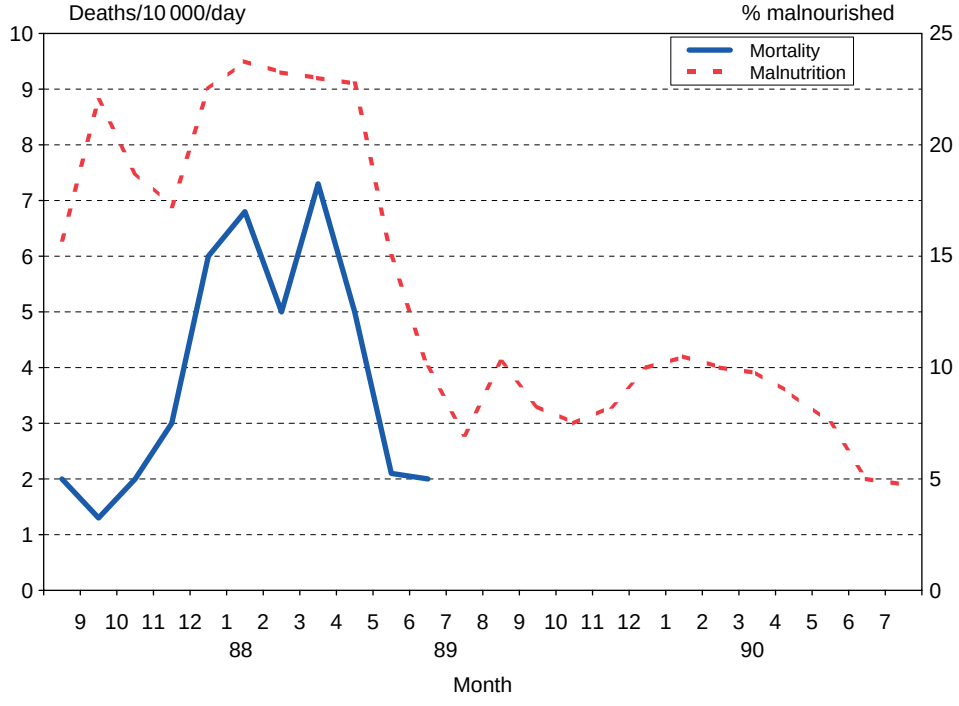


Figure 2 Relationship between malnutrition and mortality as seen in Ethiopia 1987–1990. Morbidity and Mortality Weekly Report (MMWR), July 24, 1992, vol. 41, page 10, Figure 8. Centers for Disease Control, Atlanta, Georgia.

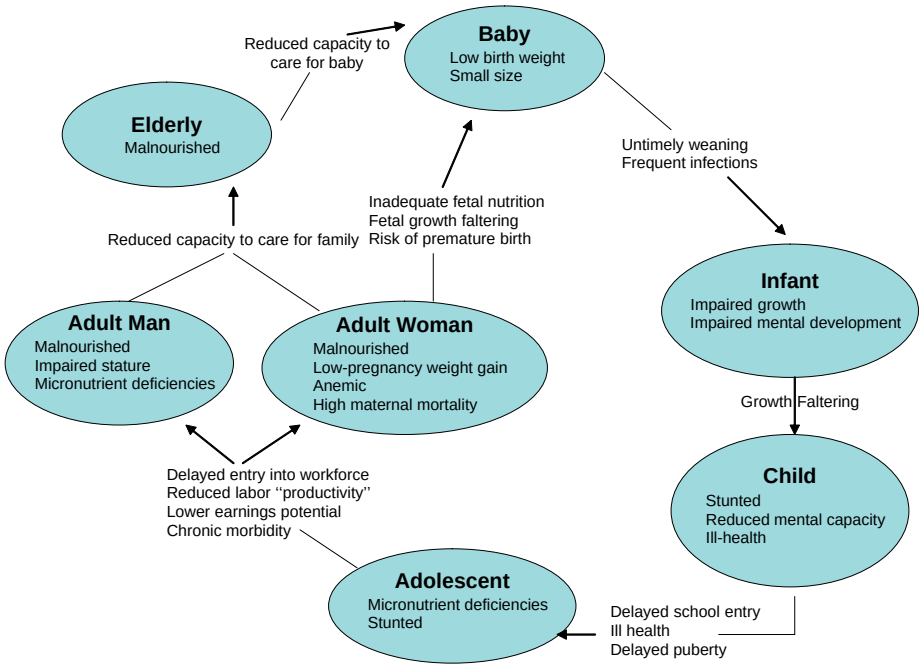


Figure 3 The life cycle and intergenerational transmission of malnutrition. (Adapted from James *et al.* (2000) The 4th Report on the World Nutrition Situation: Nutrition Throughout the Lifecycle. ACC/SCN January 2000. Geneva, Switzerland.)

Although the immediate aim of most food aid programs in refugee emergencies is to prevent excess mortality, there is also increasing evidence that malnutrition during critical periods of life has long-lasting effects. Malnutrition, or the risk of being malnourished, may be carried from one generation to another in an intergenerational cycle. Malnourished women give birth to malnourished infants who, in turn, are more likely to become malnourished adolescents and adults. Therefore, the nutritional status of refugees can have long-lasting effects on future health for individuals and generations (**Figure 3**).

Table 1 Types of malnutrition in refugees

Wasting (acute malnutrition)	Extreme thinness due to recent rapid weight loss; measured by weight-for-height *z* score
Stunting (chronic malnutrition)	Growth failure in a child that occurs over a slow, ongoing process; stunted children are short for their age; measured by height-for-age *z* score
Marasmus	A form of extreme PEM identifiable by severe weight loss or wasting; often there is good appetite and alertness
Kwashiorkor	A form of extreme PEM characterized by oedema, loss of appetite, and apathy; hair thins and may lighten in color to a light brown or red

Macronutrients

An inadequate supply of macronutrients and micronutrients (protein, fat, carbohydrates, and vitamins and minerals) results in protein–energy malnutrition (PEM), the most common form of malnutrition, especially among infants and young children. There are two types of growth failure associated with PEM: wasting (acute malnutrition) and stunting (chronic malnutrition). Wasted individuals (children, adolescents, or adults) are extremely thin, whereas stunted individuals are short for their age as a result of impaired growth during childhood. Severe PEM has a high case fatality rate and is often classified into two forms: marasmus and kwashiorkor. Both are identifiable by severe weight loss; however, the oedema associated with kwashiorkor can mask the otherwise dramatic skeletal appearance of marasmic individuals (**Table 1**).

Nutritional issues among refugees vary greatly from one region of the world to another. Rates of wasting, defined as weight for height less than −2 standard deviations of the reference population, have been as high as 50% in the Horn of Africa and as low as 5% in Southeast Asia, Malawi, and the Persian Gulf. Mortality rates in some of these populations during the acute phase of displacement have been extremely high—up to 60 times the expected rates.

Micronutrients

In addition to PEM. micronutrient deficiencies are key issues in nutrition-related morbidity and mortality. There is a misconception that people do not die of

Table 2 Micronutrient Deficiencies in refugees

Micronutrient	*Deficiency disease*	*Symptoms*
Iron	Anemia	Pallor, tiredness, headaches, breathlessness
Iodine	Goiter	Swelling of the thyroid gland in the neck
	Cretinism	Severe mental and physical disability that occurs in the offspring of women with severe iodine deficiency
Vitamin A	Night blindness	Inability to see well in the dark — an early sign of vitamin A deficiency
	Xeropthalmia	Including Bitot's spots and corneal ulceration and night blindness
Niacin	Pellagra	Affects the skin, gastrointestinal tract, and nervous system and is sometimes called the 3Ds: dermatitis, diarrhoea, and dementia
Thiamine	Beriberi	Loss of tendon reflexes; drooping of arms and feet; wet or cardiac beriberi resulting in heart failure
Vitamin C	Scurvy	Painful joints, swollen and bleeding gums, and slow healing or reopening of old wounds

micronutrient deficiencies because one does not often see the signs and symptoms as visibly as those for PEM. Nonetheless, these deficiencies can be fatal. In addition to deficiency diseases of vitamin A, iron, and iodine, conditions widely described as diseases of public health importance, epidemics of scurvy, beriberi, and pellagra, have been frequently reported among refugee populations, primarily because of limited access to a diverse diet and overreliance on one or two commodities (i.e., maize or polished rice). The importance of micronutrient deficiencies among the refugee population has only been documented since the late 1980s, and more attention is being paid to the usefulness of inclusion of micronutrient-rich foods and/or supplements in the management of refugee nutrition (**Table 2**).

Addressing Nutrition in Refugees

Due to the nature of displacement and the loss of livelihoods, refugee populations are extremely vulnerable. Often, refugees settle in camps with support from the international community and host government. In some cases, refugees may live in open situations in which they integrate into the local community. In almost all cases, refugees are dependent on outside assistance, although the level of need depends on the level of self-reliance the refugees are able to achieve. In some instances, refugees are able to bring some material goods with them when they flee and/or have some sort of income-generating activity, such as access to land and labour and employment. However, this very much depends on the policies of host governments. In these cases, refugees are not totally dependent on food aid, and nutrition management response takes these factors into account by adjusting humanitarian assistance and the food aid ration to meet the assessed needs.

There has been an evolution in the standards of food energy required for refugee populations. These were based on estimates of energy requirements from parameters such as body weight, demographic composition, environmental temperature, and activity levels. In the 1980s, the standard was approximately 1500 kcal/person/day, the minimum deemed adequate for survival. In the late 1980s, this was recalculated to 1900 kcal/person/day as a preliminary standard in order to include expenditure of energy for light activity as opposed to merely the basal metabolic rate. In the 1990s, the benchmark value was modified to a more realistic 2100 kcal/person/day. This was based on an increase in energy required for physical activity, some adjustments to the demographic composition, and an increase in the proportion of pregnant and lactating women in the population. It should be recognized that this recommended value is the average of the individual requirements based on developing nation population demographics and is not a specific provision for individual needs. In addition to the recommended kilocalorie content of the daily food ration, nutritional science has determined that the ration should have an optimal balance of fat and of protein (17 and 12%, respectively).

General Feeding Program

The sudden and massive reduction in food availability associated with displacement immediately affects the nutritional status of refugees. The first response is intervention through the implementation of general food distribution to ensure refugees have access to the required food ration. A general feeding program is the first line of intervention and the highest priority when a refugee population does not have access to sufficient food to meet its nutritional requirements. If the recommended adequate ration of 2100 kcal/person/day is not available, malnutrition levels may escalate.

Correcting Malnutrition

Even if the overall food needs of refugees are adequately met, inequities in the distribution system, disease, and various social factors may contribute to a high degree of malnutrition among certain groups. Children younger than 5 years of age, pregnant and lactating women, the chronically ill (e.g., tuberculosis and HIV/AIDS patients), and the elderly are considered vulnerable groups since they have specific nutritional requirements. A special nutrition intervention or selective feeding program targets these nutritionally vulnerable groups through supplementary feeding programs and those in need of nutritional rehabilitation through therapeutic feeding programs. Malnutrition prevalence rates, as well as an assessment of aggravating factors in the environment, are used as guidelines to determine if a nutrition intervention program needs to be initiated. Aggravating factors that influence the nutritional situation include an elevated crude mortality rate; epidemics of communicable diseases such as measles, diarrhoeal diseases, and respiratory infections; and an unstable social, political, or environmental situation.

Nutrition intervention programs are primarily managed by nongovernmental organizations (NGOs) that have specialization in the management of refugee nutrition. Other humanitarian partners have roles to play in the response to refugee situations, including host country authorities, United Nations agencies such as the World Food Programme, UNHCR, and UNICEF; and other multi-sectoral NGOs. There exist memorandums of understanding and partnership agreements among agencies to provide assistance to the affected populations. The need for partnership is essential for management of refugee nutrition and health programs. The Sphere Project was launched in 1997 to develop a set of universal minimum standards in vital sectors of humanitarian assistance. The aim of the project is to improve the quality of assistance provided to affected populations and to enhance the accountability of the humanitarian system in emergency response.

Supplementary Feeding Programs The most common nutrition intervention is supplementary feeding programs (SFPs) to address malnutrition in emergency situations. SFPs provide a high-quality food as a nutritional supplementation to the daily diet of malnourished populations. There are two main types of SFP—targeted and blanket. The goal of targeted supplementary feeding is to prevent people who are moderately malnourished from becoming severely malnourished. Blanket supplementary feeding, which provides all members of a vulnerable group with a food supplement, is intended to prevent the deterioration of nutritional status among a larger population group rather than narrowly defined individuals at specific nutritional risk. Implementation of SFPs can take two forms: prepared meals consumed on site (wet rations) or food rations issued weekly or monthly to take home for preparation (dry rations). Food supplements usually consist of a fortified blended food (FBF) mixed with oil, and sometimes sugar is included. Wet rations should provide 500–700 kcal, whereas the recommended dry ration is doubled to 1000–1200 kcal in order to account for sharing at home.

Therapeutic Feeding Programs Therapeutic feeding programs (TFPs) provide the severely malnourished with their full nutritional requirements in addition to medical care. They are initiated to reduce excess mortality among individuals facing severe malnutrition and have played an important role in reducing malnutrition-related mortality in emergencies. The first phase of a TFP focuses on treatment of infections, management of other medical complications, and metabolic stabilization. This phase has the highest mortality rate of all nutrition interventions due to the poor state of the patients and the intensive treatment required. The second phase of a TFP is a rapid weight gain period designed to rehabilitate the patient's nutritional status.

Recognition of severe acute malnutrition as a complex nutritional condition during the 1990s led to the development of certain foods defined explicitly for therapeutic treatment of malnutrition with the appropriate balance of energy and protein in order to avoid overloading the body's metabolism, which potentially may lead to cardiac shock. These products include F-75, F-100, and BP-100 biscuits and other ready-to-use therapeutic foods (RUTFs) such as 'plumpy nut.' In addition to balanced energy and protein, these products also contain several vitamins and minerals. For the first stage, TFP F-75 therapeutic milk is used, with the amount of milk given calculated according to the patient's age and weight. During the second stage, F-100 therapeutic milk, with a higher protein content required for rapid weight gain, is used. During the final stage of the treatment, a micronutrient-rich porridge is introduced and eventually, if possible, a family-type meal is introduced in order to reaccustom the patient to the kind of food eaten at home. Those who survive therapeutic treatment generally need further support under SFPs in order to reduce the likelihood of relapsing into a severely malnourished state (**Figure 4**).

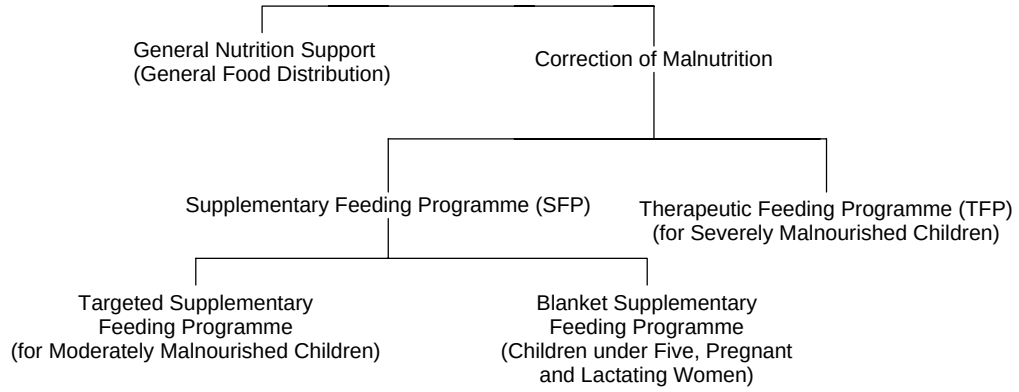


Figure 4 Nutrition interventions. (Adapted from Sphere Project (2004) *Humanitarian Charter and Minimum Standards in Disaster Response*. Sphere Project, Geneva, Switzerland.)

Community Therapeutic Care Community-based care is a recently developed public health approach to dealing with severe malnutrition and aims to treat the majority of people suffering from severe malnutrition in their homes. A community therapeutic care (CTC) programme initially is set up complimentary to traditional TFP components and represents a new approach to managing malnutrition in situations at the community level. A CTC programme has the same initial metabolic stabilization phase, and life-threatening infections are identified and treated just as in a TFP. However, once the patient is stabilized, he or she moves directly to an outpatient therapeutic programme that operates through existing health structures and, with the use of ready to use therapeutic foods (RUTFs), nutritional rehabilitation is initiated. When patients are no longer at risk of severe malnutrition, they are referred to SFPs for recuperation. This phase is followed by greater emphasis on community mobilization to increase the population's involvement and training of mothers who are selected based on their ability to raise well-nourished children even in the face of poverty. After a short training period on nutrition, these mothers team with members from their communities in order to educate them on the fundamentals of successful treatment and prevention of moderate malnutrition and the manufacturing of local RUTFs. At this point, a CTC truly becomes community owned and operated.

CTC is an innovative approach that is still under debate. Proposed benefits of this method are the improved coverage to increase the number of people treated and reduce overall mortality rates. In addition, local production of RUTFs would reduce the cost of treatment, and the shortened length of stay in centres away from the family would have economic and social benefits for the entire family. Finally, the decentralized nature of CTC would allow for earlier detection of

Table 3 Milestones in addressing nutrition in refugees

1960s:	Food response based on commodities available (donated)
	Limited recognition of relevance of nutritional content of rations
	Food provided based on resources rather than nutritional needs
1970s:	Focus on protein deficiency (in protein–energy malnutrition)
	Food ration comprised mainly cereal, pluses/beans, and oil
	Fortified blended foods (FBFs) used only in supplementary feeding
1980s:	Major relief agencies raise planning figure from 1500 to 1900 kcal/person/day
1990s:	Relief agencies raise planning figure for fully food aid-dependent populations from 1900 to 2100 kcal/person/day
	FBF included in most rations for completely dependent populations
	Basic six-commodity food basket becoming common: cereal, pulses, oil, salt, sugar, FBF
	UNHCR/WFP Memorandum of Understanding signed with clear roles and responsibilities
	Development of multi-UN agency and NGO policies and guidelines on common approaches to addressing malnutrition in emergencies
	Fortification of oil, salt, and flours, on international market
	Development of therapeutic foods for treatment of malnutrition (F100-F75)
	Local production of fortified blended foods
2000s:	Development of community therapeutic care approaches
	Development of capacity in nutrition in humanitarian staff
	Pilot testing of on-site milling and fortification in a refugee camp

Adapted from Toole M (1998) *An Overview of Nutrition in Emergencies*, paper presented to the ACC/SCN Working Group on Nutrition in Emergencies, April 11.

malnutrition, thereby reducing the incidence of severe malnutrition (**Table 3**).

Challenges

Nutrition interventions alone are not sufficient to address the multiple causes of malnutrition in refugees. The availability of public health inputs is essential in preventing increased malnutrition and in ensuring that nutrition intervention programmes have their desired effects.

Although the quality of nutritional assistance has improved considerably since the 1970s there is still room to move forward. Because so many factors – food, health, care, and environment — interact to determine nutritional well being, partnerships among agencies with different mandates is essential when trying to effectively address and correct nutritional issues. Refugee nutrition must be addressed in tandem with other services in order to ensure that underlying factors of malnutrition are being met and that nutrition interventions are effective.

See also: **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **Supplementation**: Role of Micronutrient Supplementation; Developing Countries; Developed Countries.

Further Reading

Centers for Disease Control and Prevention (1992, July 24) Famine-affected, refugee, and displaced populations: Recommendations for public health issues *Morbidity and Mortality Weekly Report* **RR-13**.

Collins S (2001) Changing the way we implement famine relief programs. *Lancet*. **358**(9280): 498–501, 2001 Aug 11.

Mason J (2002) Lessons on nutrition of displaced people. *Journal of Nutrition*. **B2**(7): 20963–21035, 2002 Jul.

Medecins Sans Frontiers (1995) *Nutrition Guidelines*. Medecins Sans Frontiers, Geneva, Switzerland.

Medecins Sans Frontiers (1997) *Refugee Health: An Approach to Emergency Situations*. Medecins Sans Frontiers, Paris, France.

Sphere Project (2004) *Humanitarian Charter and Minimum Standards in Disaster Response*. Sphere Project, Geneva, Switzerland.

UNHCR/UNICEF/WFP/WHO (2002) *Food and Nutrition Needs in Emergencies*. Geneva: World Health Organization. Joint publication UNHCR/UNICEF/WFP/WHO.

WHO/UNHCR (1988) *Nutrition in Times of Disaster*, Report of an international conference held at World Health Organization headquarters, Geneva, Switzerland, September 27–30, 1988. Geneva: World Health Organization. Joint publication WHO/UNHCR.

World Health Organization (1999) *Management of Severe Malnutrition: A Manual for Physicians and Other Senior Health Workers*. Geneva: World Health Organization.

World Health Organization (2000) *The Management of Nutrition in Major Emergencies*. Geneva: World Health Organization.

RELIGIOUS CUSTOMS, INFLUENCE ON DIET

P Fieldhouse, The University of Manitoba, Winnipeg, MB, Canada

This article discusses the nature, function, and origins of food practices associated with religious beliefs. A review of historical and contemporary dietary practices in major world religions includes Christianity, Islam, Judaism, Hinduism, and Buddhism and is followed by a brief account of dietary tenets of Jainism, Sikhism, and the Baha'i faith.

The Function and Nature of Religious Food Practices

What people need to eat to survive as biological organisms and what they choose to eat as human beings are two different matters. While practically any food combination which supplies the requisite nutrients to meet physiological needs is adequate for biological purposes, this is clearly not the case for cultural purposes. People make choices from the foodstuffs available to them which reflect a constellation of social, economic, political, and cultural influences, as well as personal preferences. Religion is one such influence, and religious adherents around the world are more or less circumscribed in their food choices by the teachings of their chosen faith.

Religious dietary laws serve a number of different functions. They can provide a way for people to demonstrate their faith—to show that they accept religious authority. As a mark of group identity they strengthen feelings of belonging, and in this way act as a material reflection of the spiritual bonds which link co-religionists.

Conversely, dietary rules may serve to demonstrate separateness by clearly demarcating cultural boundaries between religions. Foregoing food during

religious fasts is a form of self-denial, showing that one is more interested in spiritual than in worldly values. Through sacrifices or sacrificial meals, food is used as a means of communicating with God or other supernatural forces. Offerings may be made to placate the God and so forestall disaster or to seek favors and good fortune. Finally, religious practices may serve, incidentally or purposefully, to encourage ecological sustainability through conservation and judicious use of scarce resources.

Cultural food practices are rich in examples of foods which are not allowed for consumption though they be freely available, and religious codes often exclude whole categories of foods from consumption. What must not be eaten may be determined by characteristics of individuals, such as age, gender, social or physiological status, or by external constraints such as time of day or time of year (**Table 1**).

Table 1 Comparative examples of religious dietary strictures

Religion	*Dietary stricture*
Food restrictions	
Judaism	Eat only animals with cloven hooves and which chew the cud, i.e., cattle, sheep, goats, deer
	Eat only forequarters of animal
	Eat only fish with scales and fins
	No blood
Islam	No blood
	No pork
	No intoxicating liquor
Sikhism	No beef
Hinduism	Must not kill or eat any animal
Days of the year	
Christianity	No meat on Fridays during Lent (Catholics)
	Fast on Wednesday and Friday (Greek Orthodox)
Judaism	No food preparation on Sabbath
Time of day	
Islam	Foods may not be eaten between sunrise and sunset during Ramadan
Buddhism	Monks do not eat after midday
Preparation of food	
Judaism	Ritual slaughtering of animals
	Separate utensils for meat and dairy products
Islam	Ritual animal slaughter
Hinduism	Ritual bathing and donning of clean clothes by Brahmins before eating
Fasts	
Christian	40-day Great Lent fast before Easter and a 40-day Advent fast (Greek Orthodox)
Islam	Month of Ramadan
	13th, 14th, and 15th of each month

Reproduced with permission from Fieldhouse P (1995) *Food and Nutrition: Customs and Culture*, 2nd edn, p. 124. London: Chapman & Hall.

Prescriptive rules of what must be eaten, when and how, are the counterpart of prohibitions. Religious food practices often require the use of specific foods in specific situations, especially during special celebrations such as feasts or fasts, where particular foods often have important symbolic values.

Religious food customs originate in three main ways. Some are required by God and are described in scriptures; others are decreed by religious or political leaders; still others arise through adaptation or co-option of preexisting food practices. Pagan festivals were frequently assimilated and given new meanings as modern religions assumed dominance over older forms of worship. Indeed, religious food practices are far from static; they are subject to continuous adaptation and reinterpretation. Changes may occur as a result of religious reform or revisionism, acculturation, or individual, family, or community adaptations.

Immigration provides a good example of how changing circumstances may result in changing attitudes to food. Through the process of acculturation dietary practices are modified in the light of availability of foodstuffs and as an adaptation to new cultural rules, customs, and expectations.

Continued compliance with traditional rules may depend on social contexts. Adherents who are strict when with members of their own religious group may be willing to be more lax when alone or with a different social group. Even without such external forces, adherence to religious dietary laws or guidelines varies on a national, regional, community, family, or individual level. In many cases religions have developed several branches, sects, or schools of thought which make different demands on members. Such variability in practice should be kept in mind when reading the following descriptions of normative customs. **Table 2** summarizes the size and distribution of contemporary religious followings.

Dietary Practices in Selected World Religions

Christianity

Early Christians observed Mosaic dietary laws. However, the concept of uncleanliness described in Leviticus was rejected by St. Paul and, as Christianity spread across cultural and geographical boundaries, dietary laws largely disappeared (except in the Eastern Orthodox Church). Food, rather than being a way of marking separateness, became a symbol of the communality of religious experience. The celebration of the Eucharist, or Communion, with ritual sharing of bread and wine, though it varies in form

Table 2 Worldwide adherents of all religions by six continental areas, mid-1995

	Africa	*Asia*	*Europe*	*Latin America*	*Northern America*	*Oceania*	*World*	*%*	*Number of countries*
Christians	348 176 000	306 762 000	551 892 000	448 006 000	249 277 000	23 840 000	1 927 953 000	33.7	260
Roman Catholics	122 108 000	90 041 000	270 677 000	402 691 000	74 243 000	8 265 000	968 025 000	16.9	249
Protestants	109 726 000	42 836 000	80 000 000	31 684 000	123 257 000	8 364 000	395 867 000	6.9	236
Orthodox	29 645 000	14 881 000	165 795 000	481 000	6 480 000	666 000	217 948 000	3.8	105
Anglicans	25 362 000	707 000	30 625 000	1 153 000	6 819 000	5 864 000	70 530 000	1.2	158
Other Christians	61 335 000	158 297 000	4 795 000	11 997 000	38 478 000	681 000	275 583 000	4.8	118
Atheists	427 000	174 174 000	40 085 000	2 977 000	1 670 000	592 000	219 925 000	3.8	139
Baha'is	1 851 000	3 010 000	93 000	719 000	356 000	75 000	6 104 000	0.1	210
Buddhists	36 000	320 691 000	1 478 000	569 000	920 000	200 000	323 894 000	5.7	92
Chinese folk religionists	12 000	224 828 000	116 000	66 000	98 000	17 000	225 137 000	3.9	60
Confucians	1000	5 220 000	4000	2000	26 000	1000	5 254 000	0.1	12
Ethnic religionists	72 777 000	36 579 000	1 200 000	1 061 000	47 000	113 000	111 777 000	2.0	104
Hindus	1 535 000	775 252 000	1 522 000	748 000	1 185 000	305 000	780 547 000	13.7	94
Jains	58 000	4 804 000	15 000	4000	4000	1000	4 886 000	0.1	11
Jews	163 000	4 294 000	2 529 000	1 098 000	5 942 000	91 000	14 117 000	0.2	134
Mandeans	0	44 000	0	0	0	0	44 000	0.0	2
Muslims	300 317 000	760 181 000	31 975 000	1 329 000	5 450 000	382 000	1 099 634 000	19.2	184
New-Religionists	19 000	118 591 000	808 000	913 000	956 000	10 000	121 297 000	2.1	27
Nonreligious	2 573 000	701 175 000	94 330 000	15 551 000	25 050 000	2 870 000	841 549 000	14.7	226
Parsees	1000	184 000	1000	1000	1000	1000	189 000	0.0	3
Sikhs	36 000	18 130 000	490 000	8000	490 000	7000	19 161 000	0.3	21
Shintoists	0	2 840 000	1000	1000	1000	1000	2 844 000	0.0	4
Spiritists	4000	1 100 000	17 000	8 768 000	300 000	1000	10 190 000	0.2	30
Other religionists	88 000	98 000	443 000	184 000	1 068 000	42 000	1 923 000	0.0	182
Total population	728 074 000	3 457 957 000	726 999 000	482 005 000	292 841 000	28 549 000	5 716 425 000	100.0	262

from the austere to the elaborate, has retained an underlying significance as a meal shared by the followers of Jesus. The saying of a short prayer before and after eating, establishing a direct connection between God and good food, is also common amongst Christian groups.

Notwithstanding the general deemphasis of dietary laws, certain strictures persisted. Until 1966, Roman Catholics were required to abstain from eating meat on Fridays (since applied only to Fridays during Lent) in symbolic remembrance of the death of Christ. The historic consequence of meat avoidance on Fridays was the regular consumption of fish; fish on Fridays became identified with Roman Catholicism and sometimes fish was deliberately avoided by some other Christian sects who did not wish to be mistaken for Catholics.

In modern times, some Christian sects have established new dietary rules. Seventh Day Adventists are a Protestant sect who emphasize healthful living through eating the right foods and taking exercise and rest. Most Adventists are lacto-ovo-vegetarians. Tea, coffee, alcohol, and tobacco are also avoided. Eating between meals is discouraged on the grounds that the body needs sufficient time to assimilate what is eaten at mealtimes. The religiously inspired food practices, which emphasize cereals, fruits, vegetables, and pulses, has conferred nutritional benefits on the Seventh Day Adventists, who, as a group, have a lower prevelence of chronic diet-related diseases such as heart disease than the general population.

Like the Seventh Day Adventists, members of the Church of the Latter Day Saints (Mormons) assert the importance of eating a well-balanced diet in order to nourish the body as the temple in which the soul resides. Vegetables and herbs are emphasized while meat should be used sparingly. Tobacco, alcohol, and caffeine are avoided. A 24-h fast for those in sound health occurs once a month, and money or food saved is contributed to the welfare of the poor. The fast is a religious discipline and is not a dietary requirement.

In contrast to Western Christianity, Eastern Orthodoxy imposes substantial dietary strictures on its adherents. Dietary laws revolve around fasting. There are two 40-day fasts at Lent and Advent, two shorter summer fasts, and regular fasts on every Wednesday and Friday in the year, excepting the two preceding Ascension Day. Fasts do not require total abstinence but rather prohibit the consumption of all animal products and fish except shellfish. The avoidance of olive oil (historically stored in casks lined with calf stomach) during fast periods is a symbol of true sacrifice and devoutness.

The Great Lent fast commemorates Christ's 40-day fast in the desert and is replete with symbolism. In preparation for the resurrection feast following the fast, hard-boiled eggs are dyed red to symbolize Christ's blood. These eggs are considered to be tokens of good luck and are broken open on Easter morning, representing the opening of Christ's tomb. On Good Friday, lentil soup is eaten to symbolize the tears of the Virgin Mary; often it is flavoured with vinegar as a reminder of Christ's ordeal on the cross. The Great Lent fast is broken after a midnight service on Easter Saturday with a lamb-based soup, olives, bread, and fruit.

Islam

It is probable that the prophet Mohammed, the founder of Islam, adopted existing Jewish practices, for example the prohibition of pork, as a way of encouraging Jewish converts to Islam and to distinguish Muslims from their Christian rivals. The Qur'an, a Holy Book given to Mohammed by Allah, contains dietary regulations which echo those of Judaism. Flesh of animals that are cloven hooved and those that chew the cud is lawful. Pigs, blood, carrion, and foods offered to other idols are forbidden, though one who eats these foods under constraint does not sin. Carnivorous animals and birds which seize their prey with talons are forbidden, as is the flesh of the domestic ass.

Alcohol is prohibited. Fish must be alive when taken from the sea or river, and only fish which have fins and scales are allowed, thus excluding shellfish and eels. To be acceptable to a Muslim an animal must be bled to death while the words 'Bismi 'llahi. Allah Akbar' (I begin with God's name. God is great) are spoken. Such meat is 'Halal,' or lawful, and is stamped with a Halal seal.

Fasting is one of the Five Pillars of Islam and is an important duty of Muslims. It is a way of expressing piety, self-restraint, and freedom from worldly desire and is a means of reaping spiritual rewards. Except for a few holy festival days Muslims may voluntarily fast whenever they wish. Strict adherents fast on Monday and Thursday of every week and on the 13th, 14th, and 15th of each month.

Ramadan, falling in the ninth lunar month, is the major fast of the Muslim year and is one of the most strictly observed of Islamic practices. The word is derived from 'ramz,' meaning 'to burn,' and may derive either from the fact that the fast was first observed in the hot season or because it was believed that fasting would burn away sins. The Ramadan fast involves abstinence from food and water between sunrise and sunset for the whole month and is prescribed for all who have reached the 'Age of Responsibility' (12 years for girls; 15 years for boys). The day's fast should be broken as soon

after sunset as possible and this often takes place at a mosque or at house parties, for it is highly commendable to provide food to others, especially the poor. A morning meal should be eaten as late as is possible prior to sunrise. The end of Ramadan is signaled by the sighting of the new moon and is celebrated with prayers and with feasting.

Certain groups are exempted partially or totally from the Ramadan fast. Anyone who is sick, on a journey, or engaged in hard labour may break the fast, but must make up the days later. Women who are menstruating or are in childbirth are similarly exempted, while pregnant or nursing women and elderly persons in poor health may defer fasting until later in the year or may 'substitute fast' by feeding the poor. Younger children are expected to undertake short fasts in preparation for when they reach the Age of Responsibility.

Judaism

The first five books of the Old Testament, known collectively as the Torah, contain what are probably the most detailed dietary directions of any major religion. These biblical injunctions have been interpreted, elaborated, and added to by rabbis over the past 2000 years. The term 'kashruth,' meaning 'acceptable,' is used to describe anything permitted by Jewish dietary laws. While pig avoidance has become the hallmark symbol of Judaism, it is but one of many restrictions. Only animals having cloven hooves and that chew the cud are permitted. Thus cows, sheep, oxen, and goats may be eaten, whereas pigs, hares, and camels may not. Permitted fish must have fins and scales, which excludes shellfish. Other forbidden foods include teeming winged insects except locusts, certain birds of prey, and bats.

To avoid confusion and to obviate the need to make difficult discriminations between animals, the prohibition was later extended by the rabbis to include all insects and birds of prey. Neither blood nor internal organ fat of otherwise permitted animals may be eaten. The sciatic nerve may not be eaten and, as its removal is difficult, often only the forequarters of an animal is used. The rest of the meat may be sold to non-Jews. Rabbinic additions to the Biblical laws decreed that milk from non-kosher animals is forbidden as it has the same qualities as the animal from which it comes.

Animals dying of natural causes or of disease are not permitted for consumption. Meat must be obtained from animals which have been ritually slaughtered under the supervision of a rabbi. A trained butcher, or shochet, slashes the animals throat with a single cut so as to allow the blood to drain completely from the body. The animal is examined for internal irregularities which might render it unfit for consumption and, if acceptable, is given a seal of approval. Following slaughter, soaking, draining, and salting of the meat ensure that all traces of blood are removed. Historically, Jewish migration was dependent on the availability of kosher meat and thus of access to the services of a shochet.

A prohibition against mixing meat and dairy products is based on the biblical injunction 'You shall not boil a kid in its mother's milk' (Exodus 23: 19). After eating milk, hand washing and mouth rinsing is all that is necessary before eating meat; however, depending on local custom, from 1 to 6 h must elapse after eating meat and before eating milk. Margarine and milk and cream substitutes have made this particular law easier to follow: Nevertheless, many observant Jews view the use of such substitutes as being spiritually wrong. In Israel, there has been a continuing historical struggle over the banning of pig rearing and pork eating. However, these practices continue today, pork being eaten by Christians and nonobservant Jews.

There are many fasts in the Jewish calendar, some of scriptural or rabbinical origin and others which mark private events such as family deaths. Fasting is a way of showing repentance, of teaching self-discipline, or of preparing to seek divine guidance. Generally fasts are observed by boys over the age of 13 years and 1 day and by girls over the age of 12 years and 1 day.

The Sabbath being a day of rest, all food preparation is carried out on Friday. Challah is a traditional Sabbath bread and, in a modern adaptation of an historical practice, two loaves are used to symbolize the double portion of manna provided by God to the Israelites on Fridays during their 40 years in the wilderness. Other festivals, including Rosh Hashannah, the Day of Judgement, and Yom Kippur, the Day of Atonement, are rich in food symbolism, but it is the major festival of Passover that perhaps sees the most elaborate food practices. All leavened products must be removed from the house, reflecting the fact that Jews did not have time to let bread dough rise when they were driven from their homes into exile. Pieces of bread may be deliberately hidden around the house, to be 'discovered' and removed. Unleavened bread called matzah is prepared or bought commercially.

On the eve of Passover it is customary for the firstborn to fast in symbolic remembrance of the historic sparing of the first-born. On the first and second night a special family meal, the seder, is eaten, which is itself a testimony to the symbolic power of food. Seder means order, and the meal indeed has a very definite structure which gives it

its ritual character. Partaking in this food event is an important way of transmitting culturally valued knowledge from generation to generation.

Hinduism

While vegetarianism and the specific prohibition on eating cows are two of the dietary hallmarks of modern Hinduism, early Hindu writings reveal that beef was eaten freely. Beef prohibition may have arisen as a response to the challenge of Buddhism, which was critical of Brahminism and its cattle-sacrificing practices, and so Brahmins championed the sacred cow concept in order to maintain their own position. It was perhaps subsequently strengthened by the need for Hindus to distinguish themselves from their new Muslim neighbours. Certainly by the time of the Rigveda, around 1000 CE, the prohibition was firmly ensconced.

For modern Hindus, the cow is a revered animal, a symbol of motherhood and fertility. The duty to protect cows is enshrined in the Constitution of India; slaughter of cows is prohibited by religious custom and in some provinces by legal bans. In villages and towns cows wander freely, foraging where they may; there are even special government-run homes for old and infirm cows.

Observance of dietary prohibitions is strictest in the upper Brahmin caste while beef (and pork) eating is tolerated amongst lower castes, for whom it may be an important source of supplementary protein. For these castes, giving up meat eating is one way to attempt to improve their social status. The caste system is an excellent example of the way in which social structure influences food practices. Concepts of purity and pollution determine who may eat what with whom and who may accept what food from whom. Eating with, or accepting food from, members of a lower class is polluting. In the hot–cold classification system used in traditional healing raw foods are considered to be 'hot' and are therefore purer than cooked foods, which are 'cold.' Brahmins who accept food cooked by a lower caste person lose ritual purity and thus caste status; however, they may accept ghee (clarified butter) and milk—for these are products of the sacred cow and cannot be polluted by touch.

Devout Hindus are vegetarian; meat, fish, and often eggs are avoided, the latter especially by women. An exception is made for the Ksatriya warrior caste, who may consume meat without loss of status, a concession probably related to notions of meat, strength and military prowess. Most castes will eat fish, though its actual consumption varies dramatically from region to region. Where fish is eaten, white fish is preferred for it is least like meat. Abstention from eating food is a much praised virtue. Some Hindus may fast 2 or 3 days a week, during which time they may eat only pure foods such as milk, fruit, nuts, starchy roots, and vegetables. Fasts are associated with calendric, caste, family, and personal events as well as with religious celebrations.

There are literally thousands of festival days celebrated by Hindus in different parts of India. Some are national, some regional, and some purely local. On festival days there may be great processions, when people visit shrines to pray and make offerings. Only food which is pacca (contains ghee) may be offered; it thereby becomes blessed and is then distributed to the waiting crowd. Ghee, as a product of the cow, is sacred and is an important component of many rituals. For example, during marriage ceremonies a ritual flame is kept burning with ghee.

The modern Indian festivals of Dussatra and Divali illustrate the agricultural origins of many contemporary celebrations. Dussatra marks the end of the rainy season when agricultural labour must begin again, as well as commemorating the legendary hero, Rama. It includes a ritual quest for alms by people carrying small fresh stalks of barley plants. Divali, known as the festival of lights, is a new year festival celebrating the sowing of winter crops. Lamps are lit and gifts are exchanged.

Buddhism

Founded in the sixth century BCE by Siddartha Gautama, Buddhism became the state religion of India in 250 BCE, though it is now a minority religion there, Buddhism has been the most influential of religious forces in spreading the practice of vegetarianism and has developed in different ways in many parts of the world, especially Southeast Asia, Tibet, China, Japan, and Korea.

Buddhism championed the concept of Ahimsa—noninjury to all living creatures—and had a profound influence on the subsequent development of Hinduism through its stand against cruelty toward animals and animal sacrifice. Buddhists refrain from eating meat or from harming any living creatures, though the prohibition is observed strictly only by monks and the devout laity. Lay Buddhists may eat meat or may raise meat for sale to non-Buddhists. Animals found dead may be eaten, as may fish (which are not 'killed' but merely removed from the water).

In contrast to other religions, Buddhism places emphasis on wrongful killing rather than wrongful eating. Thus flesh may be eaten if it was not procured for eating purposes or supposed to have been so. Monks personify the ideal. Food is obtained through begging, and it is meritorious for the laity to voluntarily provide food to the monks, thereby assisting in their own spiritual progress.

Other Religions

Jainism is an ascetic Asian religion whose adherents advocate ahimsa, or noninjury, both as an ethical and philosophical goal, and whose example has had a strong admonitory influence on non-Jains. The Jaina monastic community has a number of characteristic practices which evince an extreme regard for life. Monks carry a small brush with which they carefully sweep the floor before sitting or lying so as to avoid crushing any insects. They may wear masks to prevent inadvertent inhalation of small creatures, and strain their drinking water. Wild honey is avoided, as bees may be killed during its collection.

Ascetics have few or no possessions and must beg for food. Some choose in old age to die through ritual fasting. Most Jainas are not ascetics, although some strive to imitate monastic ideals by pursuing a progressive path of renunciation, leading to rebirth as an ascetic. The nonmonastic Jaina community practices vegetarianism and opposes the killing of animals. Because of this, agricultural and military occupations are not suitable to Jainas, who historically have chosen instead to enter the professions or to take up business interests.

Sikh means disciple, a follower of the 10 gurus. Sikhism was founded in the fifteenth century CE by the guru Nanak who proclaimed, "There is no Hindu; there is no Muslim." Nanak rejected the social distinctions of the Hindu caste system and required his followers to eat together as a symbol of unity. Sikhs retain the Hindu reverence for cows and thus do not eat beef. Other meat may be eaten, although some Sikhs are vegetarians. Permitted animals must be killed with a single blow, 'jhatka,' literally a sudden shake or jerk. Generally, Sikhs are not rigid about adherence to dietary laws and readily adapt to the food customs of other cultures.

Food plays a part in the *Baha'i* religion through fasts and feasts and through dietary injunctions which favour vegetarianism and abstention from alcohol. Vegetarianism is held to be a compassionate practice and one which is in line with God's will, though meat is not actually prohibited. Fasting is viewed as a spiritual undertaking and is symbolic of abstinence from selfish and carnal desires. It is deemed to have both physical and spiritual benefits. A 19-day fasting period occurs in March, during which there must be complete abstention from food and drink between sunrise and sunset. Exemptions are granted for travellers, the sick, pregnant and nursing women, while those engaged in heavy work may also be excused. Children under the age of 15 years and elderly persons over 70 years of age are not required to fast but may choose to do so. Obeying the fast is a matter for individual conscience and is not enforced; if food is eaten 'unconsciously' during the fast it is deemed to be an accident rather than a breaking of the fast. Unlike in Islam, there is no making up of fast days missed as the fast can only be kept during the designated time.

Nineteen Day Feasts (so called because they are held on the first day of each of the 19-day-long Baha'i months) bring Baha'is together to consult and discuss and to offer suggestions to their Local Spiritual Assembly. At a feast, which is entered into with right thinking, the 'heavenly food' of knowledge, understanding, love, and kindness is present, providing members with a sense of spiritual restoration. The significance of providing food at these gatherings appears to be related to the Baha'i injunction to serve one's fellows—rather than to the social solidarity of sharing food. The absence of guidelines or restrictions on what may be served indicates that it is the act of serving which is symbolic, not the food itself.

See also: **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Coronary Heart Disease**: Hemostatic Factors. **Food Choice, Influencing Factors**. **Socio-economic Status**. **Vegetarian Diets**.

Further Reading

Douglas M (1970) In *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo*. Harmonsworth: Penguin.

Fieldhouse P (1995) *Food and Nutrition: Customs and Culture*, 2nd edn. London: Chapman & Hall.

Grivetti LE (1991) Nutrition past—nutrition today: Prescientific origins of nutrition and dietetics. *Nutrition Today,* Jan/Feb 13–24; July/Aug 18–29; Nov/Dec 6–17.

Khare RS and Rao MSA (eds.) (1986) *Food, Society and Culture: Aspects in South Asian Food Systems*. Durham: Carolina Academic Press.

Sherman S (1991) The Passover seder: Ritual dynamics, foodways, and family folklore. In: Humphrey TC and Humphrey LT (eds.) *We Gather Together: Foods and Festival in American Life*. Ann Arbor: UMI Research Press.

Simoons FJ (1994) *Eat Not This Flesh*, 2nd edn. Madison: University of Wisconsin Press.

Respiratory Diseases *see* **Cancer**: Epidemiology of Lung Cancer. **Lung Diseases**

Retinol *see* **Vitamin A**: Biochemistry and Physiological Role; Deficiency and Interventions

RIBOFLAVIN

C J Bates, MRC Human Nutrition Research, Cambridge, UK

Absorption, Transport, and Storage

Riboflavin (vitamin B_2) is not synthesized by higher animals. Therefore, it is an absolute dietary requirement for the synthesis of certain essential coenzymes that are needed for intermediary metabolism in nearly all living cells. Riboflavin must be transported from the food sources within the gastrointestinal tract, across the gut wall into the circulatory system, and thence into the cells of each organ. This transport process occurs against a concentration gradient, in order to ensure the efficient retrieval of the very small amounts that occur in many foods, and from the low concentrations in plasma to higher concentrations inside living cells.

Gut riboflavin transport systems have been studied by partly isolated segments of the small intestine within an anesthetized animal; by an isolated everted gut segment, or by 'vesicles', prepared from the 'brush border.'

Studies with these model systems have shown that the transport of riboflavin at low (e.g., micromolar) concentrations is temperature- and energy-dependent (it is inhibited by inhibitors of ATP production from energy substrates), it becomes saturated as the concentration of riboflavin increases, and it is sodium ion dependent. These characteristics are shared with many other types of small molecules that are actively transported across the gut wall. More specifically for riboflavin, the active transport mechanism involves phosphorylation (to riboflavin phosphate, also known as flavin mononucleotide, or FMN) followed by dephosphorylation, both occurring within the intestinal cells (**Figure 1**). This latter process is not shared by several other B vitamins, but it is one of a number of common strategies which the gut may use to entrap essential nutrients, and then relocate them, in a controlled manner and direction. A similar strategy is employed at other sites in the body, to ensure entrapment of circulating riboflavin by cells whose nascent flavin-dependent enzymes need a supply of the vitamin from beyond their borders.

Although the active transport of riboflavin across the gut wall and across other cell membrane barriers within the animal is a saturable process, if large pharmacological amounts are present then the slower and less efficient but nonsaturable process of passive absorption predominates and contributes significantly to the total mass transfer. The active transport process is increased in riboflavin deficiency and decreased if the riboflavin content of the tissues is high. The transport pathway involves calcium and calmodulin but not sodium. Specific riboflavin receptors have recently been identified, as has a role for microtubules in transport.

Although some of the available riboflavin in natural foods may be present as the free vitamin, ready for intestinal transport, a larger fraction is present in the form of phosphorylated coenzymes: FMN and flavin adenine dinucleotide (FAD), and there may also be very small amounts of a glucoside of the vitamin. These forms are all efficiently converted to free vitamin by enzymes secreted into the gut lumen, and they are therefore highly available for absorption. There are also small amounts of covalently bound forms of riboflavin, present in enzymes such as succinate dehydrogenase (succinate: ubiquinone oxidoreductase EC 1.3.5.1), which cannot be released by the hydrolytic enzymes in the gut and are therefore unavailable for absorption. Also unavailable (or very poorly available) in man is the riboflavin synthesized by the gut flora of the large bowel. Certain animal species such as rodents can utilize this riboflavin source by coprophagy.

A wide variety of artificial analogs of riboflavin have been prepared in order to explore the structural versus functional essentials of the molecule. Some of these analogs have riboflavin-like activity; others are inactive, while a number are antagonists, and can cause functional deficiency. These structural changes can affect absorption or the conversion of riboflavin to its coenzyme forms within the body. Certain drugs that are used for purposes unrelated to riboflavin function, such as the phenothiazines used as antipsychotic drugs, may have sufficient structural similarity to riboflavin to act as antagonists in some situations.

Absorption by Human Subjects

Studies of riboflavin absorption by human subjects require a combination of a test dose, usually taken by mouth, and a sampling procedure to estimate the amount absorbed, and possibly also its subsequent fate. The sampling compartment is generally the

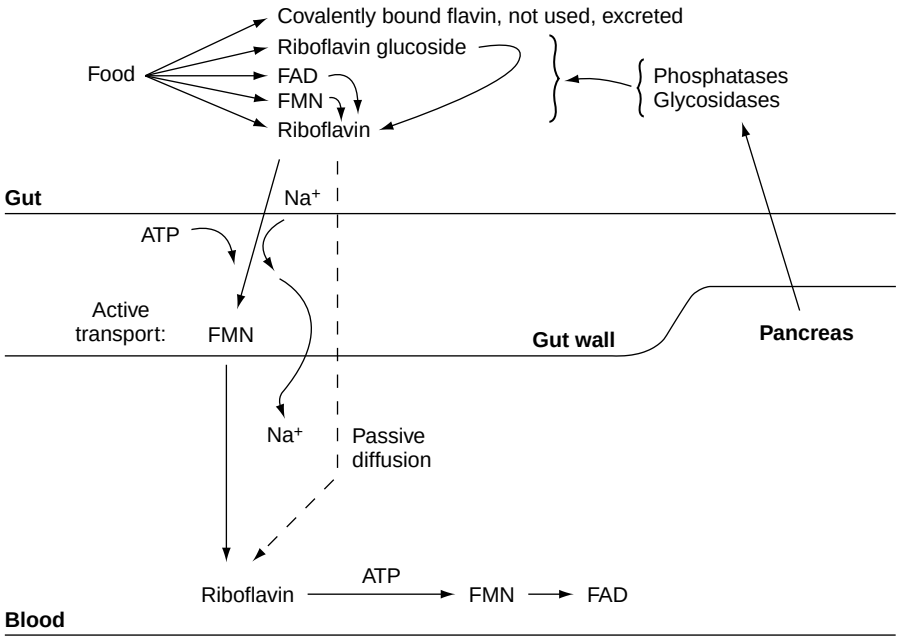


Figure 1 Characteristics of the absorption process for riboflavin and its coenzymes.

urine, since plasma has proved generally unsatisfactory. Fecal sampling is also useless because of the synthesis of riboflavin by bacteria in the large bowel. Although the use of riboflavin labeled with radioactive or stable isotopes is theoretically possible, this has not yet been applied to human studies. The majority of reported studies have relied on relatively large 'bolus' oral doses of riboflavin, comprising at least several milligrams, with urinary monitoring over the subsequent few hours. Riboflavin can be quantitated in urine by its very characteristic fluorescence, or by microbiological assay, or more accurately by high performance liquid chromatography.

For the maximum absorption of a test dose of riboflavin, the duration of exposure within the upper ileum is critical, since this is the region of greatest absorptive efficiency. There is little evidence to suggest that slow-release forms of the vitamin can enhance its absorption, but there does appear to be some absorptive advantage for certain synthetic lipophilic esters, such as the tetrabutyrate ester, which is hydrolyzed to the free vitamin during or after absorption. These have been shown to possess beneficial (e.g., antioxidant) properties in some model systems, but their usefulness in human medicine is still at a very early stage of assessment. The concomitant presence of food can enhance absorption, possibly by increasing the transit time. There is little evidence that the efficiency of absorption varies markedly with age or sex in humans.

Measurements of the plasma pool of riboflavin following test doses is not a viable method of measuring absorption, because redistribution to other tissue sites plus urinary excretion takes place too rapidly for this pool to be representative of the amount absorbed. Although the urinary response to a test dose has been the most commonly used approach to studies of intestinal absorption in humans, it suffers from the potential disadvantage that physiological intakes, and especially low intakes of riboflavin from 'poor' food sources, cannot be measured by this technique. Such studies of small doses are however needed, in order to determine the factors that modulate riboflavin absorption in developing countries, where dietary sources of riboflavin are minimal and clinical signs of riboflavin deficiency are common. A much more sensitive biochemical marker of riboflavin status at low intakes is the index known as 'erythrocyte glutathione reductase activation coefficient' (EGRAC), which will be discussed in greater detail below. It is possible to achieve a graded response to graded intakes of riboflavin, and studies of absorption efficiency using this alternative marker as the outcome measure may become feasible (and useful) in the future.

Riboflavin Transport at other Sites and Storage

As mentioned earlier, nearly all tissues require riboflavin. The free vitamin is trapped as one of its phosphorylated coenzyme forms, which then become specifically associated (and in a few cases covalently linked) to the protein chains of catalytic flavoenzymes. If not already covalently linked, the

flavin coenzyme can often be liberated by extremes of pH or by other nonphysiological maneuvers. In a few biological locations, such as the mature red cell, flavoenzymes such as glutathione reductase (NADPH: oxidized glutathione oxidoreductase EC 1.6.4.2) may exist partly in their apoenzyme form, i.e., without the flavin coenzyme and therefore without enzyme activity. An increased supply of riboflavin will permit the depleted coenzyme (in this case FAD) to be synthesized so that enzyme activity can be restored.

Different enzymes and different tissue sites differ in the tenacity with which they can retain flavin coenzymes in times of riboflavin deficiency, so there is a characteristic 'pecking order' for flavoenzyme protection, which appears to reflect the metabolic importance of the different metabolic pathways affected. Apart from this 'pecking order,' however, there is no repository of unused or nonfunctional riboflavin that can act as a 'store' in times of dietary deficiency. Although some organs (such as liver) have relatively high concentrations of flavin enzymes, all of the flavin seems to be present as coenzyme moieties of flavin holoenzymes. Each tissue has a characteristic 'ceiling' level of riboflavin at saturation, and a 'floor' level characteristic of severe depletion, and these are determined, respectively, by the total amount of apoflavoprotein, and the amount of 'resistant' holoenzyme, which cannot be depleted of its cofactor during riboflavin deficiency.

Riboflavin is secreted into milk, the concentration being species specific and to a moderate extent dependent on maternal status and intake. Riboflavin is also required by the developing fetus during pregnancy, which implies a need for active transport from the maternal to the fetal circulation; the flavin concentration being greater on the fetal side. Studies from India have identified a riboflavin carrier protein (RCP) present in bird (e.g., chicken) eggs, which is considered to be specific for riboflavin, and is essential for normal embryological development. If this protein is rendered ineffective (e.g., by immunoneutralization) by treatment of the bird with a specific antibody, then embryonic development ceases and the embryo dies. A genetic mutant lacking RCP is likewise infertile. A homologous protein, which can be rendered ineffective by the antibody to pure chicken riboflavin carrier protein, has been shown to occur in several mammalian species, including two species of monkeys, and also in humans. Very recent studies have suggested that circulating RCP levels and the immunohistochemical staining of RCP in biopsy specimens may provide new markers for breast cancer diagnosis and prognosis. Termination of pregnancy has been demonstrated by immunoneutralization of RCP in monkeys. There remains some controversy over the roles of RCP, however, and other, less specific riboflavin binders in blood, including gamma-globulins, also seem to play an important role. These studies have provided an intriguing example of the role of specific vitamin-transporting mechanisms, designed to ensure that the vitamin needs of the developing embryo will be efficiently met. Further evidence of the special needs of developing embryos has been provided by the demonstration that riboflavin analogs can cause teratogenic changes, even in the absence of any detectable damage to maternal tissues.

Metabolism and Excretion

The interconversions of riboflavin with its coenzyme derivatives are depicted in **Figure 2**. Clearly, the 'high-energy nucleotide' ATP is a cosubstrate and driving force (in energy terms) for both stages of the conversion to FAD. Some flavoenzymes specifically require

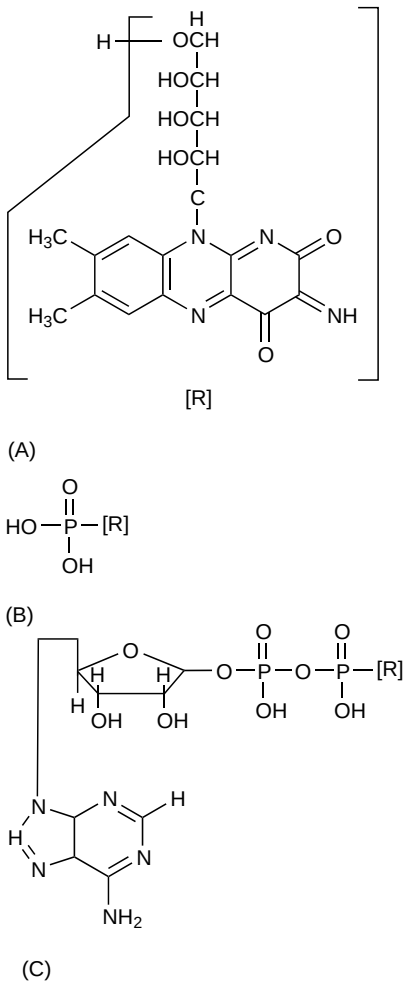


Figure 2 Structure of riboflavin and its coenzyme derivatives. (A) Riboflavin; (B) riboflavin phosphate (flavin mononucleotide, FMN); (C) flavin adenine dinucleotide (FAD).

Table 1 Categories of flavoenzymes

Category (subcategory)	*Example*
1-Electron transferases	Mitochondrial electron-transfer flavoprotein
Dehydrogenases	
Pyridine nucleotide dehydrogenases or reductases	Mitochondrial NADH dehydrogenase
Nonpyridine nucleotide dehydrogenases	Mitochondrial succinate dehydrogenase
Pyridine nucleotide-disulfide oxidoreductases	Glutathione reductase
Dehydrogenase-oxygen reductases	
Flavoprotein oxidases [$O_2 \rightarrow H_2O_2$]	Monoamine oxidase
Flavoprotein monooxygenases (1/2 $O_2 \rightarrow H_2O$)	
Internal (α-hydroxy fatty acid $\rightarrow$ fatty acid (n-1) + CO_2)	Bacterial lactate monooxygenase
External (RH $\rightarrow$ ROH)	Microsomal FAD-containing monooxygenase

Source: Merrill AH, Lambeth JD, Edmondson DE, and McCormick DB (1981) Formation and mode of action of flavoproteins. *Annual Review of Nutrition* **1**: 281–317.

FAD while others specifically require FMN; it is difficult to account for this dichotomy. **Table 1** lists the broad categories of flavoenzymes found in living tissues: the range of reaction types is considerable, but all of them clearly center around redox processes involving hydrogen transfer. This fact reflects the central biochemical reaction of the flavin coenzymes, which is the interconversion of the reduced, dihydro form of the flavin ring and the more stable oxidized form. One of the most important sites of action of flavoenzymes within higher animals is that of the electron transport chain in the mitochondria. The flavins, which form part of succinic dehydrogenase and NADH dehydrogenase, form an essential redox link between the oxidizable energy-rich substrates of aerobic metabolism, and the cytochrome chain leading to molecular oxygen, which can generate around 38 moles of energy-rich ATP per mole of glucose oxidized.

Hormone status can affect riboflavin economy in a number of important ways, and there is also some evidence that riboflavin status can affect hormone production. One important control valve for riboflavin economy is thyroid hormone status: hypothyroidism leads to reduced tissue levels of flavin coenzymes, and hence to inactivation of certain flavoenzymes, thus resembling the effects of dietary riboflavin deficiency. Both flavokinase (ATP: riboflavin 5′-phosphotransferase EC 2.7.1.26) and FAD pyrophosphorylase (ATP: FMN adenyltransferase EC 2.7.7.2) are sensitive to thyroid hormone status. In the kidney, synthesis of flavokinase and hence of flavin coenzymes is controlled by aldosterone in a similar manner.

The amount of absorbed riboflavin that can remain within the body and the circulation (in blood plasma) is strictly regulated by glomerular and tubular filtration and tubular reabsorption in the kidneys. The latter is an active, saturable, sodium-dependent transport process, with characteristics similar to those of active transport in the gastrointestinal tract. It is responsible for the very sharp and characteristic transition between minimal urinary excretion of riboflavin at low intakes, and a much higher level of excretion, proportional to intake, at higher intakes. This transition point has been extensively used to define and to measure riboflavin status and requirements (see below), and to permit studies of intestinal absorption *in vivo* (see above). Excretion of riboflavin is affected by some chemicals (such as boric acid, which complexes with it), and by certain diseases and hormone imbalances.

In addition to the excretion of unchanged riboflavin, there are also small amounts of hydroxylated breakdown products of the vitamin, which arise through normal turnover, either within the tissues of the body, or in the gastrointestinal tract from bacterial action, before absorption. The rate of destruction of riboflavin by this turnover pathway is very low in all species examined to date, and riboflavin within the mammalian body seems to be remarkably efficiently conserved, apparently throughout many cycles of cell turnover.

Metabolic Function and Essentiality

This section will deal in detail with the question: what goes wrong when deficiency of riboflavin is encountered, and how does riboflavin interact with other nutrients, or with biochemical and physiological processes, thus producing characteristic functional effects?

Fatty Acid Oxidation

The first example of serious metabolic disturbance seen in moderate riboflavin deficiency is a disturbance of fatty acid oxidation. The normal first stage in the spiral process of beta-oxidation of fatty acids within the mitochondria is the removal of two

hydrogen atoms from the two carbons located alpha- and beta- to the activated carboxyl end of the chain. The fatty acyl coenzyme A substrate is acted upon by one of several fatty acyl CoA dehydrogenase flavoprotein enzymes (e.g., long-chain acyl-CoA:(acceptor) 2,3-oxido-reductase EC 1.3.99.13), each of which is specific for a small range of acyl chains. The second stage in this process involves transfer of the electrons via another flavoenzyme, known as electron transferring flavoprotein dehydrogenase (electron transferring flavoprotein: ubiquinone oxidoreductase EC 1.5.5.1) and thence to the cytochrome chain and to oxygen. These flavoenzymes, unlike the flavoenzymes that are linked to carbohydrate oxidation, are highly sensitive to dietary riboflavin depletion. Characteristic disturbances of lipid metabolism can therefore be detected in riboflavin-deficient tissues and organisms.

Disturbances in fatty acid oxidation by isolated mitochondria, e.g., from livers of deficient animals, have been demonstrated, and one of the most characteristic metabolic changes, observed even in a mild deficiency state in experimental animals, is the appearance of abnormal dicarboxylic acids, and their derivatives, in the urine. These products seem to arise because fatty acyl intermediates become diverted away from the usual pathway of mitochondrial beta-oxidation, towards abnormal partial oxidation in the peroxisomes (which are less severely affected by the riboflavin deficiency state).

Genetically normal human subjects have not yet been shown to accumulate these abnormal urinary products but, in contrast, humans who bear an abnormal gene resulting in dicarboxylic aciduria do quite frequently respond to riboflavin supplements, showing a reduction in their excretion of abnormal fatty acid products. It seems that additional dietary riboflavin can help to overcome the inherent genetic abnormality, presumably by providing more of the coenzyme, and thereby making sure that all of the residual fatty acid oxidation machinery is working at its optimum capacity. Interestingly, the accumulation of dicarboxylic acids in urine is characteristic of riboflavin-deficient mammals but not of birds; chick embryos deprived of riboflavin via the genetic lesion of riboflavin carrier protein seem to die in hypoglycemic shock, but do not exhibit dicarboxylic aciduria.

Iron Economy

An important interaction of riboflavin with iron economy has been suspected for many years, partly because iron-deficient animals failed to respond readily to iron supplements if they were also riboflavin deficient, and also because the redox system involving riboflavin and its coenzymes has been shown to interact very readily with the redox system between ferric and ferrous iron.

Some recent studies in experimental animals have shown that not only is there evidence for some impairment of absorption of iron in riboflavin-deficient animals, and of its distribution between discrete compartments within the body, but also, more surprisingly and strikingly, a major increase in rates of iron loss from the intestinal mucosa, resulting in impaired retention of the body iron stores. This enhanced rate of iron loss is accompanied by hyperproliferation of crypt cells, and increased cellular transit along the villi, leading to an excessive proportion of immature villi, and probably also to a reduction in absorptive area. These studies begin to explain how a combination of iron deficiency and riboflavin deficiency, which is frequently encountered in human populations in many developing countries, may lead to a gradual deterioration of iron status, which is often accompanied by other intestinal lesions and by impaired gut function.

Riboflavin enhances the hematological response to iron, and deficiency may account for at least some of the anemia seen in human populations. Unlike iron-deficiency anemia, the anemia of riboflavin deficiency is reported to be normocytic and normochromic.

Malaria

Low dietary riboflavin intakes are frequently encountered in malarious areas of the world, and in a small number of studies there has arisen the apparently paradoxical observation that biochemical riboflavin deficiency is associated with a lower level of blood cell parasitemia than is encountered in riboflavin-replete subjects. Although neither animal nor human studies have indicated that riboflavin deficiency protects from the life-threatening sequelae of malaria, there does appear to be some interaction between the parasite and flavins within the blood cells, which is not yet fully understood. Interestingly, too, some of the prophylactic drugs used to prevent malaria infection have riboflavin-like structures.

Cataracts and Photoreceptors

Several micronutrients, especially those that can have antioxidant functions in living tissues, have recently been investigated in relation to possible protection against degenerative eye diseases, such as cataract. Studies in animal models have suggested, albeit indirectly, that riboflavin status may be important here, and several recent epidemiological studies, including an intervention study in one region of China, have supported the suggestion that

good riboflavin status, or riboflavin supplements, may be protective. Although the evidence must be considered as tentative and incomplete at the present time, this possibility clearly deserves further study.

Another intriguing role of flavoproteins in the eye involves a photoreceptor function that synchronizes circadian rhythms with the solar light–dark cycle, specifically via cryptocromes 1 and 2, which contain FAD and function as blue-light-sensitive photoreceptors.

Interaction with Vitamin B_6

There are several metabolic interrelationships between riboflavin and vitamin B_6. The conversion of pyridoxine or pyridoxamine phosphates to pyridoxal phosphate is catalyzed by a flavoenzyme (pyridoxaminephosphate oxidase EC 1.4.3.5), so that a deficiency of riboflavin may, at certain key sites, result in a secondary deficiency in B_6-dependent pathways. More evidence is needed to clarify the extent and importance of these interactions.

Effect on Folate Metabolism

Riboflavin in the form of FAD is an essential coenzyme for 5,10-methylene tetrahydrofolate reductase, a key enzyme of the folate pathway, which catalyzes the interconversion of 5,10-methylene-tetrahydrofolate and 5-methyltetrahydrofolate. Of the known single nucleotide polymorphisms affecting this enzyme, the best known are the C699T and A1298C variants. The former confer thermolability and potentially reduced enzyme activity in the TT homozygote. Marginal riboflavin status may, in some situations, be associated with increased plasma homocysteine levels (possibly predictive of increased vascular disease risk) that can arise as a result of reduced activity of this key enzyme of folate metabolism in TT subjects. Another example of a common gene polymorphism that affects related pathways is encountered with methionine synthase. These examples of gene–nutrient interactions, which may affect a sizeable proportion of some human populations, illustrate an area of increasing research effort, in which a synergism between a common genetic subtype and a marginal nutrient deficiency or imbalance may confer increased functional risk.

Physical Activity and Neuromuscular Function

Several recent studies have documented an apparent increase in riboflavin requirements accompanying an increase in physical exercise in human subjects. This may reflect the fact that anabolic influences and the accretion of new lean body mass creates a demand for the vitamin, for mitochondrial accretion. Intriguingly, there is also some evidence that indices of neuromuscular function, as illustrated by 'hand-steadiness, may be influenced by riboflavin status in communities where riboflavin deficiency is endemic. If confirmed, this might raise the possibility that peripheral neurological function could be affected by riboflavin status in mammals, including humans, as it is in birds.

Assessment of Riboflavin Status

Assessment of status for specific nutrients such as riboflavin is closely bound up with the estimation of requirements in human individuals and groups of subjects, and with the monitoring of human populations for evidence of the adequacy of their intakes. It is often cheaper, easier, and more accurate to collect a sample of blood or urine from an individual and carry out biochemical analyses that determine status than to carry out reliable measurements of intake over a period of time, since the latter requires considerable cooperation from the subject, and is also affected by uncertainties of food table nutrient values, in relation to specific foods and diets.

Biochemical status estimates are generally based upon urinary excretion or measurements of erythrocyte glutathione reductase (NADPH: oxidized glutathione oxidoreductase EC 1.6.4.2) and its reactivation with flavin adenine dinucleotide (FAD) in red cell lysates. Other biochemical indices, such as plasma or red cell flavin concentrations, have been less widely used, but their potential is increasing with the advent of new assay techniques such as capillary electrophoresis with highly sensitive laser-induced fluorescence detection. Functional indices directed towards the products of flavin-requiring pathways *in vivo* are not in common use except for the investigation of errors of metabolism or rare diseases. The two principal traditional status tests are considered below:

Urinary Excretion

As noted earlier, the amount of riboflavin excreted in the urine is negligible at low intakes of the vitamin. As the dietary level rises there is slow increase to a transition point, above which the slope of the excretion rate increases very sharply, and then remains proportional to intake until absorption is saturated. For population studies it has been found convenient to use the creatinine excretion rate as the denominator, and the suggested interpretation of urinary riboflavin excretion rates is: $<27\,\mu g$ riboflavin per g creatinine for deficient; $27–79\,\mu g\,g^{-1}$ for low; and $>80\,\mu g\,g^{-1}$ for acceptable. Detailed studies of the

relation between intake and excretion rate have recently shown that this index is sufficiently sensitive to distinguish riboflavin requirements between people on low-fat, high-carbohydrate diets and the slightly higher requirement associated with high-fat, low-carbohydrate diets. Metabolic states associated with general tissue catabolism can sometimes result in liberation of riboflavin during cell turnover; this increases its urinary excretion, even though dietary intake may be low.

The Glutathione Reductase Test

One rather serious drawback of the urinary excretion index for the assessment of riboflavin status is that it is relatively insensitive at low-to-moderate intakes, because the rate of excretion changes slowly and not very predictably with increasing intake in this region. Another important practical drawback is that 24-h urine samples are not easy to collect and excretion rates may fluctuate over short time periods. A more stable index was therefore sought and was identified in the degree of unsaturation of the red blood cell enzyme, glutathione reductase (NADPH: oxidized-glutathione oxidoreductase EC 1.6.4.2), with respect to its flavin cofactor, flavin adenine dinucleotide (FAD) (**Figure 3**).

As noted earlier, an inadequate supply of dietary riboflavin results in low circulating levels, and hence a gradual progressive loss of cofactor from this red cell flavoenzyme over a period of several weeks. Since the enzyme protein (apoenzyme) remains intact and reactivatable by FAD, it is possible to remove a small sample of blood, collect, wash and hemolyse the red cells, and then measure glutathione reductase activity with and without the FAD cofactor. If the individual is riboflavin replete, then the added FAD has almost no effect and the 'activation coefficient,' or ratio of FAD stimulated to unstimulated activity ('EGRAC') is between 1.0 and 1.3–1.4. If the individual is deficient, then added FAD produces a larger stimulation, and the 'activation coefficient' is higher. For people living in communities with very low intakes of riboflavin and a significant prevalence of clinically recognizable deficiency, activation coefficients as high as 2.0–3.0 are quite common. In Western countries, few values as high as 2.0 are encountered. However, recent population surveys in the UK have indicated that the proportion of values between 1.3 and 1.8 is considerable across all age ranges. Whether this apparent evidence of marginal deficiency has a technical, assay-related explanation, or is a result of decreasing intakes of riboflavin-rich foods, such as cows' milk, is uncertain.

This test has the advantage that it is highly sensitive to, and predictive of, the extent of tissue depletion in the range of severe-to-moderate deficiency; it is robust and requires only a small sample of blood, and it can be automated by modern enzyme rate reaction analyzers. Riboflavin supplements given to deficient subjects result in rapid and reproducible restoration of the saturated condition of the enzyme,

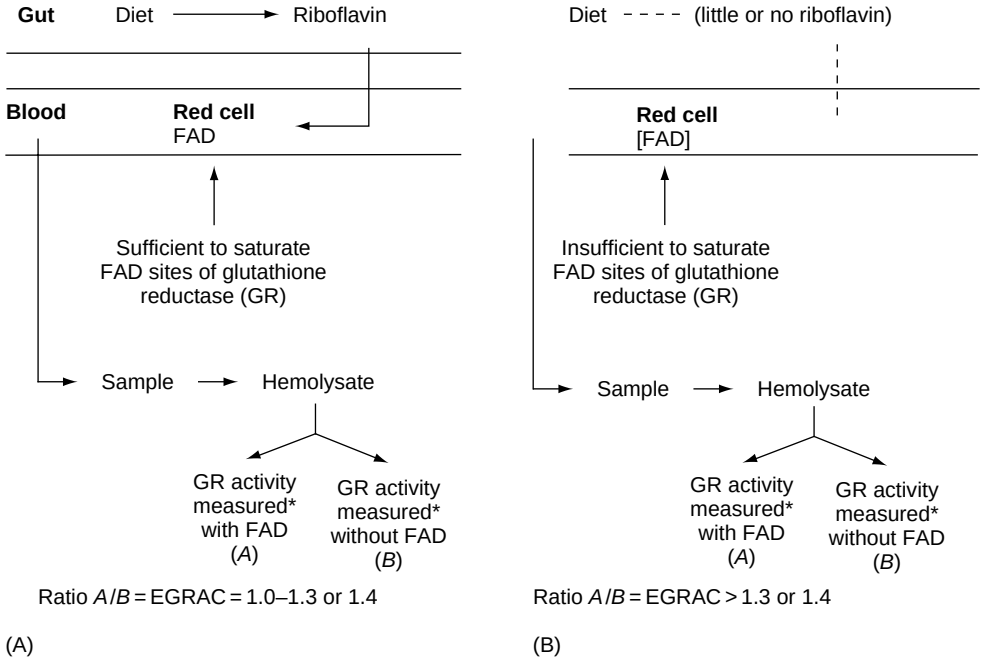


Figure 3 Basis of the glutathione reductase for riboflavin status: (A) riboflavin sufficient; (B) riboflavin deficient. *Reaction of oxidized glutathione with reduced nicotinamide adenine dinucleotide phosphate.

and graded supplements can be given to estimate human requirements.

There are minor operational differences among different published versions of the analytical procedure for 'EGRAC,' which result in small between-laboratory differences in the interpretation of the normal range and there are also some instances of specific factors that cause ambiguity of interpretation. The best known of these is the human genetic variant resulting in glucose-6-phosphate dehydrogenase (D-glucose-6-phosphate: NADP+ 1-oxidoreductase EC 1.1.1.49) deficiency. Both homo-and heterozygotes are affected, and in such subjects erythrocyte glutathione reductase becomes almost saturated with FAD, even when they are riboflavin deficient. Other tests of status, such as high-performance liquid chromatography (HPLC) measurement of riboflavin in blood fractions, are then required.

Some groups of people have increased requirements for riboflavin, which is related to special metabolic states. There is, for instance, a progressive increase in requirement during pregnancy, followed by a decrease during lactation. Babies exposed to phototherapy for neonatal jaundice also have increased requirements. In certain circumstances, the use of oral contraceptives may increase requirements, but the evidence is conflicting. The largest and most dramatic increases in requirements have been seen (as noted above) in a subgroup of people with inborn metabolic errors leading to dicarboxylicaciduria and associated clinical abnormalities. Although certain drugs are known to affect riboflavin status indices, there is no clear consensus on the question of the need for supplements by people who are prescribed such drugs – clearly this area needs further study.

Requirements

As with most micronutrients, the evidence on which requirement estimates are based can be subdivided into the following broad classes of criteria:

1. prevention of clinical (pathological) deficiency;
2. attainment of specified blood levels or tissue stores of riboflavin;
3. titration to the urinary excretion threshold;
4. tests based on cofactor saturation of one or more accessible, diet-sensitive, flavin- dependent enzymes, such as erythrocyte glutathione reductase; and
5. physiological function.

Of these five classes of criteria, the first has been useful in defining 'minimum' requirements, but as a practical test of status it has several drawbacks. First, clinical signs of deficiency in human communities tend to be nonspecific and multifactorial. Second, studies resulting in clinical deficiency in controlled trials would be ethically unacceptable. Third, the classical clinical deficiency signs such as angular stomatitis and cheilosis do not always correlate closely with, or respond rapidly to, changes in dietary riboflavin supply or biochemical evidence of deficiency. Additional factors such as local infection are likely to be critical.

Use of physiological functional indices in relation to riboflavin deficiency (analogous to dark adaptation for vitamin A; clotting factors for vitamin K, etc.) has not proved possible, because the analogous riboflavin-sensitive physiological processes are insufficiently specific or easily measurable for use in population studies. Of the biochemical indices, urinary excretion and the flavin-dependent enzyme, erythrocyte glutathione reductase, are generally considered to be the front-runners in the race for acceptance in human studies. These have already been described in the previous section.

For avoidance of clinical deficiency signs in normal healthy adults, the basic requirement for riboflavin is 0.55–0.8 mg day^{-1}. The UK reference nutrient intake for riboflavin is l.3 mg day^{-1} for men and l.l mg day^{-1} for women, rising to l.6 mg day^{-1} during pregnancy and lactation. For formula-fed infants, the reference intake is 0.4 mg day^{-1}. Requirements may increase to some extent as a result of heavy exercise or dieting, and abnormal status has been observed in the presence of anorexia nervosa.

In the US, the current recommended dietary amounts (RDAs) are 1.3 mg day^{-1} for men and 1.1 mg day^{-1} for women, rising to 1.4 mg day^{-1} in pregnancy and 1.6 mg day^{-1} in lactation, with proportional amounts, based on metabolic body weights and growth requirements, for children and adolescents. RDAs are 20% higher than estimated average requirements (EARs) for each group.

Dietary Sources and High Intakes

Table 2 lists the riboflavin contents of some commonly consumed foods in Western countries. As is the case with most other B vitamins, the richest food sources comprise items such as offal and yeast extract, with meat and dairy products also providing quite generous amounts; fruit and vegetables somewhat less, and the smallest amounts, in relation to their energy content, being present in ungerminated grains and seeds, such as nuts. There is an enormous difference in intakes and in status observed between most Western countries, where the dietary intake tends to be quite generous, and many developing countries, which depend on monotonous and riboflavin-poor staple foods such as polished rice. In developing countries, riboflavin deficiency tends to

Table 2 Riboflavin content of selected foods

Food	*mg per 100 g fresh wt*	*mg per MJ*
Meat, offal, and fish		
Stewed minced beef	0.19	0.22
Grilled pork chop	0.16	0.21
Calf liver, fried	2.89	3.94
Lamb's kidney, fried	3.10	3.95
Cod, grilled	0.06	0.15
Dairy products		
Cows' milk, full cream	0.23	0.84
Cheese, cheddar	0.39	0.23
Yogurt (whole milk, plain)	0.27	0.81
Boiled chicken's egg	0.35	0.57
Human milk	0.03	0.10
Fruits		
Apples, eating, flesh and skin	0.02	0.10
Oranges, flesh	0.04	0.25
Pears, flesh and skin	0.03	0.18
Strawberries, raw	0.03	0.27
Dried mixed fruit	0.05	0.04
Vegetables		
Potatoes, boiled, new	0.06	0.19
Carrots, boiled, young	0.01	0.11
Brussel sprouts, boiled	0.09	0.59
Cauliflower, boiled	0.04	0.34
Onions, fried	0.01	0.01
Grains, grain products, nuts		
White bread	0.08	0.09
Wholemeal bread	0.05	0.05
Rice, boiled, white	Trace	Trace
Comflakes (Kellogg)	1.3	0.81
Baked beans in tomato sauce	0.06	0.35
Peanuts, plain	0.10	0.04
Other		
Marmite (yeast hydrolysate)	11.9	15.6
Bovril (beef hydrolysate)	8.5	11.2

Compiled and calculated from data in Food Standard Agency (2002) *McCance and Widdowson's The Composition of Foods*, 6th Summary edn. Cambridge: Royal Society of Chemistry.

be widespread. Although even a severe riboflavin deficiency is less obviously life-threatening than some other types of malnutrition that are commonly encountered in the Third World, it can nevertheless cause debility, through skin lesions and metabolic dysfunctions, and riboflavin-nutriture thus deserves an important place in future public health programs.

As with most other B vitamins, riboflavin and its cofactors are remarkably nontoxic even at high intakes. The reasons for this are probably associated with limitations on absorption, once the active transport process has become saturated in the gut; coupled with very effective urinary excretion of any absorbed vitamin that is in excess of cellular requirements.

See also: **Antioxidants**: Diet and Antioxidant Defense. **Fatty Acids**: Metabolism. **Iron**. **Vitamin B_6**.

Further Reading

Bates CJ (1987) Human riboflavin requirements, and metabolic consequences of deficiency in man and animals. *World Review of Nutrition and Dietetics* **50**: 215–265.

Bro-Rasmussen F (1958) The riboflavin requirement of animals and man and associated metabolic relations. 1. Technique of estimating requirement, and modifying circumstances. 2. Relation of requirement to the metabolism of protein and energy. *Nutrition Abstracts and Reviews* **28**: 1–23; 369–386.

Friedrich W (1988) Vitamin B_2: riboflavin and its bioactive variants. In: *Vitamins*, ch. 7, pp. 403–471. Berlin: Walter de Gruyter and Co.

Institute of Medicine (2000) *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin and Choline*, pp. 87–119. Washington DC: National Academy Press.

Massey V. (2000) The chemical and biological versatility of riboflavin. *Biochemical Society Transactions* **28**: 283–296.

McCormick DB (1989) Two inter-connected B vitamins: riboflavin and pyridoxine. *Physiological Reviews* **69**: 1170–1198.

Merrill AH, Lambeth JD, Edmondson DE, and McCormick DB (1981) Formation and mode of action of flavoproteins. *Annual Review of Nutrition* **1**: 281–317.

Powers HJ (1995) Riboflavin-iron interactions with particular emphasis on the gastrointestinal tract. *Proceedings of the Nutrition Society* **54**: 509–517.

Powers HJ (1999) Current knowledge concerning optimum nutritional status of riboflavin, niacin and pyridoxine. *Proceedings of the Nutrition Society* **58**: 435–440.

Rivlin RS (2001) Riboflavin. In: Bowman BA and Russell RM (eds.) *Present Knowledge in Nutrition*, 8th edn, ch. 18, pp. 191–198. Washington, DC: ILSI Press.

Rivlin RS and Pinto JT (2001) Riboflavin (vitamin B_2). In: Rucker RB, Suttie JW, McCormick DB, and Machlin LJ (eds.) *Handbook of Vitamins*, 3rd edn, ch. 7, pp. 255–273. New York: Marcel Dekker Inc.

Ross NS and Hansen TPB (1992) Riboflavin deficiency is associated with selective preservation of critical flavoenzyme-dependent metabolic pathways. *Biofactors* **3**: 185–190.

White HB and Merrill AH (1988) Riboflavin-binding proteins. *Annual Review of Nutrition* **8**: 279–299.

Rickets *see* **Vitamin D**: Rickets and Osteomalacia

Roughage *see* **Dietary Fiber**: Physiological Effects and Effects on Absorption; Potential Role in Etiology of Disease; Role in Nutritional Management of Disease

S

Salt *see* **Sodium**: Physiology; Salt Intake and Health

Satiety *see* **Appetite**: Physiological and Neurobiological Aspects

Saturated Fat *see* **Fatty Acids**: Saturated

SEASONALITY

F Branca and P D'Acapito, Istituto Nazionale di Ricerca per gli Alimenti e la Nutrizione, Rome, Italy

Introduction

The agroclimatic characteristics of some areas of the planet with lower technological development lead to seasonal fluctuations in food production and food availability, labor demand, and incidence of disease that affect the nutritional status of rural populations. Such cyclical stress has a negative impact on the well being, productivity, and potential for development in individuals with a pre-existing marginal nutrition. Approximately 400 million adults and 500 million children in the world may be affected.

Definition and Measures of Seasonality

Seasonality is defined as the cyclical change of food availability and agricultural labor induced by climatic changes in rural areas of least developed countries (LDC). In seasonal climates a bad, lean, or slack season, also known as the *saudure* (French, meaning junction between two agricultural cycles), is present for 2–3 months a year, often coinciding with rains, leading to cyclical stress on the health and nutrition of rural populations. Agroclimatic seasonality is relevant in populations practising subsistence agriculture and among hunter-gatherers, or in other agricultural systems in which background food security is poor; such as in areas where cash crops are mainly planted. Human interventions can alter this pattern, by changing the water and sun exposure conditions, with irrigation and greenhouses, but these techniques are not accessible to the majority of peasants in LDC.

Changes in rainfall, temperature, exposure to winds, and relative humidity, with respect to the water retention capacity of the soil, are responsible for a cyclical change of water balance that may restrict the period for plant growth to some parts of the year. The proportion of dry months in a year, named absolute seasonality, can vary between 0 (sufficient rains all year long) to 1 (lack of a period suitable for plant growth). If the vegetative cycle of corn (maize) is considered (120 days), then the areas of the world can be classified as follows: low seasonality when there are more than 200 days' vegetative season per year and two harvests are possible; moderate seasonality with 120–200 days' vegetative season and one to two corn

harvests possible; and severe seasonality with less than 120 days' vegetative season and barely one corn harvest possible. With even shorter vegetative periods, agricultural production is impossible in the absence of irrigation. The different areas of agroclimatic seasonality in the world are shown in **Figure 1**.

Factors aggravating the climatic seasonal effects may be the occurrence of pests, for example, the arrival of locusts arriving with the rainy season in Sahel. Furthermore, seasonal patterns of food production are often superimposed upon longer term cycles, which leads to the periodic appearance of drought and famines in sub-Saharan Africa and in Central Asia.

Effects of Agroclimatic Seasonality on Food Availability and Dietary Intakes

Subsistence farmers store their harvest and use it progressively until the next season, so that a fluctuation of food stocks can be observed during the year. Stores decrease as a result of human consumption, animal consumption, losses due to pests, rodents and microbiological contamination, and sales, barter or donations. The market price of food staples is also subject to great seasonal fluctuations, which are usually inversely related to the size of the domestic stocks.

As a result, both energy supply and dietary quality may be affected. The dietary changes observed in slack seasons may involve eating foods that are less preferred, but are more affordable, an option that is not biologically dramatic, but perceived by people as stressful. Households that are close to exhaustion of stocks of staple foods may use alternative food sources (root crops, gathered leaves and fruits, hunted small animals), may consume immature grains, or may reduce food intake by limiting portion size, reducing the number of meals, or skipping meals for an entire day. Dietary changes may also be due to reduced time available for food collection and preparation.

In areas of high climatic seasonality the magnitude of the reduction of energy intake can be in the order of 400–500 kcal and is associated with a reduction in protein and micronutrient intake. In

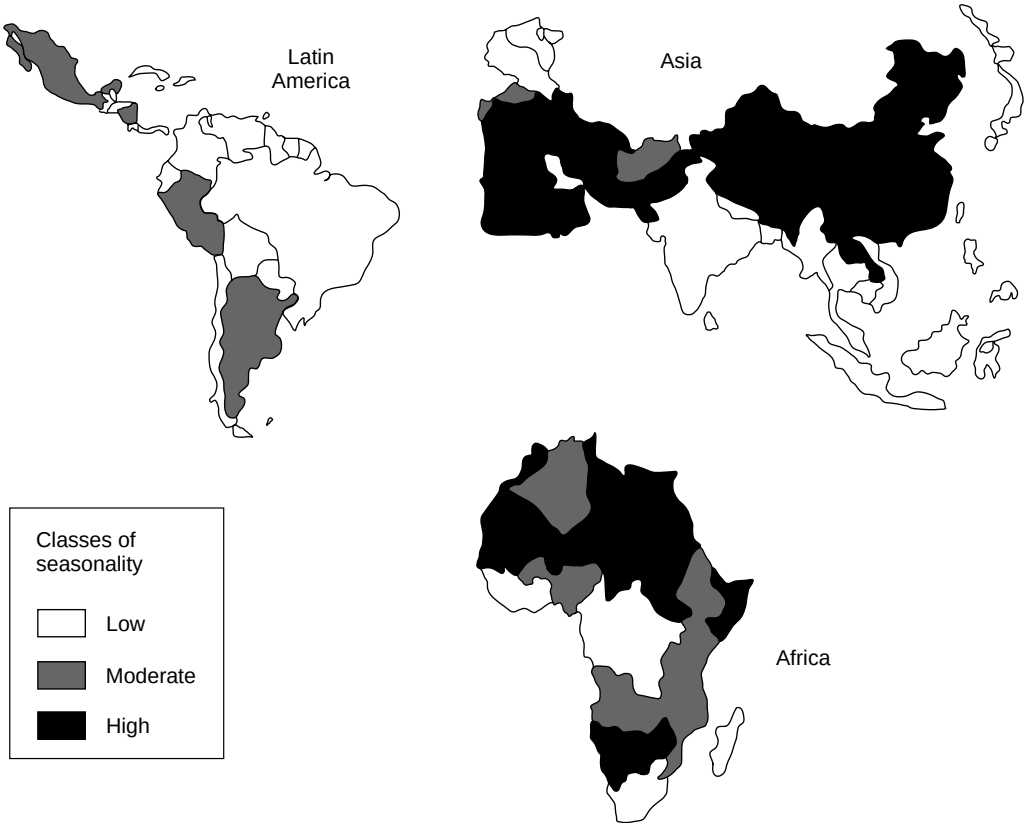


Figure 1 Agro-climatic seasonality in different world regions. Seasonality is calculated taking into account soil characteristics and water balance and is an expression of the vegetative period of a food crop allowed by such conditions. From Ferro-Luzzi A, Branca F, and Pastore G (1994) Body mass index defines the risk of seasonal energy stress in the Third World. *European Journal of Clinical Nutrition* **48**(supplement 3): S165–S178. Reprinted with permission of the Nature Publishing Group.

Cameroon, the rainy season is associated with a 20% reduction of energy intake, a 50% reduction of protein intake, and a 25% reduction of fat intake. However, the reduction in food intake is not a universal characteristic of areas prone to agroclimatic seasonality. In Mali energy intake in adults had minimal fluctuations between 11.7 MJ in the harvest season and 11.4 MJ in the rainy season.

Changes in dietary intake may affect household members unequally. For economic or social reasons some household members may be protected. In Northeastern Thailand energy intake in the rainy season is reduced in women but is constant in men. Children may be protected by maternal buffering, but may also be discriminated against in favor of male workers, who have to ensure the family's food supply. A study in Bangladesh showed that energy and proteins intakes were significantly different in male and female adults and in 1 to 4-year-old children, but were unchanged in older children. In another Bangladesh study (**Figure 2**) dietary energy intake was reduced in all age groups, including pregnant and lactating women, but the reduction in women (−28%) was far greater than that in men (−18%).

It seems that women consistently get the least food during seasonal shortages, regardless of their physiological status. In The Gambia a 12% reduction of intake has been observed in pregnant women and 29% in lactating women during the rainy season.

Seasonality does not spare weanlings, who in Bangladesh have a 33% difference in energy intake between highest and lowest intake seasons, or breast-fed babies, in whom a decrease in breast-milk intake has been observed. A decrease in breast-milk output in the early postpartum period (2–6 months) has also been observed in The Gambia during lean seasons: the daily output was $850\,g\,day^{-1}$ in a cohort of women who gave birth during the dry season and $540\,g\,day^{-1}$ in a second cohort who delivered during the wet season. Furthermore, breast-milk fat concentration was $3.95\,g\ 100\,ml^{-1}$ in the first cohort and $3.52\,g\ 100\,ml^{-1}$ in the second cohort, with a resulting decrease in energy content. The decrease in breast-milk output is not necessarily related to the deteriorating nutritional status of the lactating mothers, but rather to the limited time available for childcare,

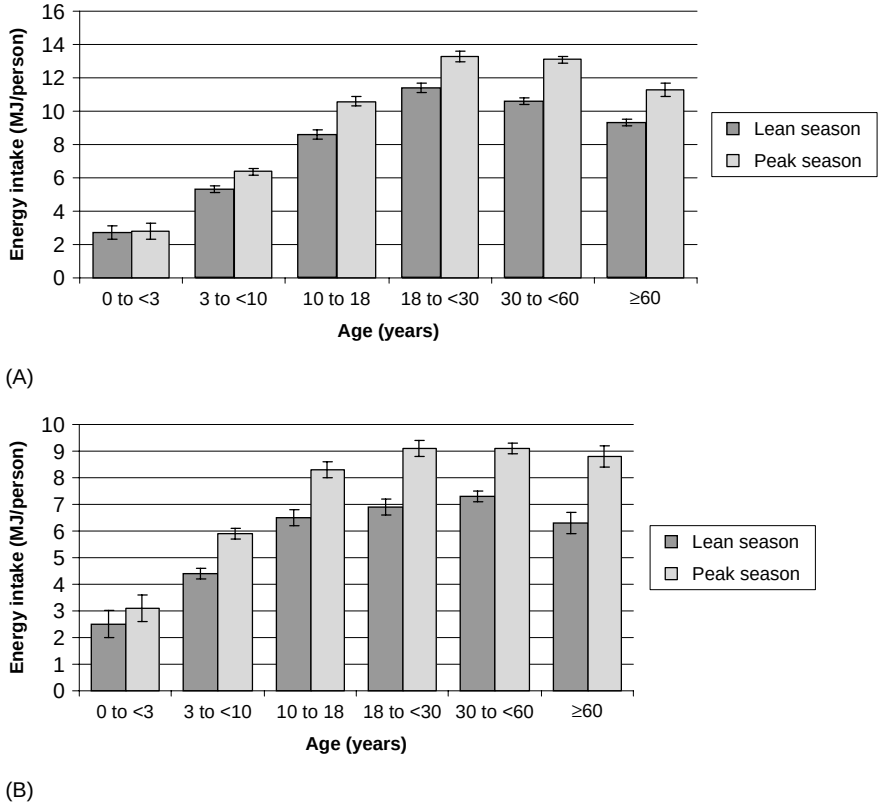


Figure 2 Seasonal changes in dietary energy intake in different age groups in Bangladesh: (A) males; (B) females. (Data from Tetens I, Hels O, Khan NI, Thilsted SH, and Hassan N (2003) Rice-based diets in rural Bangladesh: how do different age and sex groups adapt to seasonal changes in energy intake? *American Journal of Clinical Nutrition* **78**: 406–413.)

with a resulting reduction in suckling time. This latter circumstance has been specifically documented in Bangladesh.

The reduction in diet diversity may affect micronutrient intake, although not always in the same direction. In The Gambia fruit and vegetable intake was found to be seasonal and vitamin C intake varied from nil in the rainy season to about 100 mg day^{-1} in the dry season, affecting both plasma ascorbate and breast-milk ascorbate. In Cameroon, during the wet season calcium intake was reduced by 30%, iron intake by 20%, and thiamin by 15%. However, the consumption of leafy vegetables and fruit increased during the wet season, as did carotene (+100%) and ascorbic acid (+50%).

Effects of Agroclimatic Seasonality on Time Allocation and Energy Expenditure

The second important consequence of agroclimatic seasonality is the concentration of agricultural practices at certain times of the year. More time has to be devoted to often intense efforts, sometimes regardless of the actual physical capacity to perform them. Thus, although adult males are primarily involved, women and even pregnant women also have to spend more time in the field.

The increase in energy expenditure reported by different studies during the peak agricultural season amounts to 320–1050 kcal day^{-1}. In Mali, during the rainy season the energy expenditure from agricultural work is double that during harvest and triple that in the dry season, although the difference is smaller for women.

As a result of the increased demands on time for agricultural labor, time spent on activities is reallocated. Northeastern Thai men spend 2 h less resting and 1.5 h less in domestic work than in other seasons. Women not only spend less time sleeping and in leisure, but they also spend less time cooking, carrying out household tasks such as cleaning, collecting food and water, caring for the children, or carrying out income-generating activities such as handicrafts. The reallocation of women's time may have important consequences for their own health and well being and for the health of all the household members needing care. (**Figure 3**). Where possible, such household tasks may be reallocated to older children. In adolescent girls in Senegal the total energy expenditure measured by accelerometer was 5% higher in the rainy than in the dry season (100 kcal difference). In the rainy season the girls spent 1.5 h more on vigorous activities and more time in domestic activities.

Seasonal Patterns of Disease

A third factor subject to seasonal changes is morbidity. In most cases the slack season is also a wet season and the environmental changes may lead to seasonal outbreaks of diseases, such as acute respiratory infections, gastrointestinal tract infections, but also other infectious diseases such as measles, malaria, and guinea worm. Overall mortality also increases. A study in Mali showed that the duration of disease episodes in the rainy season was more than double that in other seasons and that the morbidity episodes in the rainy season accounted for more than half the yearly episodes, particularly fever, diarrhea, and respiratory illness.

Increased morbidity is a consequence of epidemic cycles and environmental factors, but is also related to increased susceptibility to infections, as a result of decreased food intake and increased stress. Impairment of immune function has been documented in a study in undernourished Gambian children aged 6½–9½ years, in whom seasonality influenced antibody responses to different vaccines. Some behaviors that are typical for the season also determine increased risk of disease. For example, food may be prepared only once a day and left over for the second meal, thus increasing the risk of contamination; there is less time for personal and household hygiene and this leads to easier spread of gastrointestinal and skin disorders.

Morbidity seriously affects the nutritional status of children and sometimes endangers their lives, but it also impairs labor capacity and imposes further time and financial burden on the households, who have to care for the sick and pay for their treatment.

Coping Strategies

People living in areas with agroclimatic seasonality have developed different strategies to cope with the environmental challenge. The earliest and most successful are aimed at maintaining an adequate level of food stocks, but when these fail, other more costly strategies are put in place and their presence indicates that a crisis is occurring.

The negative effects of seasonality can be prevented to a certain extent by selecting the appropriate crops, namely by using varieties with a shorter vegetative period and lower demand for water and by using reserve crops such as root crops, leguminous plants, and groundnuts. Manioc (*Manihot esculenta*) is used as a reserve crop in the Guinean zone, enset (*E. ventricosum*, false banana) is used in southern Ethiopia, and groundnuts (*Arachis Hypogaea*) in Sahel. Reliance on a single crop is highly risky, as failure of the crop due to a pest, a natural

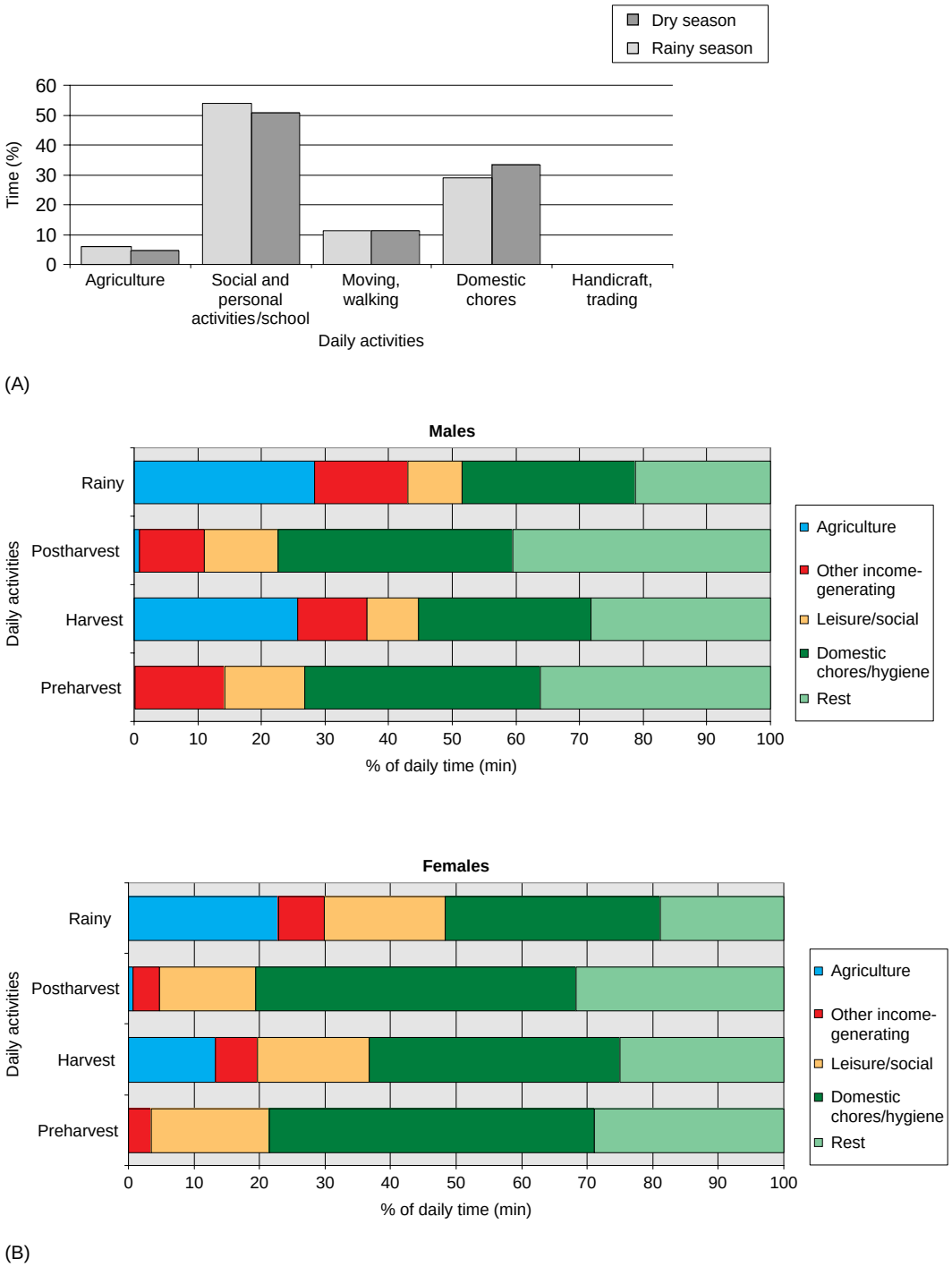


Figure 3 (A) Time allocation of Senegal rural adolescents in different seasons. (Data from Benefice E, Garnier D, and Ndiaye G (2004) *European Journal of Clinical Nutrition* **58**(2): 292–301.) (B) Time allocation of a Thai rural population in different seasons. Duration of daytime activities (night sleeping excluded) in four seasons. (Data from Murayama N and Ohtsuka R (1999) *European Journal of Clinical Nutrition* **53**(1): 39–49.)

disaster, or war may lead to permanent indebtedness and poverty. In areas where cash crops are common, the fluctuation of international market prices may also disappoint the expectations of farmers. The coffee crisis is a typical example. Furthermore, diversification allows people to shift some of the labor burden to different times of the year and to obtain a second harvest in the year.

In order to secure themselves a regular income, landless people who live on seasonal labor, will have

to seek off-farm labor opportunities. At times other than the peak agricultural season labor demand and wages fall and migration of the labor force to different climatic zones may take place. Migration may involve both adults and children, who in some cultures may be sent out to work as servants. At times of food shortage, children may also be sent out to stay with relatives with better resources.

Mutual support groups are created and donations in cash or food may be received by families who may be affected by crop failure or by the occurrence of an unexpected event, such as an illness or death of a family member. Food or money to buy food may be borrowed, although this strategy leads to permanent indebtedness. The final resort is the sale of household assets, particularly animals. This is a high-cost strategy, as the inability to maintain a permanent asset base traps people in poverty and permanent food insecurity.

Nutritional Impact of Seasonality

Body Weight, Body Composition, and Growth

People living in areas of high climatic seasonality are well aware of the nutritional impact of seasonality, as indicated by the language they use to define such seasonal stress periods. The Massa of Cameroon call the month of July in the middle of the wet season the month of 'Did you call me for food?' and they have a word to define 'hunger with threat,' when food shortage has been too long and life is in danger. However, it was not until the 1950s that the scientific community started to appreciate the presence of a nutritional impact of seasonality, and its functional significance is still a matter of discussion.

As described earlier, seasonal climates may affect nutritional status via a combination of reduced dietary intake, increased physical activity, and increased disease incidence, and may occur to a variable extent in different populations and socioeconomic groups. As a result, both energy balance and micronutrient status may be affected, and this is confirmed by the observation of changes in body weight and body composition in adults, growth performance in children, pregnancy weight gain, and birth weight as well as by changes in micronutrient status.

Body weight changes between 1 and 4 kg, corresponding to 2–5% of body weight have been observed in adults in areas of medium and high seasonality. Larger body weight changes occur in draught years or among pastoralists. Remarkable interindividual differences also feature in parallel with such mean values. Smaller changes are usually suffered by women and by people with a lower body mass index (BMI). Socioeconomic differences may go in both directions: wealthier people may either lose more weight, because they own land and have the opportunity and the need to perform more intense agricultural work, or lose less, because they rely on hired labor.

The composition of the tissue lost varies according to the size of the loss and to the initial energy stores. In a rural population in Mali the mean BMI was low ($19.8\,kg\,m^{-2}$ in men and $19.3\,kg\,m^{-2}$ in women), the weight loss in young active men (16–35 years) was 2.6 kg, corresponding to a 3.8% change in body weight, and changes in body fat were in the order of 1.5%. Arm muscle area was also reduced. In Northeast Thailand, where BMI in men was similar ($19.8\,kg\,m^{-2}$) but the weight change smaller (1 kg), a 1.5% reduction of the fat mass was also observed, but no reduction in the fat-free mass could be detected by anthropometry.

The vulnerable groups of the population are not spared by the seasonal stress. During the wet season, energy imbalance leads to the utilization of fat reserves in pregnant women and the women's own energy requirements compete with that of the fetus, leading to increased reproductive risk. In Sierra Leone, at the time of planting and harvesting, pregnant women are expected to continue working and are also more affected by malaria, anemia, and pregnancy-induced hypertension; as a result, in this season birth weights are the lowest in the year. In Taiwan predelivery skin-fold thickness of a cohort of women measured in the cold season was greater than that of a comparable cohort measured in the warm/wet season and a 150 g difference in mean birth weight was observed. In The Gambia a $0.4\,kg\,month^{-1}$ weight gain was observed among pregnant women during the rainy season, as opposed to $1.4\,kg\,month^{-1}$ during the dry season. Dry season mean birth weight was 160 g higher than in the wet season, and the prevalence of low birth weight was 13% in the dry season and 35% in the wet season. Perinatal and infant mortality were also higher in wet season cohorts.

Seasonal stress continues after birth for both mothers and children. In The Gambia lactating women lost on average $0.74\,kg\,month^{-1}$, at the same rate of nonlactating women. As shown earlier, during the wet season very young children get less attention and less breast milk from their mothers and their growth is affected. In Taiwan, children born in the hot wet summer season were smaller, but could catch-up in the following 3 months, while those born in the dry season had a larger birth weight but had a slower postnatal growth.

Seasonal impact is more evident at critical times when a more intense growth effort is required, in

order to catch-up from previous delays or at the mid-infancy growth spurt. In a rural area of The Gambia the lower weight and height gains observed in the wet season were not followed by corresponding increases at other times of the year. In Malawi the weight-for-age Z scores and height-for-age Z scores declined more rapidly during the rainy season among 1 to 6-month-old babies and among 13 to 36-month-old children, but not among the 7 to 12-month-old babies.

Weight and height increments are more sensitive indicators than achieved weight and height. In Bangladesh monthly height gain in children under 5 years ranged from a minimum of 12–20% of the reference value to 200–240% (respectively, in boys and girls), while height gain fluctuated between 52–60% and 165–180%. In Ethiopian children, height growth velocity showed a marked seasonal pattern, with values close to normal (−0.2 SD units) in July to December, a period characterized by better food availability, and lower values (−3.0 SD units) in January to June, a period characterized by intensive farm labor and heavy rains. Therefore, there was never an opportunity to recover from growth faltering and stunting was a continuous process in the first 5 years of life.

Unlike the younger children, the seasonal variables did not have a permanent effect in older children and adolescents. In the Ethiopian study, girls above the age of 10 years showed accelerated growth in the first semester of the year and delayed growth in the second semester, characterized by the wet season, so that the mean yearly growth rates were normal overall. This was also observed in Senegalese adolescents, in whom arm circumference and triceps skinfold were significantly lower during the rainy season, followed by a recovery in the postharvest season, but no change in the growth rate.

In young children, labor burden is not a critical variable, but the reduction of food availability, combined with the greater incidence of infectious diseases, especially diarrhea, leads to impaired growth. This is supported by the observation that seasonality may also affect urban children. In an urban area of The Gambia height-for-age showed little seasonal variation, but weight gain was poor during the rains and was not compensated by catch-up growth during the dry season. In older children and adolescents, the seasonal effects may instead be related to increased physical activity.

Figure 4 illustrates the magnitude of seasonal effects on mean weight changes observed in adult men and women in different regions of the world. In both sexes the observed values range between 1 and 5 kg, although the values greater than 4 kg have been documented during extreme environmental stress, as in drought.

Changes in Micronutrient Status

Few studies document the changes in micronutrient status that can be expected from the seasonal changes in dietary quality. In The Gambia both plasma ascorbate and breast-milk ascorbate had seasonal fluctuations connected with fruit and vegetable intake. Carotenoids also showed a three-fold fluctuation, while retinol was unchanged.

In another study in undernourished children in The Gambia significant seasonal changes were observed for hemoglobin levels and plasma concentrations of vitamin C and α- and β-carotene. Hemoglobin levels decreased with the wet season, while vitamin C and α- and β-carotene plasma levels were highest during the mango season (April–May) and during the rainy season (September–October) when green leafy vegetables are abundant. Zinc and retinol plasma levels were not significantly affected by seasonality.

Metabolic Adaptation

Adaptation to reduced energy intake is a possible biological mechanism to cope with seasonal energy stress. A small reduction in basal metabolic rate (BMR) has been observed. In a multicenter study carried out in India, Benin and Ethiopia the energy debt generated by seasonal changes in food intake and labour pattern was accounted for by the mobilization of fat stores in Benin, abundant in this population, and by a combination of a modest reduction of body weight in India and Ethiopia (0.3 kg and 1.6 kg respectively), together with a reduction in BMR that allowed a 30–50% saving of total energy expenditure. The adaptive response occurred at a relatively low level of energy deficit, i.e., 70 kcal day^{-1} in India and 90 kcal day^{-1} in Ethiopia.

Functional Consequences

The size of the nutritional impact is dependent on the magnitude of the seasonal stress and on the baseline nutritional status, as this sets the limit of the tolerable stress. Lean people will lose more fat-free mass than fatter people. At a BMI of 21, 50% of the weight lost is lean tissue. Therefore, in populations with a lower mean BMI there is a greater impact on their fat-free mass, with greater consequences for their productivity and fitness. A meta-analysis of the body weight change/body weight relationship indicates that farmers tend to maintain the loss at below 2% of their fat-free mass; people with lower BMI will lose less weight, but they will also have to limit their physical activity, which can have socioeconomic consequences. In some populations, body weight lost

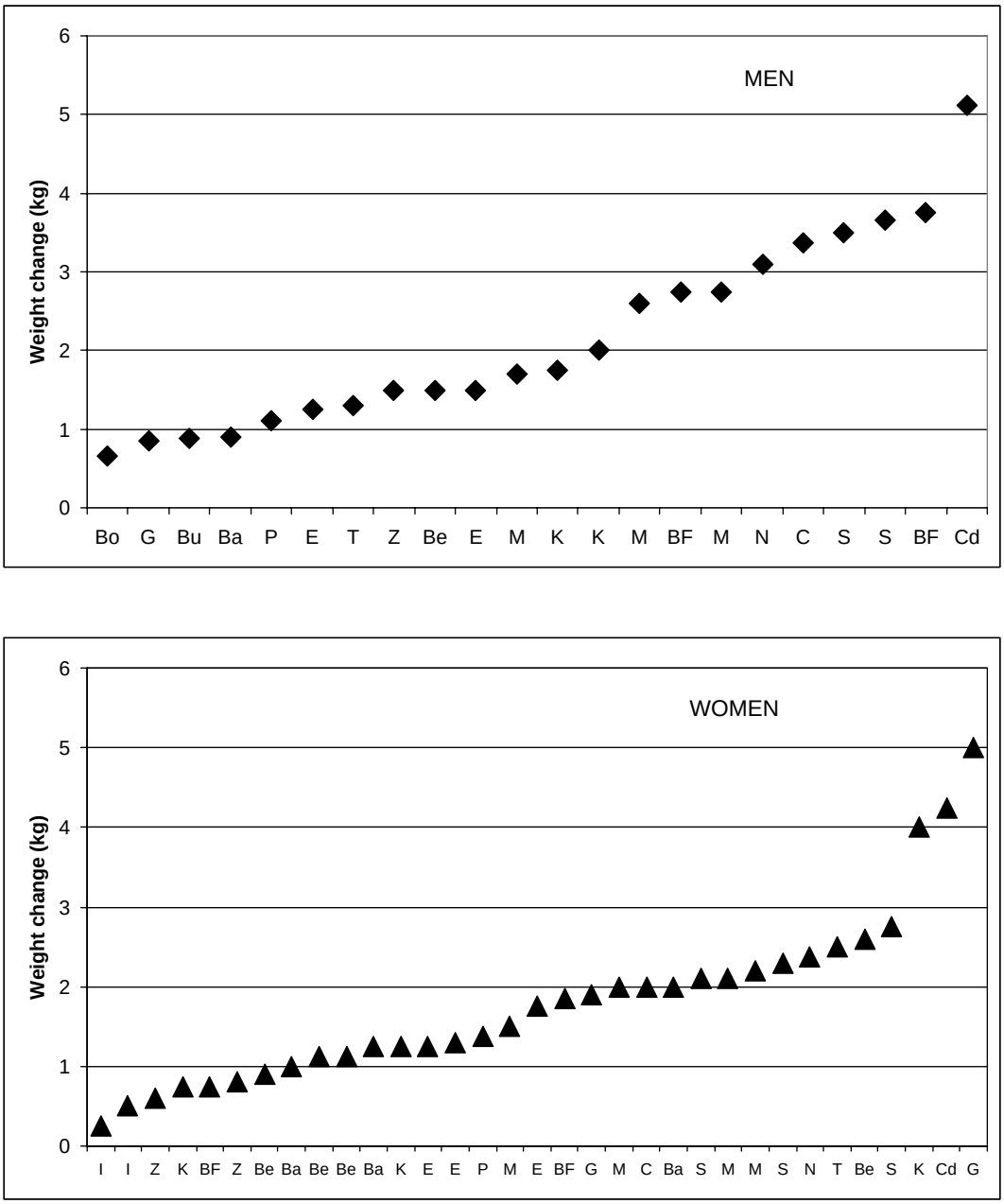


Figure 4 Range of body weight changes between lean and good seasons documented among adult men and women in different countries. Legend of countries: Ba = Bangladesh; Be = Benin; BF = Burkina Faso; Bo = Botswana; Bu = Burma; C = Cameroon; Cd = Cameroon during drought; E = Ethiopia; G = Gambia; K = Kenya; I = India; M = Mali; N = Niger; P = Papua; S = Senegal; T = Thailand; Z = Zaire. References in Ferro-Luzzi A, Branca F, and Pastore G (1994) Body mass index defines the risk of seasonal energy stress in the Third World. *European Journal of Clinical Nutrition* **48**(supplement 3): S165–S178. Additional references: Adams AM (1995) Seasonal variations in energy balance among agriculturalists in central Mali: compromise or adaptation? *European Journal of Clinical Nutrition* **49**: 809–823; Ategbo E-AD, van Raaij JMA, de Koning FLHA, and Hautvast JGAJ (1995) Resting metabolic rate and work efficiency of rural Beninese women: a 2-y longitudinal study *American Journal of Clinical Nutrition* **61**: 466–472; Murayama N and Ohtsuka R (1999) Seasonal fluctuation in energy balance among farmers in Northeast Thailan: the lack of response of energy intake to the change of energy expenditure *European Journal of Clinical Nutrition* **53**: 39–49.

can be as high as 4% of the fat-free mass, but this is probably the maximum stress cyclically tolerable.

Some authors maintain that the observed fluctuations may be regarded as an acceptable physiological response to energy imbalance or even a successful adaptive response. Indeed, having a maximum body weight at the beginning of the heavy work season may be more advantageous than keeping a constant body weight, because this minimizes the farm storage losses.

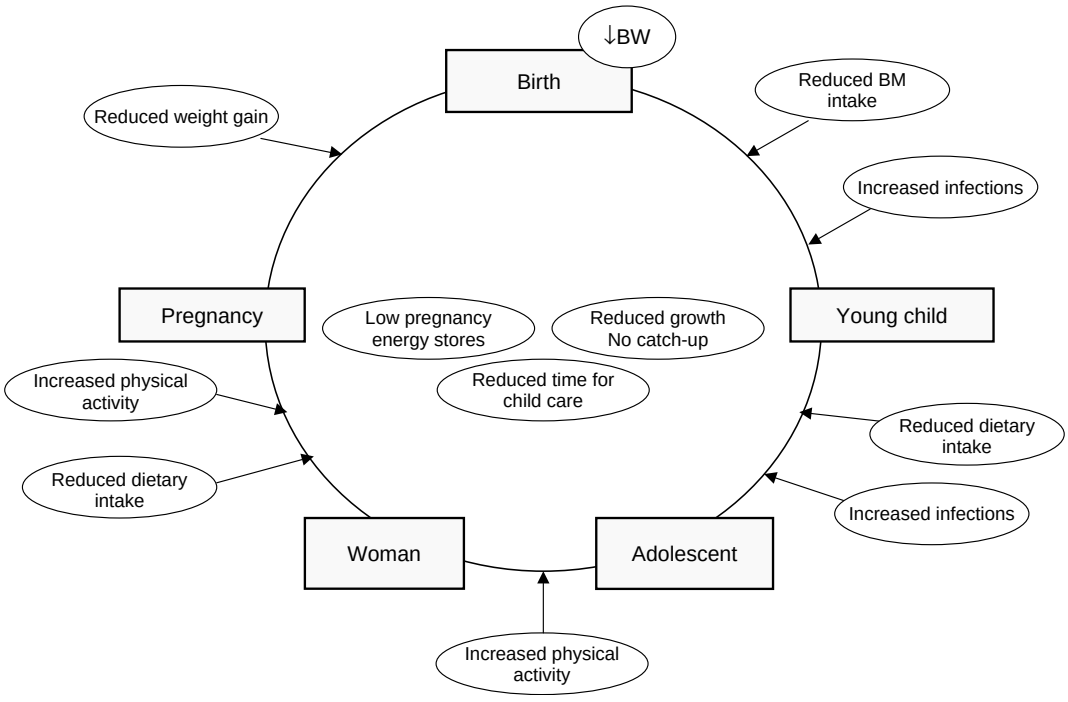


Figure 5 Model of the effects of seasonal stress on the intergenerational cycle of malnutrition. Seasonal factors are drawn outside the circle; biological and behavioral effects are drawn inside the circle. BW = Body weight, BM = Breast milk.

However, in some subgroups of the population living in seasonally prone areas such compensatory mechanisms may irreversibly affect other physiological functions such as reproductive performance and growth. While men have time to recover after the seasonal imbalance of energy, women are permanently undergoing stress and a greater impact on health, well being, and function should be expected. Furthermore, women's seasonal stress also has an impact on the early growth and development of young children, such that seasonal cycling may be considered as one of the factors responsible for the intergenerational cycle of malnutrition (**Figure 5**).

Young children suffer a double seasonal burden: the one imposed on themselves by disease and the one imposed on their mothers. Seasonal changes in children's weight and height velocity have also been observed in more developed societies. However, in high seasonal areas of LDC the periods of retarded growth in the youngest age groups are not followed by periods of adequate catch-up growth, such that height-for-age decreases progressively. Therefore, we suggest that stunting is listed among the functional consequences of seasonality, along with its correlates of impaired cognitive function, impaired metabolic function, and decreased productivity. The impact is going to be highest in younger children, and we may arbitrarily take the age of 5 years as a limit for increased seasonal vulnerability.

Extension of the Problem

Approximately one billion people live in areas of moderate and severe seasonality. Taking into consideration the BMI distribution of those populations, it is possible to calculate that 65% of the adults living in rural areas, i.e., 408 million people, are at risk of severe stress, most of whom are in Asia (90%) and the remaining 10% in sub-Saharan Africa.

Pregnant women and young children should, however, be added to this count. A rough calculation indicates that in those areas, the number of children under 5 years is about 500 million and the number of pregnant women approximately 20 million. They should also be included among the victims of seasonal climatic changes in LDC, thus bringing the estimate to about a billion people.

Nutrition interventions aimed at accelerating growth rates should then preferably be carried out at the time of the highest seasonal stress in these vulnerable population groups, particularly in younger children, in order to achieve the maximum long-term benefit.

See also: **Bioavailability**. **Breast Feeding**. **Energy**: Metabolism; Requirements; Adaptation. **Lactation**: Dietary Requirements. **Malnutrition**: Primary, Causes Epidemiology and Prevention. **Meal Size and Frequency**. **Supplementation**: Developing Countries.

Further Reading

Abdullah M and Wheeler EF (1985) Seasonal Variations and the intra-household distribution of food in a Bangladeshi village. *American Journal of Clinical Nutrition* **41**: 1305–1313.

Annegers JF (1973) Seasonal food shortages in West Africa. *Ecology of Food and Nutrition* **2**: 251–257.

Branca F, Pastore G, Demissie T, and Ferro-Luzzi A (1993) The nutritional impact of seasonality in children and adults of rural Ethiopia. *European Journal of Clinical Nutrition* **47**: 840–850.

Chambers R (1982) Health, agriculture and rural poverty: why season matters. *The Journal of Development Studies* **18**: 217–238.

Chen LC, Chowdhury AKMA, and Huffman SL (1979) Seasonal dimensions of energy protein malnutrition in rural Bangladesh: the role of agriculture, dietary practices and infection. *Ecology of Food and Nutrition* **8**: 175–187.

Dugdale AE and Payne PR (1987) A model of seasonal changes in energy balance. *Ecology of Food and Nutrition* **19**: 231–245.

Ferro-Luzzi A (1990) Seasonal energy stress in marginally nourished rural women: interpretation and integrated conclusions of a multicentre study in three developing countries. *European Journal of Clinical Nutrition* **44**(supplement 1): 41–46.

Ferro-Luzzi A and Branca F (1993) Nutritional seasonality: the dimension of the problem. In: Uliaszek SJ (ed.) *Seasonality and Human Ecology*, pp. 147–165. Cambridge: Cambridge University Press.

Ferro-Luzzi A, Branca F, and Pastore G (1994) Body mass index defines the risk of seasonal energy stress in the Third World. *European Journal of Clinical Nutrition* **48**(supplement 3): S165–S178.

Longhurst R (1984) *The Energy Trap: Work, Nutrition and Child Malnutrition in Northern Nigeria*. Cornell International Nutrition Monograph Series, No. 13. Ithaca, NY: Cornell University Program in International Nutrition.

Moore SE, Goldblatt D, Bates CJ, and Prentice AM (2003) Impact of nutritional status on antibody responses to different vaccines in undernourished Gambian children. Acta Paediatrica **92**: 170–176.

Payne P and Lipton M (1994) *How Third World Rural Households Adapt to Dietary Energy Stress. Food Policy Review 2*. Washington, DC: International Food Policy Research Institute.

Prentice AM, Whitehead RG, Roberts SA, and Paul AA (1981) Long-term energy balance in child-bearing Gambian women. *American Journal of Clinical Nutrition* **34**: 2790–2799.

Sahn E (ed.) (1989) *Seasonal Variability in Third World Agriculture: The Consequences for Food Security*. Baltimore, MD: The Johns Hopkins University Press.

Tetens I, Hels O, Khan NI, Thilsted SH, and Hassan N (2003) Rice-based diets in rural Bangladesh: how do different age and sex groups adapt to seasonal changes in energy intake? *American Journal of Clinical Nutrition* **78**: 406–413.

Seeds *see* **Nuts and Seeds**

SELENIUM

C J Bates, MRC Human Nutrition Research, Cambridge, UK

The realization that selenium (Se) may be an essential micronutrient for human diets has arisen only recently, in the second half of the twentieth century. Selenium deficiency, attributable to low soil selenium levels in farm animals, especially sheep that are afflicted by selenium-responsive 'white muscle disease,' has been recognized for at least half a century. However, the more recent identification of Keshan and Kashin–Beck diseases as endemic selenium-responsive conditions, occurring in a central 4000+-km-wide belt of central China and in areas of Russia, demonstrated conclusively that not only is selenium an essential element for man but also deficiencies occur naturally and require public health measures to alleviate them. Selenium incorporation into plants is affected by the acidity of the soil and by the concentrations of iron and aluminum present so that selenium content of human diets is modulated by these components of the environment. The very recent discovery that these diseases probably arise through the interaction of selenium deficiency with enhanced viral virulence has added a further layer of complexity, but it does not alter the fact that selenium is an essential dietary component that cannot be substituted by any other element. Another complicating factor is that moderately increased soil selenium concentrations result in the opposite condition of selenosis, or selenium overload, with equally debilitating consequences. Of all elements, selenium has a very narrow safe intake range, and unlike some other potentially toxic elements, it is absorbed efficiently by the intestine over a wide range of concentrations and across a variety of different molecular forms.

Unlike other elements, selenium can be incorporated in two distinct ways into proteins, either as a

functional active center (in specific selenoproteins) via a selective incorporation mechanism that ensures selenocysteine insertion or alternatively by a nonspecific incorporation pathway, in which selenomethionine (or selenocysteine) can replace methionine or cysteine at random, without apparently conferring any special functional characteristics on the recipient proteins. This dichotomy of incorporation complicates the task of measuring status and requirements because the different dietary forms of selenium contribute differently to these two contrasting types of incorporation. Selenomethionine and supplementary selenium in the form of selenium-enriched yeast, in which the incorporated selenium is largely present as selenomethionine, contribute to the random incorporation pathway. This is followed by a gradual turnover of selenium to enrich the specific incorporation pathway. Inorganic selenium, in contrast, feeds directly into the specific incorporation pathway via selenide. Although inorganic selenium may relieve functional selenium deficiency more rapidly than organic selenium, the inorganic forms are potentially more toxic; therefore, selenomethionine supplementation is often preferred because it is safer.

Selenoproteins seem to have a number of functions, comprising various catalytic roles (glutathione peroxidases, thioredoxin reductases, and iodothyronine deiodinases), structural roles, detoxifying functions (e.g., selenoprotein P), and storage and transport activities. Many of these functions are incompletely understood, and advances in this area should help to clarify uncertainties about human requirements and the role of selenoproteins in disease, especially in multifactorial conditions such as cancers. Several controlled intervention trials involving selenium are under way, and these should provide evidence to underpin public health programs in the near future.

Dietary Selenium, Absorption, and Mechanisms of Incorporation of Selenium into Selenoproteins

Rich food sources of selenium in human diets include Brazil nuts, offal, shellfish, and some other types of fish, although there is uncertainty about the extent of selenium bioavailability in some foods, which may in turn be linked to problems of mercury contamination. In the United Kingdom, cereal foods account for approximately 20% of total selenium intake, whereas meat, poultry, and fish account for 30–40%.

Selenium is readily absorbed, especially in the duodenum but also in the caecum and colon. Seleno-amino acids are almost completely absorbed: selenomethionine via the gut methionine transporter and selenocysteine probably via the cysteine transporter. Both selenite and selenate are >50% absorbed, selenite more readily so than selenate, and for these forms there is competition with sulphate transport. Selenite is more efficiently retained then selenate because part of the latter is rapidly excreted into the urine. Vitamins A, E, and C can modulate selenium absorption, and there is a complex relationship between selenium and vitamin E that has not been entirely elucidated for man. A combined deficiency of both nutrients can produce increases in oxidative damage markers (malondialdehyde, F_2 isoprostanes, and breath hydrocarbons) and in pathological changes that are not seen with either deficiency alone. Inorganic Se is reduced to selenide by glutathione plus glutathione reductase and is then carried in the blood plasma, bound mainly to protein in the very low-density lipoprotein fraction. Selenomethionine is partly carried in the albumin fraction.

Figure 1 summarizes the main pathways of interconversion of selenium in mammalian tissues. Selenium appears not to be an essential element for plants, but it is normally taken up readily into their tissues and is substituted in place of sulfur, forming the seleno-amino acids selenomethionine and selenocysteine, which are then incorporated at random in place of the corresponding sulfur amino acids into plant proteins. All branches of the animal kingdom handle selenium in essentially similar ways. When ingested, plant selenium-containing proteins liberate free selenomethionine and selenocysteine, either for incorporation at random into animal proteins or for metabolic turnover, to liberate inorganic selenide, which is the precursor of active selenium to be inserted at the active site(s) of the selenoproteins. Selenide is also supplied by the reduction of selenite and selenate that enters the diet from nonorganic sources (i.e., from the environment) or from dietary supplements of inorganic selenium. The inorganic forms of the element are absorbed with approximately 50–90% efficiency (i.e., only slightly less than the >90% efficiency of absorption of selenomethionine).

Selenide represents the 'crossroads' of selenium metabolism, from which it may either be committed to specific selenoprotein synthesis or be removed from the body by urinary excretion pathways that involve its detoxification by methylation to methyl selenides, of which the largest fraction is usually trimethyl selenonium. If used for selenoprotein synthesis, selenide combines with a chaperone protein, and the first metabolic step is its conversion to selenophosphate by the ATP-requiring enzyme

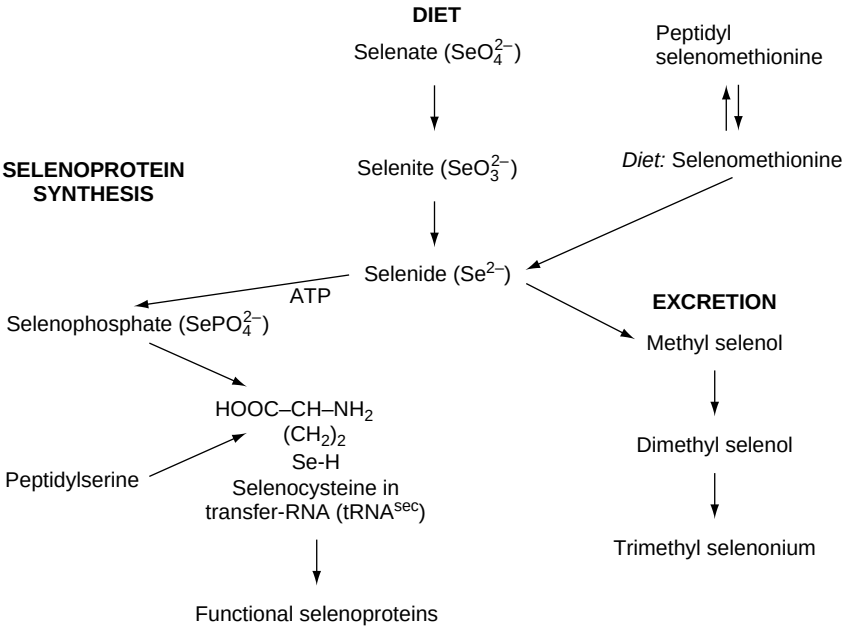


Figure 1 Interconversion of different selenium species in animal and human tissues.

selenophosphate synthetase, which is a selenoprotein. This then becomes the precursor for selenocysteinyl-soluble (or transfer) RNA, which is synthesized from a serine moiety attached to a specific soluble (transfer) RNA identified as $tRNA^{sec}$. This serine–tRNA complex is first dehydrated to aminoacrylyl–tRNA in a reaction that requires a vitamin B_6 cofactor, pyridoxal phosphate. This product then reacts with selenophosphate in a reaction that requires magnesium and the enzyme selenosynthase. The resulting selenocysteinyl–tRNA then recognizes a UGA codon in the messenger RNA sequence. This codon is also used as a stop sequence; therefore, the adjacent mRNA structure has to provide the correct 'context' (e.g., a stem-loop structure) to direct the incorporation of selenocysteine into the growing polypeptide chain of the selenoprotein. Other gene products are involved, and although the sequence of reactions and the participatory proteins have been studied in detail and largely elucidated for prokaryotes such as *Escherichia coli*, the analogous pathways are only partly understood for eukaryotes such as mammals. Specific selenoprotein synthesis is often tissue specific, with different versions of structurally similar selenoproteins being made at different tissue sites. In liver, for instance, provided that the selenium supply is generous, there is a considerable accumulation of cytosolic glutathione peroxidase type I, which can act as a storage repository of selenium for later liberation and redistribution.

Degradation of selenocysteine is catalyzed by selenocysteine lyase, which releases elemental Se, and this is then reduced to selenide by glutathione or other thiols. The urinary excretion pathway is very important for selenium homeostasis of the tissues. Urinary selenium tends to reflect recent intake rather than tissue status, but it can be a useful source of information about possible selenium overload.

Selenoproteins: Classification and Functions

Table 1 lists the selenoproteins that have been unequivocally identified in mammals, together with a summary of their main locations and known functions. Of the known glutathione peroxidases, three are tetramers and one (the phospholipid hydroperoxide-specific peroxidase) is monomeric in its quaternary structure. It appears to be this class of enzymatic activity that is critical for the action of selenoproteins in maintaining immune function, and indeed, glutathione peroxidase type I knockout mice are susceptible to viral mutation and increased viral virulence, as are selenium-deficient ones. Several other selenoproteins listed in **Table 1** also have antioxidant functions and activities. Reaction of glutathione peroxidase with peroxides yields selenic or seleninic acid at the active site of the enzyme, which is recycled by glutathione.

The three thioredoxin reductases act in conjunction with the sulfur protein thioredoxin and with NADPH to bind key transcription factors to DNA. The iodothyronine deiodinases modulate the thyroid hormones, helping to ensure an optimal supply of the

Table 1 Selenoprotein description and functions

Selenoprotein	*Molecular description*	*Function*
Glutathione peroxidases (GPx)		Removal of potentially harmful peroxides and modulation of eicosanoid synthesis
Type I	Tetramer	>50% of total Se in body; acts as Se buffer/store
Type II	Tetramer	May protect the intestine
Type III	Tetramer	Found in plasma and milk; synthesized in kidney
Type IV	Monomer	Phospholipid hydroperoxide GPx; abundant in testis; resistant to Se deficiency; involved in eicosanoid metabolism
Thioredoxin reductases (types I, II, III)		Transfers protons from NADPH via bound FAD to thioredoxin; regulates gene expression by redox control of binding of transcription factors to DNA; needed for cell viability and proliferation; can reduce dehydroascorbate and ascorbate radical to ascorbate
Iodothyronine deiodinases (types I, II, III)		Type I acts in liver and thyroid gland to convert T_4 to T_3; the other types occur in other tissues and also help to regulate thyroid hormone levels
Selenophosphate synthetase		Synthesizes selenophospate from selenide + ATP as first step in selenocysteine synthesis during Se incorporation into selenoproteins
Sperm mitochondrial capsule selenoprotein		Sperm structural protein required for integrity of sperm tail and its mobility; also an antioxidant, similar to GPx IV
Prostate epithelial selenoprotein	15 kDa	In epithelial cells; possibly redox function, similar to GPx IV
Selenoprotein P		Accounts for 60–80% of plasma selenoproteins; contains up to 10 selenocysteines per molecule; has a transport function; binds mercury; may protect the cardiovascular system and endothelial cells
Selenoprotein W	10 kDa	Small antioxidant protein found in muscle (+ heart); its loss may account for white muscle disease of sheep
18-kDa selenoprotein (SELT)		In kidney and many other tissues; not easily depleted in Se deficiency
SELR, SELN	12.6 and 47.5 kDa, respectively	
Spermatid selenoprotein	34 kDa	In sperm nuclei and in stomach; has GPx activity

There are 30–50 proteins that contain Se, as detected by ^{75}Se-labelling in mammals, only about half of which have been investigated.

most active member of the thyroid hormones, triiodothyronine. The different selenoprotein deiodinases are found at different sites in the body. If selenium and iodine are deficient in a human population, the thyroid deficiency is more severe (and goiters are larger) than if only iodine is lacking. This situation is endemic in some areas of central Africa, including Kivu province in the Central African Republic (formerly Zaire).

The sperm mitochondrial capsule selenoprotein has a structural as well as an enzymic role, and it is responsible for both the maintenance of motility and the structural integrity of the tail of the sperm. Both human and other mammals exhibit reduced sperm motility and increased sperm rupture under conditions of low selenium supply. A study in Glasgow, Scotland, recorded enhanced sperm motility and fertility in men who received a selenium supplement.

The precise functional roles of selenoproteins P and W are not well understood. Selenoprotein P contains more selenium (up to 10 atoms per molecule) than any other mammalian selenoprotein, and it can form equimolar selenium–mercury complexes, thereby probably helping to detoxify mercury. It is the major selenoprotein found in plasma and may also act as a selenium transport protein and selenium reserve. Selenoprotein W is found in muscle, and its decline may help explain the molecular basis of white muscle disease in selenium-deficient sheep.

Other selenoproteins have been characterized by their molecular size but not by their functions and health significance (**Table 1**).

Selenium Deficiency, Viral Disease and Mutation, and Immune Function

Initially, Keshan disease was thought to be a deficiency disease alone, involving inadequate intakes of

Se and also of Mo, Mg, and thiamin. However, seasonal variations in symptoms suggested that at least one other interacting factor was likely involved. Later, an enterovirus and a Coxsackie virus (strain B4) were isolated from affected individuals. The same Coxsackie virus was able to produce severe heart pathology in mice when they were fed Se-deficient grain from Keshan endemic areas.

Coxsackie virus strain B3, when introduced into Se-deficient mice, produces a myocarditis that is similar to that of human Keshan's disease, and this virus consistently undergoes mutation at six distinct amino acid (= nucleotide) sites in the Se-deficient but not the control animals. These mutated viruses are then able to produce myocarditis even in selenium-replete control mice. Thus, the virulence change has become permanent by mutation, and the increased virulence is no longer dependent on a simultaneous lack of selenium. Comparable selenium limitation-induced mutations in influenza virus have also been shown in Se-deficient mice. It is suggested that in the presence of a low selenium supply, a normally quiescent virus may become activated by increased oxidative stress and host cell apoptosis, the mutation to increased virulence being a survival strategy by the virus.

In HIV-infected individuals, the progression to AIDS and the decline in T helper (CD4) cell counts are accompanied by a parallel decrease in blood selenium levels. Selenium deficiency appears to increase the probability of mortality in HIV-infected subjects.

Human selenium supplementation (e.g., 200 μg/day), even in apparently selenium-replete individuals receiving a diet providing >120 μg Se/day, was able to stimulate the proliferation of activated T cells of the immune system. It elicited an enhanced response to antigen stimulation, an enhanced ability to generate cytotoxic lymphocytes, an enhanced ability to destroy tumor cells, and increased natural killer cell activity. Growth-regulatory interleukin-2 receptors on the surface of activated lymphocytes and natural killer cells became upregulated.

In a study in Liverpool, UK, healthy adult subjects with initial plasma selenium concentrations below 1.2 μmol/l were given placebo or 50 or 100 μg daily supplements of selenium as selenite for 15 weeks. After 6 weeks, they were given oral live attenuated polivirus vaccine, and after 9 weeks, ^{74}Se stable isotope was given intravenously to measure their body Se pool size. The Se supplements significantly increased the Se pool size, and the supplemented groups cleared the poliovirus more rapidly and their fecal viral RNA products exhibited fewer mutations. Cellular immune response (estimated by interferon-γ and other cytokines) was enhanced and there was an earlier peak of T cell proliferation and numbers in the supplemented groups. This study suggests that selenium supplements can improve a number of indices of immune function, even in individuals whose Se status is not severely deficient.

Selenium Distribution, Status Assays, and Dietary Reference Values

In an adequately supplied adult male human subject, the total body selenium content is on the order of 30–60 mg, of which one-third is found in the skeleton and two-thirds in the soft tissues. A substantial fraction of kidney selenium is retained even when selenium at other sites is severely depleted during deficiency, and renal selenium is more constant between human populations than selenium in other tissues or body fluids. Regulation of selenoprotein synthesis at the transcription level appears to ensure a hierarchy of preservation of individual selenoproteins at critical sites. The cytosolic glutathione peroxidase (GPx I) and selenoprotein P can donate selenium to other sites whenever overall depletion occurs. Selenium crosses the placenta readily, and breast milk selenium concentration is responsive to changes in maternal selenium intake. In the United States, breast milk Se concentrations are generally in the range of 0.19–0.25 μmol/l, but colostrum has levels that are two or three times higher than those of mature breast milk.

Selenium status can be measured in several ways. One recently developed and effective approach toward selenium concentration measurement is the use of inductively coupled plasma mass spectrometry. Older assays are based on the generation of selenium hydrides or fluorescent derivatives of selenium. Selenium status can be measured by its concentration in plasma or serum; in whole blood (a result that can be recalculated to provide red cell selenium concentrations); or in platelets, hair, or nails. The platelet concentration is considered to be a reliable medium-term index, whereas hair and nail concentrations can integrate selenium status, and hence intakes, over a longer term.

Glutathione peroxidase enzymatic assay in plasma or red cells is another frequently used approach to status measurement. In situations of severe to marginal deficiency, this has proven to be a sensitive and responsive index, varying consistently with variations in the selenium supply. However, once an adequate supply is achieved, there is no further capacity for increases in enzyme synthesis, and a plateau of activity is reached that does not respond

to further increases in selenium intake. Therefore, if a population exhibits a strong correlation between plasma (or red cell) selenium concentrations and glutathione peroxidase activity in blood fractions, or there is a major increase in GPx activity after selenium supplementation, this can be taken as evidence of suboptimum selenium status in the population. If there is little evidence of such a correlation or of a response to supplementation, then the population is likely to be adequately supplied. The absolute values of GPx activity are more difficult to interpret because there are many different versions of the assay in use in different laboratories, and interlaboratory harmonization has rarely been undertaken for this assay. Recent reappraisal has suggested that the plasma glutathione peroxidase (GPx III) assay may be more reliable than the blood cytosolic (GPx I) enzyme assay because haemoglobin tends to interfere with the reaction in erythrocyte extracts.

A summary of reference values and recommended intakes of selenium from three publications is presented in **Table 2**. Dietary reference values for selenium in the United Kingdom, set in 1991, were based on a number of criteria, including the facts that no evidence of deficiency was detectable in populations with intakes of 40 μg/day and that saturation of GPx in Chinese males occurred at an intake of approximately 41 μg/day (equivalent to 50 μg/day for a UK male based on a body weight comparison). On this basis, the UK Lower Reference Nutrient Intake (LRNI) was set at 40 μg/day for both male and female adults, and the corresponding RNI values were set at 75 μg/day for males and 60 μg/day for females, with lower values, proportional to body weight, for children. No extra increment was considered necessary for pregnancy, but for lactating women an additional 15 μg/day was added to both the LRNI and the RNI.

Table 2 Reference values for intakes of selenium (μg/day)[a]

Population group	*UK LRNI*	*UK RNI*	*US AI/ RDA*[b]	*WHO/ FAO RNI*
0–6 months	4–5	10–13	AI:15	6
7–12 months	5–6	10	AI:20	10
1–3 years	7	15	RDA:20	17
4–6 years	10	20	30	22
7–10 years	16	30	30–40	21–26
11–18 years, male	25–40	45–70	40–55	32
11–18 years, female	25–40	45–60	40–55	26
19–65 years, male	40	70	55	34
19–65 years, female	40	60	55	26
65 + years, male	40	70	55	33
65 + years, female	40	60	55	25
Pregnant	40	60	60	26–30
Lactating	55	75	70	35–42

[a]Where a range of values is given, the population group described in this table overlapped across more than one population group in the source table.
[b]The first two age groups are AI; the remainder are RDA.
LRNI, Lower Reference Nutrient Intake; RNI, Reference Nutrient Intake; AI, Adequate Intake; RDA, Recommended Dietary Allowance.
Sources: UK: Department of Health (1991) *Dietary Reference Values for Food Energy and Nutrients for the United Kingdom*, Report on Health and Social Subjects No. 41. London: HMSO. USA: Food and Nutrition Board, Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium and Carotenoids*. Washington, DC: National Academy Press. WHO/FAO: WHO/FAO (2002) *Human Vitamin and Mineral Requirements. Report of a Joint FAO/WHO Expert Consultation, Bangkok, Thailand*. Rome: WHO/FAO.

More recently, selenium recommendations or reference values have been slightly lower. The US committee that set Dietary Reference Intakes in 2000 interpreted the Chinese estimate of 41 μg/day needed to saturate GPx in adult men, and data from New Zealand indicating selenium intake adequacy at 38 μg/day, as supporting an Estimated Average Requirement (EAR) of 45 μg/day for adults of both sexes, and hence an Recommended Dietary Allowance (RDA) (with 10% CV of requirements) of 55 μg/day for both sexes, increasing to 60 μg/day for pregnant and lactating women. However, RNI values set by an FAO/WHO committee, published in 2002, were much lower, at only 26 μg/day for women and 34 μg/day for men, based on the premise that full saturation of GPx is unnecessary and two-thirds saturation is probably adequate. Clearly, there has been considerable divergence of opinion between different committees, and this divergence underlies the current uncertainty about the overall adequacy of selenium intakes in many European countries, including the United Kingdom. In the United Kingdom, selenium intakes have declined considerably during the past 25 years because of the substitution of North American wheat imports by European wheat with a much lower selenium content. In contrast, selenium intakes in New Zealand have increased as a result of grain imports from Australia.

There are also recommendations for the upper limit of safe intake of selenium. For the United Kingdom, it was noted that evidence of toxicity was detectable at intakes of approximately 750–900 μg/day, and the UK panel recommended a maximum safe intake of 450 μg for adults (6 μg/kg body weight/day), which was confirmed as an official safe upper level (SUL) in 2003. In the US Dietary Reference Intakes, the upper level of 400 μg/day was based on a no adverse effect level of 800 μg/day divided by

an uncertainty (i.e., safety) factor of 2. The FAO/WHO committee also set an SUL of 400 μg/day for adults. In humans, at intakes of >3 mg/day, overt signs of selenosis include damage to nails and hair, skin and nerve lesions, mottling of teeth, nausea, weakness, and diarrhea. Urinary selenium excretion is high, and a garlic odour may be apparent in the breath.

Selenium Interventions

As noted previously, neither Keshan disease nor Kashin–Beck disease are now thought to be simple dietary deficiency diseases. They probably also involve viral components and may be exacerbated by environmental toxins, including mycotoxins. Thus, they are probably multifactorial, but importantly, public health selenium supplementation interventions have had a dramatically beneficial effect on the prevalence of these diseases. The main clinical features of Keshan disease are cardiac insufficiency and enlargement, electrocardiographic changes, and fibrosis. Those of Kashin–Beck disease are osteoarthropathy and necrosis of the joints and epiphysial plate cartilage. Both diseases occur in school-age children; Keshan disease also occurs in women of child-bearing age, but adult men are less affected.

In hilly and heavily eroded areas of China where these diseases were endemic, the use of selenium-enriched fertilizers was not feasible as an intervention because of the huge geographical areas involved and hence the high cost. Instead, direct human supplementation of at-risk and affected populations was introduced during the 1970s using a 0.5 or 1.0 mg sodium selenite supplement (according to age) per person per week. In Shaanxi province, following supplementation, the prevalence of Keshan disease declined from 12 per 1000 to undetectable levels between 1976 and 1985, and in Heilongjiang province the prevalence of Kashin–Beck disease declined from 44 to 1% of the population between 1970 and 1986.

An alternative approach to intervention, by selenium enrichment of crop and grassland fertilizers, was introduced in the 1970s in Finland. Here, there was no overt evidence of selenium deficiency in the human population, but Se deficiency disease had occurred, and had been successfully eliminated in farm animals, by supplementation of animal feeds during the 1960s. Fertilizer that was Se enriched at 16 mg/kg was then applied to cereal crops for human consumption. Grassland fertilizer was enriched at 6 mg/kg. As a result, adult human Se intake increased from 25–60 to approximately 100 μg/day. Serum Se increased from 65–70 μg/l in 1975 to 120 μg/l in 1989–1991. In 1990, the selenium level was reduced to 6 mg/kg fertilizer for cereal crops as a precaution against possible overload. Selenium intervention by fertilizer enrichment was judged to be a safe, economical, and easily controlled intervention.

New Zealand, which has a similar situation of marginal intakes and status, decided not to intervene on a population or nationwide basis but instead has taken steps to ensure that particular high-risk groups, notably people receiving total parenteral nutrition or children receiving special diets for phenylketonuria prevention, are adequately supplied.

Selenium and Chronic Disease

Several supplementation and epidemiological case–control studies have suggested a possible link between increased selenium intakes or status and protection against certain cancers. First, in intercountry comparisons and studies comparing different regions of the United States having different soil selenium levels, there was a consistent correlation between lower selenium levels and higher risk of cancer. A study of 34,000 male health professionals in the eastern United States found that toenail selenium levels were inversely proportional to prostate cancer incidence, with diagnoses recorded >2 years later, which helped to reduce confounding by reverse causality (i.e., the presence of cancer causing the low selenium levels). A study in China found that the incidence of hepatocellular carcinoma was lower in a community receiving selenite-fortified salt (15 mg/kg) than in adjacent control communities, and another intervention with selenium + vitamin E + β-carotene supplements seemed to reduce total age-adjusted mortality, especially from cancer. A controlled intervention in the eastern United States recorded an approximately 50% lower cancer mortality in a high-risk group of subjects who were randomized to receive 200 μg selenium/day as Se yeast for several years. Animal model studies have demonstrated reduced susceptibility to cancer induction with increased selenium intakes. Although these studies appear promising, caveats about their design and interpretation imply that none yet conclusively prove the hypothesis that additional dietary selenium is able to reduce human cancer morbidity and mortality. Trials in multiple locations (including Europe)—PRECISE (Prevention of Cancer by Intervention with Selenium) and SELECT (Selenium and Vitamin E Cancer Prevention Trial)—are ongoing and should help to resolve this important question.

Attempts to demonstrate disease-reduction benefit with respect to cardiovascular disease by selenium intervention have proven disappointing, despite the

theoretical benefit of lipid peroxide removal by GPx and apparently beneficial changes in intermediate markers such as platelet aggregation, vasoconstriction, and thromboxane:prostacyclin ratios following supplementation. Further studies of high-risk populations are needed, including a focus on the concerted action by combinations of selenium and vitamin E.

Other conditions for which supplementary selenium has been claimed to be beneficial include rheumatoid arthritis, pancreatitis and asthma, and mood alterations. Again, further studies are required.

Conclusion

The essential role of selenium in human nutrition and its discrete biochemical functions are rapidly being characterized. Severe deficiency and selenosis (toxicity) occur in different regions and are manifested by characteristic and life-threatening human diseases. The selenoproteins have a wide variety of roles, both catalytic and noncatalytic. Interactions with redox pathways appear to be common. Selenoprotein P, in particular, appears to play an important detoxification role. Selenium appears to play an important role in cell-mediated immunity. Selenium deficiency can cause viral mutation leading to increased virulence, and such mutation and its consequences appear permanent. Optimum human intakes of selenium are still a matter of debate because some studies have reported benefits (e.g., anticancer and immunological effects) when supplements are given, even to populations that appear to be generously supplied with the nutrient. The distinction between nutritional and pharmacological benefits is unclear, and further trials to determine risk–benefit balance at different intake levels are needed in a range of populations and age and gender groups.

See also: **Antioxidants**: Intervention Studies. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements. **Cancer**: Effects on Nutritional Status. **Vitamin A**: Biochemistry and Physiological Role. **Vitamin E**: Metabolism and Requirements.

Further Reading

Allan CB, Lacourciere GM, and Stadtman TC (1999) Responsiveness of selenoproteins to dietary selenium. *Annual Review of Nutrition* **19**: 1–16.

Arthur JR, McKenzie RC, and Beckett GJ (2003) Selenium in the immune system. *Journal of Nutrition* **133**: 1457S–1459S.

Beck MA, Levander OA, and Handy J (2003) Selenium deficiency and viral infection. *Journal of Nutrition* **133**: 1463S–1467S.

Behne D and Kyriakopoulos A (2001) Mammalian selenium-containing proteins. *Annual Review of Nutrition* **21**: 453–473.

Broome CS, McArdle F, Kyle JAM *et al.* (2004) An increase in selenium intake improves immune function and poliovirus handling in adults with marginal selenium status. *American Journal of Clinical Nutrition* **80**: 154–162.

Brown KM and Arthur JR (2001) Selenium, selenoproteins and human health: A review. *Public Health Nutrition* **4**: 593–599.

Driscoll DM and Copeland PR (2003) Mechanism and regulation of selenoprotein synthesis. *Annual Review of Nutrition* **23**: 17–40.

Ellis DR and Dalt DE (2003) Plants, selenium and human health. *Current Opinion in Plant Biology* **6**: 273–279.

Hatfield DL (ed.) Journal of Trace Elements in Experimental Medicine. *Selenium: Its Molecular Biology and Role in Human Health*. Dordrecht, The Netherlands: Kluwer Academic.

Lyons G, Stangoulis J, and Graham R (2003) High-selenium wheat: Biofortification for better health. *Nutrition Research Reviews* **16**: 45–60.

Rayman MP (2000) The importance of selenium to human health. *Lancet* **356**: 233–241.

Reilly C (1996) *Selenium in Food and Health*. London: Blackie Academic & Professional.

Schrauzer GN (2000) Selenomethionine: A review of its nutritional significance, metabolism and toxicity. *Journal of Nutrition* **130**: 1653–1656.

Tapiero H, Townsend DM, and Tew KD (2003) The antioxidant role of selenium and seleno-compounds. *Biomedicine and Pharmacotherapy* **57**: 134–144.

Whanger PD (1998) Metabolism of selenium in humans. *Journal of Trace Elements and Experimental Medicine* **11**: 227–240.

WHO Task Group on Selenium (1987) *Selenium*, Environmental Health Criteria 58 Geneva: WHO.

Senescence *see* **Aging**

Skinfold Thickness *see* **Nutritional Assessment**: Anthropometry

SMALL INTESTINE

Contents
Structure and Function
Disorders

Structure and Function

D Rumsey, University of Sheffield, Sheffield, UK

The small intestine is the barrier between the external environment and the interior of the human body that all nutrients must pass; it is the organ of nutrition. Absorption of nutrients across the barrier is only possible following the complex process of food digestion. Both processes occur largely within the small intestine. Complete digestion and absorption only take place if the optimal motility patterns of the small intestine occur. Each facet of small intestinal function, absorption, digestion, and motility is dependent on the other to produce human nutrition.

Furthermore, within the physiology of the small intestine, regulatory mechanisms exist to alter the rate of nutrient absorption, signal the passage of nutrients, and change the phases of metabolism. The complexity of function in the small intestine is also related to its structure, architecture, and cellular kinetics.

Structure of the Small Intestine

The small intestine is the main site of digestion and absorption within the gastrointestinal tract. It is a hollow tube greater than 6 m in length with a lumenal diameter of approximately 4 cm. The first 20 cm distal from the pylorus is the duodenum, the next 2.5 m is the jejunum, and the final half is the ileum. There are no anatomically distinguishing characteristics along the small intestine; any alterations in architecture are gradual.

All segments of the small intestine possess a mucosa with the same sophisticated structural pattern along its length. The mucosal lining is surrounded by two muscle layers. The first innermost layer consists of circular smooth muscle sheets orientated radially around the lumen. The second thinner layer of longitudinal sheets is surrounded by a thin serosal layer. The regulation of muscular movement (motility) is achieved by the enteric nervous system, which consists of two matrices of interconnecting neurones. The first (outermost) matrix is the myenteric plexus situated between the two muscle layers. Underneath the mucosa and above the circular muscle layer is the submucosal plexus, from which extend sensory neurones into the mucosa. Interneurons connect the two plexi, which in turn receive postganglionic parasympathetic nerve fibers.

The enteric nervous system is a highly complex network exhibiting a high degree of autonomy over gut function and employing as many as 20 neurotransmitters and neuromodulators as well as the classical neurotransmitter mechanisms of the autonomic nervous system. The enteric nervous system is only surpassed by the brain and spinal chord in its capacity for information processing.

Mucosal Structure

The lining of the small intestine is remarkably adapted for the function of absorption by increasing the surface area for transmucosal transport at three levels (**Figure 1**):

- The inner surface has circular folds, which increases the area by approximately 3-fold.
- The mucosa projects from the folds into the lumen with finger-like structures called villi approximately 1 mm in length. Villi increase the surface area by an additional 10-fold. The surface of each villus is covered with epithelial cells known as enterocytes. Absorption takes place across the enterocyte barrier.
- Small hair-like filaments known as microvilli project from the luminal surface of each enterocyte into the lumen. Microvilli increase the surface area for absorption by an additional 20-fold.

The three structures combine to increase the surface area by approximately 600-fold.

Each villus is supplied with its own connective tissue support known as the lamina propria, its own arteriolar and venous microcirculation with capillaries draining the basolateral regions of all the enterocytes, and its own lymphatic system.

Cell Kinetics of the Enterocyte

The life span and kinetics of small intestinal enterocytes are particularly important in understanding the process of absorption. Enterocytes have a

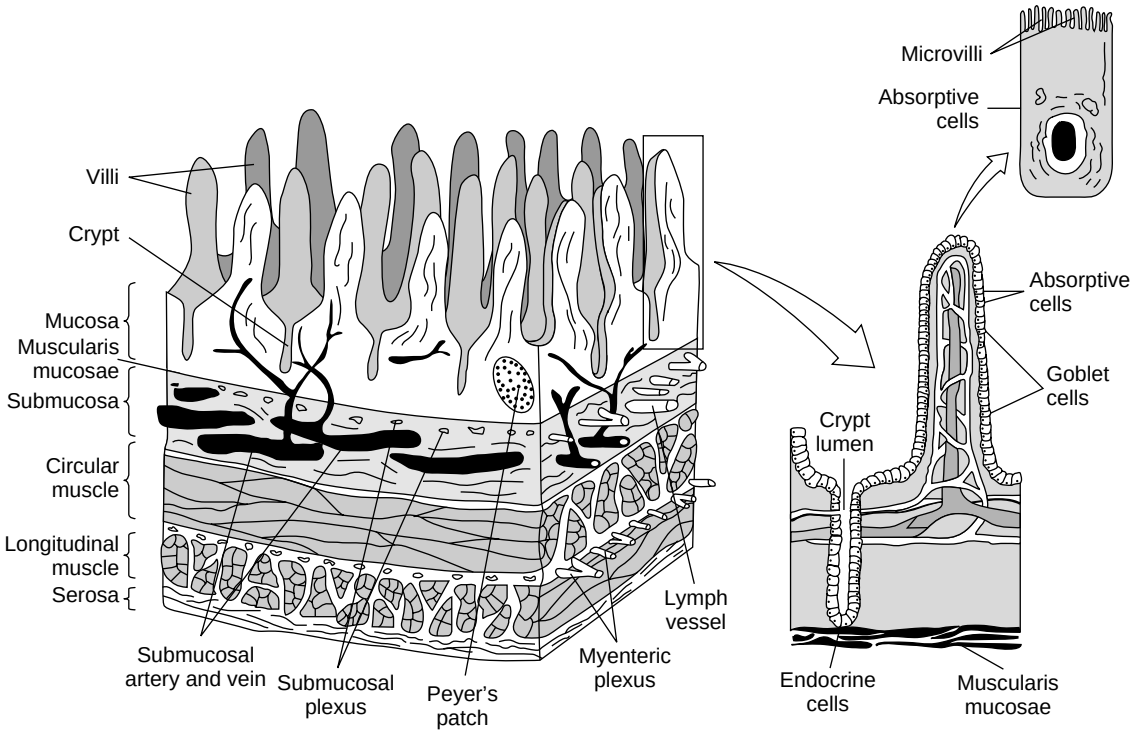


Figure 1 Wall of the small intestine. The intestinal surface area is enhanced by finger-like villi.

particularly short life and the epithelium of the small intestine is replaced every 3–6 days, making it one of the fastest growing tissues in the body. The high mitotic rates in small intestinal crypts make the tissue particularly vulnerable to irradiation and cancer chemotherapy. New cells are born in a proliferative zone within the crypts of Lieberkuhn below the villus.

After birth, new cells move in the direction of the lumen and mitosis continues for two or three more divisions while each cell remains within the crypt. As the enterocyte emerges out of the crypt, proliferation ceases and the process of differentiation proceeds so that the cell, now passing up the outer surface of the villus, reaches functional maturity with a full capacity of membrane-bound brush border enzymes toward the villus tip. On reaching the tip, enterocytes are sloughed off into the intestinal lumen and digested.

The short life span of the enterocyte bestows massive adaptive potential within the tissue since the rate of differentiation may be influenced by changes in any nutrient or other luminal factor to induce a rapid functional response.

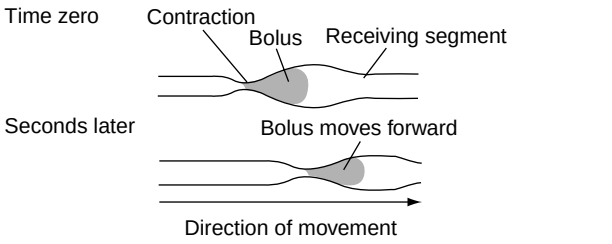
Motility

Vigorous controlled movement of the intestinal wall is essential for the two basic properties of segmentation or mixing and peristalsis or aboral propulsion (**Figure 2**). The two properties are interdependent since neither segmentation nor peristalsis alone would result in optimal digestion and absorption. The peristaltic export of chyme from the duodenal segment in an aboral direction is only effective following considerable segmentation mixing to optimise the homogeneous distribution of digestive juices digestion.

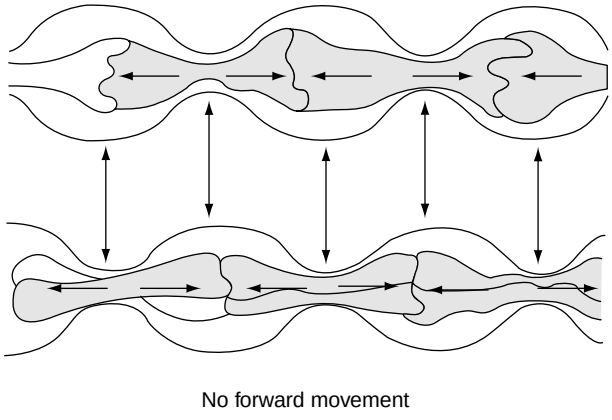
Although the neural basis of peristalsis has been acceptably modeled, answers to basic questions concerning the receptors involved and the occurrence of reverse or retroperistalsis in particular circumstances remain elusive. The precise sensory messages that are needed to switch motility from segmentation to peristalsis and vice versa are also unclear, except that the signals must be both chemical and physical since they are related to the passage and characteristics of the nutrients along the gastrointestinal tract.

Migrating Motor Complex

When the small intestine is emptied after the passage of a meal, segmentation contractions stop and is replaced by a pattern of motility called the migrating motor complex (MMC). The MMC is a series of weak peristaltic rushes in the aboral direction, but each rush only occurs along a short distance of the total length. The series begins at the level of the stomach and takes



(A)



(B)

Figure 2 Contractions in the gastrointestinal tract. (A) Peristaltic contractions are responsible for forward movement. (B) Segmental contractions are responsible for mixing.

approximately 2 h to reach the ileo-caecal valve at the distal ileum. It is considered that the MMC performs a 'housekeeping' function by sweeping debris, such as sloughed cells, residual chyme, and bacteria, aborally in the interval between meals. MMCs continue in the absence of eating. When food is introduced into the stomach, MMC patterns are replaced by segmentation contractions.

Passage of Chyme along the Small Intestine

Chyme is expelled from the stomach into the duodenum in 'packets' in response to the hormonal (cholecystokinin and secretin) influence on gastric emptying. The aqueous component of the meal is emptied first as larger solid particles are reduced by the activity of the gastric musculature. The lipid component of the meal is liable to separate within the stomach lumen and remain as one of the last fractions of the meal to be emptied.

Chyme appears to mix rapidly within the duodenum presumably by rapid segmentation action, and the early parts of the meal pass rapidly along the duodenum and jejunum possibly because the nutrient content of this initial packet is aqueous and therefore lower. Some observations suggest that certain meals produce retroperistaltic movements within the proximal small intestine.

The transit rate of chyme is reduced gradually through the second half of the small intestine, the ileum. The remnants of the meal are released across the ileo-caecal valve as the pressure of contents within the ileum rises.

Nutrient Regulation of Motility and Transit

Traditional experimental techniques in which simple nutrient substrates were introduced by tube into the 'starved' human duodenum found that, under these conditions, most nutrients were absorbed immediately within the proximal tenth of the small intestine. Real food offers a more complex problem for the small intestine since the processes of digestion have only been initiated within the lumen of the stomach. It is apparent that the major part of digestion takes place along a significant fraction of the whole small intestine, with the processes of absorption occurring, as a consequence, along the majority of segments. Certainly, enterocyte functional capacity on ideal villi is comparable to that on the jejunum.

These observations have been supported by findings that meals of a high nutrient density are more slowly emptied from the stomach and pass along the small intestine at a reduced rate. Evidence exists for a regulatory mechanism situated within the mucosa of the terminal ileum capable of sensing nutrient presence such as lipid and providing feedback restraint on motility patterns in the stomach, duodenum, and jejunum to slow the passage of meals in proportion to their nutrient density. This mechanism has been called the 'ileal brake' and, taken with the hormonal control of gastric emptying, explains the almost complete absorption of meals of the highest nutrient density.

Digestion

The chyme from the stomach is mixed in the duodenum with digestive juices from three sources: the pancreas, liver, and intestinal mucosa. The pancreas is an elongated gland situated beneath the stomach with two roles essential for human nutrition (**Figure 3**). Exocrine pancreatic function is the production of pancreatic juice for secretion into the duodenal lumen. Pancreatic juice is a potent mixture of enzymes and solutes that are essential for the process of digestion of protein, lipid, and carbohydrate and for producing the optimal pH environment within the duodenal lumen. Juice is secreted as an ultrafiltrate of plasma from acini, composed of clusters of secretory cells connected to ducts. The composition of these secretions is listed in **Table 1**.

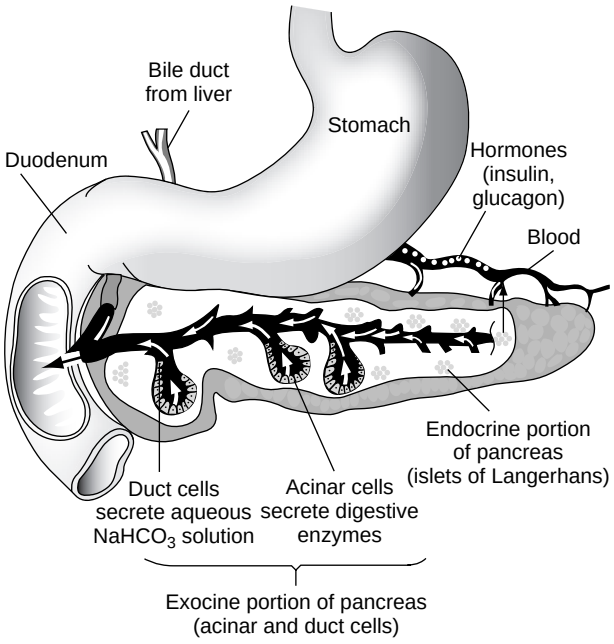


Figure 3 Exocrine and endocrine portions of the pancreas (glandular portions of the pancreas are grossly exaggerated). The exocrine pancreas secretes into the duodenal lumen a digestive juice composed of digestive enzymes secreted by the acinar cells and an aqueous $NaHCO_3$ solution secreted by the duct cells. The endocrine pancreas secretes the hormones insulin and glucagon into the blood.

The second function is the endocrine secretion of the hormones of intermediary metabolism, insulin and glucagon, into the hepatic portal vein supplying blood to the liver. The roles of insulin and glucagon in nutrition are discussed elsewhere.

Pancreatic Juice

The bulk of pancreatic secretion is a dilute solution of sodium bicarbonate ($NaHCO_3$) that neutralizes the acid effluent from the stomach and provides an alkaline environment for the enzymes present in the secretion (**Table 2**).

Table 1 Volumes absorbed by the small intestine per day

	Volume (ml)
Volume entering the small intestine	
Ingestion	
– Food	1300
– Drinks	1300
– Saliva	1500
– Gastric juice	2000
– Pancreatic juice	1500
Plasma secretions	
– Bile	500
– Intestinal secretion	1500
Total	9600
Volume absorbed by the small intestine	9000
Volume passing through the ileoaecal valve	600

Table 2 Constitutents of pancreatic juice

Secretion	*Function*
$NaHCO_3$	Establishes pH environment (pH 7–8)
Amylase	Digests complex carbohydrate
Carboxypeptidase	Digests proteins and polypeptides
Chymotrypsin	Digests proteins and polypeptides
Elastase	Digests elastin
Trypsin	Digests proteins and polypeptides
Lipase	Digests triacylglycerol
Nuclease	Digests nucleic acids

Enzymes Pancreatic proteolytic enzymes are secreted in an inactive form to avoid digestion of pancreatic tissue. Once within the duodenal lumen, the main enzymes, trypsin, chymotrypsin, and carboxypeptidase, are liberated and attack different peptide linkages in the protein molecule. Pancreatic amylase converts polysaccharides into smaller saccharide molecules, particularly disaccharide.

Pancreatic lipase is the only source for the enzymatic digestion of fat. Lipase hydrolyzes dietary triglycerides to monoacylglycerols and free fatty acids. Deficiency of pancreatic lipase results in serious fat malabsorption. Pancreatic amylase and peptidase deficiency is less serious since small intestinal enzymes would significantly minimize any loss of activity.

Regulation of Pancreatic Secretion

Pancreatic juice is stimulated to flow in response to increases in the blood concentrations of two gastrointestinal hormones, secretin and cholecystokinin. Both hormones are liberated from the duodenal mucosa in response to different components of the luminal environment. Secretin is secreted in response to acid and stimulates the pancreatic acini to secrete $NaHCO_3$ so that the acid contents emerging through the pylorus are neutralized.

Cholecystokinin (CCK) secretion is prompted by the presence of both lipid and protein in the duodenal lumen, and the effect of CCK on pancreatic acini is to promote the secretion of enzymes. All three types of pancreatic enzymes are stored in zymogen granules within the acinar cells and released into the ducts on stimulation. As the final components of the meal leave the stomach, acid and nutrient levels fall in the lumen and pancreatic secretion ceases.

Parasympathetic nerve stimulation is also recognized to produce a small stimulation of pancreatic juice flow probably as part of the cephalic and

gastric phases, preparing the duodenal environment for a meal.

The Role of Bile

The second main source of secretion is bile. Bile is formed in the liver canaliculi and passes down the bile duct into the duodenum. The opening of the bile duct into the duodenum, the sphincter of Oddi, is controlled by cholecystokinin levels. When the sphincter of Oddi is closed, bile is diverted and stored in the gall bladder.

Bile is a solution of $NaHCO_3$, similar to pancreatic juice, without enzymes and including a number of organic solutes, of which the bile salts are essential for the digestion of fat. Bile salts have a detergent action that emulsifies the large droplets of dietary fat into a water-soluble form, a micelle, rendering the fat molecules available for hydrolysis by pancreatic lipase.

Intestinal Mucosal Secretions

In contact with chyme, the intestinal mucosa secretes a mucous-containing watery fluid into the lumen that is devoid of enzymes. The mucous protects and lubricates, and water is necessary in excess for the multiple hydrolytic processes of digestion. Pancreatic juice is the sole source of digestive enzymes within the lumen of the small intestine. Enzymes are present within the intestinal mucosa but fixed in the enterocyte cell membrane.

Nutrient Absorption

Absorption is the transfer of the products of digestion together with minerals, micronutrients, and water from the gut lumen into the blood. The mucosa of the small intestine is adapted structurally to optimize nutrient absorption and enterocytes possess specific transport mechanisms to facilitate transport. The mechanisms responsible for the absorption of lipid molecules are significantly different from those governing the absorption of other nutrients.

Sodium absorption is central to the absorption of the majority of nutrients and is therefore considered first. Indeed, the interdependence of sodium, water, and nutrients, particularly sugars, is an important element in nourishment under pathological conditions, for instance, in considering the rationale for rehydration therapy. The enterocyte membrane possesses carrier proteins for the transport of specific substrates. A specific sodium transporter exists on the enterocyte basolateral membrane exporting sodium from the cell by an energy-dependent process. The effect of the sodium transporter is to maintain the low intracellular sodium concentration.

Different processes permit the passage of the sodium ion to pass from the lumen into the enterocyte across the mucosal membrane. Sodium may pass by passive diffusion down its concentration gradient or be transported by a (second) luminal transporter in association with nutrients such as glucose and amino acids. Also, some sodium may diffuse between the cells directly into the interstitial spaces. Water molecules follow sodium ions across the membranes.

Carbohydrate Absorption

The absorption of carbohydrate molecules is directly linked to the final stage of digestion (**Figure 4**). Luminal carbohydrate digestion produces largely the disaccharide molecules sucrose, lactose, and maltose. The enterocyte luminal membranes are richly endowed with disaccharidases, which hydrolyze the disaccharides to the monosaccharides glucose, fructose, and galactose. These membrane-bound enzymes are now considered to be associated with the protein transporter molecules described previously that facilitate the transport of sodium across the enterocyte luminal membrane. Sugars are cotransported with sodium ions against their own concentration gradients—a process of great nutritional value. Once concentrated within the enterocyte, the sugars pass by facilitated diffusion across the basolateral membrane into the interstitial spaces and from there diffuse into the capillaries of the intestinal villus.

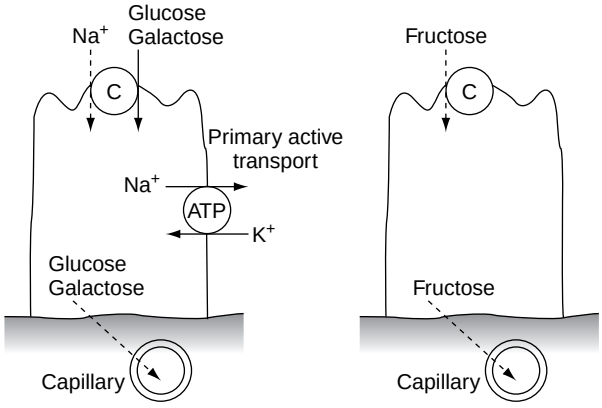


Figure 4 Carbohydrate absorption mechanisms. (Left) Secondary active transport: Glucose and galactose are absorbed using Na^+ cotransport systems. Sodium ions are removed using the (Na^+-K^+)-ATPase pump (primary active transport). (Right) Facilitated diffusion: Fructose absorption is passive but relies on a carrier molecule. Solid arrow, movement against concentration gradient; broken arrow, diffusion down concentration gradient; ATP, ATP-dependent pump; C, carrier molecule.

Protein Absorption

Similar to carbohydrate, protein digestion is incomplete within the lumen of the small intestine. However, considerably more amino acid has been produced through protein digestion, and amino acids are absorbed by secondary cotransport with sodium ion in the same manner as sugars. Some simpler dipeptides remain within the lumen, and these final peptide bonds are hydrolyzed by aminopeptidase enzymes within the enterocyte luminal membrane (**Figure 5**).

Lipid Absorption

The rate of gastric emptying is regulated to deliver packets of fat into the duodenum that may be easily digested and assimilated without overloading. The avoidance of duodenal overload is dependent on the efficacy of the neural and hormonal feedback mechanisms regulating gastric emptying and intestinal transit, the ileal brake.

The absorption of fat is a different process from that of carbohydrate and protein by virtue of the insolubility of the molecules of lipid in water (**Figure 6**). The products of lipase digestion, fatty acids and monoacylglycerols, are insoluble and so cannot diffuse through the chyme to reach the enterocyte membrane. The role of micelles produced by the action of bile salts is carriage of the lipid molecules to the enterocyte luminal membrane, where the fatty acids and monoacylglycerols diffuse from the micelle through the lipid component of the membrane to enter the enterocyte cytoplasm.

On this basis, the process of transport has always been assumed to be passive diffusion. Recently, it has been demonstrated that some carrier-dependent mechanisms are implicated. The observation that carrier dependency may be specific for certain lipids suggests that these pathways may be relevant at low lipid concentrations when the absorption of such lipids has priority. For the absorption of dietary levels of lipid, the main process of transport is probably passive diffusion, the rate being dictated by the concentration of luminal lipid.

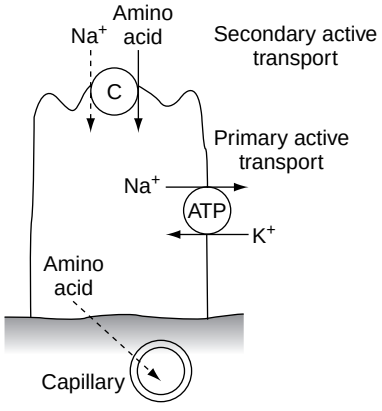


Figure 5 Amino acids (and some short peptides) are absorbed by secondary active transport using Na^+ cotransport. Solid arrow, movement against concentration gradient; broken arrow, diffusion down concentration gradient; ATP, ATP-dependent pump; C, carrier molecule.

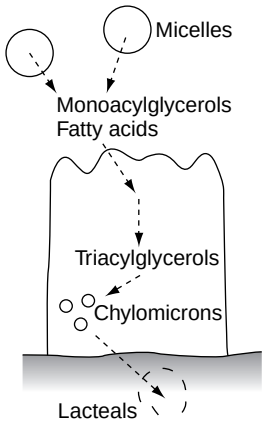


Figure 6 Lipids are absorbed by diffusion after digestion to monoacylglycerols and fatty acids. These are reconstituted into triacylglycerols and packaged as chylomicrons within the cell before entering intravillous lymphatic vessels (lacteals); these have an open endothelium (broken outline) and are therefore more permeable to larger particles than are the blood capillaries.

Lumenal micelles that have donated their fat burden across the mucosa are then free to take up more lipid from the droplet pool. Once within the enterocyte, long-chain fatty acids are bound to specific fatty acid binding proteins and transported from the cell membrane to the rough endoplasmic reticulum, where they are reesterified to reform triacylglycerols.

The intracellular triacylglycerol droplets are covered with a protein coat called apolipoprotein, which rerenders the lipid soluble once more. The coated droplets are called chylomicrons. Chylomicrons pass across the basolateral membrane of the enterocyte by exocytosis into the interstitial spaces, where they are inhibited from passing into villus capillaries by the basement membrane. Instead, chylomicrons pass into the lacteals of the villus and, from there, into the lymphatic system draining into the circulation at the great lymphatic duct behind the right atrium of the heart.

The luminal concentration of lipid, largely responsible for the rate of absorption, is more dependent on optimal intestinal mixing, peristalsis, and transit than other nutritional components.

Absorption of Minerals

Mineral absorption is normally proportional to dietary intake, with two important distinctions—the

absorption of iron and calcium, both of which can be regulated according to the needs of the body. Calcium absorption is related to the amount of specific binding protein within the enterocyte. The concentration of the calcium binding protein, which regulates calcium uptake from the gut, is secondary to vitamin D levels.

Iron absorption occurs in the duodenum and proximal jejunum. Following digestion, iron is in two forms. The first is haem iron bound to hemoglobin and myoglobin. The second form is free ionized iron in the ferrous and ferric state. Hem iron is absorbed by binding to a probable hem receptor, whereas free iron is likely to be absorbed by a specific carrier protein. Free iron is cytotoxic, so it is bound inside enterocytes to the large storage protein, apoferritin, or bound to transferrin for export to the bloodstream.

Absorption of Vitamins

Water-soluble vitamins are absorbed with water, many by sodium-dependent active transport, and fat-soluble vitamins are absorbed dissolved within the fat droplets in micelles. Vitamin B_{12} is unique in that it may only be absorbed combined with the gastric intrinsic factor by a specific mechanism in the distal ileum.

Gastrointestinal Hormones

The gastrointestinal hormones are a group of peptide molecules that increase in number every year. Most are extremely potent biological agents exerting their action at some distance from the site of secretion (endocrine) or close to the site of release (paracrine). The hormones are secreted in granules from entero-endocrine cells within the gastrointestinal mucosa in response to nutrient signals from the lumen. The precise physiologies of most of these molecules are still being clarified and the following review is restricted to those peptides on which there is broad agreement (**Table 3**).

Gastrin

Gastrin is secreted from the gastric antral mucosa in response to food in the stomach, particularly a meal high in protein. Gastrin passes into the blood and stimulates a large range of responses along the whole gastrointestinal tract involving both secretion and motility. The general physiology of gastrin is that of a promoter of gut function as food begins to pass along the gut and as a remover of the remnants of the previous meal. Gastrin secretion ceases as acidity rises within the stomach and duodenum.

Secretin

Secretin is produced by cells in the duodenal mucosa in response to the appearance of acidity and chyme. It performs at least four major functions associated with the passage of the meal through the duodenum. It stimulates the pancreatic duct and acinar cells to produce copious aqueous $NaHCO_3$ solution, it

Table 3 Functions of the major gastrointestinal hormones

Hormone	*Stimulus*	*Function*
Gastrin	Protein in the stomach	Stimulates gastric secretion Promotes gastric motility Promotes ileal motility Promotes colonic motility Relaxes ileocaecal valve Trophic action on gastric and intesintal mucosae
Secretin	Acid in the duodenum	Inhibits gastric secretion Inhibits gastric motility Stimulates pancreatic juice flow Stimulates bile flow (HCO_3) Trophic action on pancreatic exocrine tissue
Cholecystokinin	Fat/protein in duodenum	Inhibits gastric secretion Inhibits gastric motility and gastric emptying Stimulates pancreatic enzymes Gall bladder contraction Sphincter of Oddi relaxation Trophic action on pancreatic exocrine tissue Satiety signals
GIP	Duodenal chyme	Inhibits gastric emptying Inhibits gastric secretion Stimulates insulin secretion from pancreatic islets

GIP, glucose-dependent insulinotrophic peptide.

promotes a bicarbonate-rich bile flow, it inhibits gastric emptying, and it inhibits gastric secretion. The secretin secretion rate decreases as acidity production from the stomach decreases.

Cholecystokinin

Cholecystokinin is secreted from the same general duodenal mucosa as secretin but in response to the lipid and protein components of the meal. CCK has multiple roles that are not completely clear. The CCK molecule exists in a number of peptide chain lengths, each with differing properties but all with extreme potency. The effects of an elevation in CCK are as follows: Cholecystokinin is the main inhibitor of gastric emptying, slowing the release of nutrient-dense chyme from the stomach at a rate that is proportional to duodenal and pancreatic digestive function. CCK stimulates enzyme secretion from the pancreatic acinar cells and causes contraction of gallbladder wall muscle and relaxation of the sphincter of Oddi, allowing bile to flow into the duodenum.

Cholecystokinin levels directly influence levels of satiety and the perception of fullness in association with meals. Therefore, the signalling of fullness to the brain by means of CCK release is one possible pathway for the peripheral control of food intake and is of particular importance not least by virtue of the presence of CCK receptors known to be present within brain tissue. CCK secretion is inhibited by the absence of nutrients in the duodenal lumen.

GIP

Glucose-dependent insulinotrophic peptide is also known as gastric inhibitory peptide. GIP is the third duodenal peptide hormone now recognized to be the 'incretin' peptide important for the mobilization and attenuation of the insulin response to glucose as it is absorbed from the small intestine. GIP initiates the meal stimulus of the pancreatic islets, heralding the absorptive phase of metabolism. Lipid, carbohydrate, acid, and duodenal distension all serve to promote the secretion of GIP from the duodenal mucosa.

Trophic Action of Gastrointestinal Hormones

The gastrointestinal hormones gastrin, cholecystokinin, and secretin are recognized to possess trophic properties in that they are responsible for maintaining the cell populations of their target tissues. In the case of gastrin, the tissue in question is the gastric mucosa but also the intestinal mucosa; cholecystokinin and secretin are trophic for pancreatic acinar tissue.

See also: **Carbohydrates**: Chemistry and Classification. **Colon**: Structure and Function; Disorders; Nutritional Management of Disorders. **Lipids**: Chemistry and Classification. **Protein**: Digestion and Bioavailability. **Small Intestine**: Disorders. **Sodium**: Physiology. **Stomach**: Structure and Function; Disorders.

Further Reading

Johnson LR (1994) *Part V Gastrointestinal Physiology in Essential Medical Physiology*. New York: Lippincott-Raven Press.

Disorders

R D'Souza and J Powell-Tuck, Queen Mary's, University of London, London, UK

The small intestine is bathed in nutrients derived from food digested by salivary, gastric, and pancreatic enzymes and dispersed by the emulsifying effect of biliary secretions. Its mucosal cells are in intimate relationship with these nutrients, a wide variety of antigens, and the bacterial flora. This relationship is fundamental to small bowel function, which is both secretory and absorptive, providing a defense from bacteria while allowing absorption of both small and large molecules. The relationship is also crucial to the patterns of intestinal motility, which optimize mixing and surface contact during the postprandial phase and revert to a flushing, "housekeeper" role postabsorptively. Diseases of the small bowel represent a breakdown in these relationships, which characteristically result in a failure of normal absorption, inflammation, and abnormal secretory responses. The small intestine plays a fundamental role in completing digestion and absorption of carbohydrates, proteins, fats, minerals, and vitamins. Any disorder of the small intestine in its digestive or absorptive capacity or of its immune system may lead to nutritional abnormalities.

Food and Bacterial Interaction

Bacteria—anaerobes, lactobacilli, enterococci, and gram-positive aerobic organisms—colonize the gut after birth; in the upper small intestine there are 10^2–10^3 bacteria/ml and up to 10^9 bacteria/ml are present in the distal small intestine. The small intestine acts as an effective barrier, and together with the innate and acquired immune system, defends the body against foreign microorganisms. Bacteria,

derived from ingested food that is rarely sterile, colonize the gastrointestinal tract soon after birth. After eating, more than 100 species of bacteria are found in the upper small intestine but ingested bacteria are greatly reduced in number by gastric juices in the hour following the meal. The bacteria in the distal small intestine are not affected by the gastric secretions. The human gut has mechanisms to protect itself against environmental agents that are foreign to the body (immunity). Immunity can be either innate or acquired. Innate immunity comprises those defense mechanisms that a person is born with that can be mobilized on very short notice. There are a multitude of innate mechanisms that include barriers to infection (thick mucus, ciliary action, and gastric acid), phagocytic cells (neutrophils and macrophages), defense due to complement and humoral (antibody-mediated) mechanisms, and antibactericidal peptides such as α-defensins, trefoil factor peptides, and cathepsin G secreted by intestinal epithelial cells and Paneth cells at the base of the intestinal crypts. Adaptive immunity is more specialized and produces both a humoral and a cell-mediated immunity. It differs from innate immunity in that it remembers a previous infection and responds quickly to reexposure to a known foreign substance. Bacteria in the intestinal lumen are probed by dendritic cells, which reach out into the lumen to sample the bacteria and present them to lymphocytes in the lymphoid tissue present in the gut wall. Intestinal M cells achieve similar presentation by importing bacteria into the intestinal lymphoid tissue.

Infections/Food Poisoning

The vitally important area of food poisoning is dealt with in more detail elsewhere. This section outlines the principal infections affecting the small intestine. Bacteria such as *Campylobacter jejuni*, enteroinvasive *Escherichia coli* (EIEC), shigella, salmonella, and yersinia can cause food poisoning, which is dominantly colitic (dysentery) with a bloody, pus-containing diarrhea, by invading the mucosal surface. Enterohemorrhagic *E. coli* produce a bloody diarrhea secondary to adherence in the colon and toxin production. These then dominantly affect the colon. Yersinia and (non-typhoid) salmonella dominantly invade the lower small bowel (ileum), producing low abdominal pain and diarrhea. Yersinia infection can be acute or occasionally more chronic, mimicking Crohn's disease. Tuberculosis is another chronic infection that can affect especially the terminal ileum and mimic Crohn's disease. Other infective agents, such as enterotoxigenic *E. coli* (ETEC) and *Vibrio cholerae*, cause an acute watery diarrhea by producing toxins that activate intestinal cellular enzymes; they have their main effect on small bowel water and electrolyte transport systems. Viruses are common causes of gastroenteritis, with vomiting and diarrhea, particularly in young children (rotavirus, caliciviruses, enteric adenoviruses, etc.). Norwalk virus causes vomiting and diarrhea in all age groups and rotavirus is not uncommon in adults.

When a cause is found, travellers diarrhea is most commonly due to ETEC and other *E. coli*, campylobacter, shigella, and rotavirus. Persistent diarrhea sometimes arises from *Giardia lamblia* and other protozoal diseases. The source is usually contaminated food or water, and advice to avoid uncooked fruit, vegetables, and salad, uncooked meat and seafood, tap water, ice, etc. in high-risk areas is often not followed, perhaps because it is seen to detract from the very essence of most leisure travel. The problem of emerging drug resistance complicates the use of antibiotic prophylaxis.

Other infections that can affect the small bowel are fungi in immunocompromised patients and helminths and protozoans in tropical countries. These are beyond the scope of this article. *Giardia lamblia* is a protozoan that on colonizing the small bowel can cause diarrhea and malabsorption.

Patients with HIV and AIDS often have diarrhea caused by the effect of HIV on the gastrointestinal mucosa, but they are also prone to many viral, bacterial, fungal, and parasitic infections due to their immunosuppression.

Whipple's Disease

This is a rare multisystem disorder due to infection by an organism called *Tropheryma whippelii* that invariably affects the small bowel and may also be associated with skin pigmentation, arthralgias, endocarditis, and other heart problems and central nervous system abnormalities. In the small bowel, it causes malabsorption with diarrhea, steatorrhea, abdominal pain, distension, and weight loss. It is diagnosed by the presence of periodic acid-Schiff-positive granules in macrophages.

Electron microscopy shows accumulation of *T. whippelii* within these macrophages and the condition that was once fatal can now be cured by antibiotic treatment. Fluid and electrolyte replacement may be required. Most patients are malnourished until absorptive function returns to normal, and they should receive a diet high in protein and calories together with multivitamin preparations.

Anemic patients should receive iron or folate as required. Vitamin D, calcium, and magnesium should be administered until steatorrhea resolves.

Bacterial Overgrowth Syndrome

Bacterial overgrowth occurs due to stasis of luminal contents resulting from strictures, fistulas, jejunal diverticula, blind loops, pouches, or motility disorders. The intestine becomes excessively colonized with both anaerobic and aerobic organisms. The latter use vitamin B_{12} and other nutrients for their own metabolism; they deconjugate bile salts and in so doing prevent their use in fat absorption. Normal dietary fat should be substituted by medium-chain triglycerides. Hypoalbuminemia (low blood albumin concentration) may develop and cause oedema. Antibiotics should be given to eradicate the bacteria; surgery is sometimes required.

Food Allergy

Food allergy is an adverse reaction to food in which the reaction is immune mediated. The immunological response comprises food-specific antibodies (IgE mediated), immune complex production, and mucosal T cell-mediated reactions. Food allergy can be serious and lead to anaphylactic shock. This must be distinguished from food intolerance, which is a reproducible adverse reaction to a specific food or food ingredient, either as a result of abnormal absorption of a sugar due to an enzyme deficiency (e.g. lactose intolerance) or because of an exaggerated pharmacological response to chemicals in food, such as tyramine in cheese.

Food allergy is most common in infants and tends to become less of a problem as children age. Foodstuffs implicated are cows' milk (2.5% of infants reducing to approximately 0.5% after 3 years) and egg allergy, which usually disappear after the age of 5 years. Fish legumes, peanuts, soy, and cereals can all cause food allergies. True food allergy is relatively uncommon in adults, although food intolerance is not.

The history is vital in discovering food allergies and skin prick tests may also be useful. Dietary manipulation plays an important role and elimination diets with slow reintroduction of suspect foods may be necessary. Once the food allergen is discovered, treatment is mainly avoidance.

Celiac Disease

Celiac disease or gluten-sensitive enteropathy is the major small intestinal cause of malabsorption in the Western world. In the United Kingdom, it occurs in 1 in 120–300 of the population; it is more frequent in females. It may be associated with dermatitis herpetiformis, type 1 diabetes, rheumatoid arthritis, and autoimmune thyroiditis. It usually involves the upper part of the small bowel and is due to a T cell-mediated immune response to a constituent of wheat flour, gluten. This toxic component of gluten (identified as gliadin) causes damage to surface enterocytes with villous atrophy and crypt hyperplasia. There is marked reduction in absorptive capacity of the small bowel with resultant malabsorption of nutrients. Prompt clinical improvement occurs on withdrawal of the gliadin fraction of wheat gluten. Patients present with symptoms of diarrhea, steatorrhea, weight loss, and general malaise. Sensitivity to gluten occurs in other diseases such as dermatitis herpetiformis and results in a rash and similar intestinal changes to celiac disease.

The diagnosis is made by testing the blood for the presence of anti-endomysial antibodies or anti-transglutaminase antibodies and confirmed by duodenal biopsy, which shows villous atrophy and crypt hypertrophy, flattened surface epithelial cells, and a chronic inflammatory infiltrate into the lamina propria that resolve following a gluten-free diet. All cereal grains contain gluten (wheat, barley, and rye), and these should be removed from the diet. Maize flour, rice, potatoes, and gluten-free flour are available and acceptable. Response to oats is more difficult to predict, perhaps because they may be contaminated by wheat products. It often takes several weeks to respond to a gluten-free diet and response is seen by rapid gain in weight. Lifelong adherence to the diet is necessary for celiac patients.

Some patients present dehydrated, and fluid and electrolyte replacement may be required. Patients are often anemic secondary to iron, folate, or, much less often, vitamin B_{12} deficiency and should receive supplementation. Vitamin D and K deficiency is common and absorption of vitamins B_6 and C may also be impaired. Steatorrhea may also cause calcium, magnesium, or zinc deficiency. All nutritional deficiencies should be corrected until the intestinal lesion has resolved on the gluten-free diet.

In untreated celiac disease, as in other diseases that affect the absorptive capacity of small intestinal enterocytes, disaccharidase deficiency may result, with symptoms of diarrhea and abdominal distension in response to milk products that contain lactose. Once the mucosa returns to normal with the gluten-free diet, disaccharidase activity returns to normal and lactose (milk sugar) becomes tolerated again.

Tropical Sprue

Tropical sprue is a disease of unknown etiology that exclusively affects people living in or visiting specific tropical or semitropical areas, particularly Southeast Asia and the Caribbean, that resembles celiac disease in its intestinal changes. It presents with steatorrhea, abdominal cramps, bloating, and increased bowel sounds. Most patients improve or are cured with long-term broad-spectrum antibiotic treatment. Fluid and electrolyte replacement should be administered to the acutely dehydrated patient, and megaloblastic anemia is often secondary to folate or B_{12} deficiency, which should be corrected. Vitamin A deficiency may be an important feature. Sometimes, iron, calcium, or magnesium may be deficient, and these minerals should be replaced. Antibiotic treatment may be helpful.

Crohns' Disease

Crohns' disease is a chronic inflammation that can involve any part of the gastrointestinal system from the mouth to the anus. In the small intestine, Crohns' disease typically affects the ileum; however, other parts of the small intestine can be affected. It produces a segmental, full-thickness inflammation, with formation of fissures and abscesses, fistulas, and fibrotic stenoses, and it may follow a chronic relapsing course. The cause of Crohns' disease is unknown, although several hypotheses such as infections (mycobacteria, chlamydia, and viruses) have been put forward. The discovery of susceptibility gene variants in NOD2/CARD-15 suggests predisposition to the disease by abnormalities of the intracellular sensors of bacterial lipopolysaccharide and their effect on the caspase and nuclear factor-κB signalling pathways and macrophage function.

The inflammation of Crohns' disease can be treated by the use of defined formula artificial feeds or by the use of drugs, including aminosalicylates, antibiotics, steroids, and immunosuppressives such as azathioprine/mercaptopurine, methotrexate, and infliximab. There is evidence that improving nutritional status alone will reduce Crohns' disease activity. If these measures fail to control the disease, then surgical resections of the diseased segment may be necessary.

Crohns' disease presents an important model for the role of nutritional management in small bowel disease. Nutritional treatment of Crohns' disease may use defined formula feeds to reduce inflammation or may seek to correct nutritional deficiency consequent on reduced food intake (loss of appetite, nausea, abdominal pain, or diarrhea), malabsorption, or the changes in protein and energy metabolism that occur secondary to inflammation. It therefore seeks to:

- maintain adequate nutrition and correct any nutritional deficiencies;
- reduce disease activity;
- maintain nutrition against the background of intestinal failure or short bowel syndrome on a long-term basis; and
- treat or prevent growth failure (consequent on nutritional deficiency and inflammation combined) in children.

A wide range of nutritional deficiencies can arise in patients with Crohns' disease of the small bowel that result in defects in wound healing, increased susceptibility to infection, and specific nutrient deficiency syndromes. Patients become anemic from intestinal bleeding; inadequate iron, folate, and B_{12} intake; failure to absorb iron in the duodenum or B_{12} in the ileum; or because of impaired folate absorption. Vitamin B_{12} can be given by injections to avoid the problem of absorption, folic acid by mouth, and iron orally or, if necessary, by injection. Trace elements and vitamins that are deficient during relapses include zinc, ascorbic acid, calcium, and the fat-soluble vitamins. Osteomalacia, rickets, and osteoporosis may occur due to steroid treatment or malabsorption of calcium and vitamin D. Zinc deficiency can cause mouth and skin problems and results from loss of zinc in watery diarrhea.

Chemically defined artificial liquid diets can achieve induction of remission, although the response rate is slightly less than that achieved with steroids. They are particularly beneficial in children with growth failure or in patients with steroid-resistant disease.

Enteral nutrition is preferred to parenteral nutrition, but parenteral (intravenous) nutrition may be life-saving in patients with short bowel syndrome following multiple resections for Crohns' disease or in patients with treatment-resistant Crohns' disease, and it may have a place in supporting malnourished patients at the time of surgery.

Patients with Crohns' disease tend to self-select low-fiber containing foods and often feel bloated after eating foods high in non-starch polysaccharides. Exclusion diets have been tried with some success in Crohns' disease, often following initial treatment with liquid feeds, but there seems to be little consistency between which foods' exclusion benefits different patients. Some patients develop malabsorption of milk sugar (lactose) and a small minority are said to be intolerant of milk protein. Other diets such as yeast exclusion have their

advocates but there is little evidence for consistent benefit. Dietary fat reduction should not usually be recommended in underweight patients because the body gains nine calories for each gram of fat eaten, whereas nonhydrated protein and starch provide four calories each. Milk and milk products provide important, easily assimilable sources of protein, energy, and calcium and their exclusion should not be undertaken lightly.

Intolerance to lactose, the sugar of milk, increases with age. Patients who are lactase deficient (the enzyme required to split lactose into two smaller sugars that can be absorbed) pass unabsorbed lactose into the large bowel with resultant cramps and abdominal distension. Any patient with extensive inflammation in the small bowel, where lactase is located, may develop lactose malabsorption, which improves once the inflammation resolves. Evidence of lactose intolerance should be documented before withdrawal from the diet and calcium supplementation considered. Occasionally, a lactose breath test is required to detect lactase deficiency. Some patients with lactose intolerance who still wish to take milk can have lactase enzyme added to the milk or use lactase tablets that can be eaten prior to or while eating foods rich in lactose. Patients can eat aged cheese (reduced lactose content) or yoghurt made with live bacterial culture.

Patients with narrowing of the bowel (strictures) should consider a low-fiber diet to reduce the risk of intestinal obstruction. A low-fiber diet may be deficient in folic acid, ascorbic acid, calcium, and some B vitamins, and these can be supplemented.

Patients with small bowel Crohns' disease with resections may have diarrhea resulting from fat malabsorption (steatorrhea) and may have to restrict their fat intake from 100 to 70 g of fat. If steatorrhea still occurs, medium-chain triglycerides (6–10 carbon atoms) should be substituted for the normal fat diet because these are absorbed directly into the portal system. Oil, powders, or emulsions containing medium-chain triglycerides can be added to food or used in cooking and baking.

Patients with ileal resections due to Crohns' disease are susceptible to calcium oxalate kidney stones. These patients should have a low oxalate diet, modest dietary fat restriction, and dietary supplementation of calcium.

Hypogammaglobulinemia

Common variant hypogammaglobulinemia is a defect of acquired immunity. Most of these patients have diarrhea and malabsorption. Patients are prone to *G. lamblia* infections and small bacterial overgrowth. The disease may result in intestinal inflammation and is associated with celiac disease, refractory sprue, and Crohns' disease.

Stool cultures for bacteria, ova, and parasites together with duodenal aspirates may be needed. Duodenal biopsies show loss of plasma cells in the lamina propria. Patients may be treated with metronidazole for giardiasis or bacterial overgrowth and with gluten-free diet for celiac disease.

Sugar Malabsorption

The disaccharidases are localized to the apical cell membrane of the villous absorptive gut epithelial cell. Lactase, the disaccharidase that digests lactose, is the most important. Deficiencies of these enzymes may be due to a primary inherited enzyme disorder (permanent) or secondary to disorders resulting from mucosal damage or bacterial overgrowth. Disaccharidase deficiency results in an osmotic diarrhea, because the undigested sugar has a large osmotic pull, and abdominal distension.

Intestinal Pseudo-obstruction

Intestinal pseudo-obstruction encompasses several intestinal motor disorders characterized by episodes that suggest intestinal obstruction because defecation stops and abdominal distension, pain, and vomiting occur, but in which no mechanical obstruction is found. It may be due to primary abnormalities of the visceral muscle or nerves or be secondary to chronic renal failure, hypothyroidism, diabetes mellitus, amyloidosis, scleroderma, or muscular dystrophy. There is no effective treatment that is specific for intestinal pseudo-obstruction. If the patient has bacterial overgrowth, this should be treated with antibiotics. If nutrition is impaired, administration of liquid, low-residue feeds enterally is required; rarely, parenteral (intravenous) feeding is necessary.

Collagen Vascular Diseases

Collagen vascular disorders, such as polyarteritis nodosa, Churg–Straus disease, scleroderma, and systemic lupus erythematosus, may affect the small intestine, resulting in malabsorption and diarrhea and infarction. Malabsorption due to bacterial overgrowth can be treated with a course of antibiotics. Steroids and immunosuppressives are the mainstay of treatment.

Vascular Disease of the Small Intestine

Ischemic injury to the small intestine occurs due to obstruction to the superior mesenteric arterial supply by embolization of clot or local thrombosis of the vessels. It can also occur as a result of diminished blood supply to the small intestine due to conditions such as systemic hypotension or hyperviscosity states. Occlusion of the superior mesenteric artery results in ischemia or infarction of nearly all the small intestine and the right half of the colon. It may present acutely with sudden severe abdominal pain and shock consequent on the infarction with or without perforation of the affected intestine. In these circumstances, emergency surgery to remove the large segment of affected intestine is life-saving but may leave the patient dependent on long-term intravenous feeding (parenteral nutrition). Ischemia may present more chronically with poorly localized abdominal pain, typically occurring 30–60 minutes after food intake. This leads to food avoidance and weight loss. The intestinal mucosa can become ulcerated. The diagnosis depends on mesenteric angiography.

Malignant Tumors of the Small Intestine

The small bowel, like the stomach, is a common site for primary lymphoma. Tumors may originate from B lymphocytes (particularly mucosa-associated lymphoid tissue (MALT) lymphoma, which in the small intestine is termed immunoproliferative small intestinal disease (IPSID), Burkitt's lymphoma and immunodeficiency-related lymphomas, or T lymphocytes (particularly enteropathy-associated T cell lymphoma (EATL)), or it may mimic any of the lymphomas which arise in lymph nodes distant from the intestine, such as diffuse large B cell lymphoma and mantle cell lymphomas.

Intestinal lymphomas are another interesting example of disease originating from the interplay between the intestinal mucosa and the intestinal flora. In the stomach, MALT lymphoma is closely related to *Helicobacter pylori* colonization and may be cured by *H. pylori*'s eradication. IPSID may be responsive to antibiotics, also suggesting a relationship with small intestinal flora. EATL results from chronic malabsorption, most commonly celiac disease, and occurs mainly in the jejunum. It usually involves multiple segments of the small bowel and may therefore be difficult to resect surgically.

Small intestinal tumors are rare. The majority are adenocarcinomas that develop most commonly in the proximal small bowel. Metastasis by hematogenous or lymphatic spread is common with peritoneal seedlings. Rarely, secondaries from primary tumors elsewhere can be found in the small bowel.

Benign Tumors

Benign tumors are usually asymptomatic but may present with obstruction, intussusception, or bleeding. They include adenomas, leiomyomas (smooth muscle tumors), neurofibromas (isolated lesions or part of neurofibromatosis), or hematomas (Peutz–Jeghers syndrome characterized by multiple hamartomatous polyps). They can bleed or cause intussusception, in which the polyp, attached to the intestinal lining, is moved down the intestine by intestinal peristalsis and causes the intestine to telescope over itself, thereby resulting in obstruction.

Neuroendocrine Tumors

The small bowel is the most common site for carcinoid tumors that arise from cells of the neuroendocrine tissue. Carcinoid syndrome can occur if the tumor spreads to the liver, with episodes of flushing and diarrhea. Surgical resection of carcinoid tumors should occur. Once metastases have developed, various antiserotonin agents or octreotide can been used.

Congenital

The duodenum derives from the distal part of the foregut, whereas the jejunum and ileum derive from the midgut. Congenital disorders such as atresia (failure to canalize) or stenosis (constriction) or malrotation may occur during early development. Rarely, duplication of the bowel can occur with the risk of obstruction and bacterial overgrowth. Children with cystic fibrosis may develop small bowel obstruction due to meconium ileus (viscous mucus).

Meckel's Diverticulum

Meckel's diverticulum is a congenital anomaly located in the ileum in approximately 2% of the population. It can present with melaena or abdominal pain. In some patients, it can lead to intussusception and small bowel obstruction. It can be diagnosed with a radionucleotide Meckel's scan and treatment is by surgical excision.

Radiation Enteritis

Radiotherapy treatment of pelvic and abdominal malignancies can lead to radiation damage of the

small bowel and colon. Radiation damage to the small bowel and previous resections may cause decreased absorptive surfaces, bile salt malabsorption, and increased intestinal transit and hence nutritional deficiencies. Late radiation damage may cause ileal dysfunction with increased gallstones secondary to hyperoxaluria. Radiation-induced strictures and fistulas may cause stasis of bowel contents resulting in bacterial overgrowth and malabsorption.

Steatorrhea should be treated by reducing fat intake or replacing fat with medium-chain triglycerides with vitamin or mineral deficiencies corrected parenterally.

Abetalipoproteinemia

This is an autosomal recessive condition with abnormal lipid metabolism, retinitis pigmentosa, neurological deterioration, and acanthocytosis (abnormal red blood cells). The intestines and liver are unable to synthesize triglyceride-rich lipoproteins, chylomicrons, and very low-density lipoprotein. As a result, there is steatorrhea due to malabsorption of fat together with malabsorption of fat-soluble vitamins. Symptoms can be corrected by substitution of the normal fat intake with medium-chain triglycerides and polyunsaturated fat as corn oil. As they grow older, patients may be able to increase their fat intake. High-dose oral supplementation of fat-soluble vitamins, especially vitamin E, is mandatory from a young age because lack of these vitamins can cause progressive neurological and retinal deterioration.

Intestinal Lymphangiectasia

Intestinal lymphangiectasia is a protein-losing enteropathy with gastrointestinal lymphatic obstruction and excessive leakage of plasma protein into the intestinal lumen, with resultant oedema and hypoproteinemia. Patients have symptoms of diarrhea, steatorrhea, nausea, vomiting, and signs of ascites or pleural effusions. Bacterial infections due to lymphopenia and hypogammaglobulinemia occur. Patients often have malabsorption of fat-soluble vitamins and can also have hypocalcemia and hypomagnesemia. These must be supplemented.

Dietary management is often adequate to maintain remission. Patients should restrict their fat diet to less than 5 g per day. A high-protein, fat-free diet with medium-chain (C8–C10 fatty acids) triglycerides should be encouraged. Medium–chain triglycerides are absorbed directly into the portal system and are palatable and do not increase lymph flow.

Low-sodium diets, diuretics, and intravenous albumin have been used to alleviate symptoms. When these do not work, periods of fasting and parenteral nutrition can be used to alleviate very troublesome symptoms, although essential fatty acids must be added to the diet. Surgical resection is sometimes required.

Eosinonophilc Gastroenteritis

Eosinophilic gastroenteritis affects any part of the bowel from the mouth to the anus, although the stomach and small bowel are the most affected areas. Different layers of the bowel can be affected. It is characterized by eosinophilic infiltration of the affected areas of the bowel and a peripheral blood eosinophilia. It is a rare disorder for which the etiology is unknown.

In patients with mucosal layer disease (i.e., no involvement of muscle or serosal layer), food intolerance should be looked for. A trial of sequential elimination of milk, pork, beef, eggs, or gluten may be tried. Enteral elemental diets have been tried in some patients with success. Patients with mucosal layer disease who do not respond to dietary measures or patients with deeper layer involvement should try a course of steroids. Any patient who has traveled to the tropics should be considered for antihelminth therapy.

See also: **Celiac Disease**. **Colon**: Disorders; Nutritional Management of Disorders. **Lactose Intolerance**. **Small Intestine**: Structure and Function. **Stomach**: Disorders.

SOCIO-ECONOMIC STATUS

E Dowler, University of Warwick, Coventry, UK

In most societies there is an inverse relationship between socio-economic status and nutritional status, whether the former is measured as income, assets, or social class and the latter as food consumption or body dimensions. Socio-economic status is difficult to define consistently. The problems include how to define and measure basic subsistence, whether to examine the status of individuals or households, how to define and interpret social class differences, and how to understand differences in status between societies and in one society over time. People's circumstances often change, and sometimes the process of change may affect nutritional outcomes as well as the conditions themselves.

Basic subsistence usually means a level of living that enables people to survive and reproduce. Reproduction here means both physiological (women of childbearing age are menstruating and can conceive and carry children to term) and economic (a household has the means to feed, clothe, and house its members and take part in minimum social and economic exchange). Quantifying this basic level can be difficult: People disagree over the cutoff levels to define minimum energy or nutrient intakes for survival or reproduction, over the rates of child growth that are healthy, and over how much money or its equivalent is needed for a household to survive. Resolving these controversies depends partly on technical decisions and partly on agreed social value judgments. It also partly depends on who is making the decisions: experts in nutrition, health, or economics; social policy decision makers; or people living in the conditions concerned.

Disentangling what happens within the household—who owns or has access to what assets and resources—is difficult. Even someone living alone may be entitled to resources or a share of production from another household. However, most surveys use household data on consumption or asset ownership and then divide these by the number of household members to estimate individual shares. This approach ignores differences within a household over the control and management of resources, and it assumes everything is shared equally. (What would 'equally' mean for nutrients? Everyone eating the same, or everyone eating what they need?)

Social class and status are affected by social and cultural conditions as well as economic conditions; most classifications try to incorporate all three factors. Some of the work described here uses proxy measures of socio-economic status as an alternative to income. These proxies can be single indicators, such as mother's educational level or head of household's occupation. Alternatively, they can be composite indices, often drawing on census data, that combine single indicators such as crowding, occupation, car access, housing tenure, or employment status into one index (e.g., the Carstairs Index in the United Kingdom). These composite indicators usually describe households' rather than individuals' socio-economic status.

Current social and economic changes that are associated with increasing socio-economic differentials in developed countries include the growth in unemployment or insecure employment, unstable family structures, homelessness, migrancy, and reductions in welfare provision. Similar factors are in force in developing countries, with the addition of economic structural adjustment, globalisation of production and trade, and the spread of HIV/AIDS, which affects production and earnings through the illness and death of economically active young adults.

There is debate about the relationship between socio-economic status, nutrition (food and growth), and work productivity. Some argue that undernutrition contributes to poverty at the household, and potentially national, level because an inadequate diet can reduce an individual's work capacity either through a direct effect of food or indirectly over time because the individual is small and/or thin. Reduced work capacity results in lower productivity and economic return, which in turn contribute to increasing poverty and malnutrition—a 'vicious circle.' These effects may be seen in an individual and his or her immediate household or as part of the reduction in human capital in malnourished populations across generations.

Socio-economic Status and Disease Risk and Mortality

In the 1970s, the World Fertility Survey documented very large socio-economic differentials in child mortality in most of the 41 developing countries that participated. **Table 1** shows socio-economic differentials in mortality rates during the first 5 years of

Table 1 Mortality rates during the first 5 years of life (per 1000 births) for nine countries participating in the World Fertility Survey

Country	*Mortality rate per 1000 births*											
	Mother's education (years)				*Husband's occupation*				*Husband's education (years)*			
	0	*1–3*	*4–6*	*7+*	*Agriculture*	*Skilled/unskilled*	*Sales/service*	*Professional*	*0*	*1–3*	*4–6*	*7+*
Lesotho	(224)[a]	215	185	169	(157)	193	(214)	(116)	185	212	196	169
Kenya	181	164	128	111	183	172	148	101	185	208	159	127
Sudan	146	109	109	109	158	138	127	95	155	152	104	84
Peru	237	171	98	55	218	157	123	59	241	217	150	75
Mexico	153	118	87	50	143	103	98	61	155	121	96	53
Panama	134	90	52	43	84	61	53	40	100	91	60	45
Bangladesh	222	198	186	(122)	216	236	208	152	230	221	191	176
Philippines	130	118	94	53	106	82	64	46	118	122	96	62
Malaysia	67	64	56	18	72	59	48	40	88	59	60	40

[a]Rates based on fewer than 500 births are shown in parentheses.
Data from Hobscraft JN, McDonald JW, and Rutstein SO (1984) Socio-economic factors in infant and child mortality: A cross-national comparison. *Population Studies* **38**: 193–223.

life in nine of the countries that participated. During the past two decades, overall child mortality rates have declined by approximately 50%, but 10 million children younger than 5 years of age still die annually; socio-economic differentials in child survival remain and in some cases have widened. Being underweight (2 SD or more below expected weight for age) and micronutrient deficiencies are recognised as underlying causes of death from infectious diseases; fetal malnutrition, manifested as low birth weight, and not being breast-fed contribute to neonatal mortality.

The reduction in infant and child mortality rates and concomitant decline in fertility in many developing countries have led to changes in population age structure, overall health status, and mortality patterns. A shift toward an older population means adults diseases become as important as those of children. However, poorer adults certainly die younger than richer adults, whatever the causes. Many developing country populations seem to be shifting from an environmental exposure predominantly to infectious diseases (from poor water, sanitation, and insecure food supply) to one in which noncommunicable disease risks are more important (motor vehicles, unsafe workplaces, air pollution, smoking, alcohol, increased fat and sugar intakes, and decreased physical activity and social support). Current hypotheses on the fetal origins of some chronic diseases highlight the role of the early nutritional environment for the risk of adult diabetes and cardiovascular disease. Those in low socio-economic groups in developing countries, where low birth weight and inadequate infant nutrition are common, may face the double risk of infectious disease or death in childhood and, where children survive, noncommunicable diseases in adulthood.

These shifts in demographic, disease, and risk profiles, often referred to as the epidemiological transition, are occurring at different rates in most developing countries. **Table 2** shows mortality rates by major cause of death and by region, differentiating areas of high and low mortality. Understand causal determinants of health and being able to predict the consequences of changing exposures to risk are important. Expert working groups have reviewed data on the prevalence of risk factor exposure and hazard level for 14 epidemiological regions of the world. The contribution of these risk factors to global and regional mortality and burdens of disease has been calculated. **Figure 1** shows deaths and **Figure 2** shows disease burden in disability adjusted life-years on a global basis, differentiating areas of high and low mortality. One key finding of this analysis is the importance of nutrition in health worldwide. Approximately 15% of the global disease burden is attributable to childhood or maternal underweight or micronutrient deficiency, and a similar percentage is attributable to dietary risk factors such as high blood pressure, blood cholesterol, body mass index (BMI), and low fruit and vegetable intake.

The situation in developing countries is changing with time, and quite rapidly. For example, in Porte Alegre, Brazil, the poorest men have higher death rates from cancer and cardiovascular disease than wealthier men. Conversely, large population-based surveys in India suggest that the prevalence of coronary heart disease is nearly four times greater in those from higher than in those from lower

Table 2 Death by cause and mortality stratum in WHO regions[a]

Cause of death	*% total mortality*	*Africa*		*The Americas*			*Eastern Mediterranean*		*Europe*			*Southeast Asia*		*Western Pacific*	
		hc ha	*hc vha*	*vlc vla*	*lc la*	*hc ha*	*lc la*	*hc ha*	*vlc vla*	*lc la*	*lc ha*	*lc la*	*hc ha*	*vlc vla*	*lc la*
Respiratory disease	6.3	105	129	186	163	20	19	125	213	80	126	130	693	59	1513
Malignant neoplasm	12.6	241	303	645	392	74	80	199	1059	293	515	231	882	343	1859
Diabetes mellitus	1.6	20	35	78	137	15	16	36	90	30	21	62	176	17	162
Cardiovascular	29.3	482	503	1106	773	100	280	757	1760	1111	2171	571	3226	395	3350
Infection	19.3	2052	3529	65	197	117	55	957	48	64	86	373	2699	24	678
Nutritional deficiencies	0.8	71	80	8	40	11	4	39	7	3	2	36	153	2	21
Maternal conditions	0.9	97	143	0	13	8	3	66	0	1	1	22	132	0	21
Perinatal	4.4	294	283	17	124	26	26	287	10	42	18	91	952	2	351

[a]Estimates for 2001 for all member states.

Figures are deaths per 100 000 per year.

mortality stratum: vlc, very low child; lc, low child; hc, high child; vla, very low adult; la, low adult; ha, high adult; vha, very high adult.

Data from WHO (2002) *World Health Report*. Geneva: World Health Organization; Ezzati M *et al.* (2002) Selected major risk factors and global and regional burden of disease. *Lancet* **360**: 1347–1360.

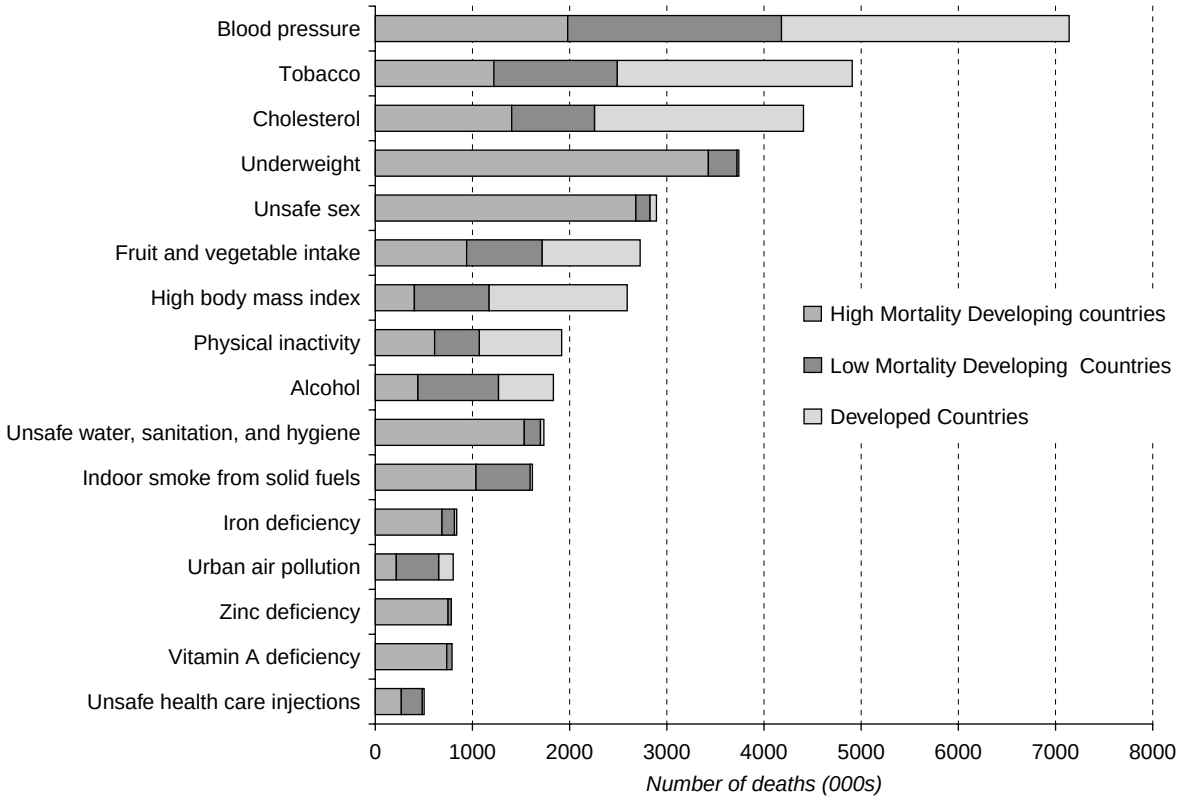


Figure 1 World deaths in 2000 attributable to selected leading risk factors. (Reproduced with permission from the World Health Organization (2002) *World Health Report*. Geneva: World Health Organization.)

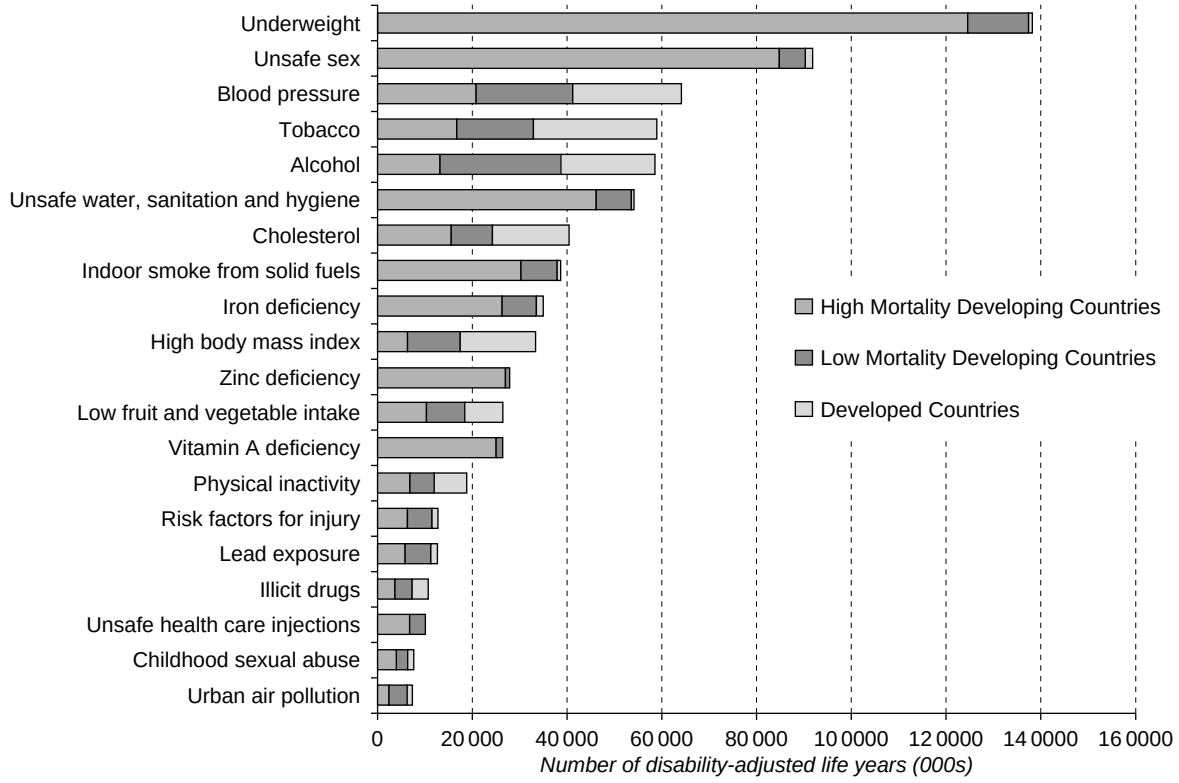


Figure 2 World disease burden in 2000 attributable to selected leading risk factors. (Reproduced with permission from the World Health Organization (2002) *World Health Report*. Geneva: World Health Organization.)

socio-economic groups. However, obesity rates are increasing throughout the world; in India, there are households with both an underweight child and an overweight adult. It seems increasingly likely that poorer people face greater risks than wealthier people from both infectious and parasitic diseases and from noncommunicable diseases.

The inverse relationship between socio-economic status and differential morbidity or premature mortality in developed countries is well documented. These differentials are seen in the main causes of mortality, coronary heart disease, and cancers, and they have particularly widened, and continue to widen, in countries where income inequality has increased. The differences seem to be cumulative: Those born into low socio-economic groups and who remain in them throughout childhood and adulthood have higher mortality risk than those who move up a class during their lifetime.

Studies have used various socio-economic indicators, such as occupational class, employment status, housing tenure, and car access. The mortality differentials are large and can be located in terms of people and places. **Table 3** shows data from the United Kingdom; the association between mortality and spatial indicators of economic and social advantage or disadvantage is very strong. However, the increasing polarization of wealth and poverty between households (a growth in two-earner and no-earner households) and places means, in fact, lower death rates occur where richer people are concentrated and higher death rates occur where disadvantaged people are concentrated. It is not just a matter of 'unhealthy places.'

Finally, there is evidence of a gradient of health and mortality rather than a threshold below which everyone does badly and above which most are reasonably healthy. The UK Whitehall longitudinal studies of civil servants showed that those in the second highest grade (not a low socio-economic group) had worse health and mortality risk than those in the highest grade. However, those in the lowest grade had the worst health and mortality risks, and studies in several countries have shown long-term unemployed to have high mortality risks.

Current thinking about why these differentials occur is that artefactual distortions or social selection cannot account for them, but that social conditions or material factors, perhaps partially mediated by life circumstances, beliefs, and behavior, offer the best explanations. Money is clearly important—having enough to meet basic needs but also for a sense of security. Psychosocial factors and the control people have over their working or social environment also play an important part. Indeed, increasing income inequality and insecurity, as much as absolute poverty, may be important predictors of mortality differentials. Biology also plays a part: The fetal experience, and nutritional challenge during the early years, may set an individual on a biological trajectory that is not able to withstand the effects of material and social poverty.

Socio-economic Differences in Food Consumption

The socio-economic status of a household can affect the members' diet in a number of ways. In general, those with higher incomes tend to spend a smaller proportion of that income purchasing food, although they may spend a larger absolute amount than those with lower incomes. Second, the kinds of foods that wealthier people buy will be different in quality (less contamination and better processed or packaged) and in nutrient density (e.g., see **Table 4**). Wealthier people are generally 'less efficient'

Table 3 Body mass index (BMI), children's nutritional status, and morbidity

Variable	*Cluster 1 (n = 178)*	*Cluster 2 (n = 190)*	*Cluster 3 (n = 124)*	*Cluster 4 (n = 67)*	p^a
	Self-employed	*Dependent self-employed*	*Casual unskilled, female head*	*Casual skilled*	
BMI, mean (SE)	19.1 (2.5)	18.8 (2.0)	16.3 (0.5)	26.1 (3.3)	<0.0001
WAZ, mean (SE)	−3.25 (3.8)	−3.58 (4.1)	−5.21 (5.2)	−3.20 (3.7)	<0.0001
HAZ, mean (SE)	−2.30 (1.4)	−2.10 (2.3)	−3.18 (3.2)	−2.25 (1.6)	<0.0001
WHZ, mean (SE)	−2.19 (4.5)	−2.30 (5.5)	−3.25 (4.9)	−2.20 (3.4)	<0.0001
Days off ill					
With fever, mean (SE)	1.50 (0.51)	1.82 (0.07)	1.88 (0.08)	1.90 (0.02)	<0.015
With diarrhea, mean (SE)	3.50 (1.02)	3.17 (0.48)	7.00 (3.61)	6.50 (1.50)	<0.010

[a]Probability that differences between groups arose by chance.
WAZ, weight-for-age *z* score; HAZ, height-for-age *z* score; WHZ, weight-for-height *z* score.
Data from National Center for Health Statistics.

Table 4 British men and women: Mean daily intake of energy and nutrients (7-day weighed intake) by socio-economic status[a]

	Women, benefit receipt (n = 150)	*Women, no benefit receipt (n = 741)*	*Men, benefit receipt (n = 110)*	*Men, no benefit receipt (n = 723)*
Energy (MJ)	6.37	6.97	8.85	9.86
Protein (g)	56.5	65.2	79.6	89.6
Total carbohydrate (g)	193	205	259	277
Fiber (g)	10.5	13.0	13.1	15.5
Total fat (g)	56.4	62.5	81.5	87.2
% food energy from total fat	34.4	35.0	35.8	35.8
Iron (mg – food sources only)	8.8	10.3	11.4	13.4
Iron (% below Lower Reference Intake, 19–50 years)[b]	50	28	2	1
Calcium (mg)	685	795	883	1025
Zinc (mg)	6.5	7.5	9.3	10.3
Folate (μg)	220	257	285	353
Total carotene (β-carotene equivalent; μg)	1406	2017	1755	2084
Vitamin C (mg)	60.4	85.1	62.7	86.6
Vitamin E (α-tocopherol equivalent; mg)	7.0	8.3	9.2	10.9

[a]Indicator of socio-economic status 'benefit receipt' = respondent or anyone in household was in receipt of a means-tested benefit from the state. Nutrients shown are from food sources only (i.e., excluding supplement).
[b]Intake of iron supplements has no effect on percentage below Lower Reference Nutrient Intake.
Data from Henderson L, Gregory J, Irving K, and Swan G (2003) *The National Diet & Nutrition Survey: Adults Aged 19 to 64 Years.* Volume 2; Energy, Protein, Carbohydrate, Fat and Alcohol Intake. London: TSO; Henderson L, Irving K, Gregory J *et al.* (2003) *The National Diet & Nutrition Survey: Adults Aged 19 to 64 Years. Volume 3: Vitamin and Mineral Intake and Urinary Analytes*. London: TSO.

purchasers of calories than poor people: Richer people can afford to buy foods that provide sufficient calories in total but in foodstuffs that are more pleasant to eat or that are less calorie dense. Poor people have to maximize the calories and nutrients they can obtain for as little cost as possible. However, many studies have shown that even the poorest do not eat a theoretical 'least-cost diet': Everyone tries to satisfy cultural demands for taste or tradition in food type or preparation methods. When low-income groups obtain more money, they are often observed to spend it on more expensive foods, which are sometimes less calorie dense than commodities bought previously but that carry higher status (such as meat) or are more tasty (such as fruit). Third, those who are poor tend to obtain most of their calories from basic staple foods, particularly roots and tubers or coarse grains (millets and sorghums); these starchy foods cost less in land or labor to produce than other foodstuffs. Poorer people also eat monotonous diets based on a few food types with little diversity; their nutrient base is restricted to a few key foods. Thus, it is difficult to meet dietary requirements for healthy living; dietary diversity is an indicator of food access and food security. As people move up the income scale, they substitute higher value grains (maize, rice, or wheat) and more processed or desirable versions of those grains (bread made from highly milled flour). They also obtain more calories from nonstaple foods and eat a more varied diet, including foods of animal origin, which in wealthier groups provide between 33 and 40% of total calories. People can also afford to express more health consciousness in food choice, buying leaner meat, more fresh produce, and expensive less-refined foods.

Variations in individual nutrient intake by income are usually less than variations between income groups in quantities and patterns of food purchased might suggest. Those who are poorest, whether by income or assets (e.g., landless labourers), tend to be very sensitive to food prices and will purchase substitute commodities as prices vary. However, people may also exhibit relative price inelasticity for a highly prized staple (e.g., they will buy rice no matter how expensive).

In developing countries, particularly those undergoing national economic changes through structural adjustment programs, there is evidence of deteriorating household food security among those whose incomes are declining and for whom reductions in welfare and food price subsidies have most significance. Those whose social and economic entitlement base is the least secure are most vulnerable to reduced food intakes following economic change. In general, energy and protein intakes, as well as micronutrient intakes, tend to vary inversely with income group, land-holding, or asset measures of socio-economic status, although the degree depends on local social and economic circumstances.

In developed countries or regions, there are fewer differences between income groups or social classes in per capita consumption of calories, proteins, or fats, although some surveys show that poorer groups eat

more energy as saturated fats or as refined sugar. However, most national and small sample surveys show marked differences in intakes of vitamins and minerals, such as iron and calcium. Those who are poorest (by whatever measure) often have very low intakes of these nutrients, well below reference levels, which indicates a high risk of poor health outcomes. Smoking can complicate this picture: Those with low incomes who smoke have lower micronutrient intakes than those who do not smoke, which is partly an income effect (they cannot afford to buy enough appropriate food) but also a direct effect of smoking (smokers have different dietary patterns—particularly lower fruit intakes—whatever their income). Those with lower incomes are more likely to be smokers.

Socio-economic status not only affects how much money people have to spend on food but also increasingly affects their social and physical environment, which has implications for food. In many countries, geographical socio-economic polarization means that poorer people tend to live in low-quality housing, with limited domestic equipment, in residential areas with few, if any, shopping facilities and insufficient public transportation. Food shops selling a wider range of better quality goods, often at cheaper prices than small local shops, tend to be located where the wealthier live or to be only accessible by car. Car ownership and housing tenure are becoming good indicators of health in developed countries; part of that relationship is mediated by economic and physical access to sufficient, healthy food.

These general findings have been examined in individual countries. For example, analyses of national survey data sets in the United States and Australia show weak relationships between nutrients and income. Educational status or household size were more predictive of low intakes. At the subnational (state) level, however, the relationship between nutrients and income is more evident, as in several European countries. A different approach from national surveys examines usage of food banks (free food provision). In many developed countries, the numbers using food banks as a main food source increased rapidly at the end of the twentieth century. Participants tend to be welfare recipients or unemployed, often with dependent children and female headed, and proportionately more are likely to be ethnic minorities. These characteristics usually indicate lower socio-economic status in developed societies.

Ethnicity is no guarantee of predictability of dietary quality, and in the United Kingdom as least, it shows a more complicated relationship with socio-economic status. For instance, members of black African, black Caribbean, or black British households in the United Kingdom, or black American or black African households in the United States, are more likely to eat a varied, healthy diet (higher micronutrients and lower proportion of energy from fat) than their white counterparts, whatever the income level. South Asian households in the United Kingdom are less homogeneous with regard to income and diet; some eat more pulses, whole cereals, and vegetables than white counterparts, but many also have high fat and saturated fat intakes.

Many studies have shown nutrient intakes to be low, dietary patterns less healthy, and dental caries higher in children from low socio-economic status households. Nonetheless, there is also evidence that parents, especially mothers, go without food, or eat less healthily, to ensure their children have enough and the best to eat in poorer households. Indeed, this is true throughout the world: Parents try to buffer their children from the worst consequences of low socio-economic status. Evidence that boys are fed better than girls is equivocal, even in south Asia.

Socio-economic Differences in Growth and Nutritional Status

Variations, both within and between populations, in body size at a given age and in the rate of maturation are partly genetic and partly environmental in origin. The most important determinants of birth weight and child growth—adequacy of the diet, quality of the environment, and access to health services—are also core elements of the standard of living. As a result, indices of child growth and birth weight have long been recognised as sensitive indicators of social inequality.

In most countries, women from poorer families produce infants of lower birth weight (with lower chances of survival) than do women from better-off families (**Table 5**). Some of this socio-economic

Table 5 Male birth weights in different socio-economic groups

Country/population	*Mean birth weight*	
	Well-off (kg)	*Poor (kg)*
Tehran, Iran	3.43	3.27
Shiraz, Iran	3.18	3.02
Lebanon	3.50	3.40
Delhi, India	3.16	2.74
Campinas, Brazil	3.41	3.18
Baltimore, USA		
Black	2.97	2.91
White	3.27	3.13

From Eveleth PB (1986) Population Differences in Growth: Environmental and Genetic Factors. In Human Growth: A Comprehensive Treatise. Faulkner F and Tanner JM (eds.) Methodology, Ecological, Genetic, and Nutritional Effect on Growth. vol. 3, New York: Plenum Press, pp. 221–239.

differential reflects the size of the mothers; women are often smaller in poorer families. However, differences in birth weight also reflect what Hytten calls a "comprehensive pattern of deprivation": a poorer quality diet, low pregnancy weight gain, mothers being underweight at conception, the absence of or poor quality prenatal care, poor housing, and behaviors such as smoking and alcohol and drug use.

Height in children and adults is positively associated with socio-economic status throughout the world. **Figure 4** shows the range of boys' heights from richer and poorer families in a number of countries. Social class differences in boys' heights varied from very little in Scandinavia to approximately 12 cm in the Indian samples. A number of studies have shown that when indicators of social position are expressed on a continuous or graded scale, the relationship with height is monotonic; that is, height increases with an increase in social position. This observation appears to be true even between groups from quite narrowly defined social strata, as shown by data from a single slum in Bangladesh (**Table 6**). A number of in-depth studies in developing countries of the relationship between socio-economic status and height of children have consistently shown height to be responsive to individual socio-economic indicators, whether of asset wealth or income, or proxy indicators, such as land-holding, grain yields, parental education, or occupation (**Table 7**).

A similar relationship can be shown in developed countries. For example, in the US National Health and Examination Survey, child height increased monotonically in relation to increments in both household income and parental education (the data were controlled for race). In Britain, the National Child Health and Development Study showed that the difference in height at age 7 between children born in social classes I and II and those in V averaged 3.3 cm; at age 16, the difference was 4.4 cm, in both instances controlling for confounding variables.

In recent years, social class gradients in height, weight, and age of peak height velocity and menarche have begun to attenuate in a few wealthy countries (Sweden, Norway, Finland, and Hong Kong). This seems to be because secular trends in growth have been faster among the lower than higher social class children, and it probably indicates general high living standards rather than growth in a 'classless' society.

Social class gradients in attained adult height are also well documented. In British adults, and Swedes born up to the mid-1950s, average height of manual and nonmanual classes differs by approximately 3 cm for men and 2 cm for women. These averages may mask the effects of upward mobility: Those who move up a class tend to be taller than those who stay within the same social class. In both Swedish and British populations, attained adult height has been shown to be related to economic conditions during childhood, independent of parental height and birth weight. Environmental influences exhibit an inter- as well as intragenerational effect on attained adult height. They also manifest in health outcomes as described previously: Social conditions contribute to nutritional deficiency in early childhood that results in delayed linear growth (indexed by shorter leg length), and this is associated with increased mortality risk for coronary heart disease in later life.

Table 6 Adult and child anthropometric status by livelihood group in a Bangladeshi slum

Variable	*Landlords/ traders*	*Petty traders*	*Laborers (male labor)*	*Laborers (female/ child labor)*	*Statistical significance (p)*
Children <5 years					
n	50	85	109	69	
Weight/age as % of NCHS median, mean (SD)	77.9 (9.8)	77.8 (8.9)	75.0 (9.9)	69.7 (11.1)	0.001
Height/age as % of NCHS median, mean (SD)	91.4 (3.8)	90.7 (3.6)	89.9 (6.5)	87.9 (6.5)	0.001
Weight/height as % of NCHS median, mean (SD)	92.3 (9.7)	92.9 (7.8)	91.5 (10.1)	87.4 (9.2)	0.001
Adults—men					
n	34	55	70	40	
BMI, mean (SD)	20.4 (2.4)	19.1 (2.8)	18.8 (2.2)	17.5 (1.6)	0.001
Adults—nonpregnant women					
n	31	47	67	41	
BMI, mean (SD)	20.1 (3.3)	19.0 (2.3)	18.7 (2.3)	17.3 (1.4)	0.001

BMI, body mass index; NCHS, National Center for Health Statistics; SD, standard deviation.
Data from Pryer J (1990) *Socio-economic and environmental Aspects of Undernutrition and Ill-Health in an Urban Slum in Bangladesh*. PhD thesis, University of London.

Table 7 Reported significant associations between household socio-economic and demographic characteristics, and indices of anthropometric status, in children

	ZHA or Ht	*ZWA or Wt*	*ZWH*	*AC*	*SF*
High caste	M	M			
Land	M,H	J,M	M	M	
Livestock	A,B	A,B		A	
Income	A,J	A,J,G	J	A	A
Grain yield	M	M		M	
House type	L,K,I	K,I,L,H			
Sanitation	I	I			
Water supply	I	I			
Father's occupation	C,E,B	E,B,H		B	
Father's education	M	M		M	B[a]
Mother's education	M,L	M		M	
Mother's age	C	G	M		
Birth order[b]	J,F,C,O,N	G,N			
Number of siblings[b]	J[c], C	I,G,N[c],H			
Family size[b] or parity[b]	I,D	I,D			

[a]Not distinguished by sex of parent.
[b]Negative association.
[c]Children from medium-size families smallest.
AC, mid-upper arm circumference; Ht, height; SF, skinfold thickness; Wt, weight; ZHA, ZWA, ZWH, standard deviation scores of height-for-age, weight-for-age and weight-for-height, compared with NCHS references; A, Bangladesh; B, Bolivia; C, Britain; D, England; E, Brazil; F, Czechoslovakia; G, Colombia; H, Costa Rica; I, Gambia; J, Guatemala; K, India; L, Jamaica; M, Nepal; N, Nigeria; O, United States.
Adapted from Strickland SS and Tuffrey VR (1997) *Form and Function A Study of Nutrition, Adaptation and Social Inequality in Three Gurung Villages of the Nepal Himalayas*. London: Smith-Gordon.

There is also consistent evidence of an association between temporal changes in socio-economic status and height or physical maturation, at both population and individual levels. For example, positive associations have been found between increased gross national product and average stature in France and decline in menarcheal age in Norway, and between per capita income and height in Italian conscripts and survivorship to adulthood. At the individual level, a Polish study showed that statural gains of sons compared to those of their fathers were significantly greater among fathers who had progressed socio-economically compared to their parents. Reversals in the secular trend in height or menarcheal age have been shown to coincide with periods of economic decline, such as occurred in Norway, Germany, and France during World War II and in three periods in Moscow (during the civil war following the October Revolution, during mass collectivization, and during the late 1930s and World War II).

Regarding indicators of 'thinness'—weight-for-height in children and BMI (weight/height^2) in adults—there is no evidence of a consistent monotonic association between thinness and social class or livelihood worldwide. In developed countries, poorer women, and increasingly men, are more likely to be overweight or obese, and this relationship is strengthening. For instance, in Britain the incidence of obesity among poorer women is twice that of women in the richest fifth of the population. In many low-income countries there is evidence of positive associations between indicators of socio-economic status and both child weight-for-height and adult BMI (**Tables 6–8**). For countries undergoing the epidemiological transition, the situation appears more complex. For example, national surveys in Brazil conducted during 1974 and 1989 indicate that at the population level, the frequency of underweight had declined, whereas obesity had increased. In both 1974 and 1989, inverse, monotonic social gradients in underweight were evident both in adults (BMI) and in children (weight-for-age), and the reduction was relatively uniform across the social strata. Among adults, although increases in the prevalence of obesity (BMI) were observed for each social strata, the proportionate increase was greatest for the poorest families. The relationship between measures of income and obesity over time was also not straightforward. For men, the social gradient in obesity was attenuated. Among women, the highest prevalence of obesity was no longer among the richest but in the middle-income group, and the greatest increases in prevalence had occurred among the poorest women. In China, where the epidemiological transition has been very rapid, national surveys in the late 1980s and early 1990s indicated an increase in adult underweight among the lowest income groups at the same time as an increase in obesity among the middle-

Table 8 Reported significant associations between household socio-economic characteristics and indices of anthropometric status in adults

Characteristic	*Height*	*Weight*	*BMI*	*SF*
House type	C	C		
Land ownership	F	C,F	C,F	
Income	A	A[a]	D	A[a]
Livelihood		E	E	
Assets	F	F	F	
Socio-economic index			B	

[a]Negative association.
BMI, body mass index; SF, skinfold thickness; A, United States; B, Congo; C, India; D, Brazil; E, Bangladesh; F, Nepal.
Adapted from Strickland SS and Tuffrey VR (1997) *Form and Function. A study of Nutrition, Adaptation and Social Inequality in Three Gurung Villages of the Nepal Himalayas*. London: Smith-Gordon.

and higher income groups. In Asia in general, there is increasing obesity in all age groups, and it seems to be fastest among middle- and lower income groups, except in the poorest countries. For instance, 30% of Indian adolescents from higher economic groups and 20% of Malaysian adults are overweight.

The association between socio-economic status and weight-for-height among children is not consistent. In Britain, a social class gradient in overweight or obesity in primary school children is emerging; a gradient in overweight and obesity was found among both boys and girls aged 11–16 years, but not among those aged 5–10 years, in Perth, Australia. Strong and persistent inverse social class gradients in weight-for-height and obesity by early adulthood have been reported for Britain, the United States, Sweden, Italy, and Australia. Childhood to adult cohort studies in the United Kingdom and Sweden indicate a strong inverse social class gradient between social conditions in childhood and obesity in young adulthood, and this is now seen in men as well as women.

The relationship between socio-economic status and indicators of thinness in developed countries probably depends on a number of factors, which not only include dietary patterns but also physical activity in work and leisure, consciousness of and attitudes about body size and shape, as well as diet-related health issues.

The Future

Socio-economic differentials seem to be increasing in a number of countries, whether developed, developing, or those with 'transitional' economies. Several agencies use child and adult anthropometric indices as indicators of socio-economic status and to monitor the effects of social change. Recent longitudinal data from Sri Lanka and Nepal suggest that adult anthropometric indices may be more reliable than those of children because children are often buffered against seasonal nutritional stress by physiological and social support mechanisms. The Nepal data also indicate that adult body size does influence fitness, productivity, and reproductive performance. However, these disadvantages do not necessarily translate into equivalent socio-economic disadvantage for the whole household. The circle is neither necessarily vicious nor resistant to social and biological mitigating effects.

See also: **Dental Disease**. **Food Choice, Influencing Factors**. **Nutrition Policies In Developing and Developed Countries**. **Nutrition Transition, Diet Change and its Implications**. **Nutritional Assessment**: Anthropometry; Biochemical Indices; Clinical Examination. **Obesity**: Childhood Obesity. **World Health Organization**.

Further Reading

Barker DJP (1994) *Mothers, Babies and Diseases in Later Life*. London: BMJ.

Caballero B and Popkin BM (eds.) (2002) *The Nutrition Transition: Diet and Disease in the Developing World*. London: Academic Press.

Dowler E (2001) Inequalities in diet and physical activity in Europe. *Public Health Nutrition* **4**(2B): 701–709.

Dowler E and Jones-Finer C (2003) *Welfare of Food*. Oxford: Blackwell. (Also special issue of *Social Policy and Administration* **36**(6)).

Ezzati M, Lopez AD, Rodgers A, Vander Hoorn S, Murray CJL, and the Comparative Risk Assessment Collabrating Group (2002) Selected major risk factors and global and regional burden of disease. *Lancet* **360**: 1347–1360.

Graham H (ed.) (2001) *Understanding Health Inequalities*. Buckingham, UK: Open University Press.

Gunnell DJ, Davey Smith G, Frankel S *et al.* (1998) Childhood leg length and adult mortality: Follow up of the Carnegie (Boyd Orr) Survey of Diet and Health in Pre-war Britain. *Journal of Epidemiology and Community Health* **52**: 142–152.

Maxwell S and Slater R (eds.) (2004) *Food Policy Old and New*. Oxford: Blackwell. (Also special issue of *Development Policy Review* **21**(5–6): 569–580).

Monteiro CA, Conde WL, Lu B, and Popkin BM (2004) Obesity and Inequities in Health in the Developing World. *Int J Obes* **28**: 1181–6.

Roos G and Prättälä R (1999) *Disparities in Food Habits: Review of Research in 15 European Countries*, FAIR-97-3096 Disparities Group (Tasks 4 and 5) Helsinki: National Public Health Institute.

Shetty PS and McPherson K (eds.) (1997) *Diet, Nutrition and Chronic Disease: Lessons from Contrasting Worlds*. Chichester, UK: John Wiley and Sons.

Strickland SS and Tuffrey VR (1997) *Form and Function. A Study of Nutrition, Adaptation and Social Inequality in Three Gurung Villages of the Nepal Himalayas*. London: Smith-Gordon.

Victora CG, Wagstaff A, Schellenberg JA *et al.* (2003) Applying an equity lens to child health and mortality: More of the same is not enough. *Lancet* **362**: 233–241.

WHO/FAO (2003) *Diet, Nutrition and the Prevention of Chronic Diseases*. Report of a Joint FAO/WHO Expert Consultation, WHO Technical Report Series No. 916. Geneva: World Health Organization.

SODIUM

Contents
Physiology
Salt Intake and Health

Physiology

A R Michell, St Bartholomew's Hospital, London, UK

Physiological, Clinical, and Nutritional Importance of Sodium

Despite the fact that the body contains more calcium and potassium, sodium is arguably the most important cation because it dictates the volume of extracellular fluid (ECF) and its concentration affects osmotic concentration of both ECF and intracellular fluid (ICF). Abnormalities of ECF sodium concentration cause movement of water into or out of cells, thus altering the osmotic concentration of ICF in parallel and causing swelling or shrinkage of cells. The main impact of this is on the brain because its cells are rigidly enclosed by the cranium.

Sodium depletion is mainly caused by enteric, renal, or adrenal disease, and sodium retention is caused by renal disease; healthy kidneys are well able to excrete excess dietary salt. However, chronic ingestion of excess salt, whether or not it increases ECF volume, is a predisposing or exacerbating factor in hypertension. Until the 1980s, knowledge of the regulation of body sodium mainly concerned defenses against depletion, while in the 1990s there was a rapid growth in knowledge of the mechanisms that excrete excess sodium. This seems appropriate since most species, especially humans, dogs, and laboratory rats, are exposed to dietary sodium intakes well above their nutritional requirement.

The nutritional requirement is a reflection of obligatory losses (maintenance) and the needs of growth, pregnancy, and lactation. Abnormal losses owing to disease, or in animals such as humans and horses which sweat extensively, raise the requirement. The impact of equine sweating is different from that in humans. Human sweat always contains sodium at concentrations well below plasma levels (and when aldosterone secretion is raised, levels of sweat sodium fall very low); horse sweat is hypertonic but this helps to offset the osmotic effect of the increased respiratory water loss during exertion, i.e., it may be a defense against hypernatremia, rather than a potential cause of sodium depletion. Similarly hypernatremia in many species induces 'dehydration natriuresis' – an appropriate defense.

Consideration of the physiology of sodium thus includes its distribution in the body, regulation of total content and concentration, causes of and responses to depletion or excess, and their nutritional implications.

Distribution

Sodium is a cation, i.e., a positively charged ion; its distribution and physiological effects are fairly independent of the negative ions (anions) that originally accompanied its ingestion though they may affect its absorption and excretion. Most sodium is in ECF (**Table 1**), kept there by the sodium pump, an enzyme system, Na^+/K^+-exchanging ATPase, which uses substantial amounts of energy (adenosine triphosphate; ATP) in maintaining a low intracellular sodium concentration and a high intracellular potassium (K^+) concentration. Sodium transport is a central issue in the physiology of sodium for a number of reasons:

1. It helps to maintain the ionic environment of ICF and the volume of ECF.
2. It prevents cell swelling (the Na^+ efflux exceeds the K^+ influx).
3. It establishes gradients which, in various tissues, allow transport of other cations in exchange, other anions in parallel or organic solutes – these are often cotransported with sodium down concentration gradients which are secondary to the low sodium environment created by the pump.
4. It establishes the membrane voltages on which excitability and secretory activities frequently depend.
5. The energy expenditure of the pump is a substantial portion of total metabolic activity and contributes to thermogenesis.
6. Sodium transport is not only a key factor in the retention and loss of sodium in the kidney, gut, salivary, and sweat glands but also influences the

Table 1 Summary of sodium (Na) distribution and requirements

Typical plasma Na concentration (mmol l^{-1})	145 (130–160)
Typical body Na content (mmol kg^{-1})	50–55
Typical proportion (%) of total Na	
ICF	10
ECF	50
Bone	40
Maintenance requirement in mammals (mmol per kg per day)	
Sheep	0.1
Cattle and goats	0.3–0.7
Pigs	0.6
Rats	0.6
Dogs	0.2–0.5
Humans	?<0.6

1 mmol = 23 mg Na^+, 58.5 mg NaCl.

excretion or retention of many other solutes. Thus, for example, diuretics intended to promote sodium excretion may also cause unintentional losses of potassium and magnesium. Similarly, when renal sodium excretion increases appropriately in response to ingestion of excess salt, there may also be unwanted losses of calcium and in postmenopausal women these may contribute to loss of bone mineral.

Bone also contains substantial quantities of sodium but, as yet, its significance is unknown since it does not appear to be mobilized during sodium depletion. Gut fluids contain considerable amounts of sodium, mostly secretory rather than dietary, and mostly reabsorbed in more distal regions of the intestine.

Extracellular Sodium

Most of the extracellular fluid is interstitial fluid (ISF) in the tissue spaces, providing the transport medium between capillaries and cells. The sodium concentration in plasma is slightly above that in ISF because plasma contains more proteins, notably albumin, which do not readily escape into ISF across the capillary membranes, and the effect of their negative charges is to hold more positively charged ions, notably sodium, in circulation (Gibbs–Donnan equilibrium).

The main effects of excess ECF volume are seen as expanded ISF, visible clinically as edema (or ascites when fluid accumulates in the abdomen rather than the tissue spaces). Mild edema is merely a cosmetic problem in itself but pulmonary or cerebral edema, or severe ascites, are potentially serious forms. Edema can result from excess ingestion or retention of sodium (overall expansion of ECF) or 'leakage' from plasma to ISF, with plasma volume continuously replenished by renal sodium retention. Such maldistribution of ECF occurs if plasma albumin is very low (renal leakage, hepatic impairment, or severe malnutrition), or with excessive capillary blood pressure (venous blockage, inactivity, heart failure, or arteriolar dilation, e.g., from heat or allergy), capillary damage, or lymphatic blockage. The latter prevents the removal of proteins that have leaked into ISF. Accumulation of protein in ISF undermines the osmotic gradient, normally favors uptake of water at the venous end of the capillary, where the pressure is lower. Since edema involves the expansion of a larger compartment (ISF) from a smaller one (plasma) it is only possible as long as the latter is replenished; hence the kidney, while seldom the primary cause of edema, is always the enabling cause; the use of diuretics is therefore appropriate in the treatment of nonrenal as well as renal causes of edema.

The main effect of inadequate ECF volume is to reduce plasma volume and thus to compromise cardiovascular function, in extreme cases by causing circulatory shock.

Regulation of ECF Sodium

In a mature, nonpregnant, nonlactating, healthy animal, sodium excretion matches sodium intake and is often used to estimate it, although this is not reliable, especially when intake is low. Dietary sodium is readily available, i.e., readily absorbed; thus the traditional view of sodium regulation emphasizes renal regulation of urinary Na^+ loss. This oversimplifies the more subtle interplay seen, for example, in herbivorous animals, where salt appetite may contribute to regulation by intensifying during sodium depletion. Moreover, in many herbivores the feces, rather than urine, may be the major route of sodium excretion and the gut may therefore be an important regulator of sodium balance. Indeed, it is interesting that sodium transport mechanisms in the small intestine show considerable similarities to those of the proximal part of the renal tubules (e.g., linked transport of Na^+, glucose, and amino acids) whereas the colon, like the distal nephron, responds to the salt-retaining (and potassium-shedding) hormone of the adrenal cortex, aldosterone. Indeed, diarrhea is essentially enteric diuresis; a failure of intestinal sodium and water reabsorbtion, which exceeds the compensatory capacity of the color.

Provided that the adrenal gland is healthy, urinary and fecal sodium loss can be reduced virtually to zero. Sweat loss can also be very low, although with severe exertion in hot climates the volume of sweat may exceed the ability of aldosterone to reduce its

sodium concentration and net loss of sodium can occur. Aldosterone also reduces salivary sodium (and raises $[K^+]$).

There are two components to the regulation of ECF sodium: the total amount of sodium retained and its concentration. The former is regulated by mechanisms that directly affect sodium, whereas the latter is essentially regulated via water balance. Thus, whatever sodium is retained in ECF is 'clothed' with the appropriate amount of water to maintain the normal plasma sodium concentration within narrow limits; deviations of less than 1% (hard to measure in the laboratory) trigger corrective responses. Thus, a raised plasma sodium concentration (e.g., after water loss) stimulates both thirst and renal water conservation; antidiuretic hormone (ADH) from the posterior pituitary reduces urine output through its effect on the renal collecting ducts. Even one of these mechanisms can defend body water; thus diabetes insipidus (inadequate production or effect of ADH) does not cause severe dehydration but polydipsia (increased fluid intake; 'thirst' is a sensation).

Excess salt intake does not raise plasma sodium concentration (hypernatremia) if water is available and the patient can drink; the excess sodium is diluted. The resulting increase in ECF volume then stimulates increased sodium excretion. Sodium also enables ECF to hold water against the osmotic 'pull' of the solutes in ICF and sodium thus functions as the 'osmotic skeleton' of ECF; it is the main determinant of its volume.

Plasma sodium concentration is therefore only indirectly related to sodium balance. When ECF volume, notably circulating volume, is severely reduced, this stimulus, rather than Na^+ concentration, becomes the main drive for thirst and ADH secretion. Until ECF volume is restored, water is retained (to protect ECF volume) even though this undermines the protection of ECF Na^+ concentration and, as a result, plasma sodium falls. Thus, during sodium depletion, contraction of ECF volume precedes significant reductions of plasma Na^+, which is therefore a poor index of sodium status.

Sodium-Retaining Hormones

Sodium depletion, by reducing plasma volume and renal perfusion, stimulates the production of renin (from the kidneys), which generates angiotensin in circulation. This hormone is a vasoconstrictor (so protects blood pressure), stimulates thirst (so helps to restore ECF volume), and, above all, stimulates sodium retention both directly (renally) and indirectly (by stimulating adrenal secretion of aldosterone); it thus reduces the sodium concentration of urine, feces, saliva, and sweat, but not milk.

Indices of aldosterone secretion (reduced sodium or potassium concentration in urine, feces, etc.) are often taken as evidence of sodium depletion or inadequate sodium intake, but the following points apply:

1. Aldosterone secretion is also stimulated directly by hyperkalemia (elevated plasma K^+) and promotes potassium excretion.
2. Such interpretations involve a subjective judgement concerning adequate or excessive sodium intake. Because physiologists and clinicians were traditionally more concerned with sodium depletion as well as its consequences and the defenses against it, elevated aldosterone secretion was readily seen as a warning signal. However, if sodium intakes associated with increased aldosterone have no other harmful effects, and especially if excess sodium intakes cause concern, low levels of aldosterone secretion might equally indicate excessive salt intake.

While sodium reabsorption in the distal nephron, influenced by aldosterone, is particularly important because it can produce sodium-free urine and promote potassium loss, the great majority of renal sodium reabsorption occurs elsewhere; about 25% in the loop of Henle and most in the proximal tubule. The loop is also a main site of magnesium reabsorption, hence the tendency for loop diuretics to cause hypomagnesemia.

The factors controlling proximal reabsorption are incompletely understood but their effect is clear: proximal reabsorption of sodium increases or decreases according to the need to enhance or diminish plasma volume. Since the fluid in the proximal tubule is similar to plasma, having been formed from it by glomerular filtration, it has the ideal composition for this purpose.

Natriuretic Hormones

Excretion of excess sodium involves not only suppression of salt-retention mechanisms but also activation of sodium-shedding (natriuretic) mechanisms. Two types of hormones are involved: atrial natriuretic peptide (ANP), produced by the cardiac atria when they are overstretched (reduction of ECF volume being an appropriate response to cardiac overload), and active sodium transport inhibitors (ASTIs), probably produced within the brain. These were probably the original molecules associated with the receptors binding cardiac glycoside drugs and are therefore also called 'endogenous digitalis-like inhibitors' (EDLIs);

their exact identity remains uncertain. Atrial natriuretic peptide has various effects that essentially oppose those of the salt retention induced by aldosterone: it increases sodium excretion, lowers arterial pressure, and promotes movement of ECF towards the interstitial compartment.

Other hormones (e.g., sex steroids, parathyroid hormone, calcitonin, thyroid hormone, prolactin) affect renal sodium retention or loss but are not thought to regulate it.

Adequate, Inadequate, and Excess Sodium

It is unlikely that adult daily maintenance requirement exceeds 0.6 mmol per kg body weight and could well be below this in many mammals. Newborn, growing, pregnant, or lactating animals have increased requirements. The appropriate sodium intake for humans remains controversial with some cultures managing on less than 1 mmol per day, while Western intakes may be in the range 200–300 mmol per day, more where processed foods are heavily consumed. There has been insufficient awareness among physicians and human nutritionists of just how high such intakes are, compared with requirements in other animals. Granted that humans are bipeds with a stressful lifestyle quite different from those of animals, there is no real evidence that human obligatory losses or sodium requirements are significantly greater. Rather, there is an ingrained tradition of regarding sodium intake as a benign pleasure, involving a harmless and healthy dietary constituent. The main warnings against this view come from the fact that hypertension is virtually unknown in low-salt cultures and that they do not even have an age-related rise in 'normal' blood pressure. Moreover, there are numerous studies that, when rigorously analyzed, indicate that human arterial pressure and salt intake are positively correlated; sufficiently to anticipate reductions in the prevalence of hypertension in response to manageable reductions in dietary sodium. Unfortunately, such reductions are still handicapped by inadequate food labeling and the fact that most sodium is added by the processor rather than the consumer. Humans, other than vegetarians, also have a very low potassium intake compared with other mammals; potassium may ameliorate the hypertensive effects of sodium.

Because obligatory losses of sodium are so low, dietary sodium depletion is hard to induce and sodium deficiency usually results from losses caused by renal, adrenal, or enteric disease; renal disease may cause either retention or loss of sodium. Globally, both in humans and animals, the most common cause of sodium deficits is acute diarrhea. Fortunately, sufficient gut usually remains unaffected for uptake of sodium and water to be stimulated by suitably formulated oral rehydration solutions. These essentially restore ECF volume (and acid–base balance), allowing natural defenses to overcome the underlying cause of the diarrhea. Despite some species variations, such solutions usually work best if their glucose:sodium ratio (in $\mathrm{mmol\,l^{-1}}$) is close to unity and they are virtually isotonic (i.e., they have a similar osmotic concentration to ECF; hypertonic solutions draw water into the gut). The function of glucose in these solutions is to promote sodium uptake; its nutritional contribution is trivial. Anions such as citrate, acetate, propionate, bicarbonate, and amino acids (e.g., glycine and alanine) may further enhance the uptake of sodium and therefore water. These sodium cotransport mechanisms are very similar to those of the proximal renal tubule. More recently, nutritional oral rehydration solutions that provide calories and glutamine (to sustain the form and function of enteric villi) have been successfully used in calves.

Sodium is thus central to the management of two of the most widespread clinical problems; hypertension (in humans) and diarrhea. Indeed, the World Health Organization (WHO) regards the discovery of oral rehydration, which depends on restoration of enteric sodium uptake, as the main life-saving development in twentieth century medicine. This powerful clinical application rests on a simple physiological observation concerning an elementary but vital dietary constituent.

Unresolved Issues

The control of renal sodium excretion is understood in great detail but the regulation of body sodium is not; key questions remaining unresolved e.g., how ECF volume is monitored, granted that most is interstitial rather than intravascular, and how the mechanisms regulating ECF volume and arterial pressure are integrated, granted that both use renal sodium excretion as their effector. The fact is that none of the common forms of general edema, i.e., excess interstitial fluid, is amenable to rigorous explanation, except via abstractions such as 'effective blood volume.'

The key nutritional concern regarding sodium is the human dietary requirement, assuming that excess intake predisposes populations to an age-related rise in arterial pressure. This is regarded as

normal but it is not seen in any population whose intake is closer to the likely requirement, compared with other mammals. For many individuals, this rise will ultimately destine them to antihypertensive therapy and predispose them to serious secondary hypertensive damage. While it is encouraging that governments are making serious attempts to reduce salt intake, it remains unlikely that it will be brought below 100 mmol day^{-1}, whereas if humans are like other mammals, requirement is unlikely to exceed 0.6 mmol kg^{-1} day^{-1} i.e., 40 mmol day^{-1} for a 70-kg human. Those who insist that human requirement is higher must provide evidence that human renal and colonic sodium conservation are uniquely inefficient or that the endocrine responses to lower salt intake, i.e., increased activity of the renin–angiotensin–aldosterone axis, diminished secretion of atrial natriuretic peptide and endogenous active sodium transport inhibitors, have pathological effects that outweigh moderation of the age-related rise of blood pressure.

See also: **Breast Feeding**. **Electrolytes**: Acid-Base Balance; Water–Electrolyte Balance. **Energy**: Metabolism; Balance; Requirements. **Energy Expenditure**: Indirect Calorimetry. **Hypertension**: Etiology; Nutritional Management. **Potassium**.

Further Reading

Avery ME and Snyder JD (1990) Oral therapy for acute diarrhea. *New England Journal of Medicine* **323**: 891–894.

Brooks HW, Hall GA, Wagstaff AJ, and Michell AR (1998) Detrimental effects on villus form and function during conventional oral rehydration for diarrhoea in calves: alleviation by a nutrient oral rehydration solution containing glutamine. *Veterinary Journal* **155**: 263–274.

Denton DA (1982) *The Hunger for Salt*. Berlin: Springer-Verlag.

El-Dahr SS and Chevalier RL (1990) Special needs of the newborn infant in fluid therapy. *Pediatric Clinics of North America* **37**: 323–335.

Field M, Rao MC, and Chang EB (1989) Intestinal electrolyte transport and diarrheal disease. *New England Journal of Medicine* **321**: 800–806; 819–824.

Hirschhorn N and Grenough WB (1991) Progress in oral rehydration therapy. *Scientific American* **264**: 16–22.

Law MR, Frost CD, and Wald NJ (1991) By how much does dietary salt lower blood pressure? *British Medical Journal* **302**: 811–815; 815–818; 819–824.

Michell AR (1994) The comparative clinical nutrition of sodium intake: lessons from animals. *Journal of Nutritional Medicine* **4**: 363–369.

Michell AR (1995) *The Clinical Biology of Sodium*. Oxford: Elsevier Science.

Michell AR (1996) Effective blood volume: an effective concept or a modern myth? *Perspectives in Biology and Medicine* **39**: 471–490.

Michell AR (1997) Pressure natriuresis, diurnal variation and long-term control of blood pressure: what is the baseline? *Perspectives in Biology and Medicine* **40**: 516–528.

Michell AR (1998) Oral rehydration for diarrhoea; symptomatic treatment or fundamental therapy? *Journal of Comparative Pathology* **118**: 175–193.

Michell AR (2000) Diuresis and diarrhoea: is the gut a misunderstood nephron? *Perspectives in Biology & Medicine* **43**: 399–405.

Narins RG (1994) *Maxwell & Kleeman's Clinical Disorders of Fluid and Electrolyte Metabolism*, 5th edn. New York: McGraw-Hill.

Rutlen DL, Christensen G, Helgesen KG, and Ilebekk A (1990) Influence of atrial natriuretic factor on intravascular volume displacement in pigs. *American Journal of Physiology* **259**: H1595–1600.

Salt Intake and Health

C P Sánchez-Castillo, National Institute of Medical Sciences and Nutrition, Salvador Zubirán, Tlalpan, Mexico

W P T James, International Association for the Study of Obesity/International Obesity Task Force Offices, London, UK

Introduction

This chapter describes the historical importance of salt use, its production and trade throughout the centuries, and its significance in food preservation, in flavor enhancement and in food processing. Over the years, man has developed complex salt mining and drying systems, which are still in use today as the demand for salt continues to grow. Humans and other animals, exposed throughout evolution to very limited salt sources, have developed an intrinsic biological drive for salt with salt-specific taste receptors and highly effective hormonal and cellular transport systems for minimizing any salt loss from the intestine, kidney, and skin. Unfortunately, the use of highly salted food then induces a series of physiopathological responses including changes in blood volume and hormonal and cellular changes, which lead, in conjunction with other dietary and environmental factors, to a range of disorders including high blood pressure with its increased risks of stroke, coronary heart disease, and heart failure. Excess salt intake also seem to promote the development of osteoporosis, gastric cancer, and bronchial reactivity. The relationship of salt intake to these conditions will be described and the options for limiting intakes will be outlined.

Occurrence in Nature

The terms salt, sea salt, or table salt relate primarily to the compound sodium chloride. Sodium is the sixth most abundant element in the Earth's crust, constituting 2.8%. Sodium is a reactive element and is always found in compound form. There is a huge variety of salts containing sodium and many of these are found in food but in most societies the dominant form of sodium is as sodium chloride. Sodium chloride is very soluble in water and in seawater comprises about 80% of the dissolved matter.

A History of Salt Intake

The fundamental drive to obtain salt can be traced back to the earliest times when humans evolved in a hot African environment with scarce sources of salt. Evidence has been found of salt use during the Neolithic Age, and the Egyptian, Babylonian and Chinese civilizations all had special culinary uses for salt that are well documented. In China, for centuries the production of salt was a major industry. Salt sources were highly valued and were often protected. A tax on salt in the form of a head tax provided the Chinese government with a reliable source of revenue from about 2200 BC onwards. For centuries the only method of extraction practiced by coastal salt-workers was to boil sea water and this technique was employed in every maritime province of China as late as 1830. Solar evaporation was also used; shallow salt fields were filled with seawater which was shifted from field to field daily until salt crystallization began. A third method used in areas either far from the sea or on higher ground involved digging wells to tap sea water or salt-enriched aquifers.

Evidence for the exploitation of saline slicks in the Austrian Tyrol dates from the Bronze Age and to this day the salt mines of Salzburg in Austria and Krakow, Poland are still in use. For the Indians in Central America salt was so precious that to please their gods they abstained from eating salt and Mexican civilizations offered sacrifices to the Goddess of Salt, Vixtocioatl. Arab cultures still offer salt to visitors as a sign that their guest is safe; even a Bedouin robber will not violate the laws of hospitality once he has tasted his host's salt.

In pre-Roman times, the principal Italian road started at the salt works near the mouth of the Tiber River and cut through the Italian peninsula towards the Adriatic. In North Africa, the caravan route linked the salt oases, while salt roads were a feature of several South American countries. From remote parts of South America, such as the Amazon and Argentina, trails of more than 1500 km were linked to form the famous 'Cerro de Sal.' In the sixteenth century, sea salt crystals were traded from the sea through the Andes, gradually becoming more expensive further from the sea so that at distances of over 300 km only tribal chiefs used it. The common people made do with salt processed from palms and human urine. Salt from springs near Bogota was traded over a distance of 200 km to the north and south. Columbus' voyages were financed by the wealthy proprietors of the Mata Salt region of Spain, and when the first Spaniards arrived in South America in 1537, they found Indians exploiting local salt reserves on a large scale.

The financial structure of Venice was also substantially affected by the salt trade, which contributed to the emergence of Venetian capitalism and the vast fortunes of some Venetian merchants. In France, salt became a political issue in the fourteenth century; the tax on salt was the most hated of all taxes and a major issue prior to the French Revolution. At that time England, Germany, and Italy also taxed salt and in Britain the control of the world salt markets was a substantial contributor to its wealth in the seventeenth and eighteenth centuries. Liverpool, a minor tobacco port in the early eighteenth century, also became a major trading city in part due to its role in the salt trade.

During the earliest period of British rule in India the supply of salt was often tightly controlled and taxed. Gandhi emphasized the essential nature of common salt for human and animal well being, especially in a tropical country like India. Gandhi's 'salt march' to the sea broke the monopoly on salt use and led to his arrest and jailing. The following revolt, with 100 000 arrests, brought a change in the law to allow people to produce salt for their own use.

The production of salt currently depends on the same range of methods that have been used for centuries with substantial amounts being obtained by dry mining and with solution mining still involving water being pumped into rock salt deposits and the resulting brine being pumped back up to the surface for purification and evaporation. Solar evaporations, the oldest of the methods, is still used in hotter climates where salt pools allow the evaporation of sea water in the sun to produce salt. Currently, world salt production is over 210 million tons a year with 60% of the production being used to manufacture chlorine, caustic soda, and synthetic soda ash. About 20% of the world's production is for food use.

Salt in Food Technology

Salt enhances and modifies flavor, controls microbial growth, and alters nutrient availability and the texture/consistency of food. It also aids extraction methods, food formulation, and helps in the malting and fermenting of foods. In the production of some foods, e.g., pickles, cheese and fermented sausages, salt induces the withdrawal of water and various nutrients from the pickled tissue and provides an appropriate environment for growing the specific salt-resistant bacteria required for the fermentation or pickling process.

Sodium is also important in forming the texture of cheese, limiting bacterial growth and dehydrating cheese, thereby helping to form the rind. Most processed meats, e.g., ham and bacon, have added salt to season and cure the meat. Salt also inhibits bacterial growth and helps to emulsify the fat in sausages.

Sodium nitrate is used as a curing agent to prevent botulism as well as to provide the cured taste and red color of such meat products. Sodium polyphosphate is often added to poultry and fish fingers to increase their water-holding capacity and to bind the product. Salt is also effective in binding meat together by altering protein structures and dissolving some proteins. Salting fish has a flavoring and preservative role and fish may be treated in brine before being smoked.

In baking, salt enhances other flavors in the product; it also controls the rate of fermentation of yeast-leavened products and prevents the development of undesirable 'wild' types of yeast, which would lead to uncontrolled fermentation rates and variable products. Salt also strengthens the gluten in bread doughs, thus helping to ensure good dough handling and reducing the rate of water absorption. Sodium acid pyrophosphates are used in many industrial baking powders for specialty products. Salting of canned vegetables is primarily for flavor, but it can be used to separate mature, starchy green beans or peas, which will sink, from the younger, fresher beans, which float.

Processed 'snacks' are often heavily salted as a marketing feature, as are processed cereals and sodium-containing ingredients are added to many processed foods (**Table 1**), with sodium chloride accounting for about 90% of the sodium used by the food industry.

Table 1 Sodium-containing additives used in food processing

Additive	*Use*
Sodium citrate	Flavoring, preservative
Sodium chloride	Flavoring, texture preservative
Sodium nitrate	Preservative, color fixative
Sodium nitrite	Preservative, color fixative
Sodium tripoliphosphate	Binder
Sodium benzoate	Preservative
Sodium eritrobate	Antioxidant
Sodium propionate	Preservative
Monosodiumglutamate	Flavor enhancer
Sodium aluminosilicate	Anticaking agent
Sodium aluminum phosphate acidic	Acidity regulatory emulsifier
Sodium cyclamate	Artificial sweetener
Sodium alginate	Thickener and vegetable gum
Sodium caseinate	Emulsifier
Sodium bicarbonate	Yeast substitute

Other Uses of Salt

The universal use of common salt has allowed it to be used as a vehicle for combating widespread iodine deficiency by fortifying the salt with iodine, and fluoride has also been added as a preventive measure against dental caries. Chloroquine or pyrimethamine salt mixtures have been used to suppress the sporozoites responsible for vivax malaria.

The Impact of Refrigeration on Salt Intakes

Salt intake varies widely across the world. Some agricultural communities, e.g., the Yanomano Indians from Brazil and the Chimbus of New Guinea, do not consume salt other than that found in natural food sources. The Kamtschadales and the Tungouses nomadic tribes from the north of Russia and Siberia are also averse to added salt, whereas the Japanese have traditionally consumed large quantities of salt in pickled salted fish and vegetables.

Without some form of food preservation it would be impossible to supply urban populations with food in any systematic way. Refrigerators were introduced on a mass scale from the 1960s onwards and this was accompanied by a fall in salt consumption in most countries (**Table 2**); refrigeration has taken over from salting as a method of preserving food. In Japan, intakes as high as the 60-g intake of a farmer recorded in 1955 and the average of 27–30 g day^{-1} had fallen dramatically to 8–15 g day^{-1} by 1988. In the US, salt intake probably started to decline in the 1920s as refrigerators became widely available.

Dietary Exposure to Salt in the Young

Table 3 shows the average daily sodium intakes from food consumed by young people in the UK.

Table 2 Salt intake as NaCl (g day^{-1})

Before 1982[a]	*Year*	*Intake*	*From 1988*[b]	*Year*	*Intake*
Communities not using added salt					
Brazil (Yanomano Indian)	1975	0.06			
New Guinea (Chimbus)	1967	0.04			
Solomon Island (Kwaio)		1.20			
Botswana (Kung Bushmen)		1.80			
Polynesia (Pukapuka)		3.60			
Alaska (Eskimos)	1961	<4.00			
Marshall Islands in Pacific		7.00			
Salt-using communities					
Kenya (Samburu nomads)		5–8	Mexico (Tarahumara Indians)		3–10
Mexico (Tarahumara Indians)	1978	5–8	Mexico rural, men[d]	1992	6.0
			Mexico rural, women[d]	1992	5.4
			Mexico urban, men[d]	1991	7.7
			Mexico urban, women[d]	1991	6.7
Denmark		9.8	Denmark	1988	8
Canada (New Foundland)		9.9	Canada		8–10
New Zealand		10.1			
Sweden (Gotenburg)		10.2			
USA (Evans County, Georgia)		10.6	USA (Chicago)		7.7
Iran		10.9			
Belgium	1966	11.4	Belgium	1988	8.4
UK (Scotland)		11.5			
UK[c]				1990	9
Australia		12.0			
India (North)		12–15	India		9–11.4
Federal Republic of Germany		13.1			
Finland (East)		14.3	Finland		10.6
Bahamas		15–30			
Kenya (Samburus, Army)	1969	18.6			
Korea		19.9			
Japan					
Japan (farmer)	1955	60.3	Japan	1988	8.15
Japan (Akita)		27–30			
Japan	1964	20.9			

[a]Source: INTERSALT Cooperative Research Group. INTERSALT and international study of electrolyte excretion and blood pressure. Results from 24 hour urinary sodium and potassium excretion. *Br Med J* 1988, **297**: 319.
[b]Source: Pietinen, P (1982) Estimating sodium intake from food consumption data. *Ann Nutr Metab*, **26**:90–99
[c]Gregoy J, Foster K, Tyler H, Wiseman M. The Dietary and Nutritional Survey of British Adults. *HMSO* (London, 1990).
[d]Sánchez-Castillo *et al.* (1996) Salt intake and blood pressure in rural and metropolitan Mexico. *Archives of Medical Research* **27**: 559–566.

Table 3 Sodium consumption in people aged 4–18 years in the UK

Age (years)	*Males*		*Females*	
	Sodium intake g day^{-1} (mmol day^{-1})	*Estimated salt intake g day^{-1}*	*Sodium intake g day^{-1} (mmol day^{-1})*	*Estimated salt intake g day^{-1}*
4–6	2.07 (90)	5.3	1.86 (81)	4.7
7–10	2.40 (105)	6.1	2.16 (94)	5.5
11–14	2.70 (118)	6.9	2.27 (99)	5.8
15–18	3.30 (142)	8.3	2.28 (99)	5.8

Source: Gregory J, Foster K, Tyler H, Wiseman M. The Dietary and Nutritional Survey of British Adults. *HMSO* (London, 1990).

Changes in Mineral Composition of Food Induced by Industrialization and Urbanization

The process of industrialization and urbanization has affected the nutritional value of many of the more traditional foods as illustrated for Mexico in **Table 4**. Although home cooked corn (maize) tortillas, together with beans, formed the staple traditional diet, tortillas are now being produced differently: both industrially and by individuals at small market stalls in the cities.

The concentrations of the major nutrients sodium, potassium, calcium, magnesium, and phosphorus in unprocessed foods vary within narrow limits, but in processed or cooked foods, where salt (NaCl) or additions of other sodium-containing ingredients are common, the concentration range of sodium is higher. A large proportion of processed food has salt added; as more processed foods are eaten, the saltier the diet becomes. **Table 4** shows that the corn in its original form contains a very small concentration of sodium but is rich in potassium. Once the grain is milled, fractionated, and processed to produce tortillas, then the nutrient composition alters. Potassium is also lost during the initial washing procedure. Limestone is added to release the niacin from its bound form; this also induces a threefold increase in calcium content. Salt is not commonly added during tortilla preparation in the country, but a remarkable 70- to 200-fold increase is found in breakfast cereals and processed corn snacks as well as substantial potassium losses. Almost no calcium is found in modern breakfast cereals whereas traditionally prepared tortillas have almost 60 times more calcium.

Table 4 Effects of industrialization on the composition of Mexican foods

Food	*Mineral content (mg per 100 g fresh weight)*		
	Na	*K*	*Ca*
Corn	4	284	55
Tortilla (traditional)	11	192	177
Processed wheat tortilla	620	73	11
Breakfast cereals	866	101	3
Processed snacks	838	197	102
Beans			
Home cooked	14	470	67
Processed	354	371	26

Source: Sánchez-Castillo CP, Dewey PJS, Reid MD, Solano ML, and James WPT (1997) The mineral and trace element content of Mexican cereals, cereal products, pulses and snacks: preliminary data. *Journal of Food Composition and Analysis*. **10**: 312–333.
Sánchez-Castillo CP, Dewey PJS, Aguirre A, Lara JJ, Vaca R, León de la Barra P, Ortiz M, Escamilla I, and James WPT (1998) The mineral content of Mexican fruits and vegetables. *Journal of Food Composition and Analysis*. **11**: 340–356.

Salt and Disease

The Roman word from which the name 'salt' is derived is Salus, Goddess of Health. Gandhi argued that salt was "essential for human well being, specially in a poor country like India where its inhabitants eat vegetables and rice which contain low salt." However, although its name evokes health, over the years a long-term excess intake of salt has come to be recognized as a major cause of hypertension and thus a risk for stroke and coronary heart disease. An excess of dietary salt may also affect gastric cancer, osteoporosis, and bronchial hyperreactivity. Evidence also suggests that high-salt intake causes left ventricular hypertrophy independently of blood pressure effects.

Salt Intake and Blood Pressure

When salt is ingested it is readily absorbed in the small intestine in association with other molecules such as glucose. The intestinal secretions also contain sodium at concentrations similar to those found in the plasma but the colon has a highly effective active transport system for absorbing practically all the sodium in the colonic contents; only about 1 mmol of sodium is normally excreted in the feces except in cases of severe diarrhea Once the sodium is absorbed the body ensures that the tonicity of the body fluids is finely maintained; so water is retained by the kidney and the blood volume tends to expand until the hormonal responses, e.g., from the atrial naturetic hormone (released in response to changes in atrial pressure) and in the renin-angiotensin system, lead to a fall in the kidney and sweat glands' reabsorption of sodium and therefore a greater sodium urinary excretion and loss in sweat. There are also adjustments in vasomotor tone and the neuronal responses as well as changes in the exchange of sodium and potassium across cellular membranes. The blood pressure then rises, as the kidney reflex demands a higher blood pressure in order to limit the body's extracellular volume expansion.

The degree to which the blood pressure rises in response to dietary salt depends on a range of interacting genetic factors and other environmental influences including the intake of potassium, magnesium, and calcium. The suppressive effects of these minerals in part explain the blood pressure-lowering effects of a diet rich in fruit and vegetables. Fat intakes have been shown to amplify resting blood pressures whereas moderately intense exercise is followed by a lower blood pressure. As fat intakes rise and physical activity falls in many modern societies the body weight and body fat of children and adults

increase. The greater storage of fat leads to changes in a range of hormonal secretions from the fat cells including angiotensinogen, a precursor of the renin-angiotensin axis affecting the kidney's excretion of sodium. Adiponectin secretion from expanding adipocytes falls thereby making the blood vessels much more sensitive to plaque formation, medial hypertrophy, and fibrosis. Salt-induced increases in blood pressure also involve an array of other hormonal responses including the potent vasocontrictor endothelin-1 and the vasodilator bradykinin, these being potentially involved in the blood pressure-independent effects of higher salt intake on arterial thickening, cardiac ventricular hypertrophy, and the synthesis of elastin and collagen in the artery. This makes them progressively thicker and less pliable.

Given this complex of interacting factors it is not surprising that the selective effect of salt intake on blood pressure has been hard to define. The role of salt in inducing high blood pressure is based on extensive animal experiments at the cellular and physiological level, on clinical studies and dietary intervention trials, as well as on major population analyses of blood pressure in relation to salt intake. Meta-analyses of longer term intervention trials to investigate the effect of salt reduction on hypertension also demonstrate that a modest reduction in salt intake has a significant effect on blood pressure in normotensive individuals and an even greater effect in those with pre-existing hypertension.

The response of neurohumoral mechanisms to salt loading varies in different individuals and for many years investigators sought to define what they termed 'salt-sensitive' individuals. There are rare genetic mutations associated with extreme salt sensitivity but within the general population there appears to be a more or less continuous variation in responsiveness consistent with multiple gene–environmental interactions. So perhaps it is not surprising that no clear cut-off points have been agreed for defining 'salt-sensitivity.' Patients with advanced renal failure do have an increased response of their blood pressure to salt loading but this is due to a loss of functioning nephrons.

Rural-urban differences in salt intake and blood pressure Migrant studies are useful in assessing the impact of environmental changes on blood pressure in different ethnic groups. Shaper's original study on Samburu men recruited from Kenyan villages to military camps was associated with a 12-mmHg increase in systolic blood pressure within weeks and similar findings were obtained in Ugandan villagers who had migrated to an urban environment. **Table 5** shows some of the differences between individuals living in their original Ugandan environment and those who had migrated to a more complex urban environment. The Ugandan analyses evaluated the rate of rise in blood pressure with age in the two communities and showed marked differences.

Table 5 Migration studies that assessed rural–urban differences in Uganda, Africa

	Villager	*Migrant*
Systolic blood pressure/age slope[a]	0.15	0.64
Urinary sodium (mmol l^{-1})[a]	82.4	108.6
Urinary potassium (mmol l^{-1})[a]	67.4	38.4

[a]Poulter, NR *et al.* (1990) The Kenyan Luo migration study: Observations on the initiation of a rise in blood pressure. *Br. Med. J.* **300**: 967–972.

Beaglehole also found that the blood pressure of Polynesian children migrating to New Zealand rose simultaneously with dietary changes and this increase was not explained simply by an increase in body weight. More recent studies, e.g., in Mexico (**Table 6**), show the effect of migration on both sexes. Blood pressure rises in association with increases in urinary sodium but pottasium excretion also rises and the men show a decrease in BMI.

Conversely, Japanese people migrating to the US showed marked reductions in the prevalence of hypertension and stroke mortality consistent with the known markedly lower salt intake in association with other environmental changes in the US.

In all these analyses, several dietary changes as well as altered salt intake have occurred, e.g., in potassium and calcium intakes together with weight

Table 6 The urinary 24-h output of electrolytes and the associated blood pressure (BP) differences in rural and urban Mexico

	Men		*Women*	
	Rural (n = 24)	*Urban (n = 19)*	*Rural (n = 54)*	*Urban (n = 58)*
Sodium (mmol day^{-1})	103.3	133.1	93.3	114.7
Potassium (mmol day^{-1})	41.6	56.7	36.9	50.4
Sodium/potassium ratio	2.64	2.51	2.67	2.44
NaCl (g day^{-1})	5.99	7.72	5.41	6.65
Systolic BP (mmHg)	110.4	114.3	104.4	113.8
Diastolic BP (mmHg)	73.3	75.6	67.0	72.8
BMI	25.5	25.1	24.1	26.6

BP, blood pressure; BMI, body mass index.
From Sánchez-Castillo *et al.* (1996) Salt intake and blood pressure in rural and metropolitan Mexico. *Archives of Medical Research* **27**: 559–566.

gain, altered intensities of physical activity, and doubtless psychosocial stress from entering an unfamiliar environment. Experimental, epidemiological, and clinical evidence suggests that dietary deficiencies of potassium or calcium potentiate the sodium induction of high blood pressure. Potassium loading prevents or ameliorates the development of sodium chloride-induced hypertension in several animal models and epidemiologically the ratio of urinary sodium to potassium (Na:K) is a stronger correlate of blood pressure than either sodium or potassium alone. Results of clinical trials also suggest that an increased potassium intake decreases blood pressure in patients with hypertension and the antihypertensive effect of potassium is more pronounced in persons consuming a high sodium chloride intake. With acculturation, primitive societies tend to increase their sodium intake and reduce the potassium content of their diet; therefore, the combination of a high potassium with a high salt diet is somewhat unusual. High potassium intakes were found, however, in the Aomori prefecture of Japan where there was a lower blood pressure and a reduced mortality from strokes despite high-salt intake.

There is also an inverse association within and among populations between dietary calcium and blood pressure. A low calcium intake may amplify the effect of a high sodium chloride intake on blood pressure, and calcium supplementation blunts this effect. High dietary calcium also preferentially lowers blood pressure or attenuates the development of hypertension in sodium chloride-sensitive experimental models.

Given all these dietary effects discerning the impact of salt intake changes as such is not easy. The migrant studies are crude compared with analyses of controlled dietary changes in the sodium intakes of volunteers. More robust analyses can also be obtained from the relationship between sodium intakes and blood pressure across a whole spectrum of different societies where account is taken of the possible effects of sodium intakes at different ages, of other dietary and environmental effects, as well as of differences in body size. The ability to reduce blood pressure by selectively limiting dietary sodium intake has also been assessed in a series of meticulous meta-analyses.

Genetic influences Primary hypertension has a well-known familial aggregation and has been calculated to be about 40% genetically determined. Children with a family history of hypertension are 30% more likely to remain in the upper quartile of systolic blood pressure than their peers. Young adults from families with hypertension have a greater rate of sodium excretion after a salt load than adults from normotensive families. Studies of twins also provide convincing evidence for a hereditary component to salt responsiveness. However, the effect of family history decreases with age as other environmental factors, e.g., weight gain, modify the risk. Studies have suggested that polymorphisms in certain genes, such as the angiotensinogen gene, might be implicated in the blood pressure response to a high-salt intake and genes whose products function prominently in the renin-angiotensin-aldosterone system are potential candidate genes contributing to essential hypertension. However, two meta-analyses assessed the relation of both insertion/deletion (I/D) polymorphisms of the angiotensin-converting enzyme (ACE) gene and the M235T angiotensinogen gene with primary hypertension and cardiovascular diseases and found no association with hypertension in ACE I/D gene polymorphism. Individuals homozygous for the deletion allele seem to have a higher risk of macrovascular and microvascular complications and the T allele encoding angiotensinogen may be a marker for hypertension, at least in white subjects, but great caution is needed before inferring that a single set of genes has a substantial impact on the development of higher blood pressures in response to increases in salt intake as so many neurohormonal mechanisms are involved.

Age-related changes in blood pressure In most populations, blood pressure increases with age but there are a few small groups who have not been exposed to modern environmental conditions and they do not show a rise in blood pressure with age. The Kuna indigenous population living on islands in the Panamanian Caribbean was among the first communities described showing almost no age-related rise in blood pressure or hypertension. Other populations in Africa, the Americas, Asia, and the Pacific region have the same characteristics. In many of these communities, the primary evidence that the protective factor is environmental rather than genetic was the blood pressure rise following migration to an urban environment. Among the many lines of evidence suggesting a role for salt intake in the pathogenesis of hypertension, particularly compelling has been the identification of these isolated communities where salt intake is low, hypertension is rare, and blood pressure does not rise with age. Salt intake in such communities generally provided less than 40 mmol of sodium per day, and typically much less. The age-related rise is rare at mean sodium excretion rates of <100 mmol per day but clearly there are many other dietary and environmental differences.

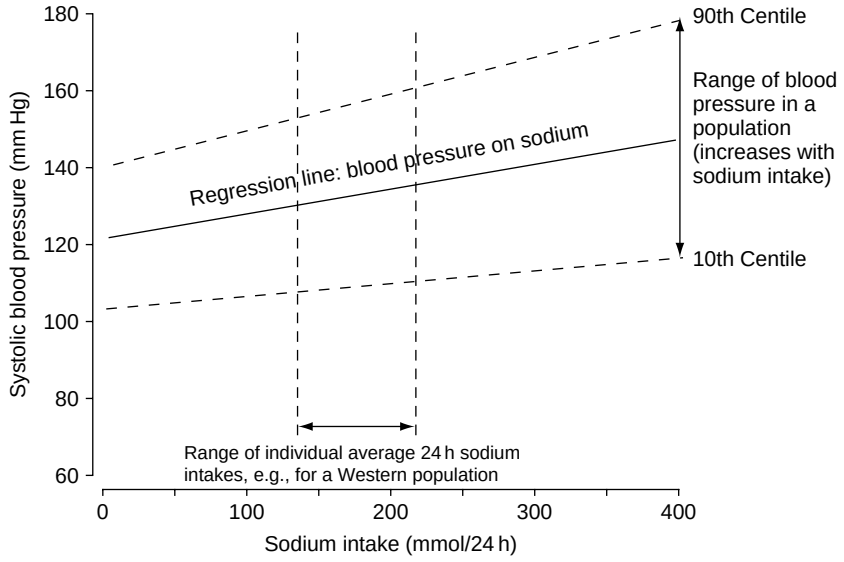


Figure 1 The relationships between sodium intake and blood pressure in the INTERSALT study. Adapted from: Frost CD, Law MR, Wald NJ (1991) Analysis of observational data within populations. *BMJ* **302**: 815–818.

Intersalt studies A major transnational study of over 10 000 men and women described the association between urinary excretion of sodium chloride (as a measure of salt intake) and blood pressure. After adjustments for body weight, alcohol intake, sex, and age, a higher sodium intake of 100 mmol day^{-1} was linked with a systolic blood pressure rise of 3–6 mmHg in adults aged 40 years but one of 10 mmHg when aged 70 years. Updated results suggest that the association between sodium excretion and blood pressure is stronger when not adjusted for body weight, but the relationship is present whether or not the adjustment is made.

Figure 1 summarizes the relationships between sodium intake and blood pressure in the INTERSALT study. Different populations may show different responses depending on the host of other environmental factors that may be involved. The figure also illustrates the fact that individuals within any population may show very different effects and that appreciable changes in salt intake may be needed before a clear change in blood pressure is evident. Part of the problem in displaying the relationship arises from the difficulty in establishing what the prevailing blood pressure of individuals is given the remarkable variation in blood pressure during the day and night; difficulty also arises because it takes many complete 24-h urinary collections to obtain a reasonable estimate of the customary sodium intakes. The age-related incline also implies longer term amplification of the pathophysiological changes in hormonal controls and in blood vessel reactivity and plasticity; thus, as the blood pressure increases the tendency to further increase is enhanced in an accelerating process. This emphasizes the potential importance of early interventions when the blood pressure is tending to rise. It also implies that interventions to alter the diet of the young may be particularly valuable. This is borne out by the observation in the Netherlands that newborn babies fed a reduced salt content in their formula milk for the first 6 months of life had very much lower blood pressures when reassessed at the age of 15 years.

Table 7 shows the estimated changes with age in blood pressures as the salt intake is increased by

Table 7 Predicted change in systolic and diastolic blood pressure (mmHg) for each 100 mmol per 24 h change in sodium intake for various centiles of blood pressure distribution

Age (years)	*Centile*				
	5th	*20th*	*50th*	*80th*	*90th*
Systolic					
15–19	3	4	5	6	7
20–29	2	4	5	6	8
30–39	2	4	6	7	9
40–49	2	4	7	9	11
50–59	4	6	9	12	15
60–69	6	8	10	13	15
Diastolic					
15–19	1	1	2	2	3
20–29	1	2	3	3	4
30–39	1	2	3	4	5
40–49	2	3	4	4	5
50–59	2	3	5	6	7
60–69	2	3	4	6	7

From Law *et al.* (1991) By how much does dietary salt reduction lower blood pressure? III. Analysis of data from trials of salt reduction. *British Medical Journal* **302**: 819–824.

100 mmol sodium per day. Epidemiologists concerned with the subtle but substantial population effects are mostly of the opinion that salt is an important causal factor in determining the steady increase in average blood pressure and the prevalence of hypertension in Western societies.

Adults with episodic high blood pressure, e.g., as a response to mental stress, have a greater tendency to develop persisting hypertension. The higher the blood pressure level becomes, the greater the further increase in blood pressure. Thus, the age-dependent increase in blood pressure may be a particularly important factor to measure in both individuals and the community.

On a population basis it has been estimated that in affluent societies, where average population blood pressures are high, a reduction of 2 mmHg in diastolic blood pressure would result in a 15% reduction in the risks of stroke and transient ischemic attacks and a 6% reduction in risk of coronary heart disease. There may also be a reduction independent of the effects on blood pressure on other conditions such as left ventricular hypertrophy.

A higher frequency of salt responsiveness has been observed in adults with hypertension. Estimates of the prevalence of this sensitivity have ranged from 29 to 60% in hypertensive populations and 15–46% in normotensive populations, although the larger studies have indicated that over 50% of a hypertensive population and approximately 25% of a normotensive population are clearly salt responsive. Longer term, e.g., 27-year-long, studies have shown that those with initially normal blood pressure but a marked responsiveness to salt had an increased risk of cardiological events and death as had those with pre-existing hypertension. In the absence of a consensus on defining either the genetic polymorphisms relating to hypertension or the parameters of salt sensitivity the greatest benefits are likely to be achieved by taking a population approach to reducing salt intake.

The most recent meta-analysis, which related to studies with modest salt reductions and a duration of at least 4 weeks, showed that there were 17 trials in hypertensives and 11 trials in normotensives for analysis. The combined and pooled estimates found significant reductions in blood pressure of 4.96/2.73 mmHg in hypertensives and 2.03/0.97 mmHg in normotensives, which on a population-wide basis are significant effects.

Recently, new diagnostic thresholds to define hypertension were made available in the Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. A new category designated 'prehypertension' is defined as systolic blood pressure values between 120 and 139 mmHg and diastolic blood pressure values of 80 to 89 mmHg. Individuals within this group require health-promoting life-style modifications to prevent cardiovascular disease since they are at increased risk for progression to hypertension. The thresholds for stage 1 hypertension are blood pressure values of 140–159 mmHg (systolic) and 90–99 mmHg (diastolic) with stage 2 hypertension being defined when blood pressure values are $\geq$160 mmHg (systolic) and $\geq$100 mmHg (diastolic) values, respectively. Both categories require life style modifications as well as drug therapy. Individuals with diabetes, who are recognized as being at greater cardiovascular risk, should keep their blood pressure below 130/80 mmHg.

Salt reduction in pre-existing hypertension Salt deprivation became the major means of treating hypertension in the early part of the twentieth century. The low-salt diets were notoriously unpalatable so patients reduced their food intake and the consequent weight loss helped to reduce the blood pressure further. A large number of trials of salt restriction have been conducted since then on both hypertensive and normotensive subjects and the overall analyses show that the greater the initial blood pressure, the more marked the fall in blood pressure, particularly if the sodium intake reduction persists. These data have been interpreted to suggest that the effect of a universal moderate reduction in dietary salt would substantially reduce a population's mortality from stroke and ischemic heart disease with an impact far greater than that achieved by drug treatment of those with high blood pressure. Thus, the World Health Organization (WHO) and most national dietary guidelines now call for a lowering of salt intake to 5–6 g day^{-1} on average or less.

More recently, two controlled intervention trials, the Dietary Approaches to Stop Hypertension (DASH) and the follow-up DASH sodium trial, compared three different types of eating patterns: (1) the 'control diet'; (2) extra fruit and vegetables; and (3) the 'DASH or combination diet,' which was lower in saturated fat, total fat, and cholesterol as well as having higher intakes of fruits, vegetables, and low-fat dairy products. All three eating plans used 3 g day^{-1} sodium. The results of the clinical trials found that the combination diet or 'DASH diet' decreased systolic blood pressure (SBP) by 11.4 mmHg below the control diet and decreased diastolic blood pressure (DBP) by 5.5 mmHg in

adults with hypertension. In adults without hypertension the decreases were 3.5 mmHg (SBP) and 2.1 mmHg (DBP).

When the selective effects of salt were examined without weight changes then reducing the salt intake from 9 to 3 g significantly reduced blood pressure by 6.7/3.5 mmHg on the controlled diet and on the higher potassium DASH diet by 3.0/1.6 mmHg. Thus, the combined effects of the DASH diet and low-salt intake on blood pressure were greater than either of the interventions alone. With this combination, mean SBP was 11.5 mmHg lower in participants with pre-existing hypertension, and 7.1 mmHg lower in participants without hypertension. The effects were observed in both sexes and across racial groups. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Pressure in the USA and the Scientific Advisory Committee on Nutrition from the Food Standards Agency, Department of Health in the UK have acknowledged that the clear and distinct effect of salt on blood pressure shown in the trial indicates that lowering salt intake as part of a healthy whole diet strategy would be most effective as a population-based approach to lowering blood pressures.

Gastric Cancer and Stroke

There is a strong geographical correlation between stomach cancer and stroke mortality, both of which correlate with salt intake. There are four recognized major etiological factors for gastric adenocarcinoma: infection with *Helicobacter* pylori, excessive salt consumption, and low intakes of ascorbic acid, carotenoids or more generically of vegetables and fruits. Sodium chloride induces atrophic gastritis and enhances the mutagenic effect of nitrosated foods. Salt may also play a role in the later steps involving the transformation of mucosal dysplasia to carcinoma. The salted pickles and salted fish of Japanese cultures appear to be strongly linked to the development of stomach cancers.

Osteoporosis

It has been known for many years that sodium intake is one of the major determinants of urinary calcium excretion. It has been estimated that urinary calcium losses increase by approximately 1 mmol per 100 mmol sodium intake. Experimentally sodium intake increases calcium excretion but also induces markers of bone resorption. It is hypothesized that trabecular demineralization may occur, leading to postmenopausal changes and an increased risk of vertebral fractures and cortical erosions. Further research is required in this area.

Bronchial hyper-reactivity

There have been no large-scale epidemiological studies, but a positive relationship between asthma mortality and regional purchases of table salt per person has been shown. In a randomized double-blind crossover trial in subjects with moderately severe asthma, the airway response to histamine was related to urinary excretion of sodium in a dose-response way, but only in men. A low-salt diet is regarded as having a potentially positive effect in patients with asthma and may help to reduce the need for anti-asthma drugs.

Sources of Salt Intake

Various approaches to measuring the daily salt intake in individuals have been tried. Salt comes from: (1) natural products; (2) salt added during industrial processing; (3) salt from catering; (4) other sodium-containing sources; (5) discretionary use of salt in cooking and table salt; and (6) sodium in drinking water. Traditional methods of estimating salt intake, e.g., with economic data, lead to marked errors and usually substantial overestimates. These have now been replaced by more modern methods.

Estimating Salt Intakes

The principal and most accurate method for estimating sodium intake is to measure sodium excretion rates in individuals who are asked to collect one or more complete 24-h urinary outputs. To measure absolute amounts a marker for completeness of collections is required. Measurement of intake from dietary assessment methods alone is considered unreliable. Until recently, however, there was no way of establishing how much of the 24-h sodium intake was derived from different sources, without the use of traditional weighing and analytical methods. A new technique involving the use of lithium has allowed a new approach; this method involves fusing Li_2Co_4 (used as a tracer) with NaCl. One preliminary 24-h collection and three full 24-h urinary collections are required.

Gains and Losses of Salt during Cooking

Only a small proportion of the salt added to water for cooking foods is eaten. A value of 24% was obtained by the lithium method for the average intake per head of the 'purchased' cooking salt used in cooking in the

UK. The only other data using traditional methods come from Hungary where 41% of the salt purchased by households was actually ingested.

Assessment of Total Discretionary Salt Use

Figure 2 compares the traditional and lithium marker techniques for assessing both total salt intake and the distribution of its sources. When table and cooking salt are combined to form a single value, the percentage contribution of these discretionary sources to the total intake measured by the lithium marker technique is significantly lower in the UK compared with that assessed by traditional methods, which do not consider salt losses during cooking and at the table. This intake in the UK seemed unusually low, but when discretionary sources (table and cooking salt) were assessed in various regions of Italy using the lithium marker technique, discretionary salt intake varied between 31 and 41% of total intake. In rural Benin the use of discretionary sources in women was higher (52%) and in rural Guatemala was as much as 77% of total intake. Thus, the more industrialized the food system the greater the proportion of nondiscretionary salt intake, which then makes it more difficult for individuals to reduce their salt intake. In Japan, salt is ingested in large amounts as pickled and salted fish and vegetables but these distinctive items may be considered discretionary sources of salt. Similarly, there are specific discretionary salted meat and vegetable extracts that are used for flavoring in Western societies.

Traditional data on table salt use are given in **Table 8** and new estimates in **Table 9**.

Pore Size in Salt Cellars

The pore size and hole number is important in determining the amount of salt actually shaken onto food. Smaller shaker hole areas lead to a marked fall in salt consumption, for example, to about a quarter of the maximum value.

Table 8 Sources of dietary salt in different countries assessed by traditional methods (grams per head per day)

Country	*Total*	*Nondiscretionary sources*			*Discretionary table/or cooking*
		Natural	*Processing*	*Catering*	
England	9.5		7.9		1.6
UK	9.7	0.9	5.7	–	3.1
UK	11.7	8.1[a]	–	1.6[a]	2.0
Sweden	11.1	1.0	5.3	–	4.8
USA	14.5	–	8.0	–	6.5
Finland	12.6	1.5	5.3	1.0	4.8
Finland	10.7	1.4	4.7	–	4.6
Finland	11.7	–	7.4	–	4.3
Mexico (State)					7.0

[a]Some cooking salt is included in this value for food consumed at home. For catering and table salt use, a value has been added.

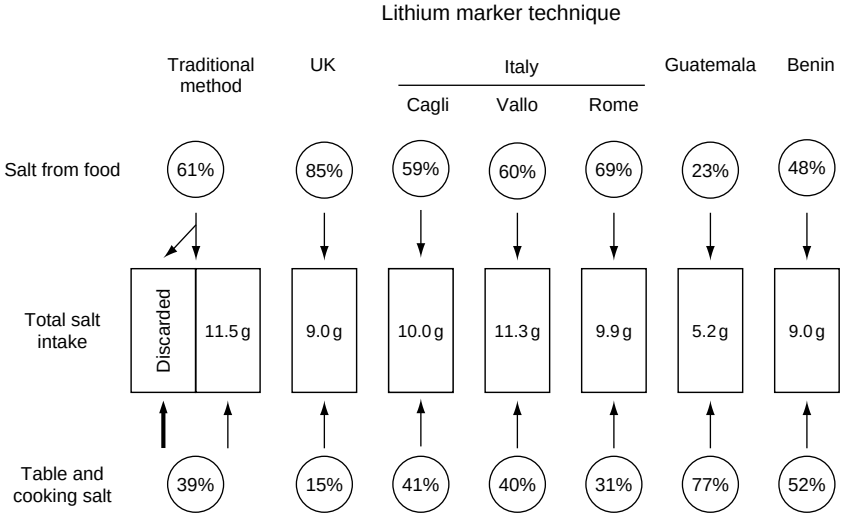


Figure 2 The assessment of total discretionary salt use. (Data from Sánchez-Castillo CP, Branch WJ, and James WPT (1987) A test for the validity of the lithium-marker technique for monitoring dietary sources of salt in man. *Clinical Science* **72**: 87–94; Leclerq C and Ferro-Luzzi A (1991) Total and domestic consumption of salt and their determinants in three regions of Italy. *European Journal of Clinical Nutrition* **45**: 151–159; Melse-Boonstra A, Rozendaal M, Rexwinkel H, Gerichhausen MJ, van den Briel T, Bulux J, Solomons NW, and West CE (1998) Determination of discretionary salt intake in rural Guatemala and Benin to determine the iodine fortification of salt required to control iodine deficiency disorders: studies using lithium-labeled salt. *American Journal of Clinical Nutrition* **68**: 636–641.)

Table 9 Estimates of the use of table salt

Subjects	*Number of subjects*	*Salt intake (grams per head per day)*		
		Total	*Table salt*	*Table salt as % total*
Men				
White and black[a]	24	9.8–16.5	1.5 (0.3)	9–15
White[b]	3	11.0	0.44	4
Black[b]	3	8.9	0.27	3
White[d]	33	10.6	1.6 (1.0)[c]	11
Women				
White and black[a]	13	4.4–4.9	0.9 (0.3)	18–20
White[b]	3	6.4	0.64	10
Black[b]	4	6.5	0.13	2
White[d]	50	7.4	0.73 (0.74)[c]	10
White adolescents	8	7.4	0.95	13
Family studies[e]	15	11.7	1.35[c]	11.6

[a]Total intake of the subjects was varied systematically in a metabolic ward.
[b]Salt intake assessed with the use of a dietary history and food model.
[c]Value based on the use of normal, not lithium-tagged salt. Where necessary values are recalculated to give the mean (SD).
[d]Study conducted in England; all other studies conducted in the US.
[e]A complex and less satisfactory approach was used.
Reproduced with permission from James WPT *et al.* (1987) The dominance of salt in manufactured food in the Sodium intake of affluent societies. *Lancet* **1**: 426–429.

Implications of the Salt-Disease Relationships in Relation to Population and Individual Strategies for Improving Health

A population-based approach to reducing disease by reducing salt intake is a public health strategy directed at the whole population rather than those individuals considered to be at high risk. Such a strategy is based on the observation that a small reduction in risk of a large number of people may result in a large reduction in risk for the entire population. This does not mean, however, that individualized strategies cannot be used to help individuals considered at high risk because of pre-existing hypertension. Indeed, the greater reductions in blood pressure in hypertensives on reducing salt intake imply that there should be a special focus on this vulnerable group. So, ideally, both strategies are needed: the whole population strategy because the risk from cardiovascular disease associated with higher blood pressures is not confined to those who are considered clinically hypertensive, but includes large numbers of people in the upper 'normal' blood pressure range. Furthermore, many surveys in different countries show that a large proportion of hypertensives are not in receipt of any treatment. This emphasizes the value of dealing with a population that overall may be at a relatively high risk of premature mortality. Such measures should cause a downward shift in the population distribution of blood pressure, which would also benefit high-risk groups.

Individualized Approaches

Hypertensives can take steps to reduce their salt intake by modifying their diets. If these individuals come from societies where a substantial amount is eaten as discretionary salt then those responsible for adding salt to the cooking either in the home or in catering establishments need to be persuaded to take progressive measures to limit salt use and substitute herbs and other flavors. Individuals can also be asked to eliminate the addition of salt at the table but this can only make a minor contribution in most cases to reducing their salt intake. In theory it is possible for patients to select foods low in salt but this usually means selecting relatively unprocessed foods. Multimineral mixes may also be found to be more acceptable for use in households as these are a mix of different salts with, for example, the addition of potassium and calcium salts to the sodium chloride thereby both diluting the amount of sodium used and adding elements that counter the sodium's effects. Theoretically, food labels can be used to choose lower salted foods but this needs far too much sophisticated understanding for most consumers. The simplest test of an individual's ability to alter their diet and reduce salt intake by avoiding salted and pickled foods, heavily salted breads, prepared meats and snacks is to check their urinary excretion of sodium. The great difficulty in permanently changing diets is shown by longer term analyses of intervention studies, which reveal very modest long-term reductions in urinary sodium.

This emphasizes the need for population approaches such as that developed in Finland where children were taught at school how to select less salted foods and to alter the use of salt in cooking within the home. There was also a multipronged drive to persuade catering organizations and restaurants to limit the salt in cooking and the food manufacturing sector was persuaded to alter their product composition and limit salt addition as well as altering their fat and fatty acid content. As a result of these measures the average systolic blood pressure of the adults of North Karelia in Finland fell by 10 mmHg over a 15-year period and this was accompanied by a dramatic fall in stroke and coronary artery disease deaths of over 75%, helped substantially by the

simultaneous falls in the average total blood cholesterol levels of the population.

The importance of altering the salt content of foods, especially bread, which is often a major source of salt, was shown in Portugal where a village baker was persuaded to reduce the salt content of his bread. Two years later the average blood pressure of the villagers was significantly lower than that of a neighboring village where no changes had been made.

Thus, governments have a major role in persuading their health services to take a systematic approach to reducing salt in hospital foods and to engage in systematic patient and public health educational initiatives. The problem is that the salt industry and other sectors of the food industry often do their utmost to contest the evidence and find reasons why they should not progressively reduce the salt content of their products. We need to see major improvements in food labels and a traffic light-type warning system so that high-salt products can readily be identified. Several countries are embarking on the exercise of defining high-salt foods and taking steps to counteract the tendency for some soft drink companies to over-salt their snack food products so that they stimulate thirst and increase demand for their drinks! Only when countries follow the Finnish lead can we expect to see an appreciable fall in salt intake and a concomitant reduction in hypertension rates and cardiovascular disease.

Conclusions

Evidence suggests that sodium intake is an important determinant of blood pressure in the population as a whole, and influences the rise in blood pressure with age. The predominant source of salt varies from country to country. In the UK, for example, the greatest potential effect involves reducing the salt content of manufactured food. A different public health approach would apply to Italy, where discretionary intake of salt is two to three times greater than that of the UK and for Guatemala where intakes from these sources are high.

See also: **Calcium**. **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Hypertension**: Etiology; Nutritional Management. **Osteoporosis**. **Potassium**. **Sodium**: Physiology.

Further Reading

Appel LJ, Moore PJ, Obarzanek E, Vollmer WM, Svetkey LP, Sacks FM, Bray GA, Vogt TM, Cutler JA, Windhauser MM, Lin PH, and Karanja N (1997) A clinical trial of the effect of dietary patterns on blood pressure. DASH Collaborative Research Group. *New England Journal of Medicine* **336**: 1117–1124.

Denton D (1982) *The Hunger for Salt*. New York: Springer-Verlag.

Dietary Sodium and Health (1997) *American Journal of Clinical Nutrition* **65**(supplement) 2.

Elliot P, Stamler J, Nichols R, Dyer AR, Stamler R, Kesteloot H, and Marmot M (1996) Intersalt revisited: further analyses of 24 hour sodium excretion and blood pressure within and across populations. *British Medical Journal* **312**: 1249–1253.

Frost CD, Law MR, and Wald NJ (1991) By how much does dietary salt reduction lower blood pressure? II. Analysis of observational data within populations. *British Medical Journal* **302**: 815–818.

Hanneman RL (1996) Intersalt: hypertension rise with age revisited. *British Medical Journal* **312**: 1283–1284.

James WPT, Ralph A, and Sanchez-Castillo CP (1987) The dominance of salt in manufactured food in the sodium intake of affluent societies. *Lancet* **1**: 426–429.

Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Pressure (2003) The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Pressure. *Journal of the American Medical Association* **289**: 2560–2572.

Law MR, Frost CD, and Wald NJ (1991) By how much does dietary salt reduction lower blood pressure? 1. Analysis of observational data among populations. *British Medical Journal* **302**: 811–815.

Law MR, Frost CD, and Wald NJ (1991) By how much does dietary salt reduction lower blood pressure? III. Analysis of data from trials of salt reduction. *British Medical Journal* **302**: 819–824.

Leclerq C and Ferro-Luzzi A (1991) Total and domestic consumption of salt and their determinants in three regions of Italy. *European Journal of Clinical Nutrition* **45**: 151–159.

MacGregor GA and He FJ (2002) Effect of modest salt reduction on blood pressure: a meta-analysis of randomized trials. Implications for public health. *Journal of Human Hypertension* **16**: 761–770.

Melse-Boonstra A, Rozendaal M, Rexwinkel H, Gerichhausen MJ, van den Briel T, Bulux J, Solomons NW, and West CE (1998) Determination of discretionary salt intake in rural Guatemala and Benin to determine the iodine fortification of salt required to control iodine deficiency disorders: studies using lithium-labeled salt. *American Journal of Clinical Nutrition* **68**: 636–641.

Multhauf RP (1978) *Neptune's Gift*. Baltimore and London: The John Hopkins University Press.

Sacks FM, Svetkey LP, Vollmer WM, Appel LJ, Bray GA, Harsha D, Obarzanek E, Conlin PR, Miller ER 3rd, Simons-Morton DG, Karanja N, and Lin PH (2001) Effects on blood pressure of reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. DASH-Sodium Collaborative Research Group. *New England Journal of Medicine* **344**: 3–10.

Scientific Advisory Committee on Nutrition: Food Standards Agency and the Department of Health (2003) *Salt and Health*. Her Majesty's Stationery Office, London.

Sánchez-Castillo CP, Branch WJ, and James WPT (1987) A test for the validity of the lithium-marker technique for monitoring dietary sources of salt in man. *Clinical Science* **72**: 87–94.

Sánchez-Castillo CP, Solano ML, Flores J, Franklin MF, Limon N, Martinez del Cerro V, Velazquez C, Villa A, and James WPT (1996)Salt intake and blood pressure in rural and metropolitan Mexico. *Archives of Medical Research* **27**: 559–566.

Sánchez-Castillo CP, Warrender S, Whitehead T, and James WPT (1987) An assessment of the sources of dietary salt in a British population. *Clinical Science* **72**: 95–102.

Sodium Chloride *see* **Sodium**: Salt Intake and Health

Spirits *see* **Alcohol**: Absorption, Metabolism and Physiological Effects; Disease Risk and Beneficial Effects; Effects of Consumption on Diet and Nutritional Status

SPORTS NUTRITION

R J Maughan, Loughborough University, Loughborough, UK

Introduction

At an international Consensus Conference held at the offices of the International Olympic Committee in 1991, a small group of experts agreed a consensus statement that began with the opening statement: "Diet significantly influences exercise performance." This is a bold and unambiguous statement, leaving little room for doubt. However, the statement went on to add various qualifications to this opening statement. These qualifications reflect the uncertainties in our current knowledge, but are also a consequence of the many different issues that arise in considering the interactions between diet and performance and the diverse needs of athletes in different sports. In the years since that statement was formulated, the world of sport has advanced, with new world records and new champions. The world of science has also moved forward and there have been some important advances in our understanding of the interactions between nutrition and sports performance.

In considering the role of diet in the athlete's life, two main issues must be considered, each of which gives rise to many subordinate questions. The first question is how the demands of training affect the body's requirement for energy and nutrients: this then has implications for body composition (including the body content of fat, muscle, and bone), for the hormonal environment and the regulation of substrate metabolism, and for various disease states that are affected by body fatness, nutrient intake, and other related factors. The second question is how nutritional status influences the responses to and the performance in competition.

Athletes should be encouraged to follow eating plans that maximize the extent of recovery between training sessions, maximize the effectiveness of the adaptations that occur in response to each training session, and minimize the risk of illness and injury that might interrupt training or prevent participation in competition. This involves identification of each athlete's nutritional goals and the formulation of an eating strategy that will allow those goals to be met. There will be special issues in the period before and during competition that will influence nutrition needs, and separate dietary strategies will be necessary for training and for competition.

Nutrition for Training

The training load of athletes varies greatly between individuals, depending on the nature of the sport and the level of competition, and it also varies over time in relation to the competitive season. Training may consist of high-intensity resistance training, brief but intense sprints, prolonged moderate intensity efforts, or technical work. Each places different demands on the muscles, cardiovascular system, and other tissues, and each has different energy requirements. The aim of training is to induce changes in body tissues and organs that will improve exercise performance, but different adaptations are required in different sports. Increasing muscle mass, strength, and power is a key objective in many sports, but in other sports, these changes would hinder, rather than help, performance. The training stimulus, therefore, must be specific to the objectives of the event. Within limits, the greater the training stimulus – consisting of the intensity, duration and

frequency of individual training sessions – the greater the adaptation that takes place. As mentioned above, nutrition is important in promoting recovery between training sessions to allow an increase in the training load that can be sustained without succumbing to illness and injury, and also in allowing more effective adaptations to each bout of training. This may be important in complex sports such as soccer, where different training objectives must be achieved and where the training must also accommodate practice of a variety of skills.

Influence of Exercise Training on Energy Balance

Energy must be supplied by the diet to meet immediate energy needs (body functions, energy for activity, and growth) and for the maintenance of body energy stores. Energy stores, consisting primarily of fat, but including the key carbohydrate stores in liver and muscle, play a number of important roles related to exercise performance, since they contribute to size and function (e.g., muscle mass) as well as providing fuel for exercise. Athletes try to manipulate these factors towards the characteristics that offer advantages to their sport: this may mean a change in body mass, a change (usually a reduction) in body fat, a change (usually an increase) in muscle mass, and optimization of muscle and liver carbohydrate stores.

Not all athletes are able to correctly identify goals that are suitable for their sport and for their individual make-up. This can lead to various problems, including excessive restriction of energy intake in an attempt to achieve an unrealistically low body mass. If energy intake is too low, and especially if carbohydrate intake is inadequate, it may not be possible to sustain the training load without the risk of chronic fatigue, injury and illness. If an energy deficit is incurred, it may lead to changes in metabolic and hormonal function, which affect performance, growth and health. One outcome of low energy availability in female athletes is a disturbance of reproductive function and menstrual regularity. Other problems are likely to occur in male athletes. There is a real danger that the focus on achieving a specific body mass and body composition, may become more important than achieving success in competition.

Monitoring of body mass can provide a useful index of energy balance in some situations, but other biomarkers are generally better. Measurement of body fat stores, usually by measurement of skinfold thickness, can be helpful in setting targets and in monitoring progress. Other markers, such as measurement of urinary ketone levels, can identify athletes who are failing to achieve an adequate carbohydrate intake. Problems are most likely to occur when the energy expenditure is either very high or very low. Athletes with very high energy demands are likely to be training at least twice per day, leaving limited opportunities for eating the large amounts of foods that are necessary. Athletes with low energy demands and who must restrict energy intake to achieve a low body mass have two main problems: they must cope with constant hunger and they must also be careful in their selection of foods to ensure that they achieve an adequate intake of essential nutrients.

An athlete's energy requirements are set primarily by the training load and by body mass, although there is also a large interindividual variability even when these factors are constant. Measurements of oxygen uptake, heart rate, and other variables made after exercise show that the metabolic rate may remain elevated for at least 12 h and possibly up to 24 h if the exercise is prolonged and close to the maximum intensity that can be sustained. After more moderate exercise, the metabolic rate quickly returns to baseline level. Therefore, it seems likely that the athlete training at near to the maximum sustainable level and who already has a very high energy demand will find this increased further by the elevation of postexercise metabolic rate: this will increase the difficulties that many of these athletes have in meeting their energy demand. The recreational exerciser, for whom the primary stimulus to exercise is often to control body mass or reduce body fat content, will not exercise hard enough or long enough to experience substantial elevations of metabolic rate after exercise.

Macronutrient Demands

Protein

The idea that protein requirements are increased by physical activity is intuitively attractive, and high-protein diets are a common feature of the diets of sportsmen and women. The available evidence shows an increased rate of oxidation of the carbon skeletons of amino acids during exercise, especially when carbohydrate availability is low. Protein contributes only about 5% of total energy demand in endurance exercise, but the absolute rate of protein breakdown is higher than at rest (where protein contributes about the same fraction as the protein content of the diet, i.e., typically about 12–16%) because of the higher energy turnover. It is often recommended that athletes engaged in endurance activities on a daily basis should aim to achieve a protein

intake of about 1.2–1.4 g kg^{-1} day^{-1}, whereas athletes engaged in strength and power training may need as much as 1.6–1.7 g kg^{-1} day^{-1}. Those who take no exercise have an estimated average requirement of about 0.6 g kg^{-1} day and the recommended intake for these individuals is about 0.8–1.0 g kg^{-1} day.

In strength and power sports such as weightlifting, sprinting and bodybuilding, the use of high-protein diets and protein supplements is especially prevalent, and daily intakes in excess of 2–4 g^{-1} kg^{-1} are not unusual. Scientific support for such high intakes is generally lacking, but those involved in these sports are adamant that such high levels of intake are necessary, not only to increase muscle mass but also to maintain muscle mass. This apparent inconsistency may be explained by Millward's adaptive metabolic demand model, which proposes that the body adapts to either high or low levels of intake, and that this adjustment to changes in intake occurs only very slowly. This means that individuals such as strength and power athletes who consume a high-protein diet over many years will find that any reduction in protein intake will result in a loss of muscle mass. This is because of an upregulation of the activity of the enzymes involved in protein oxidation to cope with the high intake: activity of these enzymes remains high when there is a sudden decrease in intake, leading to a net catabolic effect.

Protein synthesis and degradation are both enhanced for some hours after exercise, and the net effect on muscle mass will depend on the relative magnitude and duration of these effects. Several recent studies have shown that ingestion of small amounts of protein (typically about 35–40 g) or essential amino acids (about 6 g) either before or immediately after exercise will result in net protein synthesis in the hours after exercise, whereas net negative protein balance is observed if no source of amino acids is consumed. These observations have led to recommendations that protein should be consumed immediately after exercise, but the control condition in most of these studies has involved a relatively prolonged (6–12 h) period of fasting, and this does not reflect normal behavior. Individuals who consume foods containing carbohydrate and proteins in the hour or two before exercise may not further increase protein synthesis if additional amino acids or proteins are ingested immediately before, during, or after exercise.

Various high (30%) protein, high (30%) fat, low (40%) carbohydrate diets have been promoted for weight loss, and some diets even suggest almost complete elimination of carbohydrate from the diet. Some of these diets have been specifically targeted at athletes, accompanied by impressive claims and celebrity endorsements. Proposed mechanisms of action of these diets include reduced circulating insulin levels, increased fat catabolism, and altered prostaglandin metabolism, but it seems more likely that these diets achieve weight loss simply by restricting dietary choice. These diets can be effective in promoting short-term weight loss, primarily by restricting energy intake (typically to 1000–2000 kcal day^{-1}). There is no evidence to support improvements in exercise performance, and what evidence there is does not support the concept.

Carbohydrate

Carbohydrate is an essential fuel for the brain, red blood cells, and a few other tissues. Fat and carbohydrate are the main fuels used for energy supply in muscle during exercise. In low-intensity exercise, most of the energy demand can be met by fat oxidation, but the contribution of carbohydrate, and especially of the muscle glycogen, increases as the rate of energy demand increases. Carbohydrate oxidation rates of 3–4 g min^{-1} may be sustained for several hours by athletes in training or competition. When the glycogen content of the exercising muscles reaches very low levels, the work rate must be reduced to a level that can be accommodated by fat oxidation. In high-intensity exercise, essentially all of the energy demand is met by carbohydrate metabolism. Therefore, repeated short sprints place high demands on the muscle carbohydrate store, most of which can be converted to lactate within a few minutes.

Carbohydrate is stored in the body in the form of glycogen, primarily in the liver (about 70–100 g in the fed state) and in the skeletal muscles (about 300–500 g, depending on muscle mass and preceding diet). These stores are small relative to the body's requirements for carbohydrate. Carbohydrate supplies about 45% of the energy in the typical Western diet. This amounts to about 200–300 g day^{-1} for the average sedentary individual, and is adequate for normal daily activities. In an hour of hard exercise, however, up to 200 g of carbohydrate can be used, and sufficient carbohydrate must be supplied by the diet to replace the amount used. Replacement of the glycogen stores is an essential part of the recovery process after exercise: if the muscle glycogen content is not replaced, the quality of training must be reduced, and the risks of illness and injury are increased. Low muscle glycogen levels are associated with an increased secretion of cortisol during exercise, with consequent negative implications for immune function.

Table 1 Suggested carbohydrate intakes for athletes in training

Immediate postexercise recovery (0–4 h): 1 g per kg body mass per h, consisting of several small snacks
Daily recovery (moderate duration/low intensity training): 5–7 g kg^{-1} day^{-1}
Daily recovery (moderate–heavy endurance training): 7–12 g kg^{-1} day^{-1}
Daily recovery (extreme training: 4–6 h or more per day): 10–12 g kg^{-1} day^{-1}

When rapid recovery is a priority, replacement of carbohydrate should begin as soon as possible after exercise with carbohydrate foods that are convenient and appealing. Thereafter, the diet should supply sufficient carbohydrate to replace the amount used in training and to meet ongoing demands of other tissues. Some recommendations for carbohydrate intake after training or competition are shown in **Table 1**. For athletes preparing for competition, a reduction in the training load and the consumption of a high-carbohydrate diet in the last few days are recommended. This maximizes the body's carbohydrate stores and should ensure optimum performance, not only in endurance activities, but also in events involving short-duration high-intensity exercise and in field games involving multiple sprints.

The high-carbohydrate diet recommended for the physically active individual coincides with the recommendations of various expert committees that a healthy diet is one that is high in carbohydrate (at least 55% of energy) and low in fat (less than 30% of energy). However, where energy intake is either very high or very low, it may be inappropriate to express the carbohydrate requirement as a fraction of energy intake. With low total energy intakes, the fraction of carbohydrate in the diet must be high, but the endurance athlete with a very high-energy intake may be able to tolerate a higher fat intake. Recommendations, as in **Table 1**, should be framed in absolute amounts relative to body mass, i.e., grams of carbohydrate per kilogram body mass.

The type of carbohydrate eaten is less important than the amount. It is valuable to choose nutrient-rich carbohydrates and to add other foods to recovery meals and snacks to provide a good source of protein and other nutrients. The presence of small amounts of protein in recovery meals may promote additional glycogen recovery when carbohydrate intake is less than optimal or when frequent snacking is not possible. Protein taken at this time may also stimulate protein synthesis in muscles, as described above. Carbohydrate-rich foods with a moderate to high glycemic index (GI) provide a readily available source of carbohydrate for glycogen synthesis, and should be the major fuel choices in recovery meals.

Fat

Fat is an important metabolic fuel in prolonged exercise, especially when the availability of carbohydrate is low. One of the primary adaptations to endurance training is an enhanced capacity to oxidize fat, thus sparing the body's limited carbohydrate stores. Studies where subjects have trained on high-fat diets, however, have shown that a high-carbohydrate diet during a period of training brings about greater improvements in performance. Even when a high-carbohydrate diet is fed for a few days to allow normalization of the muscle glycogen stores before exercise performance is measured, the exercise capacity remains less after training on a high fat diet. It must be recognized, though, that these short-term training studies usually involve relatively untrained individuals and may not reflect the situation of the highly trained elite endurance athlete where the capacity of the muscle for oxidation of fatty acids will be much higher. For the athlete with very high levels of energy expenditure in training, the exercise intensity will inevitably be reduced to a level where fatty acid oxidation will make a significant contribution to energy supply and fat will provide an important energy source in the diet. Once the requirements for protein and carbohydrate are met, the balance of energy intake can be in the form of fat.

Fat also serves other important functions in the diet. As well as providing essential fatty acids, it acts as a vehicle for the transport of fat-soluble nutrients. Some athletes try to minimize their fat intake, but this is not wise.

Micronutrients and Physical Activity

Many micronutrients play key roles in energy metabolism, and high rates of energy turnover (up to 20–100 times the resting rate) may be required in the active muscles during hard exercise. Although an adequate vitamin and mineral status is essential for normal health, marginal deficiency states may only be apparent when the metabolic rate is high. Prolonged strenuous exercise performed on a regular basis may also result in increased losses from the body or in an increased rate of turnover, resulting in the need for an increased dietary intake. An increased food intake to meet energy requirements will increase dietary micronutrient intake, but not all athletes have high-energy intakes. Athletes who restrict food intake to control or reduce body fat

levels may have low-energy intakes over prolonged periods. Athletes may also eat monotonous diets, with a limited range of foods, thus increasing the risk of an inadequate micronutrient intake. Supplementation with micronutrients may be warranted in some instances, but normally only where specific deficiencies have been demonstrated by biochemical investigations and where dietary modification is not an option. Individuals who are very active may need to pay particular attention to their intake of iron and calcium.

Iron deficiency anemia affects some athletes engaged in intensive training and competition, but it seems that the prevalence is the same in athletic and sedentary populations, suggesting that exercise *per se* does not increase the risk. The implications of even mild anemia for exercise performance are, however, significant. A fall in the circulating hemoglobin concentration is associated with a reduction in oxygen-carrying capacity and a decreased exercise performance. Low serum ferritin levels are not associated with impaired performance, however, and iron supplementation in the absence of frank anemia does not influence indices of fitness. Routine iron supplementation is not wise, as too much may be harmful.

Osteoporosis is now widely recognized as a problem for both men and women, particularly so in women, and an increased bone mineral content is one of the benefits of participation in an exercise program. Regular exercise results in increased mineralization of those bones subjected to stress and an increased peak bone mass may delay the onset of osteoporotic fractures; exercise may also delay the rate of bone loss. Estrogen plays an important role in the maintenance of bone mass in women, and prolonged strenuous activity may result in low estrogen levels, causing bone loss. Many very active women also have a low body fat content and may also have low energy (and calcium) intakes in spite of their high activity levels. All of these factors are a threat to bone health. The loss of bone in these women may result in an increased predisposition to stress fractures and other skeletal injury and must also raise concerns about bone health in later life. It should be emphasized, however, that this condition appears to affect only relatively few athletes, and that activity is generally beneficial for the skeleton.

Water and Electrolyte Balance

Few situations represent such a challenge to the body's homeostatic mechanisms as that posed by prolonged strenuous exercise in a warm environment. Only about 20–25% of the energy available from substrate catabolism is used to perform external work, with the remainder appearing as heat. At rest, the metabolic rate is low: oxygen consumption is about $250\,\mathrm{ml\,min^{-1}}$, corresponding to a rate of heat production of about 60 W. Heat production increases in proportion to metabolic demand, and reaches about 1 kW in strenuous activities such as marathon running (for a 70-kg runner at a speed that takes about 2.5 h to complete the race). To prevent a catastrophic rise in core temperature, heat loss must be increased correspondingly and this is achieved primarily by an increased rate of evaporation of sweat from the skin surface. In hard exercise under hot conditions, sweat rates can reach $3\,\mathrm{l\,h^{-1}}$, and trained athletes can sustain sweat rates in excess of $2\,\mathrm{l\,h^{-1}}$ for many hours. This represents a much higher fractional turnover rate of water than that of most other body components. In the sedentary individual living in a temperate climate, about 5–10% of total body water may be lost and replaced on a daily basis. When prolonged exercise is performed in a hot environment, 20–40% of total body water can be turned over in a single day. In spite of this, the body water content is tightly regulated, and regulation by the kidneys is closely related to osmotic balance.

Along with water, a variety of minerals and organic components are lost in variable amounts in sweat. Sweat is often described as an ultrafiltrate of plasma, but it is invariably hypotonic. The main electrolytes lost are sodium and chloride, at concentrations of about $15–80\,\mathrm{mmol\,l^{-1}}$, but a range of other minerals, including potassium and magnesium, are also lost, as well as trace elements in small amounts. Some athletes may lose up to 10 g of salt (sodium chloride) in a single training session, and may train in these conditions twice per day. These substantial salt losses must be replaced from foods and drinks, though the use of salt supplements is seldom necessary.

Failure to maintain hydration status has serious consequences for the active individual. A body water deficit of as little as 1–2% of total body mass can result in a significant reduction in exercise capacity. Endurance exercise is affected to a greater extent than high-intensity exercise, and muscle strength is not adversely affected until water losses reach 5% or more of body mass. Hypohydration greatly increases the risk of heat illness, and also abolishes the protection conferred by prior heat acclimation.

Many studies have shown that the ingestion of fluid during exercise can significantly improve

performance. Adding carbohydrate to the fluid confers an additional benefit by providing an energy source for the working muscles. Addition of small amounts (perhaps about 2–8%) of carbohydrate in the form of glucose, sucrose, or maltodextrin will promote water absorption in the small intestine as well as providing exogenous substrate that can spare stored carbohydrate. The addition of too much carbohydrate will slow gastric emptying and, if the solution is strongly hypertonic, may promote secretion of water into the intestinal lumen, thus delaying fluid availability. Voluntary fluid intake is seldom sufficient to match sweat losses, and palatability of fluids is therefore an important consideration. It is not necessary to consume enough fluid during exercise to match sweat losses, as a body mass deficit of 1–2% is unlikely to have adverse consequences. If exercise is prolonged and sweat losses high, the addition of sodium to drinks may be necessary to prevent the development of hyponatremia. Ingestion of large volumes of plain water is also likely to limit intake because of a fall in plasma osmolality leading to suppression of thirst.

Replacement of water and electrolyte losses incurred during exercise is an important part of the recovery process in the postexercise period. This requires ingestion of fluid in excess of the volume of sweat lost to allow for ongoing water losses from the body. Re-establishment of water balance requires replacement of solute, especially sodium, losses as well as volume replacement. If food containing electrolytes is not consumed at this time, electrolytes, especially sodium, must be added to drinks to prevent diuresis and loss of the ingested fluid.

Dietary Supplements

The use of nutritional supplements in athletes and in the health-conscious recreationally active population is widespread, as it is in the general population. A very large number of surveys have been published. A meta-analysis of 51 published surveys involving 10 274 male and female athletes of varying levels of ability showed an overall prevalence of supplement use of 46%, but the prevalence varied widely in different sports, at different levels of age, performance etc., and in different cultural backgrounds.

Many different supplements are used by athletes with the aim of improving or maintaining general health and exercise performance. In particular, supplement use is often aimed at promoting tissue growth and repair, promoting fat loss, enhancing resistance to fatigue, and stimulating immune function. Most of the supplements that are sold to athletes have not been well researched, and both safety and efficacy remain open to question for many of these products. Anyone seeking to improve health or performance would be better advised to ensure that they consume a sound diet that meets energy needs and contains a variety of foods. A recent development of concern to athletes is the finding of various prohibited doping agents in what should be legitimate sports nutrition products. Supplements for which there is good evidence of beneficial effects on performance include caffeine, creatine, and bicarbonate, but the risk of an inadvertent positive doping result must always be considered.

See also: **Anemia**: Iron-Deficiency Anemia. **Bone**. **Carbohydrates**: Chemistry and Classification; Regulation of Metabolism; Requirements and Dietary Importance; Resistant Starch and Oligosaccharides. **Electrolytes**: Acid-Base Balance; Water–Electrolyte Balance. **Energy**: Balance. **Exercise**: Beneficial Effects; Diet and Exercise. **Fats and Oils**. **Fatty Acids**: Metabolism; Monounsaturated; Omega-3 Polyunsaturated; Omega-6 Polyunsaturated; Saturated; *Trans* Fatty Acids. **Protein**: Synthesis and Turnover; Requirements and Role in Diet; Digestion and Bioavailability; Quality and Sources. **Supplementation**: Role of Micronutrient Supplementation; Developing Countries.

Further Reading

American College of Sports Medicine, American Dietetic Association, and Dietitians of Canada (2000) Joint Position Statement: Nutrition and athletic performance. *Medical Science in Sports and Exercise* **32**: 2130–2145.

Ivy J (2000) Optimization of glycogen stores. In: Maughan RJ (ed.) *Nutrition in Sport*, pp. 97–111. Oxford: Blackwell.

Kiens B and Helge JW (1998) Effect of high-fat diets on exercise performance. *Proceedings of the Nutrition Society* **57**: 73–75.

Maughan RJ (1999) Nutritional ergogenic aids and exercise performance. *Nutrition Research Reviews* **12**: 255–280.

Maughan RJ, Burke LM, and Coyle EF (eds.) (2004) *Foods, Nutrition and Sports Performance*, vol. 2. London: Routledge.

Maughan RJ and Murray R (eds.) (2000) *Sports Drinks: Basic Science and Practical Aspects*. Boca Raton: CRC Press.

Millward DJ (2001) Protein and amino acid requirements of adults: current controversies. *Canadian Journal of Applied Physiology* **26**: S130–S140.

Nieman DC and Pedersen BK (1999) Exercise and immune function. *Sports Medicine* **27**: 73–80.

Noakes TD and Martin D (2002) IMMDA-AIMS advisory statement on guidelines for fluid replacement during marathon running. *New Studies in Athletics* **17**: 15–24.

Shirreffs SM and Maughan RJ (2000) Rehydration and recovery after exercise. *Exercise and Sports Science Reviews* **28**: 27–32.

Wolfe RR (2001) Effects of amino acid intake on anabolic processes. *Canadian Journal of Applied Physiology* **26**: S220–S227.

Starch *see* **Carbohydrates**: Chemistry and Classification; Regulation of Metabolism; Requirements and Dietary Importance; Resistant Starch and Oligosaccharides

STARVATION AND FASTING

J E Bines and R G Heine, University of Melbourne, Melbourne, VIC, Australia

During a normal 24-h day the body's essential cellular and organ functions remain homeostatic despite intermittent nutrient intake and changing metabolic demand. A highly sophisticated and integrated system provides the metabolic adaptation for these normal changes in substrate provision and utilization. A basic knowledge of the normal metabolic responses in the feeding–fasting cycle is pivotal to the understanding of the changes that occur during periods of prolonged fasting and starvation. In this review of the biochemical aspects of the metabolic responses to fasting and starvation, the term 'fasting' is defined as the total absence of nutrient intake, whereas 'starvation' is defined as a prolonged period of inadequate food intake. During a period of prolonged fasting, the body undergoes a sequence of changes that include the initial depletion of fuel stores and metabolic adaptation. However, if fasting continues, metabolic decompensation will occur, finally resulting in death. The extent and rate of progression through these steps depend on the amount of fuel stores at initiation of the fast, the severity and duration of nutritional deprivation, and the presence or absence of a significant catabolic stress, such as injury, sepsis, or cancer (**Table 1**).

The Feeding–Fasting Cycle

Energy Requirements

Energy is essential for many important body functions, including the maintenance of cellular integrity and function, new tissue synthesis, thermoregulation, and

Table 1 Metabolic features and laboratory parameters of starvation and fasting

	Uncomplicated protein energy malnutrition (Marasmus)	*Hypoalbuminemic malnutrition (Kwashiorkor)*	*Stressed malnutrition (mixed)*
Etiology	Decreased energy intake	Decreased protein intake associated with catabolic stress	Decreased energy and protein intakes associated with catabolic stress
Metabolic adaptation			
Weight loss	Slow (months to years)	Intermediate (weeks to months)	Rapid (weeks)
Resting energy expenditure	Decreased	Decreased or increased	Increased
Nitrogen loss	Minimal	Increased	Increased
Water, sodium	Initial loss	Retention	Retention
Hormonal	Early small increase in catecholamines, glucagon, cortisol, growth hormone; then slow decrease	Increase in catecholamines, glucagon, cortisol, growth hormone	Increase in catecholamines, glucagon, cortisol, growth hormone
	Decrease in insulin, leptin	Decrease in insulin, leptin	Decrease in insulin, leptin
Laboratory parameters			
Albumin	Normal–slowly decreased	Decreased	Decreased
Transferrin	Normal–slowly decreased	Decreased	Decreased
Total lymphocyte count	Normal–slowly decreased	Decreased	Decreased
Skin hypersensitivity	Decreased	Decreased	Decreased

physical activity. The energy requirements of an individual vary with age, sex, body composition, physical activity, and stress. In the normal adult at basal state, approximately 75% of energy requirements reflect the energy needs of major organs (brain, ~20%; skeletal muscle, 18–22%; abdominal muscles, ~25%; and heart, ~11%). However, during normal daily activity the total energy requirement and the proportion of energy needed by different tissues may vary considerably. For example, with exercise the energy requirement of skeletal muscles increases, and during a meal the abdominal organs require more energy for the process of digestion and absorption. Children require additional energy for growth.

Energy Production

The body derives energy from the metabolism of carbohydrate, fat, and protein provided exogenously in the fed state and endogenously in the postabsorptive state. A mixture of metabolic fuels, including glucose, triacylglycerols, ketone bodies, nonesterified fatty acids, alcohol, and amino acids, are present in the circulation. The proportion of these energy substrates in the blood at any one time depends on the fed or fasting state of the individual, the extent of fuel stores, and recent or current metabolic demand. In a normal, nonobese, 70-kg adult there are approximately 500 MJ (120 000 kcal) contained in adipose tissue, 100 MJ (24 000 kcal) stored in muscle and visceral proteins, and 4.2 MJ (1000 kcal) stored as liver and muscle glycogen. During a normal day, half of the total energy requirement is met by carbohydrate metabolism. At this rate, glycogen stores would be exhausted after 1 or 2 days of fasting. However, glycogen stores are maintained for a longer period due to the production of glucose from gluconeogenesis (**Figure 1**).

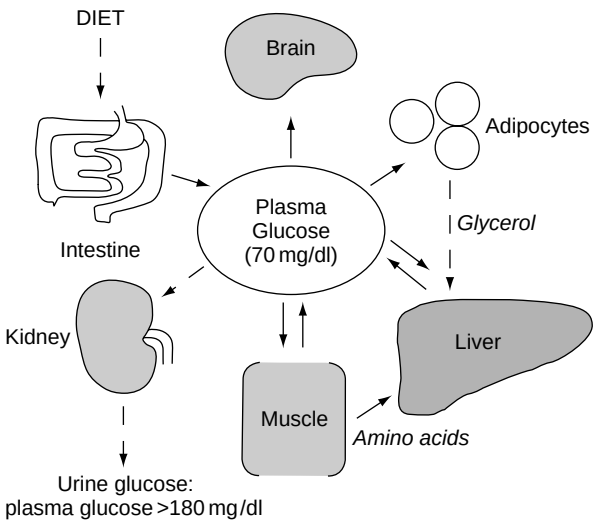


Figure 1 The metabolism of energy substrates to maintain glucose homeostasis.

Carbohydrate metabolism Glucose plays a key role in body metabolism. It is the preferred metabolic fuel for many tissues and is an essential fuel for the retina, red blood cells, the renal medulla, and the brain under normal conditions. In the fed state, glucose is derived from the digestion and absorption of carbohydrates provided in the meal. To produce energy from glucose, three metabolic pathways are involved (**Figure 2**). Glucose is first oxidized to form pyruvate via the glycolytic pathway. Pyruvate then enters the Krebs cycle and is completely oxidized to form NADH + H, $FADH_2$, and carbon dioxide. The NADH + H transports hydrogen to the respiratory chain, where it is used to reduce oxygen to water. The net yield of energy from the metabolism of 1 molecule of glucose is 38 molecules of ATP.

Because of the strict glucose requirements of the brain, the circulating blood glucose pool is tightly controlled at approximately 16 g. Three important mechanisms are responsible for this regulation:

1. Insulin enhances glucose uptake into muscle and fat and stimulates glycogen synthesis. It also inhibits lipolysis, glycogenolysis, and gluconeogenesis. High insulin levels will decrease blood glucose levels. Conversely, low insulin levels will cause a rise in blood glucose by decreased inhibition of glycogenolysis and reduced peripheral uptake of glucose.
2. Glucagon increases liver glycogen breakdown, gluconeogenesis, and ketogenesis from fatty acids. It also stimulates lipolysis from adipocytes in extrahepatic tissue. The net result of glucagon activity is an increase in blood glucose concentration that helps to maintain blood glucose levels despite the effect of insulin.
3. Neuroendocrine responses to glucose deprivation in the brain act to rapidly increase glucose release from liver glycogen.

The fed state is characterized by increased blood concentrations of glucose, amino acids, and fat. Insulin secretion is stimulated while glucagon levels remain unchanged or are decreased. As a result, there is increased glucose uptake into tissues and enhanced glycogen, protein, and triacylglycerol synthesis. Glucagon balances this effect by stimulating glycogen breakdown to maintain blood glucose levels. By this mechanism, blood glucose levels are controlled during periods of surplus carbohydrate ingestion and excess glucose is stored as glycogen or fat.

Glycogen is a complex hydrated polymer of glucose arranged in a highly branched, spherical form. It allows

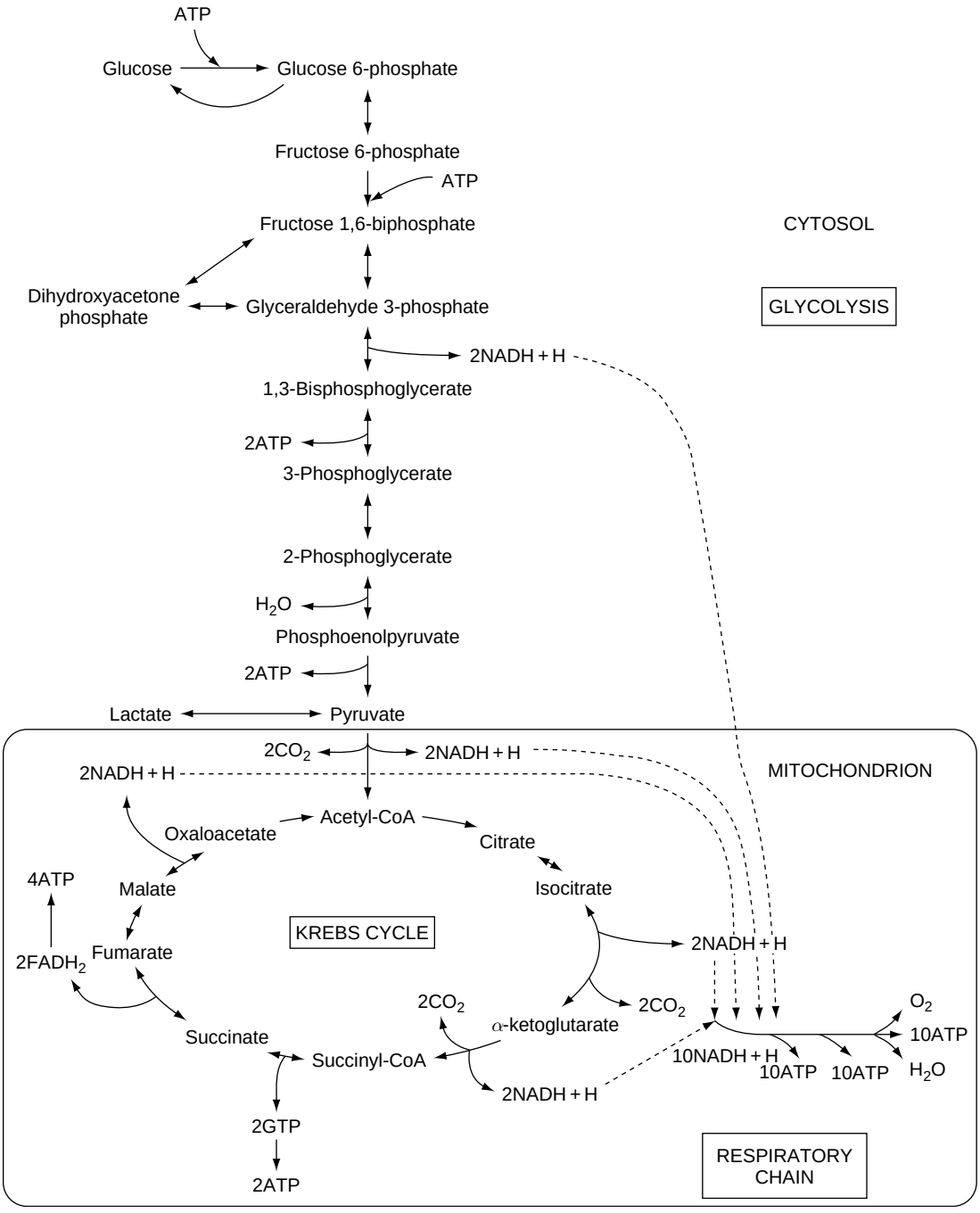


Figure 2 The production of energy from glucose via the glycolytic pathway, the Krebs cycle, and the respiratory chain.

glucose to be stored in large amounts without causing osmotic shifts. The terminal glucose molecules within this branching structure are accessible to the enzymes mediating glycogen breakdown to allow the rapid release of glucose in times of stress. The glycogen molecule expands in size after a carbohydrate-rich meal to approximately 40 nm in diameter and shrinks to 10 nm in diameter or less between meals. An adult male receiving a normal carbohydrate-containing diet has approximately 70 g of liver glycogen and 200 g of muscle glycogen. Glycogen is broken down by the enzyme phosphorylase. Glucose-6-phosphatase continues the breakdown of glycogen to glucose in the liver. Muscle glycogen is metabolized by anaerobic glycolysis to form pyruvate and lactate. Lactate is then transported to the liver, where it acts as a precursor for gluconeogenesis. This is called the Cori cycle (**Figure 3**). The Cori cycle contributes to approximately 40% of the normal plasma

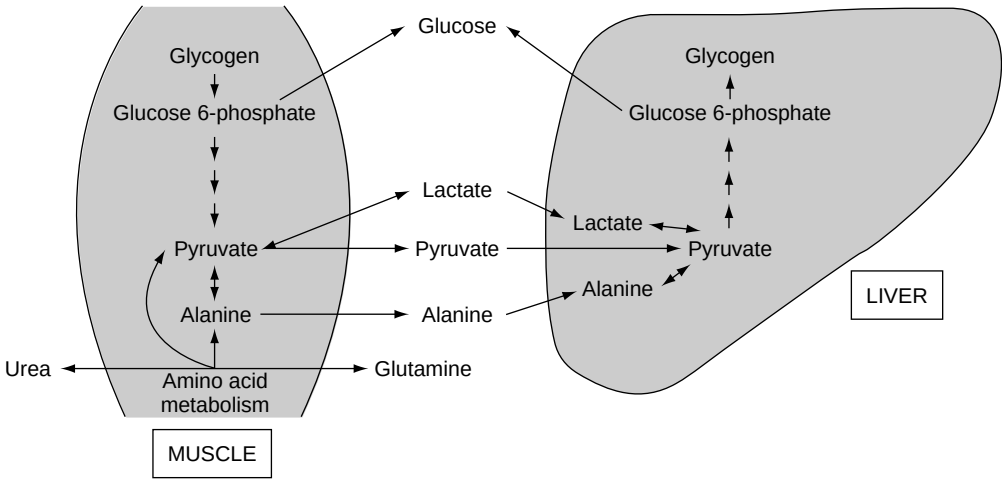


Figure 3 The metabolism of muscle glycogen and protein to form glucose involving the Cori cycle (lactate to glucose) and the glucose–alanine cycle.

glucose turnover. It has the advantage of providing energy (net 3 molecules of ATP) without the loss of glucose molecules. The energy required for the resynthesis of glucose in the liver is derived from fatty acid oxidation. The total body glycogen stores can meet the needs of the brain for approximately 3 days. After this period, alternative sources of metabolic fuel must be found.

Protein metabolism Body nitrogen resides in two main compartments. Approximately half of the body's nitrogen is contained in extracellular tissues such as collagen. The nitrogen present within these tissues is relatively fixed and does not change significantly with starvation. The nitrogen turnover within this compartment can be assessed by the measurement of hydroxyproline excretion. The remaining nitrogen is present in the lean muscle mass, comprising skeletal and visceral muscle. The proteins within these tissues are constantly being broken down and resynthesized at a rate of 3–3.5 g/kg/day in a young adult. Measurement of urinary 3-methylhistidine excretion and creatinine excretion can be used to estimate the fractional catabolic rate of skeletal muscle.

Protein synthesis and degradation involves several independent metabolic systems. The autophagic–lysosomal pathway facilitates most of the proteolysis occurring in the body. Another pathway, the proteasome–ubiquitin system, plays a central role in the degradation of specific proteins. For example, the proteasome is involved in the rapid regulation of many rate-limiting enzymes. The proteasome also mediates the loss of skeletal muscle protein in starvation and wasting disorders. The 26S proteasome consists of a proteolytic core complex, the 20S proteasome, and two 19S regulatory complexes. Substrates are conjugated with multiubiquitin chains before degradation. The autophagic–lysosomal system is regulated by plasma amino acid levels, whereas no such feedback mechanism has been demonstrated for the ubiquitin–proteasome pathway. The fatty acid, eicopentenoic acid, has been shown to downregulate ubiquitin-dependent proteolysis in mice after acute starvation. Furthermore, experimental data in starved rats suggest that ubiquitin–proteasome-dependent muscle proteolysis is ameliorated in response to insulin release and refeeding. The exact mechanisms involved in the regulation of ubiquitin–proteasome-dependent proteolysis are not completely understood.

Both autophagic–lysosomal and ubiquitin–proteasome pathways generate amino acids as their final product. Glutamine is the most abundant amino acid in humans. It is an important energy substrate for monocytes and the gut. Depletion of glutamine during starvation or chronic illness has been shown to inhibit the ubiquitin–proteasome proteolytic pathway and may contribute to impaired immune function of monocytes during starvation or critical illness.

In the fed state, amino acids digested and absorbed in excess of the body's immediate requirements for incorporation into proteins or other molecules are either oxidized for energy or metabolized to glycogen or fat. Protein provides approximately $17\,kJ\,g^{-1}$ ($4\,kcal\,g^{-1}$) of energy when metabolized as an energy source.

Prolonged fasting results in depletion of liver and muscle glycogen stores. In this clinical setting the conversion of amino acids to glucose contributes to the glucose requirements of the brain. The transition to metabolism of amino acids as an energy source is

mediated by an alteration in the balance of insulin and glucagon. The breakdown of tissue protein to provide glucose results in a sustained loss of body nitrogen of approximately 12 g per day. Experimentally, this loss of body nitrogen can be prevented by the administration of glucose. As a result of muscle protein breakdown, amino acids, predominantly alanine and glutamine, are released into the circulation. However, the amount of alanine released exceeds the alanine content of the muscle protein. This is because approximately one-third of the alanine released from muscle originates directly from the muscle protein, whereas the remaining two-thirds is derived from pyruvate. Pyruvate is formed by the metabolism of muscle glycogen or by the transamination of other amino acids contained within the muscle protein. Alanine is then transported to the liver, where it is rapidly taken up and converted to glucose: This is known as the glucose–alanine cycle (**Figure 3**). Despite the increased release of alanine from muscle, plasma alanine levels decline in early fasting. This results from the rapid uptake and conversion of alanine by the liver.

Fat metabolism Fat is an efficient store of energy providing approximately $38\,kJ\,g^{-1}$ ($9\,kcal\,g^{-1}$). Fat is predominantly stored as triacylglycerols within adipocytes. The amount of fat stores may vary substantially among individuals. In the fed state, insulin stimulates triacylglycerol synthesis. During fasting, triacylglycerol is converted to fatty acids and glycerol (**Figure 4**). Within days, glycerol and palmitate release increases by two or three times fed levels. This release is regulated by hormone-sensitive lipase. Due to the absence of glycerol kinase in white adipose tissue, glycerol cannot be completely metabolized within the adipocytes and is transported to the liver, where it is converted to glucose by gluconeogenesis. The fatty acids are either released from the adipocytes to be oxidized by the liver or other tissues or may be reesterified with glycerol 3-phosphate and reenter the cycle to form triacylglycerol (**Figure 4**). The energy cost of reesterification of fatty acids in starvation may account for 2 or 3% of the resting energy expenditure.

Oxidation of long-chain fatty acids (LCFAs) requires the mitochondrial carnitine system for transport into the mitochondrial matrix. This transport system consists of several enzymes, including the malonyl-coenzyme A (CoA)-sensitive carnitine palmitoyltransferase I (CPT-I), carnitine:acylcarnitine translocase, and CPT-II. CPT-I is regulated at

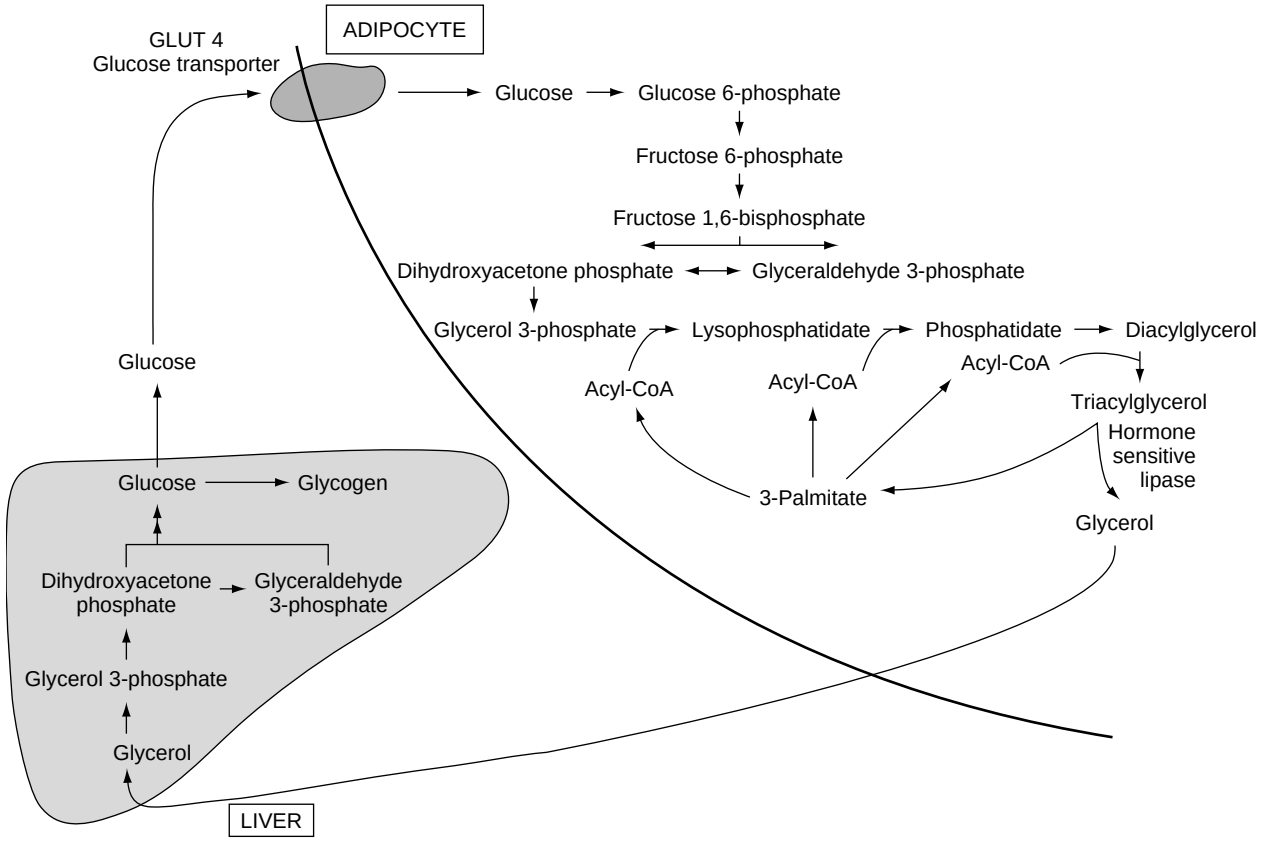


Figure 4 The triacylglycerol–fatty acid cycle.

the transcriptional level by malonyl-CoA. Although carbohydrates are the major energy source during high-impact exercise, LCFAs are the preferred substrate during endurance exercise. In human skeletal muscle, during exercise free carnitine appears to play a greater regulatory role on LCFA oxidation than malonyl-CoA. However, some aspects of the transport of LCFAs to the inner mitochondrial matrix in humans are not completely understood.

Most of the acetyl-CoA produced from fatty acid oxidation is metabolized to acetoacetate, which in turn may be converted to β-hydroxybutyrate and acetone. These products are known as ketone bodies. Acetyl-CoA is also converted to malonyl-CoA, catalyzed by the enzyme acetyl-CoA carboxylase. Although ketone bodies are produced in small quantities in the fed state, they are generally metabolized by the liver and are not released into the circulation. During fasting, the rate of production of acetoacetate and β-hydroxybutyrate significantly increases. These metabolites are released into the circulation and can be used by the brain and other tissues as an alternative energy source.

Metabolic Consequences of Fasting and Starvation

Postabsorptive State

The postabsorptive state commences when the last nutrient is absorbed from the previous meal and continues until the next meal or for approximately 12 h during a normal overnight fast. Metabolically, it is the period when there is a transition from exogenous energy consumption to reliance on endogenous energy sources. The release from the liver of approximately 200–250 g of glucose per day or 8–10 g per hour balances the rate of glucose utilization of the brain and other tissues. During an overnight fast, a significant proportion of the glucose requirements are met by the breakdown of liver glycogen. The remaining glucose is formed from noncarbohydrate sources: glycerol (from triacylglycerols), pyruvate, and lactate (from muscle).

Prolonged Fasting

In a normal adult, resting energy expenditure is proportional to lean body mass. Prolonged fasting is associated with a loss of lean body mass and in a reduction in resting energy expenditure. Fourteen to 21 days after commencing a starvation diet there is a 15% reduction in resting energy expenditure. However, at this time there is only a 5% reduction in lean body mass. Clearly, another mechanism contributes to the decreased resting energy expenditure observed in early starvation.

A number of hormonal changes occur as a response to starvation that may alter resting energy expenditure. Decreased activity of 5′-monodeiodinase in the liver and peripheral tissues resulting in a reduction in the conversion of thyroxine (T_4) to the metabolically active form, triiodothyronine (T_3), has been observed within hours to days in patients on a starvation diet. This effect is modified by both the carbohydrate intake and the total energy content of the diet. However, the mechanism linking low circulating T_3 levels to decreased resting energy expenditure in starvation is not well understood. Catecholamine secretion and turnover is decreased in uncomplicated starvation. This is clinically recognized as a reduction of core temperature, heart rate, and blood pressure in patients during starvation. Another factor contributing to the reduced resting energy expenditure observed in patients during starvation may be a reduction in the activity of the sodium–potassium pump. The activity of this pump is influenced by circulating T_3, catecholamines, and insulin—all known to be altered in starvation. There is evidence that in prolonged starvation increased intracellular sodium and decreased intracellular potassium are linked to decreased pumping of sodium from cells.

With increasing duration of the fast and depletion of liver and muscle glycogen stores, the conversion of amino acids to glucose contributes to the glucose requirements of the brain. The transition to metabolism of amino acids as an energy source is mediated by a change in balance of insulin and glucagon. Insulin levels decline, whereas glucagon levels tend to be maintained or even slightly increased. During early fasting, the muscle releases alanine and glutamine. The glucose–alanine cycle provides glucose to the muscle in exchange for alanine provided to the liver as a precursor for gluconeogenesis (**Figure 3**). Blood levels of the branched-chain amino acids double after 3–5 days of fasting but decrease if fasting is prolonged. This decrease in plasma levels results from the rapid uptake of alanine by the liver for conversion to glucose. Glutamine released from the muscle during fasting is preferentially taken up by the intestine, where it is used as an energy source, and by the kidney, where it is used for renal ammonia production. Although the metabolism of amino acids to glucose is a very important metabolic adaptation to fasting, it provides only approximately 45 g of glucose per day. This amount alone is insufficient to meet the glucose requirements of the brain and must be supplemented by energy produced from fat metabolism. Based on data from glucose balance studies, during a 24-h fast the liver contributes 80–90% of glucose production and the kidney 10–20%. However, if fasting is extended from

12 to 60 h there is a 2.5-fold increase in renal glucose production, whereas glucose production from the liver decreases by 25%.

The mobilization of triacylglycerol stores to provide energy is regulated by a number of factors. Insulin levels decrease by 35% within 24 h of fasting. This is associated with a 50–80% increase in the rate of lipolysis. Low circulating insulin levels cause a reduction in the uptake of glucose into adipocytes by altering the function of the GLUT4 glucose transporter (**Figure 4**). Lipolysis is also stimulated by glucagon and adrenocorticotrophic hormone (ACTH) during starvation. This effect is mediated by cyclic AMP-dependent protein kinase, which stimulates hormone-sensitive lipase (**Figure 5**). In prolonged starvation, cortisol increases hormone-sensitive lipase synthesis. Adequate amounts of glycerol 3-phosphate are therefore unavailable for the reesterification of fatty acids produced from triacylglycerol breakdown. Nonesterified fatty acids are released into the circulation and free fatty acid concentrations increase from 0.5–0.8 to 1.2–1.6 mmol l^{-1} within the first few days of fasting. Fatty acids circulate bound to albumin and can be oxidized in the liver or other tissues to produce energy. The switch to using ketone bodies as an energy source by the brain appears to be primarily controlled by the blood concentration of ketone bodies rather than a hormonal effect. Ketone body production by the liver peaks after 3 or 4 days of fasting. However, blood ketone body levels continue to increase rapidly for the first 7–10 days before stabilizing at approximately 6–8 mM at 2 or 3 weeks. The continued increase in blood ketone body levels despite achieving maximal liver production early in fasting is due to decreased renal excretion of ketone bodies.

As fatty acid oxidation and ketone body formation increase, there is a reduction in glucose production and oxidation. This may be mediated by inhibition of the pyruvate dehydrogenase complex activity. After a 3-week fast, a marked reduction in glucose metabolism throughout the brain is observed using positron emission tomography. Glucose uptake of the brain is more than halved after a fast of 5 weeks.

After a period of fasting longer than 2 weeks, there is a shift to conserve muscle protein and reduce body protein turnover. Urinary nitrogen losses decrease to 4–6 g per day and approximately half of this is excreted as ammonia required to buffer acid produced by ketoacids. This transition appears to be directly or indirectly dependent on increasing blood ketone body concentration. With prolonged fasting, muscles change from ketone body production to fatty acid oxidation. This may potentially conserve branched-chain amino acids, which may in turn limit proteolysis. There is evidence that leucine can stimulate protein synthesis. A small increase in insulin in response to high circulating ketone body levels may also influence protein metabolism.

With increasing ketone body production, the liver reduces the rate of gluconeogenesis. The kidney becomes the major gluconeogenic organ and produces half of the body's glucose requirements. Glutamine is the predominant substrate for kidney gluconeogenesis and the nitrogen product of this process provides the ammonia needed to buffer ketoacids in the urine.

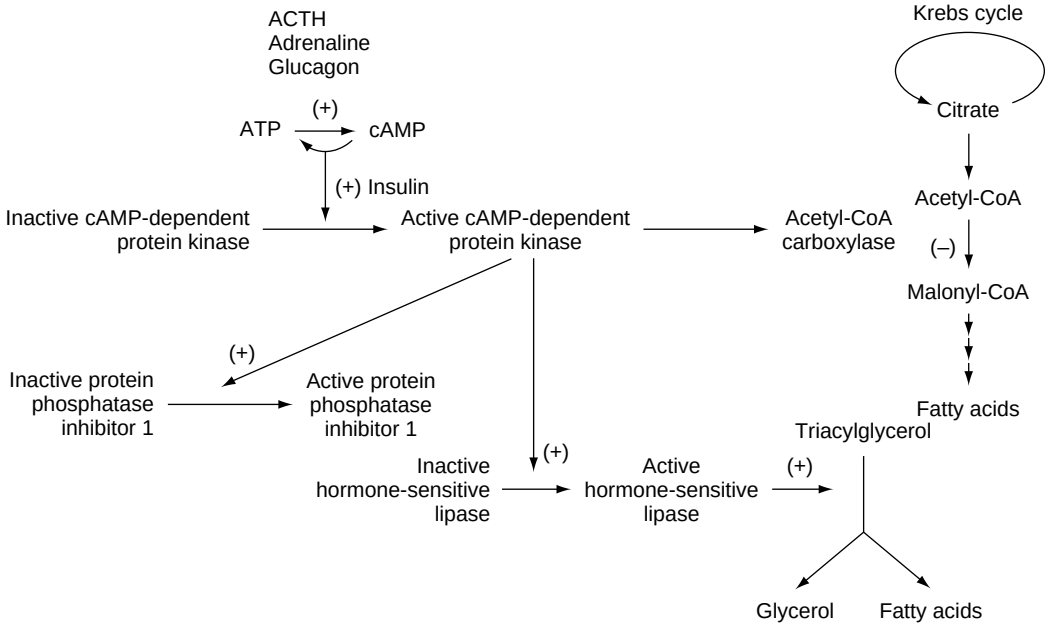


Figure 5 Lipolysis is stimulated by the action of glucagon, ACTH, and adrenaline. This effect is mediated by cyclic AMP-dependent protein kinase.

Death will occur when there is a failure to replenish fuel stores and insufficient available energy to maintain essential bodily functions. Because fat is the predominant source of energy, the time until death in uncomplicated fasting will depend on the size of the prefasting fat stores. In a normal adult, fat stores will be sufficient to sustain life for approximately 60–70 days. The extent of protein loss is also linked to survival. A loss of more than half of the lean body mass compartment (approximately half of total body protein) is predictive of death.

See also: **Amino Acids**: Metabolism. **Carbohydrates**: Regulation of Metabolism. **Energy**: Metabolism; Requirements. **Fatty Acids**: Metabolism. **Glucose**: Chemistry and Dietary Sources; Metabolism and Maintenance of Blood Glucose Level. **Lipids**: Chemistry and Classification. **Protein**: Digestion and Bioavailability.

Further Reading

Cahill GF (1970) Starvation in man. *New England Journal of Medicine* **282**: 668–675.

Cano N (2002) Bench-to-bedside review: Glucose production from the kidney. *Critical Care* **6**: 317–321.

Denton RM and McCormack JG (1995) Fuel selection at the level of the mitochondria in mammalian tissues. *Proceedings of the Nutrition Society* **54**: 11–22.

Hoffe LJ (1994) Starvation. In: Shils ME, Olson JA, and Shike M (eds.) *Modern Nutrition in Health and Disease*, vol. 2: 927–949. Philadelphia: Lea & Febiger.

Kadowaki M and Kanazawa T (2003) Amino acids as regulators of proteolysis. *Journal of Nutrition* **133**(supplement): 2052S–2056S.

Kerner J and Hoppel C (2000) Fatty acid import into mitochondria. *Biochimica et Biophysica Acta* **1486**: 1–17.

Keys A, Brozek J, Henschel A *et al.* (1950) *The Biology of Human Starvation* Minneapolis: University of Minnesota Press.

Kiens B and Roepstorff C (2003) Utilization of long-chain fatty acids in human skeletal muscle during exercise. *Acta Physiologica Scandinavica* **178**: 391–396.

Klein S, Sakurai Y, Romijn JA, and Carroll RM (1993) Progressive alterations in lipid and glucose metabolism during short-term fasting in young adult men. *American Journal of Physiology* **265**: E801–E806.

Love AHG (1986) Metabolic response to malnutrition: Its relevance to enteral feeding. *Gut* **27**: 9–13.

MacDonald IA and Webber J (1995) Feeding, fasting and starvation: Factors affecting fuel utilization. *Proceedings of the Nutrition Society* **54**: 267–274.

Pickart CM (2004) Back to the future with ubiquitin. *Cell* **23**: 181–190.

Randle PJ (1995) Metabolic fuel selection: General integration at the whole body level. *Proceedings of the Nutrition Society* **54**: 317–327.

Robinson AM and Williamson DH (1980) Physiological roles of ketones as substrates and signals in mammalian tissues. *Physiological Reviews* **60**: 143–187.

Shulman GI and Landau BR (1992) Pathways of glycogen repletion. *Physiological Reviews* **72**: 1019–1035.

Smith R and Williamson DH (1996) Biochemical background. In: Weatherall DJ, Ledingham JEG, and Warrell DA (eds.) *Oxford Textbook of Medicine*, vol. 1, pp. 1271–1278. Oxford: Oxford Medical.

Snell K (1980) Muscle alanine synthesis and hepatic gluconeogenesis. *Biochemical Society Transactions* **8**: 205–213.

STOMACH

Contents
Structure and Function
Disorders

Structure and Function

J P Pearson and I A Brownlee, University of Newcastle, Newcastle-upon-Tyne, UK

Introduction

The stomach is an organ of the upper digestive tract and its major function is to store and liquefy food to allow its further digestion and absorption by the small intestine. Liquefaction is achieved by the action of acid, pepsin, and strong rhythmic muscle contractions. The stomach is well suited to carry out its digestive functions: it is expandable to accommodate a meal; it is muscular to allow strong contractions to mix and break up the food; and the acid and pepsin needed for digestion are released on demand from specific gland structures.

Structure

The stomach is an expandable chamber between the esophagus and the duodenum. Its volume is only about 50 cm^3 when empty, yet it can expand to up

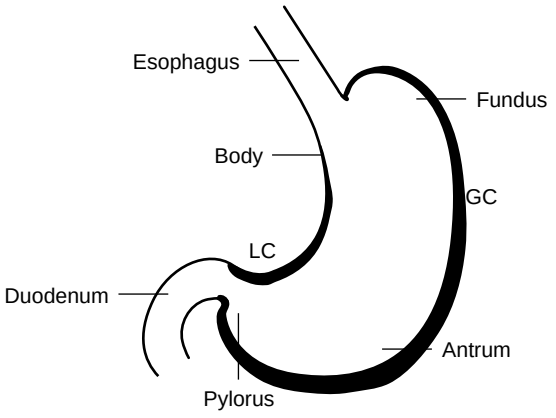


Figure 1 Stomach anatomy. The thickness of smooth muscle in the different regions of the stomach is shown by the darkened areas (■). LC, lesser curvature; GC, greater curvature.

to 1000 cm^3 when full with food. The stomach can be divided anatomically into four regions (**Figure 1**). The esophageal squamous epithelium ends at the gastroesophageal opening where the columnar epithelium of the stomach begins; the first region of the stomach is the fundus, which lies above the opening of the esophagus. The next region is the body or corpus, which together with the fundus forms a food storage reservoir. This allows control of the rate of food delivery to the small intestine, so as to synchronize it with maximal digestion and absorption. The distal part of the stomach, the antrum has a thicker muscle layer and its major function is to generate vigorous mixing of food with the gastric secretions to produce a slurry known as chyme. The stomach ends in the pylorus, a muscular sphincter that controls release of chyme into the duodenum.

Histology of the Stomach

The stomach wall consists of four types of tissue: connective, smooth muscle, neural, and epithelial (**Table 1**). Within these layers is a vascular system with arterioles supplying the external muscle layers and gastric arteries supplying a plexus of arterioles in the gastric submucosa, which in turn provide nutrients for the mucosal secretory cells.

The cells producing the gastric secretions are located in the mucosa of the stomach and are arranged in structures called gastric pits (**Figure 2**). Pits are made up of several gastric glands and are distributed throughout the mucosa. Glands in the body and fundus contain chief (peptic) cells deep within them, which synthesize, store, and secrete the inactive enzyme precursor pepsinogen. Parietal (oxyntic, from the Greek *oxys* meaning acid) cells are found more widely distributed within the gland extending towards the top. These cells secrete hydrochloric acid (HCl) and intrinsic factor into the gland lumen. Also present near the base of the glands in this region of the stomach are neuroendocrine cells, i.e., enterochromaffin cells secreting

Table 1 Layers of the stomach wall

Outside	Serosa	Connective tissue
	Muscularis	Longitudinal muscle
	arterioles distributed through this layer	Mysenteric autonomic nerve plexus
		Circular muscle
		Submucosal autonomic[a] nerve plexus
	Submucosa	Plexus of arterioles
	Mucosa	**Muscularis mucosa** – smooth muscle
		Gastric glands – endocrine cells, chief (peptic) cells, parietal cells and mucous neck cells
		Lamina propria – connective tissue and lymph nodes
		Epithelium – surface mucous cells
Inside (Lumen)		

[a]The submucosal plexus is an important structure in the human stomach, containing a number of transmitters, e.g., substance P, vasoactive intestinal peptide (VIP), nitric oxide, calcitonin gene-related peptide (CGRP), and gastrin-releasing peptide (GRP).

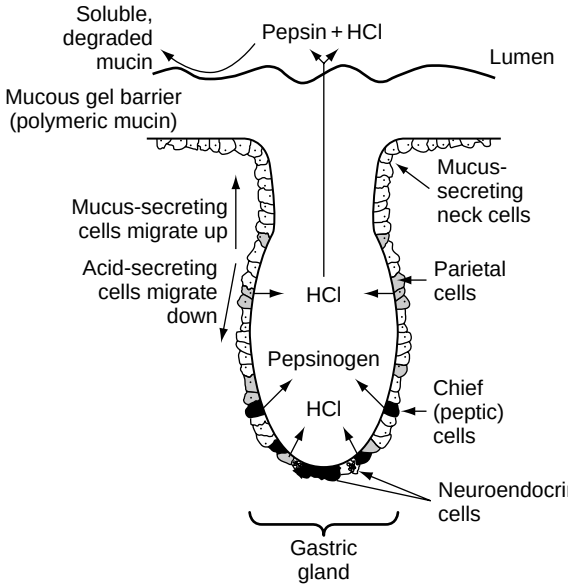


Figure 2 Structure of the body and fundus gastric mucosa. The secretory units of the oxyntic mucosa consist of apical pits with basal glands. The stem cells giving rise to all the cell types are located between the pit and the gland.

serotonin, enterochromaffin-like (ECL) cells secreting histamine, and D cells secreting somatostatin.

The antral glands lack parietal cells so no acid is secreted from this region. They do, however, contain peptic cells and therefore secrete pepsinogen. In addition, they contain endocrine G cells, which secrete the polypeptide hormone gastrin that together with histamine is involved in the control of acid secretion. G cells also contain intrinsic factor.

Throughout the gastric mucosa stem cells are continually dividing and migrating up the pits differentiating into mucus neck cells, which migrate out of the pits to replace shed surface epithelial cells. Consequently, the whole surface mucosa is replaced every 72–96 h. Stem cell progeny also migrate down the pits into the glands, differentiating into acid or pepsin secreting cells. The gastric glands do not turnover as rapidly as the surface mucosal cells with parietal cells having a life span of 150–200 days. The major secretion of the surface epithelial cells is mucus. Mucus forms a continuous surface layer and is the first line of defense against gastric juice; however, it does not extend up into the esophagus. The gastric mucus layer stops at the gastroesophageal junction, so the esophagus must have a different protection mechanism. Studies with human and animal gastric mucosa have demonstrated that the mucous layer in the stomach consists of two independent gel layers (**Figure 3**). The first is a firm adherent mucous layer (median thickness 150 μm) that protects the underlying epithelial cells against pepsin and acid; without this layer the mucosa would be destroyed. Bicarbonate secreted into this layer by the epithelial cells neutralizes the acid diffusing from the lumen in an unstirred aqueous environment. Consequently, a pH close to 7 is maintained at the epithelial cell surface. Overlying this firm gel layer is a sloppy (shear compliant) layer of variable thickness, which

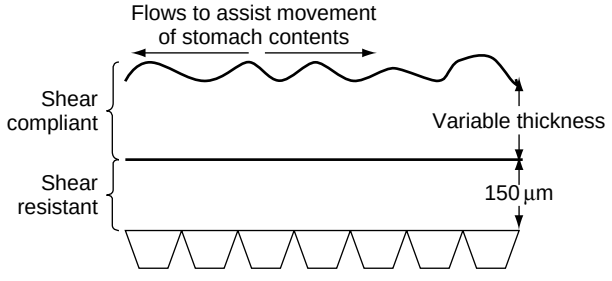


Figure 3 The gastric mucous layer. The shear resistant layer of the mucous bilayer remains at close to 150 μm, whereas the shear-compliant mucous layer thickness is always changing, due to its constant removal by shear stress and subsequent replenishment.

although a gel will flow easily when force is applied to it. This property is essential in reducing shear forces on the mucosa. When the stomach contains food the forces generated will convert the sloppy layer into a viscous liquid, the ideal lubricant reducing the shear forces between the food and the mucosal cells. The thickness of the bilayer is maintained by secretion of new mucus, balancing that lost by pepsin digestion and mechanical shear. Along with mucus the epithelial cells also secrete trefoil peptides, which have growth factor-like activities (see below).

Composition of Gastric Juice and Regulation of Production

Table 2 lists the components of gastric juice. The enzymes secreted by the gastric mucosa will be dealt with first.

Lipases

Lipases are a minor but significant enzyme secretion. Partial digestion of fat occurs in the adult stomach and preliminary digestion of triacylglycerols aids the action of pancreatic lipases. These enzymes are particularly important in infants with respect to breakdown of milk fat, 30–60% of which is lipolyzed in the gastric lumen. Gastric lipases are glycoproteins with molecular weights of about 45 000. Lipase secretion is highest in the body and fundus and very low in the antrum. Secretory granules containing lipases are located in peptic cells, particularly in the fundic mucosa. Gastric lipases have a broad pH optimum (2.5–7.0); they are stable down to pH 1.5 and therefore survive in the stomach's acidic environment and will be active during feeding when the gastric pH rises to around 5.0. Secretion of gastric lipase is coupled with pepsin secretion by peptic cells in response to pentagastrin (a functional analoge of gastrin). Lipase secretion from isolated human

Table 2 Composition of gastric juice

Mucus (salivary and gastric)
Lipases (E.C.3.1.1.3)
Pepsins 1–6 (E.C.3.4.23.1)
Urea
Intrinsic factor and haptocorrin
H_2O
Na^+; Mg^{2+}; H^+; K^+; Cl^-; HPO_4^{2-}
HCO_3^- [a]
Salivary amylase

[a]Only present when secreted into the mucous layer; once out in the lumen will react with H^+ to form H_2CO_3 which can decompose to $H_2O + CO_2$.

gastric glands is stimulated by cholecystokinin (CCK) and carbachol but histamine has no effect. In the dog secretin and prostaglandin E_2 stimulate lipase secretion. There is also a feedback loop in gastric lipase secretion: release of long-chain fatty acids from triacylglycerols by lipases stimulates CCK secretion, which in turn stimulates the secretion of pancreatic and gastric lipases.

Pepsins

Pepsins are the major enzyme secretions of the stomach, digesting protein in the diet to yield peptide fragments. Pepsins are members of the aspartate proteinase family with two aspartate residues at the active site. They are broad specificity endopeptidases with a preference for peptide bonds between hydrophobic amino acids. Human gastric juice contains two groups of pepsins: pepsin A, which contains six isoenzymes; and pepsin C, which contains two isoenzymes (**Table 3**). Once secreted into the lumen, the acidic conditions convert pepsinogens to pepsin. Unlike the activation of other proteinases, e.g., trypsinogen to trypsin via a proteolytic cleavage by another enzyme (enterokinase), pepsinogen activation is autocatalytic.

The process is understood in detail for porcine pepsinogen, which when activated produces an enzyme very similar to human pepsin 3b. Secreted pepsinogen has an N-terminal peptide of 44 amino acids blocking the active site. On exposure to pH values below 5, carboxyl groups become protonated abolishing charge–charge interactions. This allows part of the N-terminal protein into the active site where it is cleaved, releasing a 16-amino acid peptide. The enzyme is now partially activated; full activation occurs with cleavage of a further 28 amino acids from the N-terminal by either another partially activated pepsin or a fully activated enzyme.

The majority of proteinase activity in human gastric juice is due to pepsin 3 ($70.3 \pm 2.6\%$); variable amounts of pepsin 5 are also present ($16.9 \pm 2.0\%$). The only other isoenzyme found in significant amounts is pepsin 1. This differs from other pepsins in that it is an ionic complex of a 14 500 molecular weight protein and proteoglycan. Also, unlike the other pepsins, pepsin 1 is secreted only in the fundic and not the pyloric glands and is only present in significant amounts in stimulated juice. Pepsin 1 has the highest molecular weight (44 500), while the other pepsins have molecular weights around 35 000. The pH optimum against protein substrates for all pepsins is in the acidic range 1.9–3.6. Pepsin 1 secretion is elevated in peptic ulcer disease (PUD) and is therefore the pepsin associated with ulcers. Also associated with peptic ulcer disease is the bacteria *Helicobacter pylori* (HP), which colonizes the stomach at the epithelia surface under the mucous layer. *Helicobacter pylori* is believed to protect itself from the acid environment of the stomach by the action of a membrane-bound urease, which generates ammonia from the urea present in gastric juice. Interestingly, approximately 50% of all people in the UK over 50 years old are HP positive, yet only a small proportion of these will develop peptic ulcer disease. The increased secretion of pepsin 1 in PUD may be explained by HP infection.

Table 3 Pepsins of human gastric juice

Pepsin group	*Chromosome location*	*Individual pepsins*
A	11q13	1, 2, 3a, 3b, 3c, 4
C (gastricsin)	6	5 and 6

These isoenzymes can be separated by agar gel electrophoresis at pH 5.0 or anion exchange high-performance liquid chromatography.

Pepsinogen secretion The initial stimulus for pepsinogen comes from feeding. There is a basal level of secretion, which is 20% of the stimulated secretion. Maximal secretion produces gastric juice with a pepsin concentration approaching $1\,mg\,ml^{-1}$. There is a biphasic response with initial release of stored pepsinogen in a rapid phase (20–40 min) followed by a less rapid steady state of secretion. Pepsinogen secretion is stimulated by CCK, forskolin, and by insulin induced hypoglycemia mediated by the vagus nerve. Gastrin also stimulates pepsinogen secretion; however, it is much less effective than CCK. Secretion is also stimulated by the peptide hormones secretin and vasoactive intestinal peptide (VIP) and by adrenaline acting through β_2 receptors and by prostaglandins and histamine. All these agents alter intracellular cAMP levels, which in turn activates cAMP-dependent protein kinase (PKA) associated with A-kinase anchoring protein-150 (AKAP-150). Pepsinogen secretion is stimulated by cholinergic agents acting on muscarinic M3 subtype receptors; these agents also stimulate acid secretion. Acetylcholine and the CCK/gastrin family act through altering intracellular Ca^{2+} and Inositol triphosphate (IP_3) levels. The elevation of intracellular calcium is produced by release from intracellular stores and an influx of calcium via a membrane cation channel. *Helicobacter pylori* is believed to increase pepsinogen secretion via this Ca^{2+} signal transduction pathway. Recently, the G protein-coupled protease activated receptor-2 (activated by partial digestion)

on chief cells has been shown to stimulate pepsinogen secretion when activated. Pepsinogen is secreted by compound exocytosis with granules fusing together and with the plasma membrane. The enzyme is condensed within the vesicle in association with a divalent cation, e.g., Ca^{2+}. Initial fusing with the plasma membrane produces a small pore releasing the Ca^{2+}, causing the vesicle contents to swell producing a larger pore through which the pepsinogen is released.

Acid

Acid is secreted by the gastric gland parietal cells via the action of the gastric H^+/K^+ ATPase, a transmembrane proton pump. In nonstimulated parietal cells the pump is present in cytoplasmic vesicles (the tubulovesicles) separated from the apical membrane. Electron microscopy studies have shown that these vesicles are in fact small stacks of cisternae and should be called tubulocisternae. On stimulation the pumps are transported in vesicles to the apical membrane along the actin cytoskeleton where they fuse to and greatly increase the surface area of the apical membrane forming numerous microvilli. When physiological stimuli for acid secretion is removed the pumps are recycled back to the tubulocisternae. These trafficking processes involve: (1) movement to the apical membrane of vesicles containing the H^+/K^+ATPase; (2) fusion with the apical membrane and formation of actin filament-based scaffolds to form surface microvilli; (3) dissolution of the actin scaffolds; and (4) transport of the endocytotic vesicles containing the H^+/K^+ATPase back to the tubulocisternae. Numerous proteins have been reported to be associated with these processes; rab11 (a GTPase), syntaxin 3, VAMPs (vesicle associated membrane protein), c-src (a nonreceptor tyrosine kinase), clathrin, dynamin, SCAMPs (secretory carrier membrane proteins), lasp-1, actin, ezrin, coronin, myosin Vb, and myosin light chain kinases. For activity the pump needs to be associated with K^+ and Cl^- conductive pathways. There is a huge H^+ concentration gradient across the parietal cells (the lumen has $2–4 \times 10^6$ greater H^+ concentration than the blood); consequently, the cells require a great deal of energy and, as a result, mitochondria make up 34% of cell volume. The processes of HCl secretion are shown in **Figure 4**.

The H^+/K^+ pump is a noncovalent dimer of an α (catalytic subunit) and a β glycosylated subunit. The β subunit targets the pump to the apical membrane and protects the catalytic subunit from degradation. Acid secretion requires both subunits. The α subunit (mol. wt 100 000) consists of 10 membrane-spanning

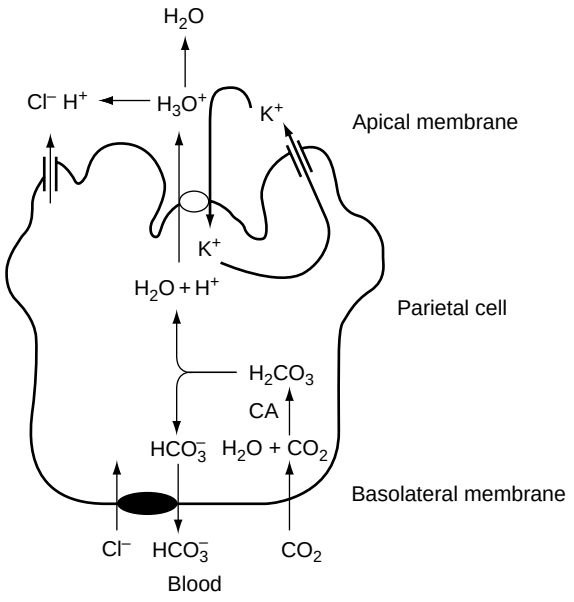


Figure 4 Ion movements in HCL secretion. Thick arrows at top show apical recycling of K^+. ‖, ion channel; ○, anion exchanger; ●, cation exchanger; CA, carbonic anhydrase. Na^+ may be transported instead of H_3O^+. HCO_3^- transport back into the blood during acid secretion is the so-called 'alkaline tide.' Cl^- entry via the basolateral membrane may be linked to Na^+ entry. Cl^- exiting across the apical membrane into the lumen may be linked to K^+ efflux.

segments with the intracellular loop between membrane spans 4 and 5 forming the ATP binding and phosphorylation sites. Hydrophilic amino acids in the membrane-spanning portions form the ion pathway. Proton pump inhibitors (PPI) are used clinically to treat acid-related diseases, e.g., gastroesophageal reflux disease (GORD) and peptic ulcer disease. PPIs are protonated in the stomach producing the active drug sulfenamide, which inactivates the H^+/K^+ATPase by binding to cysteine residues close to or in the extracellular loops between membrane segments 3 and 4, 5 and 6, and 7 and 8. The key cysteine for inhibition by omeprazole is in membrane segment 6 close to the extracellular loop between segments 5 and 6. Binding of sulfenamide to the cell surface prevents the movements of membrane domains relative to each other necessary to pump H^+. The membrane organization of the pump is shown in **Figure 5**.

Control of acid secretion Four cells are key in the control of acid secretion: the parietal, the enterochromaffin-like (ECL), the G, and the D cells. The three major stimulatory compounds are gastrin, histamine, and acetylcholine and the major inhibitory compound is somatostatin. The interaction between the four cell types and the four controlling

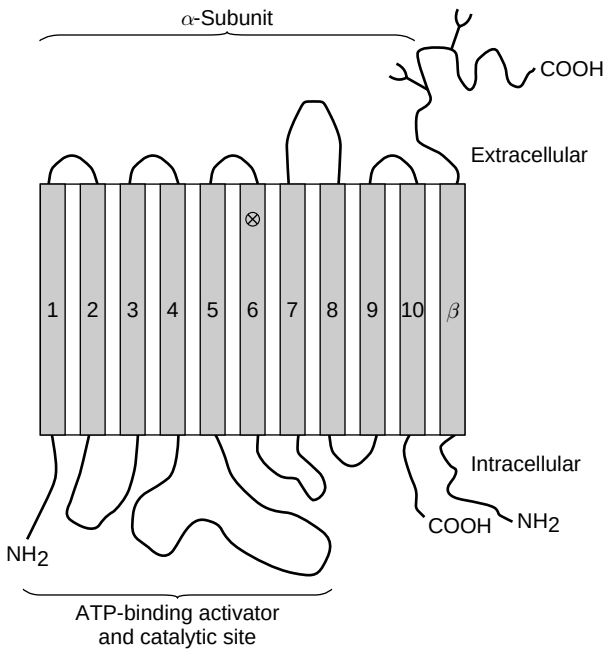


Figure 5 Organization of the H^+/K^+ ATPase in the parietal cell apical membrane. The α-subunit contains 10 transmembrane spans and the β subunit one. In addition the α-subunit has 4 intracellular loops. There is a large mass of protein in the loop between transmembrane spans 4 and 5, which contains the ATP binding and phosphorylation sites. The Activator domain is important for the conformational transitions and may work as an anchor for the ATP-binding domain. NH_2, N-terminal; COOH, C-terminal; $\otimes$, site of Cys 822, the key residue for inhibition by the PPI omeprazole.

agents and other factors are shown in **Figure 6**. Acid secretion can be divided into four phases: basal, cephalic, gastric, and intestinal. Basal phase secretion makes up only 10% of the total secretion produced by all four phases, while cephalic phase secretion accounts for 45% of the total. In response to smell, taste, sight, chewing, and swallowing, the fundic and oxyntic mucosa are stimulated by the vagus (parasympathetic) nerve. Acid is secreted by the parietal cells following direct stimulation, with acetylcholine binding to M3 muscarinic receptors. Acetylcholine also stimulates histamine release from the ECL cells, which binds to H_2 receptors on the parietal cells causing an increase in cAMP, which in turn stimulates acid secretion. In addition pituitary adenylate cyclase-activating polypeptide (PACAP) is released from the mucosal nerves, stimulating acid secretion by increasing histamine release from ECL cells. The vagus also stimulates G cells of the antrum to release gastrin into the blood, which further stimulates the parietal cells via binding to CCK-2 receptors and increasing intracellular Ca^{2+} leading to an increase in HCl secretion. During the cephalic phase gastric leptin (a hormone thought to function as a satiety signal when released from adipocytes) is also released. Gastric leptin inhibits acid secretion and is believed to act on the central nervous system presumably to suppress further food intake.

Acid secretion during the gastric phase accounts for 45% of total acid secretion. Food entering the stomach causes distension and gastrin is released from the antrum via two mechanisms: (1) local enteric and long loop parasympathetic mediated reflexes; (2) stimulation by the products of protein digestion, i.e., peptides and amino acids, of the G cells to release gastrin directly, particularly hydrophobic amino acids. This second effect is interesting considering the fact that the G cells must be covered with a mucous gel layer and the amino acids need to diffuse through it before eliciting a response. The gastrin produced by both mechanisms stimulates the parietal cells to produce HCl directly and via histamine release from ECL cells. The intestinal phase is mainly an inhibitory phase; however, there is a small stimulatory phase via amino acids and peptides in the jejunum promoting release of gastrin from the intestine. Two groups of factors lead to the inhibitory phase. First, fat, acid, hypertonicity, and distension of the duodenum cause release of secretin from duodenal S cells, which inhibits gastrin release and thereby acid secretion. This works partly through the release of somatostatin from D cells in the stomach. In addition, a smaller effect on acid secretion results from gastric inhibitory peptide, released from intestinal K cells, VIP released from nerve endings, glucagon-like peptides (GLP-1 and GLP-2) from the enteroendocrine L cells in the small intestine, and CCK released by intestinal cells, again probably via a somatostatin-mediated pathway. Second, stimulation of G cells is reduced as food leaves the stomach and pH will fall as buffering from food is lost. This high H^+ concentration will stimulate the D cells to release somatostatin further inhibiting gastrin release. As well as inhibition of gastrin secretion, gastrin is also destroyed by a neutral endopeptidase present in stomach cells. With removal of stomach distension, the vagal and intrinsic nerve stimuli for acid secretion is lost.

Regulation of Appetite

The stomach plays an important role in the regulation of appetite, in particular via the peptide hormone ghrelin (the so called hunger hormone). This hormone is produced by the neuroendocrine cells of the fundus and is upregulated in response to a fasting state, for example ghrelin levels are highest just

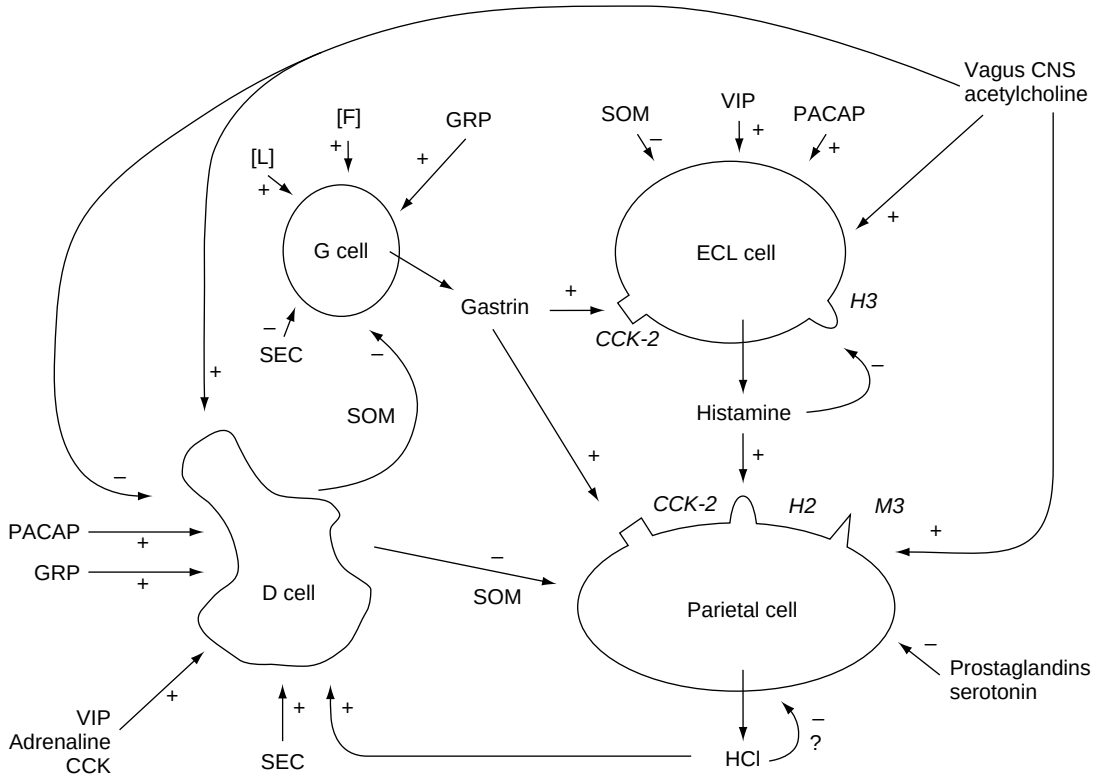


Figure 6 Control of gastric acid secretion. +Indicates a stimulatory effect on the cell; −Indicates an inhibitory effect on the cell. [F], food and aromatic amino acids; [L], local and long reflex groups; SEC, secretin; SOM, somatostatin; GRP, gastrin-releasing peptide from parasympathetic nerve fibers; VIP, vasoactive intestinal peptide; CCK, cholecystokinin; ?, suggested effect; PACAP, pituitary adenylate cyclase-activating polypeptide (a member of the secretin-glucagon-VIP family, which is present in gastric mucosal neurones and involved in vagally mediated acid secretion).

before a meal and lowest just after. Ghrelin is synthesized as a preprohormone which is cleaved by proteolysis to a 28 amino acid peptide (a 27 amino acid variant also exists). The serine residue at position 3 must be substituted with octanoic acid for activity. Ghrelin released into the circulation has several actions, stimulation of the release of growth hormone from the anterior pituitary, increasing the hunger signal by binding to receptors on the hypothalmic feeding centers and promoting gastric emptying to clear the stomach for the approaching meal. The effects of ghrelin are counter balanced by the hormones released in response to the feed state, somatostatin from the stomach D cells and GLP-1 and peptide YY from the L cells of the samll intestine.

Mucus

The mucus covering the surface of the stomach consists of a mixture of many secretions and exfoliated cells. The major viscous and gel-forming component of the mucus gel is mucin, present at approximately $40\,\mathrm{mg\,ml^{-1}}$ in the firm layer and approximately $15\,\mathrm{mg\,ml^{-1}}$ in the sloppy layer. Mucin is also present in gastric juice resulting from pepsin erosion of the surface of the gel. It is secreted by the surface epithelial cells and mucous neck cells and is a glycoprotein, about 80% carbohydrate. The core protein of mucin has large regions of amino acid tandem repeats, variable in number (VNTR), which are different for different mucins. The VNTR regions are the sites of heavy glycosylation. The N- and C-terminal regions of the protein core are globular and rich in cysteine and contain domains like those in von-Willebrand factor, a glycoprotein involved in blood clotting, which polymerizes via disulfide bridges. Mucin units are therefore polymerized via disulfide bridges into polymers with molecular weights of about 10×10^6 (**Figure 7**).

Polymerization is essential for gel formation, which results from noncovalent interactions between polymers. Also present in the gel and the gastric juice, secreted by mucin secretory cells are IgA, which combats bacterial invasion and trefoil peptides (mol. wt. 5000–10 000), which may interact with mucin to stabilize and strengthen the

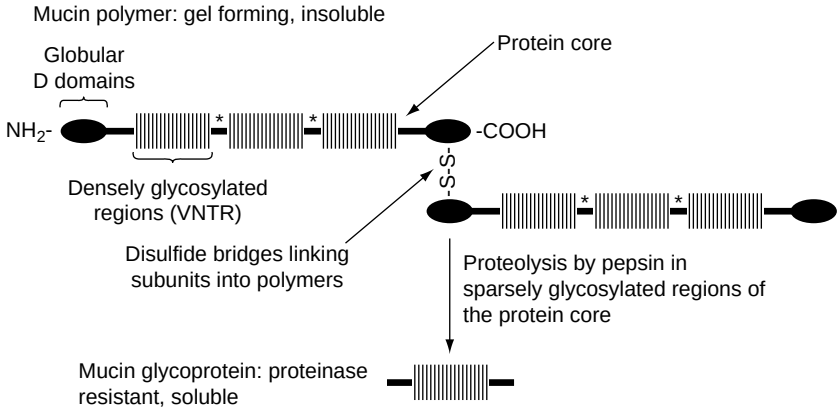


Figure 7 Model for the structure of gastric mucin. Mucin genes expressed in the stomach are MUC5AC and MUC6, both of which are encoded for on chromosome 11 on the p arm at position 15.5. Both MUC5AC and MUC6 have cysteine-rich globular domains on the C- and N-terminals of the mucin units, thus allowing end-to-end polymerization via disulfide bridges. The mucin D domains are homologous to the D domains present in the clotting factor von Willebrand factor, suggesting a common ancestor. MUC5AC has four D domains on its N-terminal and one D domain and a cysteine-rich structure called a cysteine knot on its C-terminal. MUC6 has the same N-terminal structure but only has the cysteine knot on its C-terminal. *Regions susceptible to proteolysis.

gel. IgA inhibit acid secretion and gastric motility, however their major function is as a growth factors, a promoters of gastric healing, and suppressors of tumor growth.

Mucin secretion Mucin secretion is by compound exocytosis and may be linked to Cl^- secretion, with the mucin polyanion stored inside granules condensed with Ca^{2+}-mediated charge shielding. The exocytosis mechanism is similar to that for pepsinogen. There are two routes: (1) constitutive, with a steady release from the cells; and (2) regulated, with release from storage granules. Muscarinic receptors mediate mucin secretion. Cholinergic agonists stimulate secretion via protein kinase C and IP_3 leading to an increase in intracellular Ca^{2+}. β-adrenergic agents, secretin and prostaglandins E_2 and $F_2\beta$ stimulate secretion via a cAMP-mediated mechanism and NO acts via cGMP. Inflammation and infection, mediated by the cytokines TNFα, Il-6, and Il-8 and bacterial cell wall components (e.g., lipopolysaccharides), as well as epithelial damaging agents (e.g., free radicals and mustard oil) stimulate mucin secretion.

Intrinsic Factor

Gastric intrinsic factor (IF) can be defined as a substance required for the absorption of vitamin B_{12} (cyanocobalamin, CNCb1), which is essential for the formation of red blood cells. Cyanocobalamin is a precursor for methylcobalamin, a cofactor for the enzyme methionine synthase, and 5′ deoxyadenosylcobolamin, a cofactor for methylmanonyl CoA mutase. In the absence of IF, CNCbl fails to be absorbed, erythrocyte production is defective, and pernicious anemia results. The most detrimental effect of gastric mucosal atrophy, which can result from long-term *H. pylori* infection, is the loss of intrinsic factor. Two other CNCbl binders exist in the body that are distinct from IF; these are haptocorrin (Hc), the ubiquitous binder, and transcobalamin II (TCII), the plasma binder that transports CNCbl from the terminal ileum.

Dietary CNCbl is always bound to proteins and is released by cooking and pepsin digestion. In gastric juice, free CNCbl is faced with both gastric IF and Hc mainly derived from saliva. At pH 2.0 CNCbl affinity for Hc is 50 times higher than for IF; therefore, free CNCbl binds to Hc in the stomach. However, in the small intestine trypsin degrades He and CNCbl is bound by IF. The CNCbl–IF complex is absorbed in the terminal ileum after interaction with a receptor (IFRC) specific to IF-CNCbl.

The gene for IF is found at chromosomal location 11q13. Purified IF is a glycoprotein of approximately 57 kD. Its protein component varies between 341 and 351 residues and the carbohydrate moiety, consisting of 30–37 residues, 49–68% hexoses, 27–37% hexosamines and 13–18% sialic acid, constitutes 9.2–15% of the molecule (6.1–6.6 kD). The sugar chains are either *O*-glycosidically or *N*-lactosaminically linked. *In vitro* transcription/translation studies have shown that removing 12% of the C-terminus of the molecule results in all CNCbl binding activity being lost; the receptor binding region is at residues 25–44.

IF is classically described as being secreted by stomach fundus and body parietal cells; however,

the use of immunocytochemistry, electron microscopy, and *in situ* hybridization has shown that it is also found in a subset of chief cells, enteroendocrine cells, G cells, and in secretory ducts of the salivary glands. This may explain the fact that control of IF secretion is apparently multifactorial in common with other gastric juice components and can be activated by gastrin, histamine, acetylcholine, and cholecystokinin. IF secretion in man varies between 50 and 100 nmol l^{-1} which far exceeds the daily requirement.

Absorption

Alcohol, drugs, and some fatty acids are absorbed in the stomach, whereas the products of carbohydrate and protein digestion are not. Partially digested carbohydrate and protein products are not lipid soluble and are too large to cross cell membranes. Unlike in the intestine, no specialized nutrient transporters are present in the stomach mucosa. The continuous mucous gel layer also provides a physical barrier to diffusion of anything other than low molecular weight solutes (mol. wt ~1300). There are also no specialized transport systems for ingested nonlipid soluble electrolytes such as Ca^{2+}.

Gastric lipase activity is largely overlooked. Initial digestion of dietary fat in the stomach is a prerequisite for efficient intestinal lipolysis. In infants gastric lipolysis of milk is extensive and the medium-chain fatty acids released in the stomach are absorbed through the gastric mucosa.

Ethanol and a number of drugs readily pass across the gastric mucosa. Ethanol is partially lipid soluble and can therefore diffuse through the epithelial cell membranes and into the submucosal capillaries.

Nonsteroidal anti-inflammatory drugs, e.g., acetylsalicylic acid (aspirin), oral anticoagulants (e.g., dicoumarol) and sulfonylurea oral antidiabetic agents, can all be absorbed through the gastric mucosa. These weakly acidic drugs are nonionized at gastric pH. In this form they are lipid soluble and can therefore be absorbed quickly by crossing the plasma membrane of the epithelial cells.

Alcohol and nonsteroidal anti-inflammatory drugs can cause gastric mucosal damage and have been used in models of gastroduodenal ulceration. They rapidly diffuse through the protective adherent mucous gel layer (in the case of ethanol causing its dehydration) and in these models cause epithelial damage, cell exfoliation, and, in more severe cases, vascular damage, hemorrhage, and visible lesions. Following acute damage rapid repair occurs by a process of re-epithelialization. The repairing epithelium is protected from the endogenous damaging agents of acid and pepsin by a thick gelatinous coat mainly composed of a fibrin gel formed on the mucous gel template.

Gastric Motility and Emptying: The Role of the Stomach as a Reservoir and a Churn

Motility of the stomach allows it to serve as a reservoir, act as a churn to fragment food and mix it with gastric secretions to aid digestion, and to empty gastric contents into the duodenum at a controlled rate. Gastric motility and storage are complex and subject to multiple regulatory mechanisms. The first aspect of gastric motility is gastric filling. Accommodation of large changes in volume when a meal is eaten is achieved by the plasticity of the stomach smooth musculature and by receptive relaxation. Plasticity means that smooth muscle can maintain constant tension over a wide range of lengths without changing tension. Mechanoreceptors in the proximal stomach signal the degree of distension and, beyond a certain level, a stretch-activated contraction is initiated and pacesetter cells are depolarized. These properties are augmented by receptive relaxation of the deep folds of the stomach (known as rugae), which is mediated by the vagus nerve and associated with eating, possibly via stimulation of the taste buds.

A group of pacesetter cells located high on the greater curvature of the stomach generates slow wave potentials, which sweep down the length of the stomach 3 times per minute. These spontaneous depolarizations are known as the basic electrical rhythm (BER) of the stomach. The stomach's circular smooth muscle layer may be stimulated to contract in peristaltic waves synchronized with the BER. The contractions in the thinly muscled fundus and body are weak but become stronger in the thickly muscled antrum. Food emptied into the stomach from the esophagus is therefore stored in the body of the stomach and gradually fed into the antrum where mixing takes place. The antrum can contain 30 ml of chyme but only a few milliliters of chyme are forced through the pyloric sphincter into the duodenum with each peristaltic wave. Each wave causes the sphincter to contract more forcefully, blocking passage into the duodenum. This process is called retropulsion and achieves thorough mixing of chyme in the antrum. These events are known as the gastric phase of digestion, which is initiated as soon as food enters the stomach. The dominant hormone of the gastric phase is gastrin (**Table 4**).

Table 4 Gastrointestinal peptide hormones affecting gastric motility and emptying

Hormone	*Amino acid no. (chromosomal location)*	*Cell source*	*Stimuli*	*Effects on stomach*
Gastrin	34(17q21) 17 14	G cells	Amino acids Distension pH > 3.0 Parasympathetic activity	↑ Antral motility
Gastric inhibitory peptide[a]	42(17q21.3–22) 30	K cells of intestinal mucosa	Glucose Distension Hypertonicity	↓ Gastric motility and secretion
Motilin	22(6p21)	Enterochromaffin cells of small intestine	pH > 5.0	↑ Gastric motility
Secretin	27(11p15.5)	Duodenal S cells	Hypertonicity pH < 4.5	↓ Gastric motility and secretion
Cholecystokinin	58(3p22–21.3) 39 12 8	Upper small intestinal mucosa	Fats Amino acids	↓ Gastric motility and secretion

[a]Also known as glucose-dependent insulinotropic polypeptide.

When gastrin levels are high the fundus and body of the stomach are relaxed and serve mainly to store chyme. Gastrin stimulates pyloric contraction and increases cardiac sphincter tone, preventing reflux. Gastrin also stimulates acid and pepsinogen secretion and antral motility, thus facilitating gastric digestion. These antral peristaltic contractions also provide the driving force for gastric emptying, which is regulated by multiple gastric and duodenal factors.

The main gastric factor influencing gastric emptying is the amount of chyme in the stomach. Simplistically, emptying rate is proportional to the volume of chyme. Distension triggers an increase in motility via stretching of smooth muscle as well as involvement of the intrinsic plexuses, the vagus nerve, and gastrin. The rate of emptying also depends on the degree of liquefaction of the contents.

Duodenal factors are, however, of prime importance in controlling gastric emptying rate. As soon as food begins to enter the duodenum the intestinal phase of digestion begins. Distension and the presence of acid, hypertonicity, and fat in the duodenum stimulates receptors triggering neural or hormonal factors to suppress gastric motility and emptying by reducing gastric smooth muscle excitability. Neural responses are mediated by intrinsic nerve plexuses (short reflex) and autonomic nerves (long reflex), collectively known as the enterogastric reflex. The hormonal respone is mediated by several hormones released from duodenal mucosa known as enterogastrones, including secretin, CCK, and GIP.

The volume of chyme in the duodenum is detected by mechanoreceptors and results in reflex inhibition of gastric motility and increase in pyloric tone by the vagally mediated enterogastric reflex.

The pH of duodenal contents affects gastric emptying; pH less than 4.5 inhibits further emptying by stimulating the release of secretin. This inhibits gastric motility and gastrin release while stimulating pancreatic bicarbonate secretion. Conversely, when the duodenal chyme is above pH 5.0, motilin is released, which increases the strength of the gastric contractions and the tone of the pyloric sphincter (**Table 4**).

Hypertonicity can result because digestion releases large amounts of amino acid and glucose molecules. If absorption does not keep pace, then the osmolarity of the duodenal contents increases. This results in large volumes of water entering the intestine from plasma causing circulatory disturbances and resulting in reflex inhibition of gastric emptying. Duodenal osmoreceptors trigger the release of GIP, which decreases gastric motility and secretion of both pepsin and acid.

The ingestion of fat is the most potent stimulus for inhibition of gastric motility, fat being digested more slowly than carbohydrates and proteins. Fats and fatty acids are detected by duodenal chemoreceptors stimulating the release of cholecystokinin. This hormone has multiple effects including inhibition of gastric emptying.

After a meal is completely emptied from the stomach there are no more gastric factors to enhance gastric excitability.

See also: **Alcohol**: Absorption, Metabolism and Physiological Effects; Effects of Consumption on Diet and Nutritional Status. **Cobalamins**. **Colon**: Structure and Function; Disorders. **Stomach**: Disorders.

Further Reading

Allen A, Hutton DA, Leonard AJ, Pearson JP, and Sellars LA (1989) Pepsins. In: Wallace JL (ed.) *Endogenous Mediators of Gastroinintestinal Disease*, pp. 54–69. Boca Raton, FL: CRC Press.

Allen A, Hutton DA, and Pearson JP (1998) The MUC2 gene product: a human intestinal mucin. *International Journal of Biochemistry and Cell Biology* **30**: 797–801.

Forstner G (1995) Signal transduction, packaging and secretion of mucins. *Annual Review of Physiology* **57**: 585–605.

Horn J (2000) The proton pump inhibitors: similarities and differences. *Clinical Therapeutics* **22**: 266–280.

Kraulter B, Arigoni D, and Golding BT (1998) *Vitamin B12 and B12 Proteins*. Weinheim: Pruls-Wiley-qVch.

Okamdo CT and Forte JG (2001) Vesicular trafficking machinery, the actin cytoskeleton and H^+-K^+ ATPase recycling in the gastric parietal cell. *Journal of Physiology* **532**: 287–296.

Quigley EMM (1996) Gastric and small intestinal motility in health and disease. *Gastroenterology Clinics of North America* **25**(1): 113–145.

Sands BE and Padolsky DK (1996) The trefoil peptide family. *Annual Review of Physiology* **58**: 253–273.

Schubert ML (2002) Gastric secretion. *Current Opinion in Gastroenterology* **18**: 639–649.

Silverthorn DU (2004) *Human Physiology an Intergrated Approach* 3rd ed., pp. 659–694. San Francisco: Pearson Benjamin Cummings.

Sweadner JK and Donnet C (2001) Structural similarities of Na-K-ATPase and SERCA, the Ca^{++}-ATPase of the sarcoplasmic reticulum. *Biochemical Journal* **356**: 685–704.

Yip RG and Wolfe MM (2000) GIP biology and fat metabolism. *Life Sci* **66**: 91–103.

Disorders

J A Solon, MRC Laboratories Gambia, Banjul, The Gambia

Introduction

This chapter will briefly discuss major stomach disorders such as stomach cancer, peptic ulcer disease, *Helicobacter pylori* infection, motility disorders and hypersecretory states as well as the less frequent congenital anomalies that may significantly affect early child growth and development. Epidemiological data are presented and current diagnostic and treatment modalities are discussed.

Stomach Cancer

Gastric Adenocarcinoma

The most common (85%) stomach cancers are adenocarcinomas; the rest are non-Hodgkin's lymphomas and sarcomas. The incidence of gastric adenocarcinoma has declined with time. In the early part of the twentieth century it was the leading cause of cancer-related deaths in men in the US. Although the incidence remains relatively high in Japan, China, Chile, and Ireland, a decrease in incidence and mortality has been seen in these countries as well. The risk of gastric adenocarcinoma is greater in the lower socio-economic classes. Migrants from high-incidence countries to low-incidence countries maintain the same risk of gastric adenocarcinoma but their offspring tend to have similar risks to those who are originally from the host nation. This suggests that early environmental exposures have an effect on the development of gastric cancer.

Gastric adenocarcinomas can affect different parts of the stomach. In the US, 37% of tumors arise from the proximal third, 20% arise from the mid-portion, 30% arise from the distal third, and 13% involve the entire stomach. They can be divided into two epidemiologically relevant histopathological classifications: the diffuse type and the intestinal type (**Table 1**). In the diffuse type, the cancer cells infiltrate the stomach without forming a discrete mass. In the intestinal type, the cancer cells form tubular gland-like structures. The differences in the epidemiology of these two types suggest that the etiologic factors involved may be different.

The large differences in incidence between populations and the decline over the last half-century suggest that environmental exposures play a major part in the etiology of gastric cancer. One hypothetical model for gastric carcinogenesis suggests that gastric cancer is the end result of a series of mutations and transformations that begin in the first decade of life. It also suggests that gastric hypochlorhydria (low acid output) predisposes to gastric carcinogenesis. Chronic atrophic gastritis, intestinal metaplasia, and pernicious anemia have

Table 1 Comparison of diffuse and intestinal types of gastric adenocarcinoma

Diffuse	*Intestinal*
Cells infiltrate the stomach, no discrete mass	Cells form tubular gland-like structures
Occurs throughout the stomach	Common in antrum and lesser curvature
Linitis plastica or 'leather bottle' appearance	Ulcerative
Poorer prognosis; affects younger age groups	Predominant in high-risk geographic regions

been associated with an increased risk of gastric cancer.

Dietary factors have been studied extensively, but results from various studies are inconsistent. There is an inverse relationship between the intake of vegetables and fruit and the risk of gastric cancer. Other food items such as processed meat and fish, milk, and salt have given inconsistent results. Vitamin C has been seen to be protective in observational studies in China, Italy, and Sweden, while results on tocopherol are contradictory. Vitamin A does not seem to have any effect on risk of cancer of the stomach; however, most reports indicate that beta-carotene, a vitamin A precursor, appears to be protective. Studies on protein, carbohydrate, fat, dietary fiber, and calcium have conflicting results. At the population level, dietary nitrate exposure has been correlated with risk of gastric cancer. Case–control studies on nitrate and nitrite consumption do not support their proposed role in carcinogenesis, probably because of the difficulty in measuring their levels in food constituents. It has been suggested that he decline in incidence of stomach cancer may be due to a reduction in salt intake and a departure from traditional food preservation practices such as salting and pickling with the advent of refrigeration.

Helicobacter pylori are Gram-negative, curved bacteria that live in the mucus layer of the stomach (see below). These bacteria cause a chronic infection, which results in long-standing inflammation. Infection with *H. pylori* is prevalent in populations with a high risk of gastric carcinoma. A recent study showed that the incidence and mortality rates of gastric cancer were associated with the prevalence of *H. pylori* seropositivity in 17 populations. However, there are populations (e.g., in Africa) where *H. pylori* is prevalent at young ages but gastric cancer is uncommon. Therefore, the role of *H. pylori* in the pathogenesis of cancer needs further clarification.

Clinical Features

Gastric adenocarcinoma can be asymptomatic for some time. As the tumor grows, a feeling of fullness or abdominal discomfort may be felt. Hemorrhage may occur in some cases. Difficulty in swallowing can occur in cancers near the esophagus. Anorexia, weight loss, nausea, and vomiting can also be seen. Spread of gastric cancer can occur by direct extension to surrounding tissues or through the lymphatic system to intra-abdominal or supraclavicular lymph nodes.

Diagnosis, Staging, and Prognosis

Double-contrast radiography is the simplest means of diagnosing gastric adenocarcinomas. However, it is difficult to distinguish mucosal lesions with radiographs. Endoscopy offers a greater advantage because of direct viewing plus the ability to obtain biopsy and cytological samples.

Cancers are staged according to the extent of tissue and lymph node involvement. The prognosis for adenocarcinomas that are limited to the mucosa can be good. The 5-year survival rate for cancer limited to the mucosa is 90%. Where there is invasion of the submucosa, the 5-year survival rates can drop to 60% and with nodal involvement, down to 30%. Thus, early diagnosis is key to reducing mortality rates and screening programs can be useful in countries with a high incidence of stomach cancer.

Treatment

Surgery is the best treatment option for the cure of adenocarcinomas. This involves removal of the tumor and any adjacent lymph nodes. However, this is only possible for one-third of patients. Usually, a subtotal gastrectomy is done for patients with distal stomach cancer and a total or near-total gastrectomy for more proximal tumors. In the US, the 5-year survival rate for those undergoing a complete resection of a gastric cancer is 20% for distal tumors and 10% for proximal tumors. Radiotherapy has very limited value but it can be used as a palliative measure.

Primary Gastric Lymphomas

Gastric lymphomas include those that arise primarily from the stomach (primary gastric lymphoma) and those that arise from systemic lymphomas with a gastric component. Systemic lymphomas commonly involve the gastrointestinal tract. The most common site is the stomach, followed by the small intestine, ileocecal area. and the colon.

Primary gastric lymphomas account for less than 15% of gastric cancers. They usually occur in the 6th decade of life and, as with gastric adenocarcinomas, there is a male predominance. They are difficult to distinguish clinically from gastric adenocarcinomas and definitive diagnosis can only be made through biopsy at the time of endoscopy or abdominal surgery. Histologically, these tumors can be well-differentiated superficial processes (mucosa-associated lymphoid tissue or MALT) or high-grade large cell lymphomas. *Helicobacter pylori* is particularly associated with the development of MALT lymphomas. For lymphomas without nodal

involvement, the 5-year survival rate is 80% and with nodal involvement this drops to 50%.

For primary gastric lymphomas, surgery is considered as the treatment of choice where feasible. Radiotherapy is often advocated as an adjunct to surgery for those with advanced disease, and combination chemotherapy and radiotherapy is sometimes used in those with unresectable disease. Eradication of *H. pylori* infection results in regression of low-grade MALT lymphomas. Thus, in these cases, a trial of *Helicobacter* therapy is warranted.

Gastric (Nonlymphoid) Sarcoma

Sarcomas are the least common gastric cancer. Of these, the most common are the leiomyosarcomas, which constitute between 1 and 3% of all gastric malignancies. These lesions often ulcerate and bleed, but rarely invade adjacent organs or lymph nodes. These tumors are slow growing and found deep within the stroma. The treatment is usually surgical resection with combined chemotherapy for patients with metastatic disease.

Peptic Ulcer Disease

Peptic ulcer disease refers to the mucosal lesion in the stomach (gastric ulcer, GU) or the duodenum (duodenal ulcer, DU) where acid and the enzyme pepsin contribute to tissue damage that extends into the submucosa. It is thought to occur when there is an imbalance between the protective factors and the aggressive factors in the gastric milieu.

The stomach relies on the different layers of mucosal defense (e.g., mucus/bicarbonate layer, cell membrane) to protect itself. If this is breached and epithelial cell injury occurs, then repair may occur by cellular restitution, replication, or formation of granulation tissue. A primary defect in these defense and repair mechanisms rarely causes ulcers. The hypothesis underlying the pathogenesis of peptic ulcer disease is that acid and pepsin are able to cause tissue damage if the normal defenses and repair mechanisms of the stomach are altered by *H. pylori* infection or the use of non-steroidal anti-inflammatory drugs (NSAIDs).

Helicobacter pylori-induced Ulcers

The majority of patients with peptic ulcers are infected with *H. pylori* (95–100% for DU; 75–85% for GU). Although only a small proportion of all *H. pylori*-positive individuals are found to have peptic ulcer disease (1–6%), this is a four- to tenfold increase of the number who are *H. pylori*-negative. The causal relationship between *H. pylori* infection and ulcers is further supported by the reduction of ulcer recurrence after *H. pylori* eradication. It is generally accepted, however, that other factors contribute to the pathogenesis of ulcers. It is known that smoking reduces healing and is also associated with peptic ulcer disease. Variations in bacterial strain virulence and host immune response may also be determinants of pathogenicity.

Infection with *H. pylori* results in a chronic, active gastritis in the antrum or the entire stomach. Peptic ulcers have long been associated with a diffuse antral gastritis. Gastritis is not a predominant feature in other forms of peptic ulcer (e.g., NSAID-induced) unless *H. pylori* is also present. The gastritis is present in both gastric and duodenal ulcers, but it is more severe in the former where it is also associated with gastric atrophy and intestinal cell metaplasia.

NSAID-induced Ulcers

NSAID-induced ulcers are more commonly gastric ulcers. The prevalence of ulcers is more than 15% among chronic NSAID users and less than 4% among those using NSAIDs for less than 1 year. Gastritis or *H. pylori* infection is not a prerequisite for NSAID-induced ulcers. The damaging effect of NSAIDs is thought to be due to their effect on prostaglandin synthesis. Endogenous prostaglandin aid in maintaining gastric mucosal blood flow and epithelial integrity and promote epithelial regeneration. NSAIDs reduce prostaglandin synthesis thus reducing the effect of prostaglandin in mucosal defense.

Gastric and Duodenal Ulcers

Duodenal ulcers are two to three times more common than gastric ulcers. Gastric ulcers are most frequent among those aged 40–70 years, whereas duodenal ulcers are most commonly seen between the age of 25 and 55 years. Thus, complications in gastric ulcers tend to be more severe because they tend to affect older individuals. Ulcer rates are declining rapidly for younger men and increasing for older individuals.

Acid secretion patterns differ with the location of the ulcer. Duodenal ulcers are associated with high-acid secretion while proximal gastric ulcers are associated with a low-acid output. Distal gastric ulcers can have a normal- or high-acid output. The difference in acid output is a reflection of the effects of inflammation on the underlying cell types. Inflammation of the antrum and the body is much more pronounced in gastric ulcer than in duodenal ulcer. In addition, there is usually a progression of inflammation in gastric ulcers. As a result, this could

eventually lead to hyposecretion of acid from glands in the fundus.

Clinical Manifestations, Diagnosis, and Treatment

The classic symptom of ulcer is dyspepsia, a burning epigastric pain usually occurring 2–3 h after meals and at night (between 11.00 p.m. and 2.00 a.m.) when acid secretion is maximal. Relief often occurs with ingestion of food and alkali. Although suggestive of peptic ulcer, dyspepsia is not a sensitive or specific measure of peptic ulcer. Only about 50% of DU patients have the typical symptom of dyspepsia. Some ulcer patients develop a stomach that is easily irritated by food, mechanical distention, or other chemical stimuli.

The sensitivity of radiography for the diagnosis of ulcers ranges from 50 to 90%, depending on the technical skill of the radiographer and the size and location of the ulcer. Fiberoptic endoscopy is a sensitive, specific, and safe method for diagnosing peptic ulcers. It gives the advantage of direct visualization and access to tissue for biopsy.

Data from placebo-controlled trials show that untreated peptic ulcers can heal within 4 weeks in 30% of GU and 40% of DU patients. Recurrence usually occurs in two-thirds of patients who have documented ulcer healing. Complications of ulcers might include hemorrhage, obstruction, and perforation. Treatment involves drugs that reduce acid output such as histamine-2 (H2) receptor antagonists, proton pump inhibitors, antacids, and antibiotics to treat *H. pylori* infection. In the case of NSAID-induced ulcers, treatment is targeted at reducing acid output and the cessation of NSAIDs, if possible. Surgery is reserved for those with complicated ulcer disease.

Helicobacter Pylori

Since its description in 1983, *H. pylori* has been implicated as a causative agent of gastritis, gastric adenocarcinomas, gastric B cell lymphoma, and peptic ulcer disease. In developed countries, the prevalence increases with age, whereas in developing countries, most children are infected by the age of 10 years. Transmission is believed to be by person-to-person spread, although the means of spread is unclear. Intrafamilial transmission is suggested by several epidemiologic studies.

The bacterium produces factors that allow it to colonize the gastric mucosa and, for some strains, gives it the ability to adhere to the epithelium. The flagella allow it to move quickly through the mucus layer to a region where the pH is neutral and thus favorable. The urease produced by *H. pylori* partly acts by producing ammonia from urea and locally neutralizing acid. *Helicobacter pylori* also produces factors that are associated with its virulence. The vacuolating toxin (vacA) causes vacuole formation in eukaryotic cells. Nearly half of the strains express vacA. Another protein, CagA, is often associated with more intense inflammation, duodenal ulceration, and gastric adenocarcinoma.

Acute infection with *H. pylori* induces a neutrophilic gastritis accompanied by transient hypochlorhydria. Chronic infection results in a chronic superficial gastritis characterized by neutrophils, eosinophils, and B and T lymphocytes. Inflammation is a result of bacterial products (e.g., VacA, CagA) and factors produced by gastric epithelial cells (e.g., cytokines). Most individuals with chronic infection remain asymptomatic. One in six chronically infected individuals will develop peptic ulcer disease. Atrophic gastritis, gastric lymphoma, and gastric adenocarcinoma are also thought to result from chronic infection, although these occur far less frequently.

Diagnosis of *H. pylori* infection can be made using an invasive endoscopic method or less-invasive breath tests and immunological assays (see **Table 2**). Endoscopy allows the collection of tissue biopsies that can subsequently be examined by histology, cultured for *H. pylori*, or measured for urease production. Histology with hematoxylin & eosin stain and special stains (Giemsa, Warthin-Starry) visualizes the bacteria as well as the surrounding inflammation. Culture is highly specific and it permits testing for antibiotic resistance, but its sensitivity is reduced by the risk of contamination. Urease tests on histological samples have a high sensitivity because the entire tissue sample is used to measure urease. Less-invasive tests that do not require gastric mucosal samples include serology, urea breath tests, and stool antigen tests. Serology is used to measure serum IgG antibodies.

Table 2 Diagnostic tests for *Helicobacter pylori*

Diagnostic test	*Notes*
Histology	Invasive; requires biopsy
Urease test	Invasive; requires biopsy
Culture	Invasive; requires biopsy
Serology	Not suitable for follow-up studies
Urea breath test	Noninvasive; can be used to monitor treatment
Stool antigen test	Noninvasive; can be used to monitor treatment

Table 3 Drug choices for treating *Helicobacter pylori* infections

Adjuvant	*Antibiotics*
Proton pump inhibitors	Amoxicillin
Bismuth subsalicylate	Metronidazole
Ranitide bismuth citrate	Clarithromycin
	Tetracycline

^{13}C urea breath tests require the ingestion of labeled urea, which releases labeled carbon atoms when urease is present. The labeled carbon atoms will subsequently be incorporated into carbon dioxide in the systemic circulation and measured in breath. Stool antigen detection kits offer a noninvasive way to measure infection and response to treatment. The sensitivity and specificity of the urease test on tissue samples, the breath test, and stool antigen tests range from 90–95%.

Treatment of *H. pylori* infection is clearly indicated for peptic ulcer disease (gastric and duodenal) and the rare MALT lymphomas. There may also be some benefit in treating patients with NSAID-induced ulcers who are *H. pylori* infected. The benefit for functional dyspepsia, gastric cancer, and gastroesophageal reflux disease is less certain. A consensus report by a European study group recommends a test and treat strategy for all patients less than 45 years of age with uninvestigated dyspepsia in whom there are no predominant symptoms of gastroesophageal reflux disease, no alarm symptoms, and no NSAID use.

Triple therapy is the gold standard for treatment of infections. This refers to the use of two antibiotics plus one adjunctive agent for 7 or 14 days. Fourteen days of treatment provide better cure rates. However, compliance is one of the factors reducing effectiveness of therapy. One therapeutic regime would be a proton pump inhibitor combined with clarithromycin and amoxicillin or metronidazole. Recent studies with various combinations show eradication rates mostly ranging from 80 to −95%. In cases of treatment failure, a second-line eradication therapy involves a 7-day quadruple therapy using a proton pump inhibitor, bismuth, metronidazole, and tetracycline. Eradication rates for second-line therapy range from 77 to 82%. The drug choices for treating *H. pylori* are given in **Table 3**.

Congenital Anomalies

Congenital anomalies of the stomach are infrequent. Depending on the degree of obstruction that occurs, they may present in the neonatal period or later. Gastric atresia refers to the arrest of development of part of the stomach. It usually occurs in the antrum or the pylorus as a fibrous cord remnant or as a membrane with mucosal and submucosal layers. At birth, children with gastric atresia will have signs of gastric outlet obstruction. Treatment of gastric atresia requires surgical excision of membranes or joining the stomach and duodenum (gastroduodenostomy).

Microgastria refers to a small, tubular stomach associated with a megaesophagus histologically normal. It may occur in association with other anomalies. Microgastria results in decreased gastric acid secretion that hampers iron absorption and causes iron deficiency. Decreased production of intrinsic factor may also result in vitamin B_{12} deficiency. Frequent small-volume feeding or continuous-drip feeding is used as conservative management. Surgical intervention can enable normal growth and development.

Infantile hypertrophic pyloric stenosis is caused by hypertrophy of the circular muscle layer around the pyloric channel, which results in gastric outlet obstruction. The incidence in the US is approximately 3 per 1000 live births, but it can vary among ethnic groups (highest among whites, then black Americans, then Africans, and lowest among Asians). Pyloric stenosis is the most common reason for abdominal surgery in the first 6 months of life. This usually presents as progressive projectile vomiting after feeding, and is often apparent at the 3rd or 4th week of age. The vomitus is never stained with bile but can contain signs of blood. Anorexia, dehydration, and wasting may occur with time. Owing to chronic vomiting, a metabolic alkalosis develops. Two classic physical signs in pyloric stenosis are the palpable pyloric mass (olive) and visible peristaltic waves. The typical clinical presentation may be enough for diagnosis; however, in a few infants, contrast radiography or ultrasonography may be required. Initial treatment is to correct fluid and electrolyte imbalance but definitive treatment is surgery involving a longitudinal incision in the hypertrophied pylorus. Prognosis after surgery is excellent and the infant grows and develops normally.

Gastric Volvulus

Gastric volvulus is the condition where the stomach twists upon itself. This can be transient, but it can also lead to obstruction, ischemia, and necrosis. Primary gastric volvulus occurs below the diaphragm and is found in a third of cases. Secondary gastric volvulus occurs above the

diaphragm and is associated with a defect in the diaphragm or a herniation of the stomach upwards towards the esophagus. One-fifth of cases are found in children (usually diagnosed before the age of 1 year) and are associated with a congenital diaphragmatic anomaly. Peak incidence is in the fifth decade of life.

Pain, violent retching, and inability to pass a nasogastric tube is very suggestive of acute gastric volvulus. Confirmation of the diagnosis is usually made with a plain abdominal radiograph. Treatment requires surgery. Mortality from acute gastric volvulus is around 15–20%, but in cases where blood supply to the stomach has been severely compromised, mortality rates can reach 40–60%.

Motor Disorders

Gastric stasis refers to the delayed emptying of gastric contents and this results in patients experiencing early satiety, bloating, nausea, and vomiting. Endoscopy can confirm the presence of gastric stasis by finding retained food after an overnight fast. Systemic disorders such as scleroderma and diabetes mellitus can have neuromuscular effects that can affect the stomach. Neurological disorders can likewise cause gut dysmotility.

Gastric dumping refers to the early delivery of large amounts of liquid and solids to the small intestines. Early satiety, abdominal discomfort, and hypotension are common manifestations. Hypotension is a result of fluid shifts from the plasma to the intestinal lumen. Hypoglycemia may occur secondary to the insulin surge that results from the increased glucose to the portal circulation. Gastric dumping is usually associated with resective gastric surgery. Dietary treatment alone is not really successful. Sensible advice would be to reduce fluid intake during meals. Octreotide, an analog of the hormone somatostatin, is effective in patients with dumping because it retards gastric emptying and inhibits insulin.

Zollinger-Ellison Syndrome

Zollinger-Ellison Syndrome (ZES) refers to the triad of severe ulcer disease, gastric acid hypersecretion, and non-beta islet cell tumors of the pancreas. These tumors release gastrin, a very potent stimulatory secretagogue; hence, they are referred to as gastrinomas. ZES is thought to be responsible for 1% of all duodenal ulcers. The most common presentation is that of a single duodenal ulcer that is persistent and progressive, and is less responsive to the usual therapy. Increased serum gastrin concentrations are a hallmark of ZES. Reducing gastric acid secretion via antisecretory agents is valuable in the medical management of patients with ZES while the patient is stabilized or being evaluated for definitive treatment. Surgery is the treatment of choice for the gastrinomas; successful resection of gastrinomas ranges from 20 to 45% of patients. Five-year survival rates for all patients with gastrinomas range between 60 and 75%.

See also: **Cancer**: Epidemiology of Gastrointestinal Cancers Other Than Colorectal Cancers. **Microbiota of the Intestine**: Prebiotics. **Stomach**: Structure and Function.

Further Reading

Al-Akwaa AM, Siddiqui N, and Al-Mofleh IA (2004) Primary gastric lymphoma. *World Journal of Gastroenterology* **10**: 5–11.

Correa P (2004) The biological model of gastric carcinogenesis. *IARC Science Publications* **157**: 301–310.

Huang J, Sridhar S, and Hunt R (2002) The role of *Helicobacter pylori* infection and non-steroidal anti-inflammatory drugs in peptic ulcer disease: A meta-analysis. *Lancet* **359**: 14–22.

Isakov V and Malfertheiner P (2003) Helicobacter pylori and nonmalignant diseases. *Helicobacter* **8** (supplement 1): 36–43.

Layke JC and Lopez PP (2004) Gastric cancer: diagnosis and treatment options. *American Family Physician* **69**: 1133–1140.

Malfertheiner P, Megraud F, and O'Morain C (2002) The European Helicobacter pylori Study Group. Current Concept in the management of helicobacter pylori infection. The Maastrict 2-2000 Consensus Report. *Alimentary Pharmacology and Therapeutics* **16**: 167–180.

Nakamura T, Inagaki H, Seto M, and Nakamura S (2003) Gastric low-grade B-cell MALT lymphoma: treatment, response, and genetic alteration. *Journal of Gastroenterology* **38**: 921–929.

Nardone G and Morgner A (2003) Helicobacter pylori and gastric malignancies. *Helicobacter* **8** (supplement 1): 44–52.

Perri F, Qasim A, Marras L, and O'Morain C (2003) Treatment of *Helicobacter pylori* infection. *Helicobacter* **8**: 53–60.

Plummer M, Franceschi S, and Munoz N (2004) Epidemiology of gastric cancer. *IARC Science Publications* **157**: 311–326.

Tahara E (2004) Genetic pathways of two types of gastric cancer. *IARC Science Publications* **157**: 327–349.

STROKE, NUTRITIONAL MANAGEMENT

S McLaren, London South Bank University, London, UK

Introduction

Stroke is a common and devastating event, the incidence rising with age. It has been estimated that circa 125 000 and 500 000 new or recurrent strokes affect individuals each year in the UK and the US, respectively, creating a significant burden of long-term disability in survivors. Stroke is a syndrome that is sudden in onset, featuring signs of cerebral dysfunction of more than 24 h duration that are vascular in origin. In first strokes, thromboembolic infarction is the underlying pathophysiological event in approximately 80% of cases. Of the remainder, 10% have been attributed to primary intracerebral hemorrhage, 5% to subarachnoid hemorrhage, and 5% to uncertain types. Resulting neurological and functional impairments vary in range, combination, and severity. These can include altered consciousness, motor paralysis, and somatosensory, auditory, olfactory and visual loss. Impairments of memory, speech, language, and continence can also occur. Overall, the impairments and disabilities that result from stroke can exert a variable, negative impact on oral food ingestion, mobility, mood, social role-function, and quality of life.

The challenge of nutritional management is apparent in the scope and complexity of issues presented. These include the physical, psychological, and social impact of stroke on appetite and food ingestion, as well as prestroke nutritional status. Potential effects of metabolic injury responses, immobilization, and infective complications on energy, nitrogen, and micronutrient requirements must be considered. In individuals in whom oral feeding cannot be established, intravenous hydration and artificial nutritional support by the enteral route, utilizing either nasogastric catheters or percutaneous endoscopic gastrostomy, may be necessary. Finally, the presence of disorders such as obesity, diabetes mellitus, and hypertension, which are known risk factors for stroke, can require dietary modifications as part of the overall plan of treatment.

Effective nutritional management of stroke requires a multidisciplinary approach in which the skills and knowledge of physician, dietitian, nurse, speech and language therapist, physiotherapist, and occupational therapist are effectively and efficiently integrated to the benefit of the patient. Involvement of patient and carers in all aspects of a rehabilitation plan is vital. Ethical and legal issues relating to hydration and nutritional support can arise in stroke management.

Risks of Protein-Energy Malnutrition

The reported frequency of malnutrition after acute stroke has ranged from 8 to 34%. A number of factors may interact to impair the nutritional status of individuals who have suffered a stroke. These include factors that have led to a deterioration in prestroke nutritional status, direct physical and psychosocial effects of stroke on the consumption of food and fluids following hospital admission, and organizational factors that can hinder efficient, effective meal delivery and consumption in institutional and longer term residential settings.

Prestroke Nutritional Status

The prevalence of protein-energy malnutrition at the time of hospital admission following stroke has been variably reported as 8–30%. More detailed information obtained within 4 days of hospital admission has classified 9% of stroke patients as undernourished, 16% as overweight, and 75% as of normal nutritional status, based on a range of observational and nutritional assessment techniques. In relation to stroke outcomes at 6 months, the presence of undernutrition shortly after admission has been independently associated with a significantly greater mortality and increased likelihood of developing pneumonia, other infections, and gastrointestinal bleeds before hospital discharge. Furthermore, patients of normal nutritional status have been found to be less likely to develop pressure sores than those who were undernourished or overweight. Specifically, a low serum albumin concentration has been found to be a significant predictor of post-stroke functional impairment, morbidity, and mortality. Risk factors for malnutrition identified at the time of hospital admission have included increasing age, living alone, dementia, and inadequate dental status. These findings are consistent with those found in wider surveys of elderly populations where risk factors for protein-energy malnutrition also included social isolation, bereavement, poverty, physical disability, inadequate facilities for preparing and cooking food, and the impact of multiple medications on appetite. The presence of malnutrition in older adults may be associated with impaired

immune responses, increasing vulnerability to pneumonia and sepsis following hospital admission. The significance of a low serum albumin concentration in predicting clinical outcomes has been noted in other geriatric populations, where it may be more a reflection of disease severity than an indicator of nutritional status.

Poststroke Eating Problems

Neurological and functional impairments can result in eating problems following stroke, which can lead to an increased risk of protein-energy malnutrition or exacerbate prestroke undernutrition. Specifically, eating disability has been associated with an inadequate consumption of food and fluids and a deterioration in body mass index, triceps skin fold thickness, mid-arm muscle circumference, and serum protein concentrations during the acute phase of recovery. Specific eating problems contributing to this have included anorexia, impaired lip closure leading to oral repulsion of food and fluids, dysphagia, an inability to manipulate utensils linked to loss of motor skills in eating, and difficulties in maintaining an upright posture to aid food ingestion at mealtimes. The presence of visual field and/or perceptual deficits can result in an inability to see or perceive the contents of a meal tray, while aphasia, dysphasia, or dysarthria can hinder or prevent the expression of dietary preferences. Loss of concentration and short-term memory impairment are common sequelae to stroke and can make it difficult for individuals to sustain the sequence of activities necessary to complete a meal, or even to remember to eat and drink. A number of assessment instruments have been developed to enable health professionals to identify the extent of eating disability and ingestive skills, together with the nature of support needed.

Organizational Factors

A number of organizational factors can hinder dietary provision and consumption in institutional settings where dependent elderly people are recovering from a range of disorders including stroke. These include an inadequate mealtime environment, marked by poor lighting, noise, sterile decor, temperature extremes, and lack of facilities for socializing. Lack of manpower or involvement of untrained personnel in dietary selection, meal delivery, and supervision may result in the delivery of a meal is that inappropriate in relation to texture, portion size, and poststroke swallowing capacity, or lead to inadequate assistance being provided in relation to the level of physical dependency.

Management of Psychosocial and Physical Problems Impairing Food Consumption

Guidelines and Standards

The provision of effective nutritional care requires a concerted approach by health professionals in developing evidence-based standards and guidelines for nutritional screening, assessment, and all stages in the provision of dietary support that are linked to clinical audit. Recognition of the need for such guidance in the nutritional management of stroke patients has led to the development and dissemination of evidence-based guidelines by interprofessional expert groups, designed to inform professional judgment and bench-mark care within wider approaches to stroke management. Implementation of guidelines can be best achieved by multifaceted strategies, for example, combining education of health professionals with leadership and sensitive change management approaches.

Following acute stroke, guidelines emphasize the need for screening and assessment for nutritional risk to be undertaken within 48 h using a valid, reliable instrument by appropriately trained personnel. Individuals who are already malnourished at the point of admission or likely to become so can then be referred to dietitians for further assessment and an institution for prompt nutritional support. It is vital at an early stage to identify individuals who are capable of ingesting sufficient food via the oral route without risks of aspiration and those with abnormal swallowing, who require referral to a speech and language therapist for a more in-depth assessment. In order to achieve this, swallowing screening using a validated, reliable, and safe method should be undertaken within 24 h by appropriately trained personnel before patients are given food or drink. Following speech therapy referral, modification of dietary textures may be advised to ensure safe food ingestion in those with some degree of swallowing impairment. In others, artificial nutritional support using enteral routes may be necessary, owing to the severity of dysphagia or cognitive and functional impairments. In individuals who are capable of taking food orally, the provision of support for psychosocial and physical problems is a vital aspect of nutritional management. In order to achieve this, referral for specialist assessment, i.e., occupational and therapy, physiotherapy, psychology, and social therapy, should emanate where appropriate from other assessments.

Psychosocial Problems

In the acute phase following stroke, 25–30% of patients develop clinical signs of depression, 30% are anxious, and a similar proportion report loss of

confidence as a major psychological problem. Depression may result from an interaction of several factors including left frontal lobe damage, reactions to physical loss, and impaired performance of activities of daily living. Comparatively little is known of interactions between depression, anorexia, and nutritional status in the early stages of recovery following stroke in individuals with and without physical eating problems. However, patterns of behavioral disturbance characterized by verbal expressions of depressed mood, anorexia, and insomnia have been identified and associated with weight loss. Anxiety-evoking experiences relating to being fed, or choking in the presence of dysphagia, may also result in avoidance or withdrawal from eating. General approaches to the treatment of poststroke depression can involve the use of antidepressant drugs and behavioral and psychotherapeutic techniques. Specifically, the exercise of therapeutic skills in communication, assisting eating, and providing emotional support are vital in alleviating mealtime anxiety and increasing interest in food.

The enjoyment of eating as a social activity can be affected adversely by poststroke depression and impairments of speech, language, lip closure and, manual dexterity. Severely disabled individuals who are relearning swallowing techniques initially require a quiet environment with privacy. As swallowing and other difficulties abate, social integration at mealtimes can he achieved in conjunction with sensitive assistance.

Communication Problems

The aphasia/dysphasia present in 20–30% and anarthria/dysarthria in 40% of acute stroke patients result in difficulty in expressing thoughts in language or a total inability to do so (expressive motor aphasia/dysphasia) as well as difficulty in comprehending language (receptive aphasia/dysphasia) or producing speech (anarthria/dysarthria). Expressive aphasia, also known as Broca's aphasia, results from strokes affecting the prefrontal gyrus, while Wernicke's receptive aphasia results from lesions of the central sulcus. In contrast, dysarthria results from neurological damage affecting the neuromuscular systems that control the mechanisms of speech production. Since these systems are also concerned with swallowing, it is common to find difficulty with swallowing (dysphagia) present in dysarthric individuals.

The communication problems result in inabilities/difficulties in expressing meal preferences (aphasia/dysarthria), reading a menu, or writing preferences down (aphasia can occur in conjunction with dyslexia and dysgraphia). In contrast, receptive aphasia can impair the ability to comprehend instructions at mealtimes and thus affects compliance with rehabilitative advice. If paralysis and visual field and perceptual deficits are combined with expressive dysphasia and dysarthria, then nonverbal communication can also be limited, resulting in an inability to denote assent or dissent by nodding the head or to use gestures to convey meaning or point to food items/utensils. Early involvement of speech therapists is vital to enable individuals to regain lost functions in speech and language. In selected patients use of visual material, i.e., charts or pictures of food items and symbols, can he helpful. Use of short sentences, normal volume speech, avoidance of jargon, and patience in allowing individuals to respond to questions are also helpful in general communication.

Impairments of Arm Movement and Posture

Stroke can affect one or many of the areas and neural mechanisms controlling voluntary movement and posture. These include the motor cortex and associated pathways, cerebellum, basal ganglia, and brain stem. The impact on eating skills can be considerable, since weakness or paralysis affecting the arm occurs in 80% of strokes. Loss of coordination, spatial impairment, abnormal muscle tone, and sensory loss may also occur. Common problems resulting from this are difficulties manipulating cutlery, lifting/loading utensils, cutting food, inserting food in the mouth, drinking from a cup, or discerning the spatial relationships between objects.

If one arm is unaffected, then some degree of compensation is possible, particularly if this is dominant. Use of the unaffected hand is important in detecting temperatures of food and liquids where sensation is impaired. An occupational therapy assessment is necessary to identify appropriate aids to feeding. Lightweight plastic cups with molded handles and cutlery with built up indented handles may be useful, while plateguards and nonslip mats can be provided. Where upper limb impairment is severe, individuals may require continuous assistance with food preparation and ingestion.

Postural impairment following stroke can result in an inability to maintain an upright sitting position for effective food preparation, insertion, chewing, and swallowing. A physiotherapy assessment can identify the most effective postural techniques to counteract abnormal muscle tone (spasticity) and appropriate aids to seated balance. The latter can include molded seating, tilting chairs with table attachment, and limb stabilizers.

Visual Field Loss and Visual Neglect

It has been estimated that 30–60% of individuals who sustain an acute stroke suffer from visual field loss due to partial or complete hemianopia. Neurological damage affecting the parietal or temporal lobes and involving the sensory pathway between the optic chiasma and visual cortex underlies this problem. The impact of loss in up to half the visual field is that food items on a meal tray may not be seen and therefore may remain uneaten. Compensatory interventions include instruction in scanning the visual field, or placing items within it for those who are unable to do this. Consistent placement of items on a meal tray and verbal identification of contents using a clock system is also helpful.

Neurological damage affecting the visual cortex of the occipital lobe, common following right hemisphere strokes, can result in neglect of half the visual space. A classic feature of this problem is failure to eat food on the left side of a plate. Affected individuals need reminding to focus on food items in the neglected space; placing a colored marker on one side of the plate can be helpful. This problem may occur in conjunction with visual field loss.

Attention Span, Short-Term Memory

Impairment of attention span and short-term memory of a few minutes duration are common following stroke. Attention deficits result in an inability either to focus on immediate events or to establish a new focus unless a current stimulus is removed. As a consequence, an activity that requires a sequence of steps, such as eating a meal with two or three courses, cannot he completed. Lack of concentration is also unhelpful in relearning eating patterns. Removing or minimizing distractions at mealtimes, simplifying the complexity of information necessary to regain eating skills, and providing verbal, written or auditory alarms as reminders to eat are important in overcoming this problem.

Swallowing

Difficulty in swallowing (dysphagia) affects approximately 27–50% of the stroke population. Variable in severity, it is characterized by sensory and/or motor loss affecting one or more of the stages of swallowing, i.e., oral preparation, oral transport, pharyngeal transport, and reflex swallowing (**Table 1**). The effects of stroke on esophageal peristalsis are not fully known. It has been

Table 1 Stages of swallowing: effects of stroke

Stage	*Effects of stroke*
1. Oral preparation	Inadequate lip seal causes leakage of food/fluid chewing slower, food impacts in oral sulci
Duration variable	Hyposecretion of saliva
Lip closure forms anterior seal	Paralysis of tongue impairs bolus formation
Comminution of food by mandibular, maxillary teeth; chewing of food	Sensory loss leads to impaired bolus lateralization
Salivation evoked by parasympathetic nervous system	
Bolus formation controlled by tongue	
Sensory feedback from oral mucosa on volume and consistency determine timing of bolus ejection	
2. Oral transport	Slowed transport
Duration 1 s	Bolus localization, separation and formation impaired; can lead to food retention in oral cavity
Bolus of 5–15 cm^{-1} separated, moved to tongue midline	Lack of fine motor coordination may lead to loss of liquid bolus control; risk of aspiration
Oral cavity sealed, mandible raised, pressure exerted by tongue against palate propels bolus to posterior oral cavity	Abnormal positioning of bolus; diminished tongue elevation; inadequate bolus propulsion
3. Pharyngeal transport/reflex swallowing	Events may occur in abnormal sequence/timing
Duration 0.5–0.6 s	Impaired sensation, delay/absent swallowing reflex
Bolus impacts on sensory receptors in tissues of soft palate, pharynx, tongue, fauces	Velopharyngeal closure impaired; food regurgitated through nose/mouth
Swallowing reflex stimulated; elevation/closure of velopharyngeal mechanism, elevation of larynx, closure of vocal cords, pharyngeal peristalsis, relaxation of esophageal sphincter	Incomplete laryngeal elevation/vocal cord adduction Swallowing reflex delay/absence leading to coughing, aspiration
Respiration transiently ceases as bolus enters esophagus; breathing resumed; soft palate returned to resting position	
4. Esophageal transport	Effects of stroke little investigated
Duration 8–20 s	Aging results in slight impairment of peristaltic amplitude
Peristalsis moves bolus to stomach	

estimated that approximately 50% of dysphagic patients either die or recover spontaneously within the first 2 weeks of stroke onset, leaving half with swallowing impairment, which can exert a negative impact on functional, recovery and, thereby, quality of life.

Typical clinical features of dysphagia can include delayed oral and pharyngeal transit, oral repulsion of food/fluids, impaction and retention of food in the cheek cavity, choking or regurgitation of food/fluids through the nose and mouth, delay or absence of laryngeal cartilage elevation, abnormal gag, delayed or absent triggering of swallow, poor tongue coordination, wet, 'gurgly' voice after eating or drinking, dysarthria, and dysphonia. Complications resulting from dysphagia can be life threatening, i.e., aspiration of food/fluid into the respiratory tract, pulmonary infection, and dehydration. Longer hospital stay, strong inverse correlations with functional capacity, and an increased mortality have been associated with the presence of dysphagia. Early identification of the problem is therefore vital, encompassing screening as soon as possible after admission, clinical bedside assessment, videofluoroscopy and, if necessary, esophagoscopy.

Screening is a procedure intended to identify patients with potential swallowing problems, who can then be referred if necessary for more detailed assessment to provide information on phases of swallowing, together with judgment of extent of dysfunction and risk of aspiration. Systematic reviews have identified a number of screening methods of varying validity and reliability, which combine the identification of clinical features of dysphagia with or without swallowing water. Prescreen checking of conscious level and oromotor and laryngeal function, together with signs of respiratory aspiration and the extent to which the patient can safely cooperate with the screening examination are necessary. In the first few days following hospitalization, screening should be repeated to identify rapid recovery.

Currently, the definition of clinical bedside assessment (CBA) varies and no standardized tests are available for use by speech and language therapists or other specially trained health professionals. Bedside assessment methods of tested validity and reliability are available. General approaches can encompass the medical history relating to onset of swallowing problems: oral sensory and motor testing; laryngeal and pharyngeal assessment; presence/absence of swallowing, cough and gag reflexes; cognitive and language function; and alertness, attention span, and ability to follow instructions. Aspiration of food or fluid into the respiratory tract is not always accompanied by choking; it may be silent, or indicated by voice changes (wet, hoarse, gurgling) and breathlessness. Loss of swallowing and protective gag reflexes or the presence of features of dysphagia or aspiration when attempting to swallow a teaspoon of clear fluid at an initial screen are indications that nil should be given by mouth. Further detailed investigation is then necessary.

CBA is limited in its ability to detect some cases of oropharyngeal dysphagia and individuals who are aspirating food, fluids, or saliva into the respiratory tract. Recognition has been given to the need to develop and validate CBA as a standardized test that reaches acceptable levels of inter-rater reliability for use by trained health professionals.

Videofluoroscopy (VSS) using the modified barium swallow has been long regarded as the 'gold standard' in the assessment of dysphagia. It provides a more detailed radiological assessment of the oral, pharyngeal, and upper esophageal phases of swallowing and can detect functional impairments resulting in aspiration, evaluate the optimal head/neck position during swallowing, and determine the impact of food textures on the process. Limitations of videofluoroscopy include its labor-intensive nature, exposure to radiation, problematic accessibility in transporting disabled stroke patients to radiology departments, lack of standardization across centers with respect to volumes, consistencies, or textures of food and fluids used in the process, together with variable approaches to screening positions adopted. Some variability in the reliability of reporting has been identified both between and within raters judging the outcomes of VSS.

Fiberoptic endoscopic evaluation of swallow (FEES) in which a flexible endoscope is passed into the nares, over the velum and into the pharynx has been shown to provide an inexpensive and reliable alternative to VSS in the detection of laryngeal penetration and aspiration. Although useful in the observation of bolus transit through the hypopharynx, a limitation is that it cannot be used to investigate the oral stage of swallowing. This accessible, portable technique can be performed in a range of settings, overcoming the transport problems associated with VSS.

Where a specialist assessment has confirmed that an oral diet can be attempted with modified food textures and viscosity of fluids, the intent is to alter the rate at which nutrients pass through the pharynx, to assist the patient in control of swallowing and to reduce the risk of aspiration. Variations in

Table 2 Food textures suitable for different levels of dysphagia

Level	*Dysphagia severity*	*Designation*	*Food textures omitted*	*Food textures allowed*
I	Severely impaired oral preparatory stage Reduced pharyngeal peristalsis Reduced swallowing reflex[a]	Patients unable to chew solids or swallow thin liquids safely	Coarse, hard, or brittle textures Unsuitable foods: fruits, nuts, raw vegetables, sticky foods, all foods requiring bolus formation or controlled oral manipulation No thin liquids	Thick, homogeneous, smooth textures or semisolids; cold liquids thickened with commercial agent Suitable foods: poached eggs, soft puddings, thick vegetable/fruit purée
II	Moderately impaired oral preparation Reduced pharyngeal peristalsis	Patients who cannot swallow thin liquids and have minimum tolerance of chewed food	Coarse, hard, brittle textures No thin liquids	Thick, homogeneous purées, semisolids Suitable foods: cold thickened beverages, eggs, yogurt, cottage cheese, custards, puréed vegetables, fruits, tender meats
III	Mild impairment of oral preparation Mild chewing problems	Patients who can tolerate soft foods and liquids	Coarse, hard, brittle textures Unsuitable foods: nuts, minced meat, raw, stringy, crisp vegetables	Soft foods, not puréed or blended Suitable foods: soft bread, eggs, cottage cheese, casseroled small meat pieces, macaroni, hot/cold thick puddings
I	Mild oral preparation impairment Can chew soft textures	Patients who can tolerate soft food textures and liquids	Coarse, fibrous, hard foods Unsuitable foods: nuts, deep fried or raw, crisp foods	Soft foods not requiring grinding or chopping Suitable foods: soft, toasted bread, cold cereals, milk, pasta, eggs, cooked, canned or over-ripe fruit, fine moist meats, tender vegetables, soft desserts

[a]Videofluoroscopy screening essential for safety.
Adapted from Martin (1991) Dietary management of swallowing disorders. *Dysphagia* **6**: 129–134.

the degree of modification and textures suggested are apparent in published dysphagia diets. The aspiration-risk reduction diet outlined in **Table 2** was devised for the management of oropharyngeal dysphagia secondary to neurological impairment. This recognizes the potential risks of aspiration posed by purees and thin liquids, which are difficult to control in the oropharynx of individuals suffering from this type of dysphagia; therefore, the diet offers three choices based on no fluids, thickened fluids, and unaltered fluids. It is the responsibility of the speech and language therapist to recommend the most appropriate food textures for an individual. Dietitians should ensure that texture modified meals are adequate to meet nutrient requirements, offer choice, and are palatable. Skilled nursing interventions at mealtimes can also be of assistance in food ingestion. Impaired oral preparation can be compensated for by positioning food on, and tilting the head towards, the unaffected side; posterior positioning of food on the tongue can promote oral transport and in impaired pharyngeal transport/reflex swallowing, ensuring posture is upright, head stable in the midline with a slight forward flexion to protect the airway and synchronizing the sequence of inspiration, breath-holding, swallowing, expiration and coughing on expiration to clear food debris can be helpful. Maintaining an upright posture for at least 30 min after meals is advisable to prevent regurgitation/aspiration.

Many nondietary, therapeutic approaches to the management of dysphagia have been identified including: use of palatal training devices to assist triggering of the swallowing reflex; insertion of prostheses to lower the palatal vault to improve bolus formation; oral electrical stimulation; oral thermal stimulation; drug therapy (nifedipine); high-intensity swallowing therapy; exercises to improve where appropriate laryngeal closure, labial/mandibular closure, tongue elevation and lateralization; use of biofeedback involving mirrors and VSS. Although benefits have been described for many of these interventions, the lack of randomized, controlled clinical trials with adequate power has limited conclusions relating to effectiveness.

Nutrient Requirements

A variety of approaches to the estimation of energy requirements can he used. Predictive equations can

provide an estimate of resting energy expenditure based on body weight, height, age, and sex. Modifications to equations can be applied that incorporate activity and injury factors. Additional requirements for tissue repletion should be considered in malnourished individuals. Estimation of resting energy expenditure by indirect calorimetry using a portable metabolic monitor provides more accurate estimates derived from the respiratory quotient. Values obtained do not consider periods of activity, pyrexia, pain, or energy increments necessary for nutritional repletion. Further corrections are therefore necessary.

Following stroke a number of factors may affect energy requirements. Inactivity caused by paralysis will reduce energy expenditure, but the presence of pulmonary infection, a common complication of stroke, will increase it (a 1 °C rise in core temperature raises energy expenditure by 13%). The impact of the cerebral injury on poststroke metabolism, notably on resting energy expenditure at different levels of stroke severity, has not been fully investigated. Evidence for metabolic injury responses based on hormonal profiles and changes in blood glucose concentration in the acute phase is limited. Hyperglycemia is common following stroke and has been associated with an increased morbidity and mortality. Hyperglycemia can be attributed to overt or latent diabetes mellitus, stress responses, and effects of glucose intolerance in elderly subjects. Elevated plasma cortisol and catecholamine concentrations representing a transient stress response have been reported in the first 72 h following stroke.

Nitrogen requirements following stroke can be estimated using reference ranges based on body weight, or using nitrogen balance studies. The chronic disorders diabetes mellitus and hypertension are common in the stroke population, which is predominantly elderly. Both disorders require dietary manipulation.

Fluid balance requirements also require careful assessment, since dehydration is a serious risk in individuals with dysphagia and physical disabilities that impair fluid consumption. Oral intake of fluid is contraindicated where the swallowing and gag reflex are lost or swallowing and/or level of consciousness are impaired. Intravenous fluid replacement therapy is then necessary, usually in the short term. Fluid requirements can be calculated on the basis of 35 ml per kg body weight daily in adults, but sepsis and fever can increase needs. Fluid intake and output in conjunction with insensible losses should be monitored on a daily basis together with the symptoms of dehydration, i.e., thirst, dry mucous membranes, and loss of skin turgor.

Artificial Nutritional Support

The presence of severe dysphagia and cognitive and complex physical impairments may render oral feeding unsafe or insufficient to meet nutritional requirements. If the gastrointestinal tract is functional, the options for delivering enteral nutritional support are either via a fine-bore nasogastric tube or via a catheter inserted by percutaneous endoscopic gastrostomy (PEG). Decisions concerning the choice of route are influenced by the anticipated duration of dysphagia and benefits versus risks. Impact on nutritional status, rehabilitation, quality of life, safety, tolerance, flexibility, ease of use, costs of insertion, removal, and maintenance are all important considerations. In the majority of cases dysphagia resolves within the acute phase of stroke, i.e., approximately 2–3 weeks. For relatively short time periods, enteral feeding by fine-bore nasogastric catheter is usually undertaken. This can be preferable because of the technical simplicity of intubation, maintenance and removal, low cost, and ease of use. Discomfort and risks of aspiration, accidental endotracheal intubation, and displacement or perforation of the esophagus are the potential complications, most of which are uncommon, although aspiration may occur in up to 10% of patients.

For feeding over longer time periods, a PEG tube inserted under local anesthetic offers greater comfort, toleration, ease of use, and reported improvements in nutritional status. However, costs are greater and this more invasive procedure carries a technique-related fatality of 1–2%. Minor complications include stomal sepsis, leaking, and outlet blockage. Peritonitis, perforation, gastrointestinal bleeding, and intestinal obstruction can occur, but are rare. Most enterostomy catheters are made from nonacid-hardening polyurethane or silicone and can be left *in situ* for up to 6 months. No consensus exists concerning the time period within which gastrostomy feeding should be initiated following stroke, but it should he considered where dysphagia is likely to persist beyond 14 days, and earlier for those intolerant of nasogastric tube feeding. In a small number of cases, enteral nutrition may be contraindicated following stroke owing to gastrointestinal bleeding resulting from severe stress ulceration. Rarely, nonstroke-related contraindications may also be present, i.e., gastrointestinal failure, ascites, Crohn's disease, bleeding, and clotting disorders.

In considering the relative merits of PEG versus nasogastric feeding poststroke, a systematic review found that PEG was superior in terms of weight maintenance and other indicators of nutritional status. PEG was also associated with significantly

lower end-of-trial case fatality rates and treatment failures in the limited evidence derived from only two randomized, controlled clinical trials. Issues relating to the optimum timing of commencement of artificial feeding have not been clarified in systematic reviews; however, some guidelines emphasize the need to consider the institution of nutritional support within 5 days of a nil oral regime. Failure to provide nutritional support for patients who have not met or are unlikely to meet their nutritional requirements for >7 days is viewed as unethical. In the presence of dysphagia, those unable to meet their nutritional requirements by the oral route should be assessed for nasogastric feeding within 7 days.

Evaluation of Nutritional Support

It is vital that nutritional status is monitored in the acute phase of recovery and that dietary intakes are readjusted accordingly. Appropriate dietary, anthropometric, and clinical assessments, which can be performed on a weekly basis, are discussed (*See* 00201 and 00202). Other important components of monitoring include recovery of physical functions related to independence in eating including swallowing capacity and the complications of enteral support techniques. Providing effective nutritional management following stroke requires the coordination of the professional skills of doctor, nurse, speech therapist, dietitian, occupational therapist, and physiotherapist, ideally within the context of a nutrition support team. Dynamic leadership and clear accountability, lines of communication, and referral policies are essential for the team to deliver effective support. Follow-up services in the community are also necessary to prevent deterioration in nutritional status in the later stages of rehabilitation.

See also: **Diabetes Mellitus**: Dietary Management. **Energy**: Balance; Requirements. **Energy Expenditure**: Indirect Calorimetry. **Hypertension**: Nutritional Management. **Malnutrition**: Secondary, Diagnosis and Management. **Nutritional Assessment**: Anthropometry; Biochemical Indices. **Older People**: Nutritional Requirements; Nutrition-Related Problems; Nutritional Management of Geriatric Patients. **Protein**: Requirements and Role in Diet.

Further Reading

Bath PMW, Bath FJ, and Smithard DG (2002) Interventions for dysphagia in acute stroke (Cochrane Review). In: *The Cochrane Library,* Issue 1. Oxford. Update Software.

Doggett DL, Tappe KA, Mitchell MD, Chapell R, Coates V, and Turkelson CM (2001) Prevention of pneumonia in elderly stroke patients by systematic diagnosis and treatment of dysphagia: an evidence-based comprehensive analysis of the literature. *Dysphagia* **16**: 279–295.

Ekberg O, Hamdy S, Woisard V, Wutte-Hanig A, and Ortega P (2002) Social and psychological burden of dysphagia: its impact on diagnosis and treatment. *Dysphagia* **17**: 139–146.

Finestone HM and Greene-Finestone LS (2003) Rehabilitation Medicine: 2. Diagnosis of dysphagia and its nutritional management for stroke patients. *Canadian Medical Association Journal* **169**(10): 1041–1050.

FOOD Trial Collaboration (2003) Poor nutritional status on admission predicts poor outcomes after stroke. *Stroke* **34**: 1450–1456.

Green CJ (1999) Existence, causes and consequences of hospital and community malnutrition; clinical and financial benefits of nutritional intervention. *Clinical Nutrition* **18**(supplement 2): 3–28.

Greenwood R (2003) *Handbook of Neurological Rehabilitation* Hove, Sussex: Psychology Press.

Hudson H, Daubert CR, and Mills R (2000) The interdependency of protein-energy malnutrition, aging and dysphagia. *Dysphagia* **15**: 31–38.

Logeman JA (1997) *Evaluation and Treatment of Swallowing Disorders* San Diego: College Hill Press.

Logeman JA (1999) A screening procedure for oropharyngeal dysphagia. *Dysphagia* **14**: 44–51.

Martin AW (1991) Dietary management of swallowing disorders. *Dysphagia* **6**: 129–134.

McFie J (2001) Ethics and nutritional support: a clinician's view. *Clinical Nutrition* **20**(supplement 1): 87–99.

Miller A (2004) *The Neuroscientific Principles of Swallowing and Dysphagia.* Singular Publications.

O'Neill PA, Davies I, Fullerton K, and Bennett D (1991) Stress hormone and glucose response following acute stroke in the elderly. *Stroke* **22**(7): 842–847.

Penman JP and Thomson M (1998) A review of the textured diets developed for the management of dysphagia. *Journal of Human Nutrition and Dietetics* **11**: 51–60.

Perry L and Love CP (2001) Screening for dysphagia and aspiration in acute stroke: a systematic review. *Dysphagia* **16**: 7–18.

Royal College of Physicians (2002) *National Clinical Guidelines for Stroke: Intercollegiate Working Party on Stroke.* London: The Royal College of Physicians. http://rcp.ac.uk.

Wolfe CD (2000) The impact of stroke. *British Medical Bulletin* **56**(2): 275–286.

SUCROSE

Contents

Nutritional Role, Absorption and Metabolism

J Brand-Miller, University of Sydney, Sydney, NSW, Australia

Sucrose plays a unique role in human diets. It satisfies our instinctual desire for sweetness and contributes an average of 10% of the energy in modern Western diets. Sucrose has many functional roles in foods which extend beyond its sweetness, including preservative, textural, and flavor modifying qualities. Unfortunately, sucrose has a 'bad reputation,' especially in respect of dental caries. In the past, refined sucrose was suggested to cause diabetes, overweight, heart disease, micronutrient deficiencies, and even hyperactivity in children. But within the last decade a wealth of new research on sugars in the diet has shown most of these assumptions to be false. We now know that refined sucrose consumption is much lower than we originally estimated (45–65 g per day instead of 125 g per day in industrialized countries), that intake of sugars correlates inversely with the fat content of the diet (the higher the sugar intake, the lower the fat), and that high-sucrose diets are associated with *lower* body weight. In addition, new research shows that moderate intake of sugars is associated with the highest intakes of micronutrients and that sucrose-containing foods do not raise the plasma glucose level any more than most starchy foods. While dental caries is still associated with high sucrose consumption in nonindustrialized countries, there is no relationship in developed nations. The intake of fluoride, frequency of food intake, and dental hygiene are more important factors influencing the incidence of dental caries in these countries.

What Is Sucrose?

Sucrose is a pleasant tasting substance that contributes most of the sweetness in our diet. It has played a role in human diets ever since primates began evolving on a diet of fruit and berries in the tropical forests of Africa 50 million years ago. Sucrose is chemically classified as a carbohydrate and a simple sugar, specifically a disaccharide composed of glucose and fructose (**Figure 1**). Its proper scientific name is β-D-fructofuranosyl-α-D-glucopyranoside. The natural sweetness of fruit and honey comes from mixtures of sucrose, glucose, and fructose. The mild sweetness of milk comes from another disaccharide, lactose, composed of glucose and galactose.

Because sweetness comes from a mixture of sugars (not just sucrose) in many sources, we use different terms to define the original source, e.g., naturally occurring sugars, refined sugars, added sugars, concentrated sugars, intrinsic sugars, and extrinsic sugars. Refined sucrose is also known as table sugar, cane sugar, or beet sugar. Unfortunately, the term 'sugar' means different things to different people, and the literature can be confusing. In this article, as in everyday language, the word 'sugar' refers to refined sucrose, unless otherwise indicated.

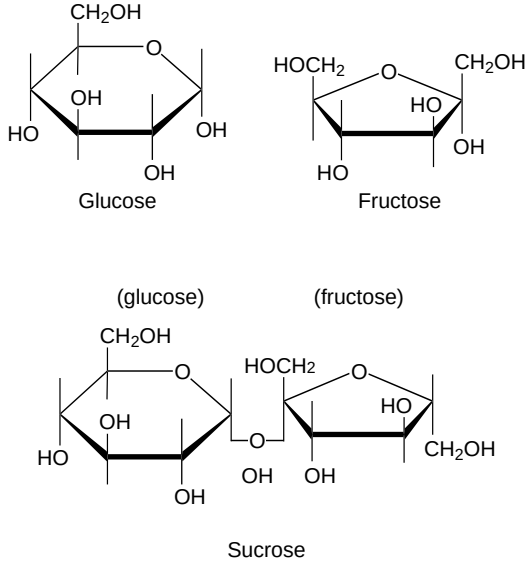


Figure 1 The chemical structure of glucose, fructose, and sucrose.

The Role of Sucrose in the Diet

Historical Perspective

Sugar cane and sugar beet have a naturally high content of sucrose and have been commercially exploited as a concentrated sources of sucrose since 1600 AD. Sugar cane was first cultivated in Papua New Guinea 10 000 years ago, and the practice spread gradually to Egypt (2300 years ago), Arabia (1300 years ago), and Japan (1100 years ago). Sugar beet was first grown in Europe 500 years ago. Prior to this, refined sucrose was still a rare and expensive commodity and honey was much cheaper. When the industrial revolution began 200 years ago, sucrose consumption increased dramatically, replacing honey as the major source of concentrated sweetness. Intake of refined sucrose peaked in about 1900 and consumption has remained, with minor variations, much the same over the past century. Since 1970 high-fructose corn syrup solids (glucose–fructose syrups made from hydrolyzed corn starch) have partially replaced refined sucrose in manufactured products, particularly in the USA.

Instinctual Liking for Sweetness

The appreciation of the sensation of sweetness runs deep in the human psyche. In literature and mythology sweetness is linked with pleasure and goodness, and in everyday language we use terms associated with sweetness to describe those we love (sweetie pie, honeybun). Our first food, breast milk, is sweet—in fact it is the sweetest of all mammalian milks. Newborn human infants drink more of a sweet solution than of plain water or of a salty, acidic, or bitter solution. It is not a learned taste: Everyone could be said to be born with a 'sweet tooth.' The reason for this sweet preference is not known. We could speculate that the brain's dependence on glucose as its sole source of fuel has coevolved in an environment where glucose or its precursor was not evenly distributed throughout the food supply. Perhaps those early primates who were able to detect sweetness best were most likely to survive.

Table 1 Sources of sweetness in human diets

- Honey
- Manna
- Honey ants
- Grape sugar
- Dates
- Maple sugar
- Sorghum
- Corn syrup solids
- High-fructose corn syrup solids
- Sugar beets
- Sugar cane
- Sugar alcohols (e.g., sorbitol)
- Intense sweetners (e.g., saccharin)

Hunter–gatherers of the past relished honey and other sources of concentrated sweetness such as maple syrup, dried fruit, and honey ants (**Table 1**). Wild honey was so highly prized that we went to great lengths to obtain it. Australian Aboriginals would attach a tiny feather to a bee and follow it all the way back to its hive. The Wild Men (Veddas) of Sri Lanka esteemed honey so highly that they regularly risked their lives to obtain it.

The Contribution of Sucrose to Energy Intake

The contribution of macronutrients and individual sugars to total energy intakes in industrialized nations is shown in **Table 2**. Sucrose is at the top of the league table for sugars, contributions coming from both the naturally occurring sources of sweetness such as fruit and vegetables and also from refined sucrose. Sucrose, like all carbohydrates, is burned (oxidized) in the body to yield energy, specifically 16 kJ g^{-1}. This is only half the energy yield of a gram of fat (37 kJ g^{-1}) and much less than that

Table 2 The contribution of different types of carbohydrate and sucrose to energy intake in industrialized nations

Macronutrients	*% energy (adults)*	*Men (g per day)*	*Women (g per day)*	*Children (g per day)*
Carbohydrate	45			
Starch	21–23			
Total sugars	21–24			
– Added (refined) sugars	9–11	60–70	40–50	40–50
– Naturally occurring sugars	11–16	60–80	60–75	60–70
Individual sugars				
Sucrose	9–10			
Glucose	4–5			
Fructose	4–5			
Lactose	2–4			
Maltose	<0.3			

of alcohol (29 kJ g^{-1}). Other carbohydrates such as starch, glucose, and fructose have the same energy content per gram as sucrose. The new 'intense' sweeteners such as aspartame contribute virtually no energy, hence their use in 'low joule' products.

In Western countries, total carbohydrate intake amounts to about 200–280 g per day for the average man and woman, or about 45% of all the food energy eaten (protein contributes 15–20%, fat 35–40%). About half of our carbohydrate is in the form of starch (100–140 g per day), the other half representing a mixture of sugars, with sucrose predominating (**Table 2**). A decade ago, the only figures we had for consumption of refined sucrose were derived from apparent consumption statistics (production plus imports minus export, wastage, and nonfood usage). These statistics suggested that the average intake was around 50 kg per head per year or 125 g per day for every man, woman, and child, equivalent to 25% of total energy intake. Only recently have we been able to estimate accurately intake of refined or added sugars (most of which is refined sucrose but includes small amounts of honey, molasses, high-fructose corn syrup solids, etc.). These more direct estimates show daily intakes of 40–50 g in women and 60–70 g in men in developed countries. Children ingest 40–50 g per day refined sugars. Thus apparent consumption or per capita calculations overestimated consumption by as much as three times. Refined sugars today contribute about 10% of total energy in industrialized nations in Europe, North America, and Australasia. Japan is an exception, with intakes of only 10–20 g per day. Intakes of refined sucrose in nonindustrialized countries are also small, averaging 1–20 g per day.

Some countries have made a distinction between the sugars that are consumed with intact cell wall structure (intrinsic sugars) and those which have been released from nature's original packaging (extrinsic sugars). In this classification, the sugars in an orange are intrinsic sugars while those in orange juice are extrinsic. The lactose in milk is also extrinsic because it is not inside a cell wall. The belief is that nonmilk extrinsic sugars are less desirable because they will be absorbed more quickly into the bloodstream and that the fiber and micronutrient content of the food or diet will be lower. However, recent research has shown that many of these assumptions are incorrect.

Functional Roles of Sucrose in Foods

Refined sucrose is added to foods for more than just its sweetness. The difficulties inherent in producing low-joule products using intense sweeteners attest to this. For example, sucrose contributes to the bulk and texture of cakes and cookies and it provides viscosity and mouth feel in liquids such as soft drinks and fruit juices. Sucrose is also a powerful preservative and contributes the long storage life of jams and confectionery. In frozen products like ice cream, sucrose has multiple functions: It acts as an emulsifier, preventing the separation of the water and fat phases; it lowers the freezing point, thereby making the product more liquid and 'creamier' at the temperature eaten. The presence of sucrose retards the crystallization of the lactose in dairy foods and milk chocolate (tiny crystals of lactose feel like sand on the tongue). In canned fruit, sucrose syrups are used to prevent mushiness caused by the osmotic movement of sugar out of the fruit and into the surrounding fluid. Because sucrose masks unpleasant flavors, sugar syrups are used as carriers for drugs and medicines, especially for young children who cannot swallow tablet formulations. In products like yogurt and coffee it masks the acidity or bitterness and balances the sugar–acid ratio in fruit juices and cordials. Lastly, sucrose is a substrate for fermentation. It is added as food for the yeast in bread-making and beer-making. But it is converted to alcohol and other products in the process and therefore not consumed as sucrose. For all these reasons, when manufacturers design a low joule–low sugar product, they find that many substances need to be added to perform all the roles that sucrose did alone.

Patterns of Consumption

Honey versus Sucrose

In preindustrial times, honey was the main source of concentrated sweetness in the diets of many peoples. There are no precise figures for historical consumption because honey was part of either a hunter–gatherer or subsistence economy. Until recently historians and food writers have proposed that it was a scarce commodity available only to a wealthy few. However, a reappraisal of the evidence in the Stone Age, Antiquity, the Middle Ages, and early Modern times suggests that ordinary people ate much larger quantities of honey than has previously been acknowledged. The Ancient Egyptians, for example, made frequent use of honey in their spiced breads, cakes, and pastries, and for priming beer and wine. In Roman times, half the recipes in a famous cookery book call for honey, and in Ancient Greece, those who died some distance from home were sometimes preserved in honey. These details give an impression of plenty. Even the poorest people could own a beehive because bees often made their homes

in a hollow log or a broken pot. Wealthy landowners might own dozens of beautifully constructed beehives and employ a beekeeper. During medieval times we know that honey was sold in large volumes (gallons and even barrels), units unlikely to be used for a scarce commodity. It was present in sufficient abundance to make mead a common alcoholic drink made from honey.

It is therefore possible that intakes of honey at various times during history may well have rivalled our current consumption of refined sugar. There are implications therefore for the role of sugar in modern diets. Refined sugar may not have displaced more nutrient-rich items from our present-day diets but only the nutritionally comparable food, honey.

Changes in Sucrose Consumption

It is a much more straightforward business to enquire about sugar (refined sucrose) consumption than honey consumption in preindustrial times. All sugar supplies, in Europe, came from imports, so customs records constitute a readily accessible record of national consumption. In the 1520s, the Dissolution of the Monasteries reduced demand for bees-wax for church candles and brought about a small decrease in the production of honey. Almost simultaneous with this came an increase in the supply of refined sucrose, imported from the new European colonies. Sugar was still considerably more expensive than honey, but this combination of events gained it a more complete following among the wealthy. Cookery books were used exclusively by the well-to-do at this time and clearly illustrate that, for this section of society, sugar had, by the 1550s, usurped honey's place in the diet.

It was not until the early 1700s, however, when the supply of sugar boomed, its price fell, and coffee, tea, and chocolate entered the British diet, that ordinary people finally began to buy significant amounts, and the per capita consumption reached 1.8 kg per year. The changeover from honey to sugar occurred more gradually in rural areas than in the cities. From this point sugar consumption rose inexorably, while honey consumption declined. Beekeeping ceased to be the general custom that it had been in former years—there was no longer a hive in every garden. By the beginning of the twentieth century the availability of refined sugar reached about 50 kg per head per year in most industrialized nations. Surprisingly, it did not continue to increase but remained at approximately this level or declined throughout the next 100 years. The 'steady state' suggests that the market and the taste buds have reached saturation.

Added versus Naturally Occurring Sugars

Recent studies also show that the intake of naturally occurring sugars from fruits and vegetables (including juices) is about 40–80 g per day, roughly equivalent to the intake of refined sugars. Thus all sugars contribute about 20–22% total energy intake in developed countries, or about half of all the carbohydrate eaten.

The Sugar–Fat Seesaw

In industrialized nations the intake of sugar varies from person to person, but there is a consistent relationship between sugars and fat intake. As refined sugar intake rises, fat falls, and vice versa. As the total sugars intake rises, most of the increase is due to refined sugars (refined sucrose, corn syrup solids, etc.). This relationship has important implications. The proportion of individuals achieving dietary guidelines for lower intakes of total and saturated fats is much higher among subgroups with higher intakes of total sugars and refined sucrose. In the past, we believed that sugar and fat went 'hand-in-hand' in foods and that a diet high in sugar was likely to be high in fat as well. But it turns out that the reverse is true: A high-sugar diet is more likely to be low in fat. It seems that very few people achieve a low-fat diet without also increasing sugars intake. Perhaps humans, consciously or not, strive to eat calories in their most concentrated energy-dense form.

One of the most important implications of these findings is that recommendations to reduce *both* sugar and fat may be counterproductive. A reduction in fat intake is certainly more likely to result in desirable changes in body weight, blood lipids, insulin sensitivity, and cardiovascular risk factors. Trying to reduce sugar intake as well may not only compromise the effort to reduce fat but reduce the palatability of the diet and hence long-term compliance with a low-fat diet.

Effect on Micronutrient Intake

Refined sucrose and other added sugars are regarded by many people as undesirable because they are 'empty calories' (energy without micronutrients). It was reasoned that sucrose would 'dilute' the micronutrient (vitamin, mineral) content of the diet and that high sugar intake would increase the likelihood of micronutrient deficiencies. When these assumptions were tested, diets containing moderate amounts of added sugars were found to be no less nutritious than diets low in sugar. This has been shown to be the case in both adults and children. Very high consumption of refined sugars is associated with low micronutrient intake, but so too is very low intake of sugars. Moderate consumption of sugars is associated with the highest intake of micronutrients.

One of the reasons for this unexpected result is the sugar–fat seesaw, i.e., low-sugar diets are higher in fat, which is essentially a poor source of micronutrients too. Another reason is that sweetened breakfast cereals and dairy products such as flavored milk, yogurts, and ice cream are a good source of micronutrients. Fruit juice drinks are a source of vitamin C as well as sugars. Thus refined sucrose aids the consumption of some nutritious but fairly unpalatable or bland foods.

Digestion and Absorption of Sucrose

Digestion

In the mouth, food is mixed with saliva and masticated. The physical matrix that encases the sucrose, e.g., the plant cell wall, is partially disrupted in the process. Mixing occurs very effectively in the stomach and sucrose is dispersed throughout the gastric contents. Peristaltic movements drive the semifluid material, called chyme, through the pyloric sphincter into the duodenum. The rate of stomach emptying varies as a function of the volume and acidity of the stomach contents as well as the osmolality of the chyme in the small intestine. Solutions that are acidic and hyperosmotic are emptied more slowly. Thus an acidic, high-sucrose food such as sweetened yogurt will be emptied relatively slowly.

Once in the small intestine, sucrose is too large to cross the epithelial cell membrane and must therefore be hydrolyzed for absorption to take place. The enzyme responsible for sucrose digestion is an α-glucosidase called sucrase, located in the microvillous brush border lining the small intestine (**Figure 2**).

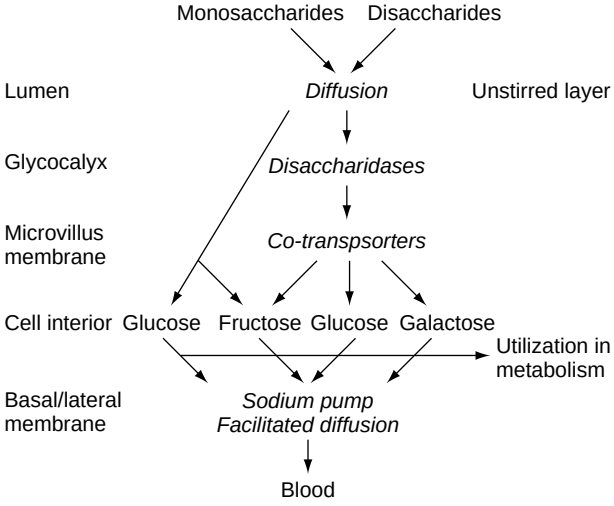


Figure 2 Disaccharidases such as sucrase are located on the luminal side of the brush border membrane of the small intestine. From Southgate D (1995) Digestion and metabolism of sugars. *American Journal of Clinical Nutrition* **62**: 203S–211S.

It digests some of the breakdown products of starch digestion as well as sucrose. Sucrose in foods is also easily hydrolyzed under weakly acidic conditions at room temperature such that considerable hydrolysis occurs even before ingestion, e.g., in soft drinks. Hydrolysis is accelerated by heating so that much of the sucrose used in food is actually swallowed as an equimolar mixture of glucose and fructose. Conditions in the stomach are also likely to hydrolyze sucrose, further increasing the monosaccharide concentration in the chyme.

Hydrolysis within the brush border is extremely rapid and the rate-limiting step is not digestion, but the transport of monosaccharides across the enterocyte. Sucrase, however, is located physically close to the monosaccharide cotransporter systems in the microvillus membrane.

Absorption of Glucose and Fructose

The glucose product of sucrose digestion is transported across the epithelial cell membrane more rapidly than is free glucose, a phenomenon that may be related to specific glucose transporters that are not dependent on sodium. The fructose released by sucrose digestion is absorbed more slowly across the brush border membrane by a process called facilitated diffusion (carrier-mediated). Fructose appears to be absorbed better when ingested with glucose (separately or combined in sucrose) than it is by itself. This explains why 100 g fructose gives rise to osmotic diarrhea, whereas 250 g sucrose does not. The absorption of free fructose is incomplete when intakes exceed about 35 g per day.

Once inside the epithelial cell, glucose and fructose are presumed to traverse the enterocyte by diffusion. The basal-lateral membrane acts as a barrier preventing the free movement of monosaccharides into and out of the enterocyte. Movement across the membrane appears to be energy-dependent but sodium-independent.

Metabolism of Glucose

From the epithelial cell, the monosaccharides pass into the portal circulation and to the liver where some of the glucose and virtually all the fructose is removed. The remainder of the glucose passes into the systemic circulation, entering the peripheral tissues by mass action and/or under the influence of the hormone insulin.

Metabolism of Fructose

When fructose reaches the liver, most is removed from the bloodstream and converted to glucose, lipid, or lactate. Some fructose is converted in the

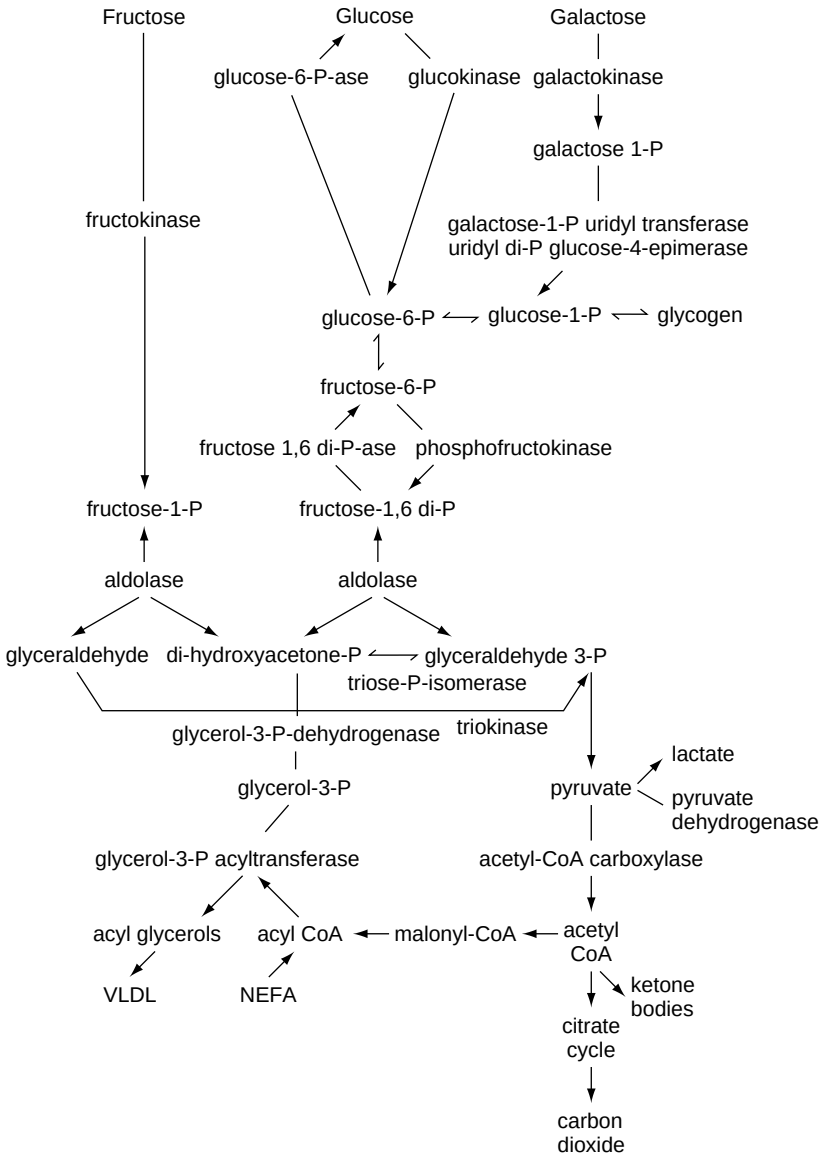


Figure 3 The monosaccharide products of sucrose digestion are metabolized to carbon dioxide and water via the tricarboxylic acid (TCA) cycle. From Southgate D (1995) Digestion and metabolism of sugars. *American Journal of Clinical Nutrition* **62**: 203S–211S.

liver to fructose-1-phosphate, which is spilt into two C_3 products, glyceraldehyde and dihydroxacetone phosphate. The latter is an intermediate for both the glycolytic and gluconeogenic pathways. The glyceraldehyde is phosphorylated and recombines with dihydroxyacetone phosphate to produce fructose-diphosphate and ultimately glycogen. The major metabolic effect of fructose is to increase the production of pyruvate and lactate. This has the effect of depressing fatty acid oxidation and increasing the esterification and the synthesis of very low-density lipoprotein (VLDL). There is only a small rise in plasma glucose (and insulin) levels because gluconeogenesis is strongly inhibited. A rise in plasma lactate is seen after fructose consumption.

Metabolic Sequelae

Consumption of sucrose (and other carbohydrates such as starch) produces a range of metabolic and hormonal responses, all of which serve to limit the rise in plasma glucose levels to within acceptable levels. Blood glucose levels above 10 mM will result in glycosuria (glucose in the urine), a waste of valuable energy, and levels below 3–4 mM will impair brain function. Insulin plays an important role in bringing blood glucose levels back to normal after a meal. It does that by promoting glucose uptake in the liver and muscle cells and inhibiting gluconeogenesis in the liver. Dietary carbohydrate, including sucrose, has four main metabolic fates: (1) oxidation in tissues,

Table 3 Glycemic index of foods (glucose = 100)

Food	*Index*	*Food*	*Index*	*Food*	*Index*
Breakfast cereals		**Sweet biscuits**		Raisins	64
Kellogg's AllBran™	30	Arrowroot	69	Rockmelon	65
Kellogg's Cocopops™	77	Morning coffee	79	Sultanas	56
Kellogg's Cornflakes	77	Oatmeal	55	Watermelon	72
Kellogg's Mini Wheats™	58	Shredded wheatmeal	62		
Museli		Shortbread (commercial)	64	**Dairy foods**	
– Toasted*	43			Milk	
– Untoasted	56	**Cakes**		– Whole (av)	27
Kellogg's Nutrigrain™	66	Apple muffin	44	– Skim	32
Porridge (av)	50	Banana cake	47	– Chocolate flavour	34
Puffed Wheat	80	Sponge cake	46	– Custard (made with powder)	43
Rice bran	19	Waffles	76	Ice cream (av)	61
Kellogg's Rice Bubbles™	89			– Low-fat	50
Kellogg's Special K™	54	**Vegetables**		Yoghurt (flavoured, low-fat)	33
Kellogg's Sultana Bran™	54	Beetroot	64		
Kellogg's Sustain™	68	Carrots	49	**Beverages**	
Uncle Toby's Vita Brits™	61	Parsnip	97	Apple juice	41
Sanitarium Weetbix™	69	Peas (green)	48	Cordial (diluted)	66
		Potato		Fanta™	68
Grains/paste		– Baked (av)	85	Lucozade™	95
Buckwheat	54	– New (av)	62	Orange juice	53
Bulgur wheat	48	– Pontiac	56		
Rice		– French fries	75	**Snack and convenience foods**	
– Calrose (white)	83	Pumpkin	75	Corn chips	72
– Doongara/basmati	59	Sweet corn	48	Fish fingers	38
– Calrose (brown)	76	Sweet potato	48	Peanuts	14
Noodles (instant)	47	Swede	72	Popcorn	55
Pasta		Yam	51	Potato crisps	57
– Egg fettucine	32			Sausages	28
– Ravioli (meat)	39	**Legumes**		Soup	
– Spaghetti (av)	41	Baked beans (av)	48	– Lentil	44
– Vermicelli	35	Broad beans (av)	79	– Pea	66
Taco shells	68	Butter beans (av)	31	– Tomato	38
		Chickpeas (av)	33		
Bread		Haricot beans (av)	38	**Confectionery**	
Bagel	72	Kidney beans (av)	27	Chocolate	49
Croissant*	67	Lentils (av)	28	Jelly beans	80
Crumpet	69	Soya beans (av)	18	Life Savers™	70
Fruit loaf (white)	47			Mars™ bars	68
Kibbled barley bread (av)	45	**Fruit**		Muesli bars	61
Mixed grain bread (av)	45	Apple (av)	36		
Oat bran bread (av)	44	Apricot (dried)	31	**Sugars**	
Pita bread	57	Banana (av)	53	Honey	58
Rye bread		Cherries	23	Fructose	20
– Kernel, e.g., pumpernickel	50	Grapefruit	25	Glucose	100
– Flour, e.g., blackbread	76	Grapes	43	Lactose	57
White bread (av)	70	Kiwifruit	58	Maltose	105
Wholemeal bread (av)	77	Mango	51	Sucrose	65
		Orange (av)	43		
Crackers/crispbread		Papaya (paw paw)	56		
Jatz	55	Peach			
Kavli	71	– Canned in juice	30		
Puffed crispbread	81	– Fresh	28		
Ryvita	69	Pear (av)	36		
Sao	70	Pineapple	66		
Watercracker	78	Plum	24		

mainly as glucose molecules; (2) storage as glycogen in liver and muscle cells; (3) storage as triacylglycerol (TAG), mainly in the liver; and (4) conversion of glucose into C_3 storage as precursors in the liver where they are used as substrates for gluconeogenesis. The latter is a futile cycle, but it seems that large amounts of glucose undergo this transformation.

During the 4 h after a typical meal, the amount of ingested carbohydrate far exceeds the amount of glucose that can be oxidized in the cells. As a result, most of the dietary glucose is stored as glycogen in liver and skeletal muscles and is subsequently released and oxidized within the next 12 h.

Insulin stimulates glucose oxidation by enhancing glucose transport into insulin-sensitive cells. Inside the mitochondria, insulin stimulates glycolysis at several steps and activates pyruvate dehydrogenase complex, the port of entry of glucose-derived acetyl-CoA into the tricarboxylic acid (TCA) cycle (**Figure 3**). Insulin also inhibits lipolysis in adipose tissue, thereby lowering plasma free fatty acid levels (FFA) and reducing lipid oxidation in the muscle cells. The inhibition of fat oxidation is directly related to reduced plasma FFA and mirrors the stimulation of glucose oxidation.

The body's glycogen reserves are small (250–500 g in a 50–70 kg adult human), although the capacity to store more can be developed by exercise, training, and diet. A normal diet provides about 200–280 g carbohydrate a day. Thus within any 24 h period, there is total oxidation of absorbed dietary carbohydrate, including sucrose. Other metabolic pathways for disposal of dietary carbohydrate, such as conversion into TAG or nonessential amino acids, are not quantitatively important.

Plasma Glucose and Insulin Responses

After a meal containing sucrose, the plasma glucose rises, reaching a peak within 15–30 min, and returns to baseline within 2 h. The classification of carbohydrates as sugars or starches does not predict the magnitude of this response. In the past it was assumed that refined sucrose caused a more rapid rise in blood glucose levels than starchy foods or naturally occurring sources of sugars like fruit. This view has been shown to be incorrect. Most starchy foods, including potatoes, bread, and many packaged breakfast cereals, are digested and absorbed rapidly and the blood glucose response is almost as high as that seen with an equivalent load of pure glucose. Foods containing refined sucrose, such as soft drinks and ice cream, have been shown to give moderate rises in blood glucose. Furthermore, the glycemic response to foods containing refined sugars is similar to that of foods containing naturally occurring sugars.

The 'glycemic index' approach has been used to classify foods according to their ability to raise the level of glucose in the blood. Foods are tested in equivalent carbohydrate portions according to standardized methodology. On a scale where glucose = 100, the glycemic index of refined sucrose (= 65) is similar to that of white bread (= 70). **Table 3** shows the glycemic index of a range of common foods. Refined sucrose elicits an insulin response commensurate with its glycemic response, i.e., it does not stimulate inappropriately high insulin secretion.

See also: **Carbohydrates**: Chemistry and Classification; Regulation of Metabolism; Requirements and Dietary Importance. **Dental Disease**. **Diabetes Mellitus**: Classification and Chemical Pathology; Dietary Management. **Fructose**. **Glucose**: Chemistry and Dietary Sources; Metabolism and Maintenance of Blood Glucose Level. **Glycemic Index**.

Further Reading

Allsop K and Brand Miller JC (1996) Honey revisited: A reappraisal of honey in pre-industrial diets. *British Journal of Nutrition* **75**: 513–520.

Baghurst KI, Baghurst PA, and Record SJ (1992) Demographic and nutritional profiles of people consuming varying levels of added sugars. *Nutrition Research* **12**: 1455–1465.

Bolton-Smith C and Woodward M (1994) Dietary composition and fat to sugar ratios in relation to obesity. *International Journal of Obesity* **18**: 820–828.

Brand Miller J, Pang E, and Broomhead L (1995) The glycemic index of foods containing sugars: Comparison of foods with naturally occurring versus added sugars. *British Journal of Nutrition* **73**: 613–623.

COMA (1989) *Dietary Sugars and Human Disease*, Report of the Panel on Dietary Sugars, Report on Health and Social Subjects No. 37 London: HMSO.

Davis EA (1995) Functionality of sugars: Physicochemical interactions in foods. *American Journal of Clinical Nutrition* **62**: 170S–177S.

Gibney M, Sigman-Grant M, Stanton JL, and Keast DR (1995) Consumption of sugars. *American Journal of Clinical Nutrition* **62**(supplement): 178S–194S.

Gibson SA (1993) Consumption and sources of sugars in the diets of British schoolchildren: Are high-sugar diets nutritionally inferior? **6**: 355–371.

Glinnsmann WH, Irausquin H, and Park Y (1986) *Evaluation of Health Aspects of Sugars Contained in Carbohydrate Sweeteners*, Report of the Sugars Task Force 1986. Washington: U.S. Food and Drug Administration.

Jenkins DJA, Wolever TMS, Taylor RH *et al.* (1981) Glycemic index of foods: A physiological basis for carbohydrate exchange. *American Journal of Clinical Nutrition* **34**: 362–366.

Southgate D (1995) Digestion and metabolism of sugars. *American Journal of Clinical Nutrition* **62**: 203S–211S.

Truswell AS (1994) Food carbohydrates and plasma lipids—An update. *American Journal of Clinical Nutrition* **59**: 710S–718S.

Wolever TMS and Brand Miller JC (1995) Sugars and blood glucose control. *American Journal of Clinical Nutrition* **62**: 212S–227S.

Dietary Sucrose and Disease

B Caballero, Johns Hopkins University, Baltimore, MD, USA

Introduction

For decades, sucrose (the disaccharide of glucose and fructose) has been the main sweetener added to human diets, particularly for sweetened beverages. But in the past 15 years, sucrose use has been progressively displaced by high-fructose corn syrup. Furthermore, the increasing popularity of diet sodas has also displaced some of the sweetened beverages, in favor of artificially sweetened ones. Still, sucrose continues to be an important dietary energy source in many parts of the world.

The term sugars includes a variety of refined carbohydrates, and its use in the literature is not always consistent, and it may refer to all refined carbohydrates present in a food item, or only sucrose, or only added sugars. The latter is particularly important for assessing health outcomes, because it is the fraction of dietary refined carbohydrates that can obviously be manipulated over a wide range, similar to the case of added sodium. These added sugars, contrary to natural sugars present in fruits or milk, contribute no essential nutrients to the diet, but do add to energy intake, and are thus 'empty' calories.

Several mechanisms have been suggested for the possible health effects of refined sugars: (1) favoring an increase in total energy intake, possibly mediated by changes in ingestive behavior or by metabolic effects on energy regulation; (2) displacement of nutrient-rich foods, with the consequent decrease in intake of essential nutrients; and (3) metabolic effects adversely affecting glucose homeostasis. It should be noted that few studies have evaluated sucrose's effects separate from that of other refined sugars that may also be present in the diet. Furthermore, some studies do not separate added sugars from other oligosaccharides naturally present in foods.

Effects on Energy Intake and Body Weight

The preponderance of evidence does show that increased consumption of added sugars is associated with increased total energy intake. A study by the US Department of Agriculture using national survey data found that 60% of individuals in the upper quartile of sugar intake exceeded their recommended energy intake, compared with 22% in the other quartiles. Individuals who did not exceed their energy allowance did so by reducing intake of fruits, vegetables, and milk, evidence of an undesirable displacement of nutrient-rich foods by added sugars. A similar displacement effect of sugars has been shown in 6–13-year-old children. However, it should be noted that several other cross-sectional studies have found no correlation between added sugars intake and total energy intake, but these studies did not adjust for physical activity, which is an important factor determining total energy intake and possibly fluid intake as well.

Sweetened Beverages

This issue is of interest given the dramatic increase in consumption of sweetened beverages over the past decade, which have become the main contributors to added sugars intake in the US. There is evidence that sweetened beverages may be the main source through which added sugars result in higher total energy intake. For example, adolescents who consumed at least 2 cans (12 oz) of soda per day had a total energy intake of 2600 kcals, compared to 1980 in those who did not consume sodas. Similarly, a randomized trial comparing sweetened and artificially sweetened beverages in overweight individuals showed a significantly higher ad libitum dietary energy intake in those consuming sweetened beverages.

In spite of consistent evidence of an effect of added sugars on total energy intake, linking that effect with excess weight gain has been more difficult, possibly due to the many factors affecting body weight homeostasis. Furthermore, there are still relatively few well-designed and adequately powered longitudinal studies assessing body weight and adiposity in relation to added sugar consumption. Some of the few available studies report relatively modest but significant effects on body weight, specifically for sweetened beverages. An observational study in 7-year-old children reported an increase of $\sim$0.20 in BMI for each additional serving of sweetened drink over a 19-month period. Another 9-year follow-up of adolescents found a significant correlation of sweetened beverage intake with BMI, but not with body fat. Other studies in children found no correlation. Almost all cross-sectional studies found no association between sugar intake and body weight, but it is well recognized that these types of studies are not the most appropriate to test this association.

Nutrient Dilution

Adding sugars to foods will increase their energy density (kcal per unit weight), and since they do not add nutrients, this will result in a net reduction

in nutrient content per kcal. This dilutional effect is evidenced in the increasing per cent of individuals who do not meet their RDA for one or more micronutrients as their consumption of added sugars increase. In adolescents, soft drinks tend to displace nutrient-rich milk and juices. Data from the US population indicate that this dilutional effect starts at around 20% of sugars in the diet, and becomes significant at around 25%. Thus, this level has been defined as a 'maximum' acceptable intake.

Some investigators have suggested that high-fructose corn syrup (HFCS), the most widely used sweetener, may have specific adverse effects on health. The rationale is that because of its rapid absorption, HFCS generates a steeper postprandial plasma insulin response, which may have adverse long-term consequences. But whether the dramatic switch from sucrose to HFCS that occurred in the US and in other countries has played a causal role in the increase in obesity or type 2 diabetes remains unclear. Similarly, while short-term, controlled experiments suggest that humans may not fully compensate for calories consumed in fluid form, there is insufficient data to assess the potential implications of such response on total energy intake and body weight regulation in the general population.

Dental Caries

Sucrose, glucose, lactose, and fructose are excellent substrates for the first step of caries formation, which involves bacterial fermentation. This process results in acidification and subsequent demineralization of the tooth surface, allowing bacterial invasion. The more substrate there is available, the more fermentation and subsequent enamel invasion.

In spite of that clear relational pathway, the precise contribution of sucrose intake to dental caries is not simple. Several experts consider that dental hygiene is a more powerful determinant of cavity prevalence than sucrose intake. For example, the National Health and Nutrition Examination Survey III (NHANES III) from the US showed no correlation between sucrose intake and dental caries in people under the age of 25 years, who were born after widespread use of fluoride. Conversely, the association is found in older people, before fluoridation was common. In studies in the UK, the correlation between socioeconomic status and caries was found to be three times that between sugar intake and caries, indicating a strong effect of dental hygiene and health practices in general. Thus, while the role of sucrose in the causative pathway of dental caries is unquestionable, it seems clear that there are other modulating factors that have come to the forefront in contemporary society, namely the use of fluoride and better oral hygiene practices. Nevertheless, sucrose continues to be one of the factors involved in caries formation, perhaps with more relevance for younger children. This is also affected by the food source, for example, hard candies that remain in the oral cavity for some time may allow longer periods of exposure than other sources of sucrose.

Type 2 Diabetes

Several prospective, observational studies have found no association between sugar intake or total carbohydrate intake and type 2 diabetes. Similarly, US national survey data (NHANES III) show no correlation between carbohydrate intake and HgbA1C. However, a recent analysis of the longitudinal Nurses' Health Study in the US reported an association between consumption of sweetened beverages and risk of type 2 diabetes in women. At least one study found an inverse association, i.e., higher sugar intake linked to reduced rates of diabetes.

Summary

While some of the putative adverse effects of sucrose require further study, there is consistent evidence that added sugars facilitate excess energy intake, particularly when consumed as sweetened beverages, while contributing no essential nutrients to the diet. Given that obesity is a major global public health problem, most experts advise to reduce consumption of added sugars as a means to avoid excess weight gain. Even acknowledging that the issue is still a matter of controversy, a recent WHO report has recommended that added sugars should not exceed 10% of total calories. The US dietary reference intakes established a maximum of added sugars of 25% of calories, based on the dilutional effect that added sugars have on micronutrient intakes, and on the obligatory glucose requirements of the brain and other organs. Nevertheless, all the experts are consistent in recommending that most of the carbohydrates in a healthy diet should be of the complex type, for which there is solid evidence of benefits to health maintenance and disease prevention.

See also: **Adolescents**: Nutritional Requirements. **Beverages**. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Fructose**. **Glucose**: Chemistry and Dietary Sources. **Sucrose**: Nutritional Role, Absorption and Metabolism.

Further Reading

Britten P, Basiotis PP, Davis CA, and Anand R (2000) Is intake of added sugars associated with diet quality? Center for Nutrition Policy and Promotion, US Department of Agriculture. *Nutrition Insight* **21** (available online at http://www.usda.gov/cnpp/Insights/Insight21.PDF).

Colditz GA, Manson JE, Stampfer MJ, Rosner B, Willett WC, and Speizer FE (1992) Diet and risk of clinical diabetes in women. *American Journal of Clinical Nutrition* **55**: 1018–1023.

US Department of Agriculture (2004) *Dietary Guidelines for Americans 2005*. Washington, DC: US Department of Agriculture. (available at http://www.usda.gov/cnpp/DG2005).

(2002) *Dietary Reference Intakes for Energy, Carbohydrates, Fiber, Fat, Fatty Acids, Cholesterol, Protein and Amino Acids*. Washington DC: National Academies Press.

Gibson SA and Williams S (1999) Dental caries in preschool children: association with social class, toothbrushing habit, and consumption of sugars and sugar-containing foods. *Caries Research* **33**: 101–113.

Harnack L, Stang J, and Story M (1999) Soft drink consumption among US children and adolescents: Nutritional consequences. *Journal of the American Diet Association* **99**: 436–441.

Heller KE, Burt BA, and Eklund SA (1995) Sugared soda consumption and dental caries in the United States. *Journal of Dental Research* **103**: 42–45.

Janket SJ, Manson JE, Sesson H, Buring JE, and Liu S (2003) A prospective study of sugar intake and risk of type 2 diabetes in women. *Diabetes Care* **26**: 1008–1015.

Ludwig DS, Peterson KE, and Gortmaker SL (2001) Relation between consumption of sugar-sweetened drinks and childhood obesity: a prospective, observational analysis. *The Lancet* **357**: 505–508.

Mrdjenovic G and Levitsky DA (2003) Nutritional and energetic consequences of sweetened drink consumption in 6–13-year-old children. *Journal of Pediatrics* **142**: 604–610.

Newby PK, Peterson KE, Berkey CS, Leppert J, Willett WC, and Colditz GA (2004) Beverage consumption is not associated with changes in weight and body mass index among low-income preschool children in North Dakota. *Journal of the American Diet Association* **104**: 1086–1094.

Sugar *see* **Carbohydrates**: Chemistry and Classification; Regulation of Metabolism; Requirements and Dietary Importance; **Galactose**. **Glucose**: Chemistry and Dietary Sources; Metabolism and Maintenance of Blood Glucose Level; Glucose Tolerance. **Sucrose**: Nutritional Role, Absorption and Metabolism; Dietary Sucrose and Disease

SUPPLEMENTATION

Contents

Dietary Supplements

S S Percival, University of Florida, Gainesville, FL, USA

In 2004, global sales of dietary supplements represented a significant business. Worldwide sales have been estimated at $70–250 billion. The demand for herbal products worldwide increased at an annual rate of 8% from 1994 to 2001, although this growth has slowed in recent years.

Issues and controversies in the dietary supplement market are related to defining exactly what is a dietary supplement, understanding how sales and marketing data are derived, defining the regulatory environment, safety issues, product quality issues, labeling and health claim issues, and scientific evidence for benefit. This article describes some of these controversies and provides examples to illustrate these issues.

How Is the Sales Data Derived?

Global sales have been estimated to be between $70 billion and $250 billion. This approximately 3-fold difference in estimates is due to the variation in what products are actually included in product sales results. As will be discussed, the definition of dietary supplements varies greatly from country to country; therefore, deriving sales data is complex.

Another difficulty in assessing sales of dietary supplements is the source from which sales data are gathered. Many business surveys rely on only one or two of the following sales outlets to derive their results:

- Supermarkets and mass merchandisers
- Natural food and health food stores
- Direct sales from Internet, mail order, practitioners, and multilevel marketing
- Pharmacies and drugstore chains

What Is a Dietary Supplement? How Are They Regulated in Different Countries?

Each country has developed regulatory definitions and systems that place dietary supplements, particularly botanicals, into categories of drugs, traditional medicines, or foods. However, in the late 1980s, many countries launched major changes in regulations that may or may not have been approved at the time of this writing. Many regulations are still in draft form.

The US Congress defined the term 'dietary supplement' in the Dietary Supplement Health and Education Act (DSHEA) of 1994. A dietary supplement is a product, taken orally, that contains a 'dietary ingredient' that is intended to supplement the diet. The dietary ingredient includes vitamins, minerals, herbs or other botanicals, amino acids, a dietary substance for use by man to supplement the diet by increasing the total dietary intake (e.g., enzymes or tissues from organs or glands), or a concentrate, metabolite, constituent, or extract. Dietary supplements may be found in many forms, such as tablets, capsules, softgels, gelcaps, liquids, or powders. They may also be produced in other forms, such as a beverage, spread, or bar, in which case information on the label must clearly state that the product is a dietary supplement and it is not represented as a conventional food or a sole item of a meal or diet.

Whatever their form, DSHEA places dietary supplements in a special category under the general umbrella of 'foods,' not drugs, and requires that every supplement be labeled a dietary supplement and carry a Supplement Facts Label.

In the United Kingdom, there is a distinct separation of food supplements and herbal medicines. The Food Standards Agency developed the Food Standard Act of 1999 and is responsible for protection of public health. The Food Supplement Directive 2002/46/EC, which harmonizes European Community legislation on food supplements, was published in 2002. This directive is stricter than existing UK standards and regulations but is relatively more liberal than what exists in other European countries. The directive defines the term 'food supplements,' contains a list of vitamin and mineral sources that may be used in the manufacture of food supplements, states labeling requirements, and, in the future, will provide a framework for maximum and minimum levels for vitamins and minerals in food supplements. Herbals and botanicals are not discussed in this directive.

The Foods Supplement Directive defines a food supplement as any food the purpose of which is to supplement the normal diet and which is a concentrated source of a vitamin or mineral or other substance with a nutritional or physiological effect, alone or in combination, and is sold in dose form. Dose form means capsules, pastilles, tablets, pills, and other similar forms, and also powders, ampoules, drops, or other similar forms of liquids or powders, designed to be taken in small measured quantities. Because the directive defines a food supplement as something to supplement the diet, products that are not meant to supplement the diet (e.g., a weight loss product) are outside the scope of the regulations. There remains a complex legal area between food supplements and medicinal products, although the directive indicates that if a product is used for treating or preventing disease, or restoring, correcting, or modifying a physiological function, then it falls under the Medicines Directive 2001/83/EEC, Medicines Act 1968, or Medicines for Human Use Regulations 1994.

The *Trans*-Atlantic Business Dialogue (TABD) approved a position statement regarding dietary supplements in 2002. The TABD is a group of corporations that promote closer commercial ties between the European Union and the United States. This position statement established industrywide consensus on standards and definition of permissible claims, as well as defining what is necessary for substantiation of those claims. In keeping with the Foods Supplement Directive, the TABD dealt only with vitamins and minerals, with the understanding that some of the conclusions may be revisited when warranted for herbals, botanicals, or other dietary supplements.

Herbal medicines, on the other hand, are regulated by the Medicine and HealthCare Products Regulatory Agency based in London. A herbal remedy is defined as

> a medicinal product consisting of a substance produced by subjecting a plant or plants to drying, crushing or any other process, or of a mixture whose sole ingredients are two or more substances so produced, or of a mixture whose sole ingredients are one or more substances so produced and water or some other inert substance.

There are two alternative regulatory routes in the United Kingdom for herbal medicines:licensing and exemption from licensing requirements:

- Licensed herbal medicines:To receive a product license prior to marketing, herbal medicines are required to meet safety, quality, and efficacy criteria in a similar manner to any other licensed medicine.
- Herbal remedies exempt from licensing requirements:The exemption applies to herbal remedies meeting certain conditions set out in Section 12 of the Medicines Act 1968. Section 12 allows a person to make, sell, and supply a herbal remedy during the course of his or her business provided the remedy is manufactured or assembled on the premises and that it is supplied as a consequence of a consultation between the person and his or her patient. Section 12 also allows the manufacture, sale, or supply of herbal remedies where the processing of the plant consists only of drying, crushing, or comminuting; the remedy is sold without any written specification as to its use; and the remedy is sold under a designation that only specifies the plant and the process and does not apply any other name to the remedy.

Canada has been estimated to have approximately 3% of the market share of the global nutritional market. Health Canada established the Office of Natural Health Products. Premarket assessment, labeling, licensing, and monitoring of herbal supplements are items in its mandate. The definition of a natural health product includes products for the use in "diagnosis, treatment, mitigation, or prevention of a disease, disorder, or abnormal physical state or its symptoms in humans; restoring or correcting organic function in humans; or modifying organic functions in humans, such as modifying those functions in a manner that maintains or promotes health." These products include homeopathic preparations, substances used in traditional medicine, a mineral or trace element, a vitamin, an amino acid, an essential fatty acid or other botanical-, animal-, or microorganism-derived substance. Foods are not included in this product category called natural health products. Canada's Food and Drugs Act of 1953 regulates foods and drugs but does not specifically deal with natural health products. Therefore, these types of products are regulated as either a food or a drug depending on the type and concentration of active ingredient and whether claims are made on the products.

Germany regulates vitamins and minerals as food if they are sold to complement the nutritive value of the diet and do not exceed safe levels. However, if the vitamin or mineral is used for disease treatment or prevention and is used at pharmacological levels, then it is considered a drug. Safety and efficacy of drugs must be established by clinical research. Medicinal plants are regulated differently depending on what plant and in what form it is sold. In general, extracts of plants are considered drugs and must be prescribed. Teas, on the other hand, are sold over-the-counter in pharmacies. Other teas, such as those that contain alkaloids, must be sold by prescription only. Beginning in 1980, an extensive analysis of the literature on more than 300 herbal remedies was undertaken by the German Kommission E. Approximately two-thirds of the herbals were listed as safe and at least minimally effective. The results were published as a series of monographs by the German Kommission E, and this body of work was summarized and translated into English by the American Botanical Council. These substances are generally purchased at the pharmacy and are reimbursable through health insurance. One caveat regarding the German herbal preparations is that they are not likely to be the same preparations that are produced by other countries; thus, the safety and efficacy statements in the Kommission E are only for the preparations that are prepared in German pharmacies.

Australia regulates therapeutic goods under the Therapeutic Goods Act of 1989. Therapeutic goods include vitamins, minerals, plants and herbals, nutritional food supplements, naturopathic and homeopathic preparations, and some aromatherapy. The Therapeutic Goods Administration (TGA) developed the Office of Complementary Medicine to evaluate new substances and products. Basically, the TGA regulates these therapeutic goods as they do pharmaceutical products, and thus their criteria are more rigorous than the criteria of other countries. Most of the therapeutic goods are 'generally listed' rather than regulated. Listed medicines are considered to be relatively harmless, so the regulations allow for manufacturers to 'self-assess' their products in some situations. The majority of listed medicines are self-selected by consumers and used for self-treatment, and they are all manufactured with well-known established ingredients, such as vitamin and mineral products or sunscreens. These are assessed by the TGA for quality and safety but not efficacy. This does not mean that they do not work; rather, it means that the TGA has not evaluated them individually to determine if they work. It is a requirement under the act that sponsors have information to substantiate all of their product's claims.

The Japanese Ministry of Health and Welfare does not define or recognize a distinct category known as dietary supplements. Instead, there are only two

Table 1 Regulatory categories of different countries

Country, act	*Definition*
United States, DSHEA	Vitamins, minerals, herbal, other botanical, amino acid, enzymes, organs, glands
Europe, Food Supplement Act	Vitamin and minerals
United Kingdom, Medicine and Health Care	Medicinal plants
Canada, Office of Natural Products	Mineral; trace element; vitamin; amino acid; essential fatty acid; botanical-, animal-, or microorganism-derived substances; homeopathic preparation; traditional preparations
Germany, Kommission E	Vitamin and mineral as both foods and drugs, botanicals (approved and not approved), teas as prescription and as over-the-counter
Australia, Therapeutic Goods Administration	Vitamin and mineral, plants, herbs, nutritional food supplements, naturopaths and homeopathic preparations, aromatherapies
Japan, Ministry of Health and Welfare	No definition of dietary supplements, regulations for foods, drugs, and Kampo

classifications, food and drugs. In 1993, Japan defined a group of foods known as Foods for Specific Health Use (FOSHU). As of 2004, approximately 342 foods had been approved as FOSHU. The dietary ingredients are sold in the form of foods, not in the form of capsules, tablets, or powders.

The herbal supplements market in Japan has been strongly influenced by the practice of Kampo. Kampo (or Kanpo) is the adaptation of Chinese herb formulas to Japanese medicine. Approximately 25 years ago, the Japanese Ministry of Health formally recognized that certain traditional Chinese herb formulas (and a few formulas of similar nature developed in Japan) were suitable for coverage by national health insurance. These formulas are prepared in factories under strict conditions.

In summary, developing global data on dietary supplement sales depends on how they are defined. **Table 1** summarizes the differences in regulatory categories of different countries.

Product Quality and Safety Issues

Product quality is an issue derived from the explosive growth of the industry in the post-DSHEA world. Quality issues revolve around products that contain wrong ingredients, incorrect claims, contamination, or incorrect amounts—either too much or not enough.

An example plant misidentification was published in 1998 by Slifman *et al*. Two patients were admitted to hospital emergency rooms with palpitations, vomiting, nausea, and chest pressure, among other symptoms. Both individuals, having been admitted 1 month apart, had each consumed a program of dietary supplements, one containing 14 herbs, a tablet containing 11 herbs, liquid clay, a bulking powder, and capsules containing microorganisms. Of the five supplements, the one made up of 14 herbs tested positive for cardiac glycosides. The investigators determined that *Digitalis lanata* was present in the supplement. *Digitalis lanata* contains cardiac glycosides, which resulted in the cardiac symptoms. Further investigation revealed that raw material labeled as plantain (genus *Plantago*) had been contaminated with *D. lanata* due to misidentification in the field.

Another quality issue that has safety manifestations was an incorrect claim on a product. PC-SPES, a combination of eight herbs, is claimed to be a nonestrogenic treatment for prostate cancer. However, several of the herbs used in this preparation do in fact have estrogenic activity. In 1998, DiPaola *et al.* showed a significant amount of estrogenic activity in both *in vitro* (yeast) and *in vivo* studies (mice and humans) with PC-SPES. Use of the supplement by men with prostate cancer resulted in similar side effects as would develop with estrogen therapy and theoretically could confound the results of standard therapy.

By law (DSHEA), the manufacturer is responsible for ensuring that its dietary supplement products are safe before they are marketed. Unlike drug products that must be proven safe and effective for their intended use before marketing, there are no provisions in the law for the US Food and Drug Administration (FDA) to 'approve' dietary supplements for safety or effectiveness before they reach the consumer. Also unlike drug products, manufacturers and distributors of dietary supplements are not required by law to record, investigate, or forward to the FDA any reports they receive of injuries or illnesses that may be related to the use of their products. Under DSHEA, once the product is marketed, the FDA has the responsibility to show that a dietary supplement is 'unsafe' before it can take action to restrict the product's use or remove it from the marketplace.

In 2003, the FDA banned all products containing ephedra alkaloids. Ephedra-containing products were, until the ban, marketed in conjunction with enhancing athletic performance

and/or promoting weight loss. Recent studies provided enough additional evidence that ephedra presents a significant and unreasonable risk of illness and injury that the FDA banned all ephedra-containing products from the market and advised consumers to stop taking such supplements. Strong statements were issued cautioning about the use of ephedra-containing products, especially when strenuously exercising or in combination with other stimulants, such as caffeine.

Interactions

An issue that has become of concern is the interaction of dietary supplements with herbs and other dietary supplements, drugs, foods, lab tests, and diseases or other conditions. There are literally hundreds of potential interactions that have not yet been recognized. Both practitioners and consumers must be aware of the possibilities. In some cases, knowledge about interactions comes from documented reports. However, in other cases, the knowledge is theoretical, based on the pharmacological profile or mechanism of action of the supplement and the drug, food, test, or condition. For example, ginkgo biloba contains ginkgolides in the leaf that competitively inhibit platelet-activating factor (PAF). PAF inhibition decreases platelet aggregation among other many other physiological effects. Inhibition of PAF may increase cardiac contractility and coronary blood flow. Concomitant use of herbs and supplements that affect platelet aggregation could theoretically increase the risk of bleeding in some people due to ginkgo's effects on platelet aggregation. Spontaneous hematomas (broken blood vessels) and hemorrhaging in the anterior chamber of the eye have been reported in ginkgo users, although it is not known what other drugs or supplements these individuals were taking.

Herbs and supplements that promote platelet inhibition include angelica, anise, capsicum, celery, chamomile, clove, fenugreek, feverfew, fish oil, garlic, ginger, horse chestnut, horseradish, licorice, meadowsweet, onion, Panax ginseng, red clover, vitamin E, and willow. Similarly, concomitant administration of drugs, including aspirin, clopidogrel (Plavix), dalteparin (Fragmin), enoxaparin (Lovenox), heparin, indomethacin (Indocin), ticlopidine (Ticlid), and warfarin (Coumadin), may increase the risk of bleeding in some people. This is just one example of the interactions between drugs with herbals and herbals with other herbals. There may be an infinite number of interactions.

Currently, there are no mandated US federal guidelines to report adverse events or consumer health complaints associated with the use of dietary supplements. MedWatch reporting is voluntary. In 2004, the Life Sciences Research Office published a report, *Recommendations for Adverse Event Monitoring Programs for Dietary Supplements*.

Label Claims

Label claims regarding dietary supplements are a complex issue that varies from country to country. Yet no matter what specific claims are allowed or disallowed by a country, it is reasonable to assume that any global regulation requires that the claim be true, not misleading, and be clear to the consumer. A summary of US label claims follows.

The Nutrition Labeling and Education Act (NLEA) was passed in 1990 as a result of a pre-1984 FDA position that prohibited making any therapeutic or disease-related claims on a food or dietary supplement label. The NLEA permits certain claims describing a positive relationship between a supplement and a health-related condition (or disease). These claims are considered 'health claims' in order to distinguish them from nutrient content claims. A health claim must be authorized by the FDA, and the FDA can only authorize a claim if there is "significant scientific agreement among qualified experts" or by the 1997 amendment that permits a manufacturer to rely on a statement from an "authoritative scientific body" of the US government or the National Academy of Sciences. This is a rigorous assessment and only 14 claims have been authorized to date.

In addition to health claims, dietary supplement labels are permitted to have qualified health claims or structure-function claims. The rationale behind the development of a qualified health claim was the idea that the First Amendment should allow disclaimers to be considered as solutions to making claims nonmisleading (Pearson v. Shalala). In other words, the First Amendment does not allow the FDA to reject health claims unless it shows that disclaimers would fail to remedy harm from misleading statements. The criteria for a qualified health claim were released in 2003 and in this context the FDA will not take enforcement action against a manufacturer using the following specified qualifiers provided the FDA is satisfied that the qualifiers are not misleading:

- "Although there is scientific evidence supporting the claim, the evidence is not conclusive."
- "Some scientific evidence suggests .... However, FDA has determined that this evidence is limited and not conclusive."

- "Very limited and preliminary scientific research suggests .... FDA concludes that there is little scientific evidence supporting this claim."

Qualified health claims for dietary supplements recognized by the FDA as part of its enforcement discretion include such examples as the relationships between phosphatidylserine and cognitive function, B vitamins and cardiovascular disease, omega-3 fatty acids and cardiovascular disease, selenium and cancer, and antioxidant vitamins and cancer.

Dietary supplements are not permitted to carry labeling statements that imply such issues as 'cure,' 'mitigate,' 'treat,' or 'prevent disease' because these statements are considered within the definition of a drug and drugs are subjected to a rigorous premarket approval process. However, under DSHEA, structure–function claims are permitted on dietary supplements because dietary supplements may have effects on the structure or function of the body without the implication that they act as a drug and/or are related to disease. Structure–function claims include those that describe the role of the dietary supplement in affecting the structure or function in humans or the documented mechanism in which a dietary supplement acts to maintain such structure or function. In addition, dietary supplement label claims allow statements of benefits related to classical nutritional deficiency or statements regarding the general feeling of well-being derived from consumption.

Potential Benefits of Dietary Supplements

The 2000 *Dietary Guidelines for Americans* (new release due 2005) emphasizes choosing foods sensibly, maintaining a healthy weight, and exercising regularly. It acknowledges that some people may need a vitamin–mineral supplement to meet specific needs. Similarly, the Food and Nutrition Board and the American Dietetic Association also recognize that dietary supplements may be desirable for some nutrients and for some individuals. The following is a compilation of recommendations by these groups:

- Folic acid supplements for women of childbearing age due to the risk of neural tube defects
- Vitamin B_{12} supplements for people older than age 50 years due to inefficient absorption
- Vitamin B_{12} supplements for vegans who eat no animal products
- Calcium for people who seldom eat dairy products
- Vitamin D for elderly people who do not consume fortified dairy products and for others with little exposure to sunlight
- Iron supplementation for pregnant women
- Multivitamin–mineral supplement for people who are following a severely restricted weight-loss diet

Specifically for athletes, the position of the American Dietetic Association, Dietitians of Canada, and the American College of Sports Medicine is that physical activity, athletic performance, and recovery from exercise are enhanced by optimal nutrition. These organizations recommend appropriate selection of food and fluids, timing of intake, and supplement choices for optimal health and exercise performance. In sports, athletes who are at greatest risk of micronutrient deficiencies are those who restrict energy intake or use severe weight-loss practices, eliminate one or more food groups from their diet, are sick or recovering from injury, or consume high-carbohydrate diets with low micronutrient density. In practice, athletes should consume diets that provide at least the RDAs/DRIs for all micronutrients from food. It follows that, in general, no vitamin and mineral supplements are required if an athlete is consuming adequate energy from a variety of foods to maintain body weight. Supplementation may be necessary under conditions of inadequate food intake. Athletes, as for the general population, should follow supplementation recommendations unrelated to exercise, such as folic acid in women who may become pregnant.

Conclusions

One of the difficulties in assessing the nature of the worldwide dietary supplement industry and its regulations is largely in understanding what products are considered dietary supplements. In the United States, only pills, capsules, tablets, and the like are considered dietary supplements. Globally, it is sometimes difficult to discuss dietary supplements without discussing functional foods or nutraceuticals. Functional foods are similar in appearance to conventional foods but have demonstrated physiological benefits beyond the traditional nutritional value. Nutraceuticals may to go so far as to declare not only health benefits but also medical benefits that reduce the risk of chronic disease beyond basic nutritional functions. Canada regulates functional foods, nutraceuticals, and dietary supplements under one regulatory agency. The United States clearly distinguishes between foods and dietary supplements, although both fall under the category of food, which is distinct from drugs. The United

Kingdom distinguishes between herbal medicines and dietary supplements containing vitamins and minerals. Japan regulates functional foods as FOSHU and has no regulatory definition for dietary supplements as defined in the United States. Moreover, these regulations are in a constant state of flux as the industry changes and develops over time. Issues that must be monitored regarding dietary supplements consumption are product quality and potential harmful interactions among supplements, foods, and drugs. Health claims that have been approved by regulatory agencies worldwide stress that the claims be truthful, clear, and not misleading to the ultimate consumer. Current scientific expertise acknowledges that dietary supplements, specifically some of the vitamins and minerals, have potential benefits in certain populations.

See also: **Folic Acid**. **Functional Foods**: Regulatory Aspects. **Pregnancy**: Dietary Guidelines and Safe Supplement Use. **Supplementation**: Role of Micronutrient Supplementation; Developing Countries; Developed Countries.

Further Reading

Anonymous (2000) Nutrition and athletic performance. Position of the American Dietetic Association, Dietitians of Canada, and the American College of Sports Medicine. *Journal of the American Dietetic Association* **100**: 1543–1556.

Huang SM and Lesko LJ (2004) Drug–drug, drug–dietary supplement, and drug–citrus fruit and other food interactions: What have we learned?. *Journal of Clinical Pharmacology* **44**(6): 559–569.

Percival SS and Turner RE (2000) Applications of herbs to functional foods. In: Wildman R (ed.) *Handbook of Functional Foods*. Boca Raton, FL: CRC Press.

Turner RE, Degnan FH, and Archer DL (2005) *Nutrition in Clinical Practice* **20**: 21–32.

U.S. Department of Agriculture (2000) *Dietary Guidelines for Americans*. Washington, DC: U.S. Department of Agriculture. Available at www.health.gov/dietaryguidelines/dgac.

Role of Micronutrient Supplementation

R D W Klemm, Johns Hopkins University, Baltimore, MD, USA

Introduction

Globally, almost two billion people (one-third of the human race) are affected by vitamin A, iron, iodine, and/or zinc deficiencies that put them at an increased risk of poor growth, morbidity, intellectual impairment, and/or mortality. Since the mid-1980s micronutrient supplementation has been a major public-health strategy in developing countries to prevent and control deficiencies in vitamin A, iron, and, to a lesser extent, iodine. More recently, zinc supplementation has come to be considered as an efficacious adjunctive therapy for diarrhea in populations with an elevated risk of zinc deficiency. This article will define micronutrient supplementation, examine the role of supplementation as a strategy for the prevention and control of micronutrient deficiencies, and examine evidence for vitamin A, iron, iodine, and zinc supplementation interventions with respect to efficacy, recommended dose, frequency of administration, safety, and program effectiveness.

Definition of Micronutrient Supplementation

Supplementation refers to the provision of added nutrients in pharmaceutical form (such as capsules, tablets, or syrups) rather than in food. Micronutrients are substances required by the body in small amounts for vital physiological functions. They cannot be synthesized by the body and therefore must be consumed in foods and/or in supplements.

Choice of Interventions

Micronutrient supplementation is one of three major categories of nutrition intervention strategies – the other two being fortification and dietary change. The choice of strategy or mix of strategies will depend on multiple factors including the magnitude, severity, and distribution of the micronutrient deficiency in the population, the relative intervention efficacy, the in-country resources available to deliver the intervention to the target group effectively, the target groups' acceptance of the intervention, and the ability to sustain the intervention.

The theoretical relative advantages of micronutrient supplementation over fortification and dietary-improvement interventions include rapid coverage of a high-risk population, the ability to provide directly a controlled and concentrated dose of the micronutrient(s) to the target group, an immediate impact on micronutrient status and associated functional outcome(s), the relatively low cost of training workers compared with nutrition counselling for diet improvement, and high coverage if supplements are delivered using existing services that already reach a high proportion of the target group. Most supplementation

programs have been shown to be cost-effective in achieving their nutritional goals and health impacts, although sustaining large-scale programs over the long-term may be more costly than either fortification or dietary improvement.

Generally, prophylactic micronutrient supplementation is intended as a short-term means of rapidly preventing nutrient deficiency in high-risk individuals and populations until adequate and sustainable food-based programs become effective. However, in many cases, supplementation programs may be the only effective means of reaching specific vulnerable groups, particularly those who have limited or no access to processed fortified foods or those, such as young children and pregnant women, who have high micronutrient requirements that may not be met even with fortification and dietary-improvement interventions. In these situations and populations, supplementation should be sustained over a longer period until nutrient intake from fortified and non-fortified food is adequate.

Based on experiences from vitamin A, iron, and iodine supplementation programs, the key limitations of supplementation are inadequate targeting or coverage (where deficient individuals are missed or reached irregularly), an inability to sustain high coverage over long periods of time as financial, political, or other health priorities change, and poor compliance by target individuals who are expected to take a daily supplement for extended periods of time (e.g., iron supplementation during pregnancy). As illustrated in **Figure 1**, in many countries – particularly those with high regional variability in socio-economic status, food availability, and market-access – a mix of strategies, rather than any single strategy, is more likely to reach a greater proportion of the at-risk population.

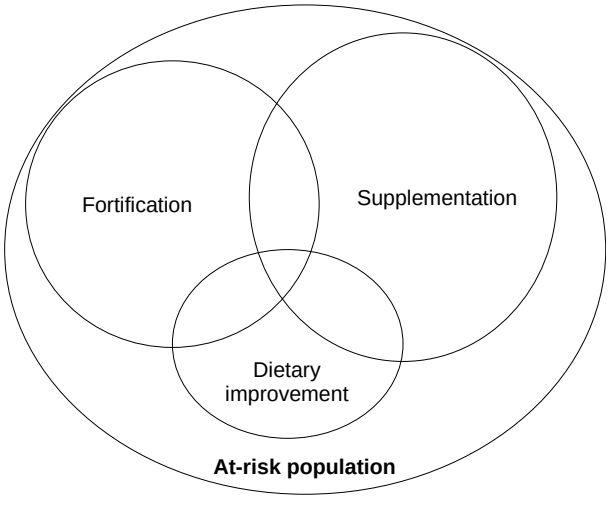


Figure 1 Micronutrient supplementation in combination with other nutrition-focused interventions to prevent micronutrient deficiencies in a target population.

Cost of Micronutrient Interventions

The World Bank's World Development Report 1993 found micronutrient programs to be among the most cost-effective of all health interventions. The cost of micronutrient supplementation needs to be balanced against the cost of other food-based and public-health interventions as well as against the cost of not addressing the insidious effects of micronutrient deficiencies. Costs are likely to vary depending on the scope of the program, existing delivery mechanisms, the nutrient involved, and other factors. Based on World Bank estimates, the costs of vitamin A, iron, and iodine supplementation programs are relatively modest, ranging from US$0.20 to US$1.70 per beneficiary per year. These costs are slightly higher than the estimated relative costs of fortification (US$0.05–0.15 per beneficiary per year) but are considerably lower than the unit costs of education programs (US$5–10) and feeding programs (US$70–100 per beneficiary per year).

Prophylactic Micronutrient Supplementation

Micronutrient supplementation has been the method of choice for the treatment of severe clinical nutrient deficiencies for several decades. Prophylactic supplementation, however, gained wider acceptance only in the late 1980s with the publication of results from a randomized trial in Aceh, Indonesia, showing a 34% reduction in young-child mortality among preschoolers given vitamin A supplements. The introduction of routine vitamin A supplementation to preschool children in developing countries has encouraged this approach and the development of other micronutrient supplementation programs. Each single-nutrient or multiple-micronutrient supplementation strategy should be evaluated separately for efficacy, feasibility, safety, cost, and appropriateness for the cultural and political context in which it will be implemented.

Vitamin A Supplementation

Periodic distribution of high-dose vitamin A supplements, either universally to all preschool children or to targeted high-risk groups, has been the most widely practiced intervention for the prevention and treatment of vitamin A deficiency throughout the world. Giving a high dose of vitamin A every

4–6 months is based on the assumption that vitamin A is stored in the liver and mobilized, as needed, to meet the demands of target tissues. This is in contrast to iron supplements, which need to be given on a daily or weekly basis.

Efficacy of Prophylactic Supplementation

Preschool children Giving children living in areas where vitamin A deficiency is prevalent a large dose (i.e., 200 000 IU or 60 mg Retinol Equivalent (RE)) of vitamin A every 4–6 months has been shown to reduce the risk of both noncorneal and corneal xeropthalmia by 90%, increase serum retinol levels for 1–2 months following supplementation, and reduce young-child mortality by an average of 23% when coverage levels of at least 80% are achieved in deficient populations. In populations deficient in vitamin A, prophylactic supplementation of preschool children is one of the most cost-effective public-health interventions to improve the survival of children aged between 6 months and 6 years.

Post-partum women High-dose vitamin A supplementation is also recommended for post-partum mothers within 6 weeks of delivery, when the chance of pregnancy is remote, because the physiological demands of pregnancy and lactation deplete the mother's vitamin A stores. The provisional recommendation of the International Vitamin A Consultative Group is to give two doses of 200 000 IU at least 24 h apart to all women living in areas where vitamin A deficiency is prevalent and to give the supplement as soon as possible after delivery in order to maximize the beneficial effects on maternal vitamin A status, breast-milk vitamin A concentrations, and subsequent infant vitamin A status. Spacing the two doses by at least a day minimizes the risk of raising breast-milk retinoic acid (a short-lived metabolite of vitamin A) to potentially toxic levels.

Newborns and young infants Two randomized controlled trials in Asia that provided 50 000 IU of vitamin A to infants in the first 2 days of life found significant reductions in infant mortality among the vitamin A supplemented newborns. A third, confirmatory, trial is currently underway in Bangladesh, which, depending on the findings, may lead to newborn dosing recommendations. The current World Health Organization (WHO) recommendation is to provide 50 000 IU of vitamin A to infants with each of the three doses of diphtheria, pertussis, and tetanus at 6, 10, and 14 weeks of age to improve vitamin A status. Further trials, however, are needed to confirm the benefit of implementing this recommendation for early infant vitamin A status, morbidity, and survival.

Pregnant women Where maternal night blindness or biochemical vitamin A deficiency is highly prevalent, prophylactic supplementation with up to 10 000 IU daily or 25 000 IU weekly has been given safely. The efficacy of low-dose maternal supplementation is still under investigation, but a recent community-based randomized placebo-controlled study in Nepal reported a 40% reduction in pregnancy-related maternal mortality among women given low doses of vitamin A or β-carotene, provided as a weekly supplement, during pregnancy and for 3 months post-partum. This is the first trial to link vitamin A supplementation and maternal survival. Confirmatory efficacy trials are underway in Bangladesh and Ghana and should provide guidance on future maternal vitamin A supplementation recommendations.

Form of Supplement

Vitamin A in the form of a gelatinous capsule is the overwhelming choice of delivery mode used in large public-health programs, although there are reports of the successful use of liquid vitamin A in a bottle using a calibrated dispenser. Vitamin A has also been dispensed from an inhalation device in children with parasitic infections as an alternative delivery mode in the very small number of children who exhibit extreme intestinal malabsorption of vitamin A.

Safety

High doses of vitamin A are safe and well accepted by preschool children, although evidence from program evaluations and a randomized trial in the Philippines suggests that up to around 9% of preschool children may experience acute transient side-effects including nausea, vomiting, headache, and/or fever after dosing. Most episodes begin within 24 h of capsule receipt and resolve spontaneously within 12–24 h of onset.

Earlier animal experiments linked high doses of vitamin A to birth defects; however, experimental data proving the teratogenic effect of vitamin A in pregnant women are limited and, for ethical reasons, very difficult or even impossible to obtain rigorously. Nonetheless, high doses of vitamin A should be avoided during pregnancy because of the theoretical risk of teratogenesis.

Delivery Mechanisms

There are a variety of ways to deliver high-dose vitamin A supplements to at-risk populations including restricting delivery to clinic settings for

treatment purposes, integrating delivery with existing services such as immunization contacts or routine growth monitoring, or universal delivery, to attain the widest coverage of preschool children, through semi-annual campaigns that specifically promote vitamin A capsule distribution or that are combined with other national programs such as national immunization days or child-health weeks. Each delivery mode has advantages and disadvantages. Restricted delivery targets those most likely to be deficient in vitamin A and requires few additional resources (apart from the supplements); however it may result in poor coverage if those who are most at risk do not regularly access health clinics. 'Piggybacking' vitamin A distribution onto existing community services can be cost-effective but may also miss the children at greatest risk of vitamin A deficiency if their access to and use of these services is limited. Finally, universal distribution requires strong community mobilization and social marketing to attain coverage levels of at least 80%. Sustaining this coverage level every 4–6 months can be challenging, but there are numerous examples of countries where such levels have been sustained for at least 5–8 years.

Iron Supplementation

Globally, supplementation with iron tablets is the most widely used strategy for the prevention and control of iron-deficiency or anemia in pregnancy. Pregnant women require nearly three times as much iron as non-pregnant women owing to the physiological demands of pregnancy (expanded red-blood-cell volume, the needs of the fetus and placenta, and blood loss at delivery). This high requirement is unattainable by most pregnant women in developing countries, especially those who struggle to meet the $1.5\,\mathrm{mg\,day^{-1}}$ requirement when not pregnant, and therefore iron supplementation is recommended during pregnancy.

Efficacy

Pregnancy The rationale for iron supplementation during pregnancy in developing countries is based on a combination of considerations including the high prevalence of anemia in pregnancy (the majority of which is probably associated with iron deficiency), carefully conducted trials that show that consuming iron tablets during pregnancy improves maternal iron status, the higher maternal mortality risks associated with severe anemia, and the postulated risks of iron deficiency in pregnancy (i.e., increased risk of fatigue, cardiovascular stress, impaired resistance to infection, and poor tolerance to heavy blood loss and surgical interventions at delivery) and for fetal development. Although evidence supports the efficacy of iron supplementation in improving the iron status of pregnant women, no trials have examined the impact of iron supplementation on maternal mortality in severely anemic women. Also, there is a lack of causal evidence from controlled studies linking mild-to-moderate iron-deficiency anemia – which is much more prevalent than severe anemia – with an increased risk of low birth weight, preterm delivery, or obstetrical or perinatal complications.

Infancy Iron supplementation in infants is sometimes advised to prevent iron deficiency, even in populations with a relatively low prevalence of iron-deficiency anemia. The US Institute of Medicine, for example, recommends iron drops for exclusively breast-fed infants between 4 and 6 months of age. There is ample evidence from well-designed and controlled studies to show that iron supplementation in infancy significantly improves hemoglobin and ferritin levels, and studies are currently investigating the impact of iron supplementation on dimensions of cognitive development. The benefits and risks of infant iron supplementation, however, remain controversial, particularly in iron-replete children. This is because, although iron is an essential nutrient for adequate infant growth, immune function, and development, it may also contribute to a greater risk of infection if the excess iron increases a pathogen's access to free iron for its own growth and reproduction. Some studies have reported a higher prevalence of diarrhea in iron-supplemented infants, which calls into question the appropriateness of existing hemoglobin and ferritin cut-offs for defining true deficiency in infants and points to the need to clarify the cut-off issue in order to determine an appropriate age for starting iron supplementation.

Low-birth-weight infants Low-birth-weight infants are born with low iron stores and have higher iron requirements for growth. Their iron needs cannot be met from breast milk alone, and, therefore, they are a priority target for iron supplementation.

Preschooler and school-age children Several, but not all, placebo-controlled supplementation trials have demonstrated that iron supplements improve hemoglobin concentrations in preschoolers in developing countries, and there is substantial evidence that iron supplementation of anemic children improves their school performance and verbal and other skills.

Dose

The WHO has published global guidelines for iron supplementation and recommends daily prophylactic iron supplementation with 60 mg of iron for all women in developing countries in the second and third trimesters of pregnancy (**Table 1**). In other countries, iron supplementation is recommended only for women with proven iron-deficiency anemia (in Great Britain) or for women with low pre-pregnancy iron stores (in Canada). The efficacy of maternal iron supplementation increases with daily iron doses of up to 60 mg. The WHO also recommends providing low-birth-weight infants with supplemental iron drops from 2 months of age.

Table 1 Micronutrient supplementation: target groups and prevention schedules

Micronutrient	*Target group*	*Dosage*	*Frequency and duration*
Vitamin A[a]	Children at risk of vitamin A deficiency		
	<6 months	50 000 IU	One dose at 4, 10, and 14 weeks[e]
	6–11 months	100 000 IU	One dose every 4–6 months[f]
	1–5 years	200 000 IU	One dose every 4–6 months
	Post-partum women	400 000 IU[g]	One dose before 8 weeks post-partum
Iron (plus folate)[b]	Pregnant women (living in areas where anemia prevalence is less than 40%)	60 mg iron and 400 μg folic acid[i]	Daily for 6 months[h] in pregnancy
	Pregnant women (living in areas where anemia prevalence is at least 40%)	60 mg iron and 400 μg folic acid[i]	Daily for 6 months[h] in pregnancy, and continuing to 3 months post-partum
	6–24-month-old children of normal birth weight (living in areas where the prevalence of anemia in children is less than 40%)	12.5 mg iron and 50 μg folic acid	Daily to 6–12 months of age
	6–24-month-old children of normal birth weight (living in areas where the prevalence of anemia in children is greater than or equal to 40%)	12.5 mg iron and 50 μg folic acid	Daily to 6–24 months of age
	2–24-month-old children of low birth weight (less than 2500 g)	12.5 mg iron and 50 μg folic acid	Daily to 2–24 months of age
Iodine[c]	Pregnant women in areas where iodine deficiency is endemic[j]	300–480 mg	One dose annually
	Non-pregnant fertile women[j]	400–960 mg	One dose annually
	Children in areas where iodine deficiency is endemic[j]	240 mg iodine	One dose annually
Zinc[d]	Children with persistent diarrhea[k]	10–20 mg	Daily for 14 days
	Children with an elevated risk of zinc deficiency; children who are severely stunted, or have low plasma zinc, or both	Further research needed on relative efficacy of different frequencies and doses	

[a]Adapted from World Health Organization (1997) *Vitamin A Supplements: A Guide to Their Use in the Treatment and Prevention of Vitamin A Deficiency and Xerophthalmia*, 2nd edn. Geneva: World Health Organization.
[b]Adapted from Stoltzfus RJ and Dreyfus JL (1998) *Guidelines for the Use of Iron Supplements to Prevent and Treat Iron Deficiency Anemia*. INACG ILSI Press, Washington DC.
[c]Adapted from World Health Organization (1996) Safe Use of iodized oil to prevent iodine deficiency in pregnant women – a Statement by the WHO. *Bulletin of the World Health Organization* **74**: 1–3.
[d]Adapted from the recommendations of an expert group: Fontaine O (2001) Effect of zinc supplementation on clinical course of acute diarrhoea. *Journal of Health Population and Nutrition* **19**: 339–346.
[e]Give at the time of each of the three diphtheria–pertussis–tetanus vaccinations.
[f]Immunization against measles provides a good opportunity to give one of these doses.
[g]Provisional recommendation of the International Vitamin A Consultative Group, Annecy, France, 30 October–2 November, 2000.
[h]If 6 months' duration cannot be achieved in pregnancy, continue to supplement during the post-partum period for 6 months or increase the dose to 120 mg iron in preganancy.
[i]Where iron supplements containing 400 μg of folic acid are not available, an iron supplement with less folic acid may be used. Supplementation with less folic acid should be used only if supplements containing 400 μg are not available.
[j]Where access to iodine-fortified salt is limited and immediate attention is needed.
[k]In areas where there is an elevated risk of zinc deficiency in the population.

Multiple Micronutrient Supplements with Iron

Currently, folic acid is added to most iron supplements for women of fertile age because it reduces the risk of neural-tube defects and because lack of folic acid may limit the hemoglobin response to iron supplements. In the absence of these nutrients such as vitamin A, vitamin B_{12}, and riboflavin may also limit the efficacy of iron supplements, and studies are underway to assess the effect of multiple micronutrient supplements on anemia.

Safety

Iron supplements can cause unpleasant gastrointestinal symptoms (e.g., nausea, constipation, vomiting, and diarrhoea), which may contribute to poor compliance, but these usually occur at higher doses. When iron tablets are taken with meals or if slow-release tablets are used, any side-effects may be mitigated. Complications of excessive iron storage, including hemochromatosis and hemosiderosis, are possible but uncommon in women consuming iron tablets. Another potential danger of iron supplements is accidental overdosing by children in the home, and therefore supplements should be kept out of the reach of children.

Frequency

A perceived concern about the side-effects, compliance, and potential toxicity of a daily regimen of iron supplementation generated research to assess the relative efficacy of weekly versus daily supplementation in pregnant women, adolescents, and children. A review of these studies concluded that both daily and weekly iron supplementation reduces the prevalence of iron deficiency and anemia, daily supplementation is more effective than weekly supplementation in increasing hemoglobin and ferritin, and while daily supplementation produces only a slightly higher average hemoglobin response (approximately $2\,g\,l^{-1}$) than weekly supplementation, its relative impact on reducing anemia risk is 34%, largely because daily supplementation is more effective at increasing low hemoglobin concentrations.

From the results of two other studies, in Bangladesh and Indonesia, that carefully monitored the number of iron tablets consumed, it appears that the size of the hemoglobin response to iron appears to depend on the total amount of iron consumed. In these studies, most of the hemoglobin response was produced by the first 20–50 tablets consumed. But more research is needed before recommendations can be made about consuming a fixed number of tablets over a defined period of time while permitting flexibility about the consumption interval (i.e., daily, two or three times per week, weekly, etc.).

Taken together, the available evidence suggests that iron supplements should be taken daily to treat iron-deficiency anemia, especially in pregnant women. Weekly supplementation may offer a more feasible preventive strategy, particularly if it reduces costs, improves compliance, and reduces side-effects; however, more information is needed to assess the relative effectiveness of daily versus weekly supplementation under program conditions.

Form of Iron

In tablets, the most common form of iron is ferrous sulphate (which contains 20% iron), but ferrous fumarate (33% iron) and ferrous gluconate (12% iron) are also used. Infant supplements are usually in liquid form and more costly, but when crushed or mixed with food, tablets can also be used.

Effective Iron-Supplementation Programs

Reviews of large-scale iron-supplementation programs in developing countries have reported limited effectiveness in reducing maternal anemia. The limited effectiveness is often attributed to implementation constraints including low compliance, short intervention duration, inadequate supplement supply, or poor coverage of pregnant women. For iron-supplementation programs to achieve improved effectiveness, careful attention must be given to ensuring an adequate supply of iron tablets at the distribution points, access of pregnant women to the distribution points, promotion of the benefits of iron supplementation, counselling about managing possible side-effects, and communication strategies to encourage pregnant women to consume the supplement.

Iodine Supplementation

Salt iodization is the recommended means of population-based intervention to prevent and control iodine deficiency disorders, but, in isolated communities with an urgent need for iodine prophylaxis, direct supplementation of priority groups can be rapidly implemented as an interim measure while salt iodization is being established.

Efficacy

Intramuscular iodine injections Numerous studies have confirmed that iodine supplementation by injection before a woman becomes pregnant can prevent endemic cretinism and that a single injection can prevent goitre for up to 3–4 years. Other documented benefits of maternal supplementation observed

in several controlled studies include reduced infant and young-child mortality, improved birth weight, and better manual function in children born to iodine-supplemented mothers.

Oral iodized oil Although efficacious, injections of iodized oil have largely been replaced by oral iodized oil owing to the concern over the AIDS pandemic and use of needles as well as the higher cost of supplies (syringes) and personnel (skilled injectors). Oral delivery of iodized oil appears to be as effective as intramuscular injection but is less costly, carries no infection risk (through a contaminated needle), is painless, and can be administered by untrained personnel. Oral iodized oil is considered to be safe for pregnant women and can be given at any time during pregnancy; however, it appears to protect against moderate and severe neurological abnormalities in the infant only when given during the first two trimesters. The best outcomes are likely to occur when supplementation is given during the first trimester, but even if it is given in late pregnancy or to the infant after birth slight improvements in brain growth and developmental quotients, but not neurological status, are evident.

Dose

Owing to the fact that damage to the developing brain is the most severe consequence of iodine deficiency, women of child-bearing age and children are the first priorities for receiving iodized oil. Recommendations are to give a single dose of 460 mg of iodine to all females below the age of 40 years. A single annual dose of 240 mg of iodine appears to be adequate for children. Larger doses do not necessarily provide longer protection because of increased urinary loss after administration. It is possible that smaller more frequent doses may be more effective, although this issue requires further study. Evidence from studies using 200–500 mg of iodized oil suggests that the protective effect lasts for between 6 months and 2 years.

Safety

Oral iodine supplementation is safe, although side-effects can include transient submandibular swelling and subclinical hypothyroidism.

Zinc Supplementation

During the past decade, zinc supplementation has received increasing attention as results from research trials reveal the extent of zinc deficiency in developing countries and the role of zinc supplementation in reducing intrauterine growth retardation and disease incidence and severity and improving children's cognitive development, growth, and survival. However, most of what is known about zinc supplementation is based on the results of research trials and not of large-scale programs that deliver zinc supplements.

Efficacy

Diarrhoea and pneumonia Short-term treatment of diarrhoea and possibly pneumonia with zinc supplementation has proven efficacious in numerous randomized controlled trials. A pooled analysis of nine randomized controlled trials reported an 18% lower incidence and a 25% reduction in prevalence of diarrhoea in zinc-supplemented children regardless of age, baseline zinc status, wasting prevalence, or sex, suggesting that zinc supplementation may benefit many subgroups of children living in areas at high risk of zinc deficiency. Studies have also investigated the efficacy of a combination of micronutrients given together with zinc and have shown that zinc alone is just as efficacious as a multiple micronutrient that includes zinc in reducing the severity of acute diarrhoea.

There is growing evidence that zinc supplementation reduces the risk of pneumonia. A pooled analysis of five randomized controlled trials reported a 34% reduction in the incidence of pneumonia in zinc-supplemented children, but evidence from three short-course zinc-supplementation trials suggests a non-significant reduction in the incidence of pneumonia and in hospitalization rates for acute lower respiratory infection.

Other outcomes Evidence from a limited number of trials suggests a potential benefit of zinc supplementation on morbidity related to *Plasmodium falciparum* infections, child survival, weight gain in low-birthweight infants and severely malnourished children, length gain in stunted children, and a host of maternal health and pregnancy outcomes; however, more research is required to determine the benefits of a large-scale zinc-supplementation program targeted at groups of infants, children, and pregnant women.

Dose

There is a need for systematic studies to determine the appropriate dose of supplemental zinc for preventing zinc deficiency in different age groups and under different clinical conditions. However, based on therapeutic studies, giving zinc supplements in doses ranging between one and four times the recommended dietary allowance per day (15 $mg.day^{-1}$ for children aged less than 1 year, and 20–30 $mg.day^{-1}$ for children aged more than 1 year) for 14 days is efficacious in reducing the severity of

diarrhoea and the duration of the episode significantly. In hospitalized children, zinc can be given as two or three divided doses each day, although in community interventions a single dose of $20\,mg.day^{-1}$ appears both safe and efficacious. Studies are underway to determine the feasibility and efficacy of adding zinc to oral rehydration solution.

Form of Zinc

There are many zinc compounds that can be used to produce zinc supplements. They differ in color (from colorless, to white, grey, or yellowish white), taste (bitter, astringent, slightly sour, or bitter), odour (odourless, vanilla odour, or faint odour of acetic acid), solubility in water (insoluble, slightly soluble, or soluble), cost, side-effects, and safety. Water-soluble compounds (e.g., zinc acetate, zinc sulfate, and zinc gluconate) are more easily absorbed and therefore preferred. Zinc supplements have been prepared in flavored syrups, chewable tablets, single-dose 'sachets' to be added to food, as a high-fat spread to be consumed alone or with other food for infants and young children, and as dry supplements (tablets, capsules, or powders) alone or with other nutrients. The choice of supplement form will depend on the age of the target group, preference, and whether other nutrients will be included.

Effectiveness of Zinc-Supplementation Programs

There is little information about the effectiveness of zinc-supplementation programs implemented on a large scale. There is a need to conduct such studies to assess the best ways to deliver zinc supplements to children with diarrhoea, paying particular attention to the feasibility, sustainability, and cost-effectiveness of different zinc-delivery mechanisms.

Summary

The feasibility of micronutrient supplementation and the degree to which it should be pursued in combination with other strategies to prevent and control micronutrient deficiencies depend on the local needs, resources, capabilities, commitment, and evidence of benefit. The successful prevention and control of vitamin A, iron, and zinc deficiencies will probably result from a combination of repetitive distribution of high-dose nutrient supplements, fortification of staple foods, and behavioral change, whereas fortification of salt alone with iodine has already achieved much success in combating iodine deficiency disorders. The adoption of supplementation approaches should be guided by evidence of a need for targeting, impact potential, costs, and potential sustainability.

See also: **Folic Acid**. **Iodine**: Physiology, Dietary Sources and Requirements; Deficiency Disorders. **Iron**. **Pregnancy**: Nutrient Requirements; Prevention of Neural Tube Defects. **Supplementation**: Developing Countries; Developed Countries. **Vitamin A**: Deficiency and Interventions. **Zinc**: Deficiency in Developing Countries, Intervention Studies.

Further Reading

Allen LH (2002) Iron supplements: scientific issues concerning efficacy and implications for research and programs. *Journal of Nutrition* **132**(suppl 4): 813S–819S.

Angermayr L and Clar C (2004) Iodine supplementation for preventing iodine deficiency disorders in children (Cochrane Review). In: *The Cochrane Library*, Issue 2, 2004. Chichester: John Wiley & Sons, Ltd.

International Zinc Nutrition Consultative Group (IZiNCG) (2004) Assessment of the risk of zinc deficiency in populations and options for its control. Hotz C and Brown KH (eds.) *Food and Nutrition Bulletin* **25**(suppl 1): S91–S204.

Mahomed K (2004) Iron and folate supplementation in pregnancy (Cochrane Review). In: *The Cochrane Library*, Issue 2, 2004. Chichester: John Wiley & Sons, Ltd.

Mahomed K and Gülmezoglu AM (2004) Maternal iodine supplements in areas of deficiency (Cochrane Review). In: *The Cochrane Library*, Issue 2, 2004. Chichester: John Wiley & Sons, Ltd.

Sommer A and West KP Jr (1996) *Vitamin A Deficiency: Health, Survival and Vision*. New York: Oxford University Press.

Stoltzfus RJ and Dreyfuss ML (1998) *Guidelines for the Use of Iron Supplements to Prevent and Treat Iron Deficiency Anemia*. International Nutritional Anemia Consultative Group (INACG)/WHO/UNICEF. Geneva: World Health Organization.

World Health Organization (1996) Safe use of iodized oil to prevent iodine deficiency in pregnant women: a statement by the World Health Organization. *Bulletin of the World Health Organization* **74**: 1–3.

World Health Organization (1997) *Vitamin A Supplements: A Guide to Their Use in the Treatment and Prevention of Vitamin A Deficiency and Xerophthalmia*, 2nd edn. Geneva: World Health Organization.

Developing Countries

R Shrimpton, Institute of Child Health, London, UK

Micronutrient supplementation is the distribution of specially formulated preparations of one or more nutrients, usually in the form of a pill, a capsule, or syrup. It seems to be the Cinderella of nutrition

interventions, more than capable of dancing but not quite good enough to be invited to the Ball. It is often described as a 'short-term' option and a 'medical' approach and considered more appropriate for the treatment of severe micronutrient deficiencies in those most affected than to prevent deficiencies in whole populations. However, for the half of humanity affected by micronutrient deficiencies, the overwhelming majority of whom are the poor concentrated in the developing world, solving these problems through food-based approaches is only likely to happen in the very long term. The immune system is compromised by vitamin A deficiency in 40% of children younger than 5 years old in the developing world, leading to approximately 1 million deaths each year. In the 6- to 24-month-old age group, mental development is impaired due to iron deficiency in 40–60% of the developing world's children. Severe iron deficiency also causes more than 60 000 deaths of women during pregnancy and childbirth every year. Approximately 18 million infants per year are born mentally impaired as a result of iodine deficiency during pregnancy. Providing vulnerable groups, such as children and women of childbearing age, with low-cost vitamin and mineral supplements is the least that governments can do to protect the growth and development of the next generation as a first step toward realizing the right of every individual to be adequately nourished.

Experience in achieving high coverage of those most at risk with micronutrient supplements is quite varied, with both successes and failures. A good communication strategy is an essential part of achieving high levels of adherence in micronutrient supplementation programs, but these aspects are not particular to nutrition programs and are not considered here. Deficiencies of iodine, iron, vitamin A, and folate are the most commonly recognized deficiencies for which there are programs, but in practice most of those affected have multiple vitamin and mineral deficiencies that overlap and interact at great cost. This article reviews the policy dimensions of the efforts to establish programs aimed at eliminating iodine deficiency, iron deficiency anemia, and vitamin A deficiency through supplementation, and it provides a perspective on zinc supplementation and multiple micronutrient supplementation as future components of nutrition programs in developing countries.

Iodine Supplementation

Today, approximately 70% of the world's salt is iodised, compared to just 10% in 1990, and therefore iodine supplementation programs are greatly reduced. Until universal salt iodization is guaranteed in the third of the world in which iodized salt is not yet available, especially in remote populations in which goiter is endemic, supplements should be used during pregnancy and early childhood. In the past, it was common to provide annual intramuscular injections of iodized oil to women of reproductive age in order to ensure iodine status during the first months of pregnancy when the risk of cretinism is greatest. In more recent years, oral iodized oil capsules have proven to be as efficacious and more effective in controlling iodine deficiency in both women of reproductive age and schoolchildren. Oral iodine supplements initially based on expensive poppy seed oil have since been replaced by cheaper rapeseed and peanut oil preparations, which are equally effective.

Vitamin A Supplementation

The use of supplements to eliminate vitamin A deficiency is a notable success, with remarkable advances achieved within the past decade. Although the elimination of vitamin A deficiency by year 2000 was one of the goals set at the World Summit for Children in 1990, little progress was evident at mid-decade. Clinical vitamin A deficiency was estimated to affect approximately 3.3 million children younger than the age of 5 years in 1995, with an additional 100 million subject to subclinical deficiency. The periodic distribution of high-dose vitamin A supplements, originally employed in Indonesia during the 1970s for the prevention of blindness in children, was shown in the 1980s to also impact young child mortality. The supplements have the advantage of ensuring requirements for 4–6 months after administration, such that two or three capsules per year can meet vitamin A requirement of preschool children.

The lack of perception of vitamin A deficiency as a problem was a substantial barrier to establishing large-scale preventive supplementation programs. The prevalence of clinical signs of frank vitamin A deficiency, such a Bitot's spot and corneal lesions, that make it a 'public health problem' is very small at just 0.5%. Since clinical signs are often more common in rural populations, a significant vitamin A deficiency problem can easily go undetected. National representative surveys were thus a prerequisite for taking action. Another barrier is the voice of those who advocate for food-based approaches and view supplements as technical fixes or golden bullets that are of questionable sustainability promoted by the pharmaceutical sector. In reality, of course, these are not either/or options.

Convincing proof of the efficacy of vitamin A capsules for child mortality reduction in the early 1990s helped to create increased momentum for populationwide preventive supplementation programs. The turning point for increasing the coverage of vitamin A supplements was undoubtedly the publication of a meta-analysis of the efficacy trials of massive-dose vitamin A capsules. The analysis of eight mortality trials indicated that improving the vitamin A status of children aged 6 months to 5 years by massive-dose capsule distribution reduced child mortality rates by approximately 23%. The important conclusion of the meta-analysis was that increased risk of mortality from vitamin A deficiency was not just limited to those portions of the population with severe vitamin A deficiency problems but was present across the whole population distribution.

What consisted of 'the justification' for carrying out vitamin A supplementation programs evolved rapidly during the latter half of the 1990s. Many of these discussions were held at the meetings of the International Vitamin A Consultative Group and the working group on vitamin A of the Standing Committee on Nutrition of the United Nations. A broad technical consensus was finally accepted that even in the absence of survey data, it was highly likely that the benefits of vitamin A supplements would be evident in populations in which the mortality rates for those younger than 5 years old were higher than 70 per 1000. Prior to this, vitamin A supplements were targeted to those children with illnesses such as measles and diarrhea. The most recent programmatic recommendations are that if mortality for those younger than 5 years old is higher than 50 per 1000, then supplements should be employed routinely as a preventive measure for all young children. Subsequent to this consensus, a global policy to integrate vitamin A capsule distribution into regular immunization schedules, and also to incorporate vitamin A capsules into the national immunization campaign days being promoted to achieve the eradication of polio, was rapidly adopted.

Programmatic vitamin A interventions received considerable impetus from the Vitamin A Global Initiative, an informal interagency advocacy group that worked to promote the adoption of vitamin A supplementation programs. The initiative included WHO and UNICEF, together with CIDA from Canada, DIFID from the United Kingdom, USAID from the United States, and the Micronutrient Initiative (MI). Through their networks, these various organizations worked together to convince governments with high mortality rates for children younger than age 5 years to introduce periodic vitamin A capsule distribution programs. Vitamin A capsules were made available by CIDA through UNICEF to any developing country that wanted them, and UNICEF and MI with USAID and DIFID funds developed a global communication campaign.

By the end of the 1990s, vitamin A supplementation programs had seen a remarkable expansion. Most countries with high mortality rates for children younger than 5 years old adopted vitamin A supplementation programs, with the most notable exception being India. The number of countries with vitamin A programs increased from 10 in 1995 to 72 in 2000. The ways in which the vitamin A capsule programs were developed and implemented varied by country, but the most common strategy was to use national immunization days for polio eradication to piggyback vitamin A supplements. The use of this approach doubled from 30 countries in 1997 to 60 in 1999. Because the polio eradication strategy requires two nationwide campaigns not more than 2 months apart, some countries also promoted separate micronutrient days, or child health days, so that children would get at least two capsules during the course of a year, 6 months apart. UNICEF procured through its central warehouse in Copenhagen and supplied through its country programs an average of 289 million capsules per year from 1993 to 1998, which was estimated to be only 38% of the worldwide need.

Estimates of the coverage of vitamin A capsules indicate a remarkably high coverage of supplements by the turn of the century, with remarkable saving of life. Based on multiple sources, UNICEF estimates that in 1999 half of all children aged 6–59 months in developing countries outside of China, and 80% of such children in the least developed countries, received a vitamin A capsule within the past 6 months. Coverage was highest in sub-Saharan Africa, where 70% of children aged 6–59 months received a capsule in the past 6 months. Extrapolation of the protective effect of a 23% reduction in child mortality shown by the meta-analysis to the increased coverage of capsules achieved between 1998 and 2000 suggests that 1 million lives were saved in this short period.

The challenge that remains for vitamin A supplementation is one of sustainability. Although supplements are traditionally viewed as a short-term solution, in reality they need to be maintained during at least the medium term if continued gains in mortality reduction are to be realised. Increases in other sources of vitamin A, be it through diet and/or fortification, are unlikely to be achieved in the short

term. The eventual phasing out of national immunization days, as polio eradication becomes a reality, will cause problems for maintaining the high coverage of vitamin A capsules. Alternate strategies are needed and are being put in place in many countries. Bangladesh and Nepal are two examples of countries that successfully promote biannual micronutrient days with large-scale social mobilisation efforts. Sustaining the provision of the vitamin A capsules is also likely to become a problem. Until now, supplements have predominantly been provided by the Canadian government and supplied through UNICEF, and how long this will be sustained is not known. The costs for individual governments to take on are small, however, and the benefits in terms of lives saved will likely remain enormous for many decades.

Iron/Folate Supplementation

Although iron deficiency is the most widespread of nutritional problems, supplementation with iron has not proven to be a very successful intervention. Global policy recommendations to routinely provide iron/folate supplements for women during pregnancy and lactation have changed little in almost three decades, and all anemic pregnant women should receive such supplements in almost all contexts. Approximately half of the developing countries in the world are reported to have national iron supplementation policies. The World Summit for Children's goal to reduce anemia in women by one-third was given little or no priority by the principal actors involved such that no progress was made during the past decade. Anemia still affected 44% of nonpregnant women and 56% of pregnant women in developing countries at the end of the twenty first century.

Although there is ample evidence that iron deficiency and the anemia associated with it are a great burden on society, especially the poor, the advocacy base for pushing for program implementation is still weak. The link of iron deficiency to maternal and child survival has not been concretely proven. The effect of iron deficiency on cognitive deficits in children and on adults later in the life course has been established. The absolute losses in Southeast Asia are estimated to be approximately $5 billion annually, and for India the median value of productivity losses due to iron deficiency alone is approximately $4 per capita or 0.9% of gross domestic product. The efficacy of iron/folate supplements for controlling anemia is well documented, and there is a considerable amount of descriptive evidence linking maternal anemia to both low birth weight and maternal mortality.

Despite high cost-effectiveness, little or no priority has been given to iron deficiency anemia reduction programs. At $0.002 per tablet, the iron supplement is relatively cheap, and the cost per disability adjusted life year of $13 makes the supplementation of pregnant women with iron a very cost-effective intervention. At the national level, despite the existence of national policies, rarely is there a budget for the provision of supplements and/or supervision of iron deficiency anemia programs. Although UNICEF is a major supplier of iron/folate supplements to the developing world, the level of supply is far lower than that believed to be needed. In the period 1993–1996, 2.7 billion tablets were shipped to 122 countries at a cost of $7.5 million as part of UNICEF assistance to programs aimed at eliminating maternal anemia. This was less than 5% of that needed to cover all pregnancies in developing countries. There have been few, if any, attempts to gauge the coverage of iron/folate supplements at any level, be it district, national, or international. Neither has there been any effort put into creating political accountability to ensure high coverage.

Many meetings and publications during the past few decades that have examined the causes and solutions of iron deficiency anemia conclude that lack of effectiveness of iron supplementation programs for anemia control is largely related to problems with supply of the supplement. Although the side effects of iron pills are often cited as the reason why iron supplementation programs do not work, this rarely seems to be the case. One of the major causes of nonadherence seems to be lack of understanding of the benefits the supplements can bring among health staff that deliver the tablets. Most of the program reviews have concluded that where supportive community-level delivery mechanisms are put in place that encourage adherence, and the supply of supplements is ensured, high levels of coverage can be achieved and sustained. It is often the case, however, that in health systems in developing countries, nutrition is everybody's business and nobody's responsibility, and iron supplements have ended up low on the list of things to do.

Despite an international consensus that supplementation has a key role to play in the control of iron deficiency anemia, there are still those that question such programs. In 1998, a technical consensus meeting on what was needed to solve the problem of iron deficiency made the recommendation that although the interventions already existed for reducing both iron deficiency and iron deficiency anemia, more work was needed to develop large-

scale programs with packages of interventions delivered through multiple sectors. Despite the consensus, there are still those who question the advisability of iron supplementation programs, suggesting that hemoglobin cutoffs may be set too high and/or that receiving excessive amounts of iron is dangerous for those who are iron replete. Another complicating factor is undoubtedly related to the fact that adequate coverage of iron supplements alone will not ensure anemia control in many settings. A global review of anemia causality revealed that perhaps only half of anemia is solely due to iron deficiency, with other micronutrient deficiencies such as vitamin A contributing as well. Infections such as malaria and helminth are also important causes. Programs to eliminate anemia thus require packages of interventions, of which iron supplements are but a part.

Zinc Supplementation

There is strong evidence for the efficacy of therapeutic zinc in improving the prognosis of children being treated for diarrheal disease, and a new WHO/UNICEF recommendation is to give supplemental zinc for 10 days as part of the treatment of diarrhea. A pooled analysis of randomized controlled therapeutic zinc trials in children with diarrhea showed that zinc-supplemented children with acute diarrhea had a 15% lower probability of continuing diarrhea on a given day, and in those with persistent diarrhea there was a 24% lower probability. In addition, children with persistent diarrhea had a 42% lower rate of treatment failure or death if zinc supplemented. The WHO/UNICEF recommendation is to give zinc in the form of a tablet for 10 days to all children that are treated for diarrhea. Given that even the current interventions included in child health programs for diarrheal disease treatment, such as oral rehydration therapy, face enormous barriers to achieving and maintaining high levels of coverage, the challenge for achieving high levels of coverage of zinc supplements in the treatment of diarrhea is likely to be considerable. If these efforts are successful, however, then the impact is likely to be great. The most effective way to give preventive zinc supplements is an ongoing research question.

Multiple Micronutrient Supplementation

In recent years, the case has increasingly been made for providing multiple micronutrient supplements instead of iron supplements for young children and women of reproductive age in developing countries. A woman's or an infant's diet that is deficient in iron is likely to be deficient in many other micronutrients. Outside of emergency situations, such as natural catastrophes, famine, and civil strife, poor dietary quality rather than quantity is the determinant of inadequate micronutrient status among infants and women. The nutrient-to-energy ratios of iron, zinc, folate, vitamins B_6 and B_{12}, vitamin A, riboflavin, and calcium are commonly below the recommended levels needed, assuming energy needs are met.

Although the incremental cost of distributing a multiple micronutrient supplement is likely to be small, the increased benefits may be large. The main cost of the delivery of a nutrient supplement for women of reproductive age is not the cost of the supplement but the cost of the delivery system. Although it may not be working very well, a delivery system already exists for the iron/folate supplements that could be used to provide these other micronutrients. There has been much speculation about the costs of a multiple micronutrient supplement compared to the iron/folate tablet currently procured and provided by UNICEF. Adding the extra nutrients to the iron/folate tablets will not add more than 20% to the cost of the tablet, as long as they are procured in bulk on the international market, as is the case for the current iron/folate tablets supplied by UNICEF.

The potential benefits for pregnant women from improving not only iron and folate but also zinc, vitamin A, and other antioxidant nutrient status are likely to be great. Providing women with vitamin A supplements together with iron supplements has been shown to improve the hematinic effects of the iron supplements. The findings that vitamin A supplements to women of reproductive age reduced maternal mortality by approximately 40% in Nepal and that zinc supplementation improved birth weight among poor women in the United States point to the possible multiple benefits, beyond anemia reduction, of introducing a multiple micronutrient supplement for use by women in developing countries.

The composition of a multiple micronutrient supplement for use in trials among pregnant and lactating women in developing countries has recently been agreed on. The proposed formulation contains physiological doses of the micronutrients based on the recommended daily allowance (RDA). The US/Canadian recommendations were used since they are the most recent and best documented. The selection of nutrients included in the supplement was based on evidence of deficiencies, possible consequences of deficiencies for mother and child, weighing of risks and advantages, and interaction between

nutrients. Furthermore, information about toxicity levels, cost of nutrients, the size of the resulting supplement, and possible side effects related to supplement intake were considered. The formulation agreed upon includes 15 micronutrients (vitamins A, D, E, B_1, B_2, B_6, B_{12}, and C, niacin, and folic acid and minerals Fe, Zn, Cu, I, and Se), all at the RDA level, except for folic acid, which was included at the 400-μg level—considered sufficient to prevent neural tube defects if taken periconceptually.

The multiple micronutrient tablet formulation for supplementation during pregnancy was developed with various users in mind. For pregnant women, the tablet should be taken on a daily basis for as long as possible during pregnancy. For lactating women, the supplement should be taken daily until at least 3 months postpartum. The tablet can also be taken by adolescent girls on a once-a-week basis as a way of improving micronutrient status before getting pregnant. Another possible target group is refugees, who can take the supplement according to their biological state (pregnant or not) and, in case of severe deficiency, can take two tablets per day.

Although the supplement is not considered a dangerous product, it was still recommended for use in trials, with special attention to monitoring and evaluation. Tablets of similar composition are regularly prescribed by physicians and/or purchased by mothers in developed countries, and they can be found in the pharmacies of the capitals of most developing countries and are widely consumed by the richer segments of the population. Despite the relative safety of the supplement, it was recognized that many issues related to multiple micronutrient supplements remain to be investigated. Research topics identified included the assessment of risks as opposed to benefits of regular supplement intake in environments in which many disease agents are present. Factors that influence adherence to tablet intake were also considered a crucial area for investigation.

The need to carry out both efficacy and effectiveness studies of the multiple micronutrient supplements in various different populations is well recognized. The need for micronutrient supplementation in pregnancy is likely to be great because of widespread maternal malnutrition. However, it has to be recognized that public health resources are always limited and priority is given to interventions that are both efficacious and effective. Proving the efficacy of multiple micronutrient supplements is thus essential for being able to advocate for their widescale use. However, multiple micronutrient supplements are almost always going to be part of a package of interventions, especially in developing countries. Indeed, the micronutrient supplements will likely be most effective as part of a package that also seeks to control the major diseases afflicting the mothers, be that malaria, sexually transmitted diseases, or intestinal parasites. Being part of a package obviously makes it difficult in field conditions to test the relative merits of the individual pieces, including the micronutrient supplement. For these reasons, both 'plausibility'- and 'probability'-based approaches are recommended for measuring performance of the multiple micronutrient supplements so that experience can be gained on how to develop programs that promote fetal and infant growth and it can be determined whether micronutrient supplements are truly efficacious in developing countries.

UNICEF has incorporated the multiple micronutrient supplements into ongoing programs as part of an initiative to prevent low birth weight. A total of 18 million infants are born low birth weight (<2.5 kg) every year, accounting for approximately 14% of all live births. These infants are at increased risk of infections; to have weakened immunity, learning disabilities, and impaired physical development; and of dying soon after birth. UNICEF has started promoting programs to reduce the prevalence of low birth weight in Indonesia, the Philippines, Vietnam, China, Bangladesh, India, Pakistan, Nepal, Tanzania, Mozambique, and Madagascar. Low birth weight prevention programs are being fashioned to fit local circumstances, in accordance with the nutrition strategy approved by the UNICEF executive board. The approach being developed in each country elaborates on the 'care for women' element of the care initiative that has been developed to facilitate improvements in caring practices among families and communities. In addition to the multiple micronutrient supplements, other interventions that form part of low birth weight prevention efforts include promoting increased child spacing, increased rest and food for the mother during pregnancy, improved reproductive health, deworming, and malaria control during pregnancy as appropriate. Reducing teenage pregnancy rates is also part of the package, but it is a major challenge since, although the mortality risks for both the teenage mother and her child are known to be considerably increased, teenage pregnancy rates are very high in most developing countries. Prepregnancy weight, weight gain during pregnancy, and birth weight all receive special attention as the principal evaluative indicators of program success.

WHO and UNICEF recommend the use of syrup and/or tablets containing iron for the treatment of anemia in young children, and such products are available through UNICEF supply division in Copenhagen. These products have very little penetration considering the size of the infant anemia problem in most developing countries, where half of all children are commonly affected. Despite the recognition that iron deficiency often coexists with zinc deficiency, together with inadequate intakes of other B vitamins (B_6, riboflavin, and niacin) in infant dietaries, there is no multiple micronutrient supplement available for infants. UNICEF has also been testing the efficacy of a foodlet (a large crumbly pastille that is a cross between a tablet and a food) containing multiple micronutrients during infancy through the Infant Research on Infant Supplementation trials. Trials of multiple micronutrients as preventive supplements are also being carried out by many different groups using supplements provided in the form of sprinkles, tablets, and even as a beverage. Preliminary results of these trials point to a greater impact on anemia and enhancement of multiple micronutrient status by the multiple micronutrient supplements than that of iron supplements, as well as small improvements in growth.

There is a need to bring all of this broad spectrum of experimental and programmatic work together to reach conclusions and achieve consensus before policy and program recommendations can be made on how best to include multiple micronutrient supplements in programs to improve maternal and child health in developing countries. Whether this happens will depend on the continued efforts of the agencies interested in and responsible for promoting maternal and health, to champion the importance of micronutrient supplementation in their programs.

See also: **Anemia**: Iron-Deficiency Anemia. **Folic Acid**. **Iodine**: Physiology, Dietary Sources and Requirements; Deficiency Disorders. **Iron**. **Pregnancy**: Nutrient Requirements. **Supplementation**: Role of Micronutrient Supplementation; Developed Countries. **Vitamin A**: Biochemistry and Physiological Role. **World Health Organization**. **Zinc**: Physiology; Deficiency in Developing Countries, Intervention Studies.

Further Reading

Gillespie S, Kevanny J, and Mason J (1991) *Controlling Iron Deficiency*, ACC/SCN State of the Art Series, Nutrition Policy Discussion Paper No. 9. Geneva: SCN.

Gross R, Dwivedi A, and Solomons NW (2003) Supplement: Proceedings of the International Workshop on Multimicronutrient Deficiency Control in the Life Cycle, Lima Peru May 30–June 1, 2001. *Food and Nutrition Bulletin* **24**(3):S3–S61.

Huffman SL, Baker J, Shumann J, and Zehnaer ER (1998) *The Case for Promoting Multiple Vitamin/Mineral Supplements for Women of Reproductive Age in Developing Countries*. Washington, DC: The Linkages Project, Academy for Educational Development.

Institute of Medicine (1998) *Prevention of Micronutrient Deficiencies: Tools for Policy Makers and Public Health Workers*. Washington, DC: National Academy Press.

Mason JB, Lotfi M, Dalmiya N, Sethurman K, and Deitchler M (2001) *The Micronutrient Report: Current Progress and Trends in the Control of Vitamin A, Iron, and Iodine Deficiencies*. Ottawa, Ontario, Canada: Micronutrient Initiative, International Development Research Centre.

Ramakrishnan U and Huffman SL (2001) Multiple micronutrient malnutrition: What can be done? In: Semba RD and Bloem MW (eds.) *Nutrition and Health in Developing Countries*, pp. 365–391. Totowa, NJ: Humana Press.

Shrimpton R and Schultink W (2002) Can supplements help meet the micronutrient needs of the developing world? *Proceedings of the Nutrition Society* **61**: 223–229.

UNICEF/UNU/WHO (2001) *Iron Deficiency Anaemia: Assessment, Prevention and Control. A Guide for Programme Managers*. Geneva: World Health Organization.

UNICEF/UNU/WHO/MI (1999) *Preventing Iron Deficiency in Women and Children: Background and Consensus on Key Issues. Report of a Technical Workshop, UNICEF, New York 7–9 October 1998*. Ottawa, Ontario, Canada: Micronutrient Initiative and International Nutrition Foundation.

WHO/UNICEF/IVACG (1997) *Vitamin A Supplements: A Guide to Their Use in the Treatment and Prevention of Vitamin A Deficiency and Xerophthalmia*, 2nd edn. Geneva: World Health Organization.

Developed Countries

M F Picciano, National Institutes of Health, Bethesda, MD, USA
S S McDonald, Raleigh, NC, USA

A dietary supplement is a product that is intended to supplement the diet and contains at least one or more of certain dietary ingredients, such as a vitamin, mineral, herb or other botanical, or an amino acid. These products may not be represented as conventional foods; rather, they are marketed in forms that include capsules, tablets, gelcaps, softgels, and powders. Although manufacturers must have evidence to support their claims of a dietary supplement's safety and efficacy, US Food and Drug Administration (FDA) approval is not required before a product is marketed. Micronutrient dietary supplements (vitamins and minerals for purposes

of this discussion) are commonly purchased and consumed in developed countries, even though taking greater quantities of micronutrients than recommended may not have proven benefits for the general population and, for some micronutrients (e.g., vitamin A), may have harmful effects. It may seem logical to assume that the majority of people who live in developed countries can use food sources to obtain the amounts of micronutrients required to maintain overall good health. However, the possibility of helping to prevent chronic diseases through micronutrient supplementation is attracting the interest of many people. A rigorous research approach must be used to determine in what circumstances micronutrient dietary supplements can have beneficial, including preventive, health effects. Special attention must be given to possible differences in micronutrient requirements at different life cycle stages. These stages include infancy (birth to 12 months), childhood (1–18 years), adulthood, and older adulthood (70 years and older). Evidence supporting dietary supplementation at different lifestyle stages is summarized here for several micronutrients.

Prevalence of Micronutrient Supplement Use

In the United States, vitamins and minerals are the most widely used dietary supplements. Between 1993 and 2003, total retail sales of vitamins and minerals more than doubled, increasing from approximately $3 billion to $6.7 billion for vitamins and from approximately $0.6 billion to $1.8 billion for minerals. These figures include sales of multivitamin and multimineral combinations as well as individual vitamins and minerals. Multivitamin/mineral preparations, accounting for almost half of micronutrient purchases, consistently have been the best-selling micronutrient supplements, with sales increasing from $2.64 billion in 1997 to $3.68 billion in 2003.

Findings by several research groups show that micronutrient supplement use generally is more common among people with higher education levels, higher incomes, and better diets. Survey results in The Netherlands indicate that micronutrient supplements are used by approximately 20% of adults in that country, fewer than in the United States. Data collected in the 1988–1994 National Health and Nutrition Examination Survey (NHANES III) showed that approximately 40% of the US population 2 months of age and older (44% females versus 35% males) were taking a vitamin, mineral, or other type of dietary supplement during the month before the NHANES III interview. Data from NHANES 1999–2000 (post-DSHEA) indicate that 52% of U.S. adults were taking at least one dietary supplement. Supplement users were more likely to be toddlers and preschool-aged children and middle-aged and older adults. Across all age groups, vitamin/mineral combinations and multivitamins were the most common types of supplements used by individuals who took only one supplement. Collection of these type of data is important to monitor use, identify usage trends, and help understand the popularity of micronutrient supplement use.

Motivation for Micronutrient Supplement Use

People choose to use micronutrient supplements for various reasons. Survey data indicate that many individuals decide to take micronutrient supplements based on advice from health professionals, family, and friends. A majority of supplement users regard micronutrient supplements as 'insurance' against general poor health or becoming ill, even though they recognize that scientific evidence for this belief may be lacking. Generally, people report that they use supplements either because they think that it is difficult to consume a balanced diet or because they believe that even consuming a balanced diet cannot supply the quantity of micronutrients they need for optimal good health.

Major health reasons given for taking supplements include a sense of well-being and 'feeling better' (especially multivitamins/minerals), preventing colds and flu (especially vitamin C), preventing chronic disease (especially vitamin E and calcium), increasing 'energy,' coping with stress, and improving the immune system. Many vitamin E users believe that the vitamin helps prevent heart disease, and most calcium users know that calcium use helps prevent osteoporosis. Using micronutrient supplements is one way by which people who may be at high risk for certain diseases try to gain some degree of personal control over their health outcomes. Ironically, many individuals who take supplements regularly report that they do not discuss the supplement use with their physicians because they believe that physicians are biased against supplements and are not knowledgeable about the products.

Research Approach for Determining the Health Impact of Micronutrient Supplements

A micronutrient supplement will be beneficial to a person's health only when the person's normal

dietary micronutrient intake is lower than the amount required for maximum biological benefit. Every person does not have the same micronutrient requirements. The amount of micronutrients required by any person is determined by metabolic, genetic, and environmental factors unique to that person. It may not be readily apparent when micronutrient supplements are needed by certain groups of people. Therefore, as new information becomes available, recommendations for supplementation must be revised. For example, it was observed that pregnant women with periconceptual folate intake at the low end of the range of recommended intake, which was still considered adequate, had an increased risk of giving birth to an infant with neural tube defects (NTDs) such as spina bifida. NTDs originate during the first 4 weeks of pregnancy, before a woman may even realize that she is pregnant. United States survey data (1988–1994) indicated that typical dietary folate intake by women of reproductive age was less than the 400 μg/day believed to be required to reduce the risk of NTDs. Therefore, in 1992, the Centers for Disease Control and Prevention recommended that all women who could become pregnant should take a daily 400 μg folic acid supplement as a preventive measure. In addition, the FDA mandated that, as of January 1998, enriched grain products must be fortified with folic acid, adding an estimated 100 μg folic acid/day to the average diet of US women. Fortification refers to adding nutrients to commonly consumed foods at levels greater than the levels that are part of the standards of identity for the foods; other examples of fortification are vitamin D in milk and calcium in orange juice.

Any recommendations for supplementation must be based on scientific evidence that the supplements are both effective and safe. Ideally, a rigorous systematic research approach (**Table 1**) is carried out and the results are evaluated to assess whether a micronutrient supplement is beneficial to health and whether its recommendation is warranted. All available evidence, including epidemiologic and survey data, as well as preclinical evidence from in vitro laboratory research and in vivo animal studies, is reviewed thoroughly and objectively to determine whether the evidence regarding effectiveness and safety justifies proceeding to clinical trials. If so, the trials are normally conducted in three phases: (1) human safety trials; (2) small efficacy trials, usually in defined target groups; and (3) large-scale trials that are essential in moving from the basic science to evidence-based recommendations that have human health benefits. In fact, the large-scale, double-blind, randomized, placebo-controlled clinical trial, which is designed to eliminate all possible bias, is considered to be the gold standard of scientific intervention research. In such trials, some people receive the substance being tested (e.g., drug, micronutrient, or other dietary constituent) and some receive an inactive placebo. These trials may not be possible in all circumstances, however, because of ethical issues that make it inappropriate to withhold the substance being tested from any trial participants. For example, now that it is established that low periconceptual folate intake by women is linked to NTDs, a placebo-controlled intervention trial to test the minimum effective supplemental amount would be unethical. In such cases, all available evidence from in vitro laboratory research and in vivo animal studies, as well as epidemiologic studies and surveys, must be reviewed systematically and objectively to draw conclusions about the possible effectiveness and safety of the substance of interest and to make recommendations for supplementation. However, convincing evidence is currently unavailable to indicate that lowering homocysteine through folate and other vitamin (vitamin B-6 and B-12) supplementation will reduce risk of CVD. A number of randomized, placebo-controlled clinical trials are on-going to test the effects of vitamin supplementation on primary and secondary prevention of CVD and stroke.

Research to determine a possible impact of micronutrient supplements on the nutritional status and health status of people has been under way for many

Table 1 Components of a research approach to evaluate dietary micronutrient supplements

Basic biomedical laboratory research
In vitro experiments (e.g., in cell culture and tissue culture)
In vivo animal experiments (e.g., in mice and rats)
Human observational epidemiologic studies to identify possible links between micronutrients and nutrition/health status (includes surveys of micronutrient intake)
Hypothesis development: Evaluation of existing laboratory and epidemiologic evidence on micronutrient safety and effectiveness as related to human health benefits (decision point: proceed or do not proceed)
If proceeding
Human safety trials to identify adverse side effects and determine safe doses
Small trials in defined populations to measure micronutrient effectiveness at various safe doses (e.g., vitamin D supplementation in elderly Scandanavians with low serum 25-hydroxy-vitamin D)
Large-scale, double-blind, placebo-controlled, randomized clinical intervention trials to test whether micronutrient supplementation has the hypothesized human health benefit
After health benefits are confirmed, develop recommendation for supplementation

years. Considerable preclinical evidence related to human health effects from in vitro laboratory research and in vivo animal studies exists for many micronutrients. In addition, many epidemiologic studies throughout the world have focused on the possible relationship between specific micronutrients and chronic disease. Small clinical studies related to chronic disease also have been carried out for many micronutrients, and human safety data are available for most micronutrients. A comprehensive review of epidemiologic studies and randomized controlled trials of vitamin supplementation to prevent either cancer or cardiovascular disease (CVD) was conducted by the US Preventive Services Task Force. The Task Force concluded that findings did not demonstrate a consistent or significant effect of any single vitamin or combination of vitamins on either incidence of CVD or death from this disease. Also, the Task Force concluded that β-carotene supplements and combinations including β-carotene appeared to be harmful to those at risk for lung cancer but not to the general population.

Important issues to be addressed in research aimed at determining the effects of micronutrient supplements on health include developing better methods to measure the contribution of micronutrient supplements to total micronutrient intake for various population groups and to monitor these contributions over time to identify usage trends. Having accurate data for micronutrient supplement intake and intake trends is essential to help identify possible associations between supplements and health outcomes; such associations can then be tested for validity in future randomized, controlled trials. Collecting data to measure and ultimately monitor consumer use of micronutrient supplements can be expensive and time-consuming, however, particularly if detailed data are required. Currently, in the United States, NHANES interviewers collect the most detailed information about micronutrient supplement intake, including data on supplement brand, labeled ingredients, dose, and frequency of dose. Available dietary supplement databases are based on values declared on product labels rather than direct analysis. Evidence suggests, however, that supplement labels may not always give the true supplement content; this can decrease the accuracy of survey results.

A major concern associated with clinical trials designed to evaluate the health effects of micronutrients (as well as other dietary supplements and drugs) is that participants might take additional micronutrient supplements, which could influence trial outcomes. In the Prostate Cancer Prevention Trial (PCPT) of the drug finasteride, for example, almost half of the participants reported using a multivitamin/mineral supplement, about one-third used single supplements of either vitamin C or E, and one in five used calcium supplements. Very little evidence is available on how individual micronutrient substances may interact with one another to influence health outcomes. For minerals, particularly, supplementation with one mineral may compromise the bioavailability of another. Also, much remains to be learned about how individual genetic susceptibilities may influence the health-related effects of micronutrient supplements. This issue also must be addressed when designing clinical trials.

Evidence Supporting Recommendations for Micronutrient Supplement Use

Importance of Life Cycle

Evaluation of existing evidence related to effects of micronutrient supplements on nutrition and health, aimed at formulating recommendations for supplementation, must take into account the influence of a person's stage of life and general health status on the absorption, usefulness, and need for any particular micronutrient. Physiological needs for specific micronutrients and, consequently, for micronutrient supplements differ at various stages in the life cycle. For example, infants require additional iron after 6 months of age, women who may become pregnant benefit from additional folate, and elderly people who lose their ability to absorb naturally occurring vitamin B_{12} in food require an alternative source of the vitamin. When studies are designed to investigate the relationship between micronutrient supplements and specific health outcomes, the outcomes that are chosen to be measured usually depend on the specific life cycle stage of the study participants. For any life cycle stage, a person's genetic makeup and lifestyle behaviors will also influence his or her individual micronutrient requirements (**Figure 1**).

Infants

Iron Iron is a component of a number of proteins including hemoglobin, which is essential for transporting oxygen to tissues throughout the body for use in metabolic processes. The most well-known consequence of iron deficiency is anemia. A full-term infant normally has a high hemoglobin concentration and a large amount of stored iron. Based on research evidence, this stored iron plus the iron provided in human milk is assumed to be adequate for solely breast-fed infants during the first 6 months after birth. Even though the amount of iron in

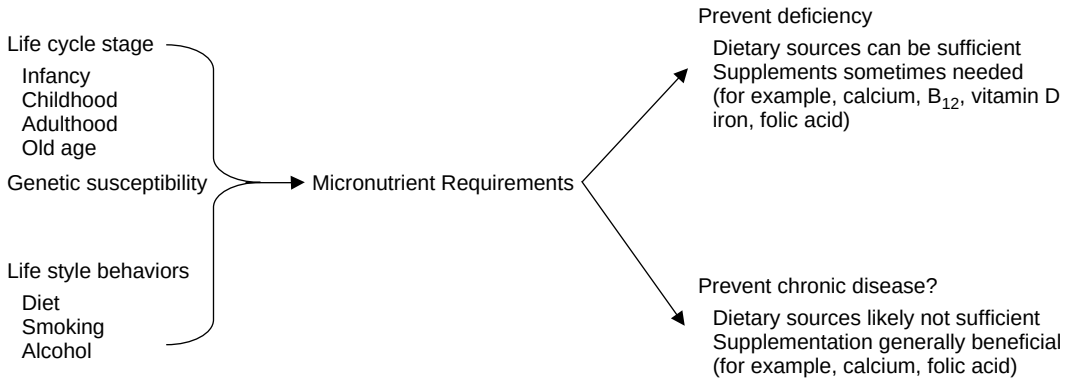


Figure 1 Factors that influence micronutrient requirements.

human milk is low, its bioavailability is greater (>50%) than that of the iron in infant formula (<12%). The body stores of iron in infants decrease during the fourth through sixth months after birth. After 6 months of age, most of the infant's iron needs must be met from food intake. In Western countries, the primary food introduced after 6 months is infant cereal, usually fortified with iron that has low bioavailability. Evidence suggests that infants benefit from iron supplementation after 6 months, and that administration of iron drops between 6 and 9 months has a significant influence on iron status. The American Academy of Pediatrics (AAP) discourages using low-iron infant formulas. AAP recommends that infants who are not breast-fed or who are only partially breast-fed should receive an iron-fortified formula from birth to 12 months of age.

Vitamin D Vitamin D enhances the efficiency of the small intestine to absorb calcium and phosphorus from the diet and thus helps to maintain normal serum levels of these minerals. Vitamin D deficiency in infants and children results in inadequate mineralization of the skeleton, causing rickets, which is characterized by various bone deformations. The major source of vitamin D is its formation in the skin as a result of exposure to sunlight. Dietary sources include fortified foods, such as milk and cereals, and certain fish. Infant formula is fortified with vitamin D in many countries. Because human milk contains only low amounts of vitamin D, breast-fed infants who do not receive either supplemental vitamin D or adequate exposure to sunlight are at risk for developing vitamin D deficiency. Subclinical vitamin D deficiency can be assessed by measuring serum 25-hydroxyl-vitamin D; deficiency occurs months before rickets is obvious on physical examination. Rickets in infants continues to be reported in the United States as well as in other countries. Epidemiologic evidence indicates that African American infants and children are more likely to develop nutritional rickets than Caucasian infants and children. In the US, the AAP recommends that all breastfed infants receive a daily supplement of 200 IU vitamin D/day, beginning within the first two months of life, unless they are weaned to at least 500 mL per day of vitamin D-fortified formula (<1 year old) or milk (>1 year old).

Children

Calcium Bone is a dynamic tissue that is constantly being formed and resorbed; in children, bone formation is greater than resorption. Adequate calcium intake during childhood is essential for bone mass development. Data for calcium intake, presented in **Table 2**, indicate that for children in the United States, only those younger than 8 years of age are meeting their recommended intake. Factors that may contribute to low calcium intake are restriction of dairy products, low vegetable consumption, and high intake of low-calcium beverages such as juices and sodas. The highest calcium intake levels are required during the preteen and adolescent years to support the rapid growth and bone mineralization associated with pubertal development. In girls, peak calcium absorption and deposition takes place at or near menarche; at this life cycle stage, the bone calcium deposition rate is five times greater than that in adults. During peak bone mass development, calcium intakes of less than 1000 mg/day are associated with lower bone mineral density. Epidemiologic studies have found a direct correlation between calcium intake and bone density in children. Evidence suggests that low intake of dairy products during childhood and adolescence may result in less bone mass and greater risk of fracture as an adult. In addition, evidence from randomized trials suggests that increasing the calcium intake of girls is

Table 2 Average calcium intake and recommended adequate intake levels for US children

	Age/gender					
	1–3 years/ M and F	*4–8 years/ M and F*	*9–13 years/ F*	*9–13 years/ M*	*14–18 years/ F*	*14–18 years/ M*
Average intake (mg)	793	838	918	1025	753	1169
Adequate intake (mg)	500	800	1300	1300	1300	1300

F, female; M, male.
From Institute of Medicine (1997) *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride*. Washington, DC: National Academy Press.

associated with increased bone mineral deposition, especially during prepuberty. Although it is best to obtain as much calcium as possible from foods, because calcium-rich foods also provide nutrients involved in calcium utilization, calcium supplements may be necessary for children who do not eat calcium-rich foods.

Adults

Vitamin E Vitamin E (α-tocopherol) functions as an antioxidant that promotes normal formation of red blood cells and normal function of the nervous and immune systems. The main dietary sources of vitamin E are vegetable oils; normally, it is possible, unless people consume a very low-fat diet, to obtain amounts of vitamin E intake from foods that are sufficient to prevent signs of deficiency. However, vitamin E is a commonly consumed supplement, likely because of its hypothesized role in decreasing risk of CVD, prostate cancer, and various other chronic diseases.

Evidence from epidemiologic studies suggests that vitamin E supplementation is beneficial for reducing CVD risk. Nevertheless, data from randomized clinical trials, in populations both with and without a history of CVD, generally do not support the epidemiologic findings. The review of evidence by the US Preventive Services Task Force included five well-designed, large cohort studies that investigated the association between vitamin supplementation and CVD mortality, three clinical trials of primary prevention of CVD, and seven clinical trials of secondary prevention of cardiac events. As stated earlier, the Task Force concluded that findings did not demonstrate a consistent or significant effect of vitamin E on either incidence of CVD or death from this disease. Four large clinical trials are currently in progress in the United States to study the effect on CVD of vitamin E supplements alone or combined with other antioxidants: the Women's Health Study, the Women's Antioxidant and Cardiovascular Study, the Physicians' Health Study II, and the Heart Protection study.

Laboratory studies suggest that vitamin E can inhibit the growth of human prostate cancer cell lines. Results of epidemiologic studies, however, do not consistently support a beneficial effect of vitamin E on risk for prostate cancer. Findings from the Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC) study, a large, randomized clinical trial conducted in Finland, suggest a substantial benefit of vitamin E in decreasing prostate cancer risk. This study reported a decrease of 32% in prostate cancer incidence and a decrease of 41% in deaths from prostate cancer among current and former male smokers who received supplemental vitamin E (50 mg/day). Additional information on the relationship between vitamin E supplementation and prostate cancer likely will be available from the Prostate Cancer Prevention Trial (PCPT), which was stopped in June 2003 when analysis showed that the test drug, finasteride, reduced the risk of developing prostate cancer by 25%. In PCPT, 35% of the study population took vitamin E supplements, and study analyses will include interactions between vitamin E and other supplements and between vitamin E and finasteride. The Selenium and Vitamin E Prevention Trial (SELECT), described later, is also expected to help clarify the association between vitamin E and prostate cancer.

Selenium Selenium, a strong antioxidant, also shows other biological activity, such as enhancing the immune response and inhibiting cell growth. Laboratory and epidemiologic studies support a beneficial effect of selenium on cancer risk. In a large clinical trial, selenium supplementation did not prevent the recurrence of nonmelanoma skin cancer, but it did significantly decrease the total number of deaths and deaths from cancer. In addition, the incidences of prostate, colorectal, and lung cancers all were significantly decreased in the group that received selenium supplements. These findings and the results of the ATBC study linking vitamin E supplementation with decreased prostate cancer risk led to the development of SELECT. Started in 2001, SELECT is a randomized,

double-blind trial designed to test whether selenium (200 μg/day) alone, vitamin E (400 mg/day) alone, or selenium and vitamin E combined reduce the risk of prostate cancer among healthy men. Men who join SELECT are required to stop taking any purchased vitamin supplements that contain either selenium or vitamin E. An ongoing intervention trial in France, the Supplementation en Vitamines et Mineralaux AntioXydants (SU.VI.MAX) study, is testing nutritional levels of both selenium and vitamin E, as well as vitamin C, β-carotene, and zinc, for reducing incidence of cancer and CVD. In addition to cancer and CVD, French researchers are investigating a possible beneficial role for selenium in arthritis and HIV/AIDS.

Folate Folate, a B-complex vitamin, includes the naturally occurring form found in foods as well as the synthetic form (folic acid) found in fortified foods and supplements. The rationale for the recommendation that all women who may become pregnant should take a daily 400 μg folic acid supplement, a preventive measure to reduce the risk of NTDs, has already been discussed. Folate intake is important throughout pregnancy because of its key role in nucleic acid synthesis, which is essential for cell growth and replication.

A deficiency of folate, vitamin B_{12}, or vitamin B_6 may increase the level of homocysteine, an amino acid normally found in the blood. Evidence indicates that a high homocysteine level increases the risk for CVD and stroke, possibly by either damaging coronary arteries or making it easier for blood platelets to clump together and form a clot. However, no evidence is available to suggest that lowering homocysteine through vitamin supplementation will reduce the risk of CVD. Clinical intervention trials to test the effects of vitamin supplementation on CVD and stroke are needed.

Because folate is involved in the synthesis, repair, and functioning of DNA, some have hypothesized that a deficiency of folate may result in DNA damage that can lead to cancer. A comprehensive review of epidemiologic, preclinical, and clinical evidence linking folate deficiency with increased cancer risk concluded that the evidence is strongest for colorectal cancer. Also, it has been suggested that folate deficiency may increase the effects of other cancer risk factors. Researchers are continuing to investigate whether increasing folate intake from foods or folic acid supplements may reduce cancer risk.

Folate is important for cells and tissues that divide rapidly; therefore, high-dose methothrexate is often used to treat cancer because this compound interferes with folate metabolism. Methothrexate, however, has undesirable side effects, including inflammation in the digestive tract. It is not known whether folic acid supplementation can help control these side effects without decreasing the effectiveness of methrothrexate. Low-dose methothrexate is used to treat a variety of diseases, such as rheumatoid arthritis, lupus, psoriasis, asthma, and inflammatory bowel disease. Low-dose treatment can deplete folate stores and cause side effects similar to folate deficiency. In this case, supplemental folic acid may help reduce the undesirable effects of low-dose methothrexate without decreasing treatment effectiveness.

Calcium Bone formation and resorption are balanced in healthy adults, but formation becomes slower than resorption after menopause and also with aging in both men and women. In menopausal women, decreased estrogen production is associated with accelerated bone loss in the first 5 years after menopause, particularly from the lumbar spine. Evidence indicates that although increasing calcium intake at menopause does not prevent this bone loss, it is beneficial for reducing bone loss in compact bones (e.g., hips, legs, and arms). Furthermore, data suggest that calcium supplementation also reduces lumbar spine bone loss in women who are more than 5 years beyond menopause. In the United States, the recommended calcium intake is 1000 mg/day for men and women ages 19–50 years and 1200 mg/day for men and women ages 51–70 years. People who are not able to obtain this amount of calcium from foods should consider taking calcium supplements to help decrease the risk of reduced bone mass and osteoporosis.

Elderly

Physiological changes that may occur during the natural course of aging can affect micronutrient requirements. Given the same amount of sun exposure, the skin of young adults synthesizes much more vitamin D than the skin of the elderly; thus, choosing good dietary sources of vitamin D becomes essential. Vitamin D deficiency can be a factor in reduced calcium absorption in the elderly. Furthermore, it is estimated that atrophic gastritis, a change in gastrointestinal physiology that results in low-acid conditions in the stomach, is present in approximately 20% of elderly people. Atrophic gastritis has been related to infection with the bacterium *Helicobacter pylori* and is not necessarily a result of normal aging. The low-acid conditions, however, can decrease the absorption of vitamin B_{12} from food and of folate and calcium in general.

Vitamin D Vitamin D is important in the elderly for enhancing calcium absorption, inhibiting cellular growth, and activating lymphocyte function. Vitamin D deficiency may lead to osteoporosis and osteomalacia and possibly increase the risk for some cancers; it has been associated with increased incidence of hip fractures. More than 50% of elderly people have been reported to be vitamin D deficient in some studies. In addition to the skin's decreased ability to synthesize vitamin D as people age, the kidneys, which help to convert vitamin D to its active form, sometimes do not function as well when people age. All elderly people, particularly people with limited sun exposure, such as those who are either homebound or live in northern latitudes, should include vitamin D-fortified foods and fish in their diets. If elderly people are unable to meet their vitamin D needs using dietary sources, they may require a supplement. Evidence suggests that vitamin D supplementation may reduce the risk of osteoporotic fractures in elderly people with low serum levels of vitamin D.

Vitamin B_{12} Vitamin B_{12} is essential for proper brain and nerve development and for DNA synthesis; also, it improves learning and supports methylation metabolism. Dietary vitamin B_{12} must be separated from food proteins before the vitamin can be bound to intrinsic factor and then be absorbed by the body. Under low-acid conditions in the stomach, neither the separation from protein nor the binding to intrinsic factor can take place, significantly decreasing the bioavailability of vitamin B_{12}. Elderly adults with atrophic gastritis and low stomach acid should consume a source of unbound vitamin B_{12} such as that found in supplements or food that has been fortified with the vitamin to ensure adequate intake. In addition, evidence suggests that the use of antibiotics can improve vitamin B_{12} absorption in these elderly adults.

Folate Atrophic gastritis greatly reduces the ability of elderly people to absorb folate. This problem can be corrected by administering folic acid with dilute hydrochloric acid to increase stomach acidity and thus increase absorption. There is concern, however, about the possibility that supplemental folic acid could mask the signs of vitamin B_{12} deficiency. Folic acid can remedy the anemia that results from vitamin B_{12} deficiency, its key diagnostic sign. It cannot, however, correct the permanent nerve damage that is possible if vitamin B_{12} deficiency is not treated. Intake of supplemental folic acid should not be greater than 1000 μg per day to prevent the masking of signs of vitamin B_{12} deficiency.

Calcium Adequate calcium intake is required to maintain bone mineral density and reduce the risk of osteoporosis in the elderly. In addition to the reduced absorption of calcium by elderly people that results from age-related changes in vitamin D metabolism, the elderly also show a reduced ability to increase the efficiency of calcium absorption as an adaptive response to low-calcium diets. Also, as noted earlier, the low-acid conditions resulting from atrophic gastritis can reduce calcium absorption. Dietary calcium reacts with hydrochloric acid in the stomach to form soluble calcium chloride, which is absorbed in the small intestine. In the United States, the recommended calcium intake is 1200 mg/day for men and women older than age 70. Many elderly people may benefit from calcium supplements.

See also: **Adolescents**: Nutritional Requirements. **Aging**. **Children**: Nutritional Requirements. **Folic Acid**. **Infants**: Nutritional Requirements. **Older People**: Nutritional Requirements. **Supplementation**: Dietary Supplements; Role of Micronutrient Supplementation; Developing Countries.

Further Reading

Blendon RJ, DesRoches CM, Benson JM, Brodie M, and Altman DE (2001) Americans' views on the use and regulation of dietary supplements. *Archives of Internal Medicine* **161**: 805–810.

DeJong N, Ocké MC, Branderhorst HAC, and Friele R (2003) Demographic and lifestyle characteristics of functional food consumers and dietary supplement users. *British Journal of Nutrition* **89**: 273–281.

Ervin RB, Wright JD, and Kennedy-Stephenson J (1999) Use of dietary supplements in the United States, 1988–94. National Center for Health Statistics. *Vital Health Statistics* Series 11, No. 244.

Institute of Medicine (1997) *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride*. Washington, D.C.: National Academy Press.

Institute of Medicine (1998) *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin, and Choline*. Washington, D.C.: National Academy Press.

Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. Washington, D.C.: National Academy Press.

Institute of Medicine (2001) *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc*. Washington, D.C.: National Academy Press.

Kim Y-I (1999) Folate and carcinogenesis: Evidence, mechanisms, and implications. *Journal of Nutritional Biochemistry* **10**: 66–88.

Morris CD and Carson S (2003) Routine vitamin supplementation to prevent cardiovascular disease: A summary of the evidence for the U.S. Preventive Services Task Force. *Annals of Internal Medicine* **139**: 56–70.

Ritenbaugh C, Streit K, and Helfand M (2003) *Routine vitamin supplementation to prevent cancer: A summary of the*

evidence from randomized controlled trials for the U.S. Preventive Services Task Force. Rockville, MD: Agency for Healthcare Research and Quality. Available at www.ahrq.gov/clinic/3rduspstf/vitamins/vitasum.htm.

Russell RM (2001) Factors in aging that affect the bioavailability of nutrients. *Journal of Nutrition* **131**: 1359S–1361S.

Special Supplement (2003) Dietary supplement use in women: Current status and future directions. *Journal of Nutrition* **133**(6): 1957S–2013S.

U.S. Preventive Services Task Force (2003) Routine vitamin supplementation to prevent cancer and cardiovascular disease: Recommendations and rationale. *Annals of Internal Medicine* **139**: 51–55.

Zeisel SH (2000) Is there a metabolic basis for dietary supplementation? *American Journal of Clinical Nutrition* **72**: 507S–522S.

SURGERY

Contents

Perioperative Feeding

E Kelly, Harvard Medical School, Boston, MA, USA

Introduction

Resumption of normal oral feeding following surgery can be complicated or delayed. Many factors can contribute to the inability to maintain sufficient oral intake in the postoperative period. Some of these factors are attributable to complications of surgery, some are attributable to the use of medication for anesthesia and analgesia, such as narcotics, some are attributable to mechanical swallowing impairment (dysphagia) or mechanical bowel obstruction, and some are attributable to host factors in advanced disease or critical illness. Our review will examine the expected normal resumption of oral feeding as well as focus on the above complications of postoperative feeding using physiologic principles and discuss the standard alternative nutritional support strategies for these complications.

Normal Resumption of Oral Feeding

With minor surgical procedures, involving only local anesthetics or sedation, normal oral feeding can be safely resumed once the sedative effect has worn off. Even in procedures involving general anesthesia, most patients can start a clear liquid diet once the postoperative nausea has passed, and can rapidly advance to a regular diet if clear liquids are tolerated.

Postoperative Dysphagia

Oropharyngeal dysphagia can occur in the postoperative setting for a number of reasons. Procedures involving the upper aerodigestive tract for either benign or malignant disease can result in varying degrees of dysphagia with or without aspiration due to alteration in the anatomical mechanism of swallowing. In addition, procedures in the neck can be complicated by inadvertent injury to the aerodigestive tract or to the nerves innervating the muscles of deglutition, resulting in dysphagia. The recurrent laryngeal nerves, which control the motion of the vocal chords, have a long and tortuous course that goes into the chest and back into the neck. These nerves are especially vulnerable to injury during operations in the neck or chest, particularly esophagectomy, as the nerves run parallel to the esophagus. The superior laryngeal nerves have a course that is not so tortuous and is entirely within the neck; however, due to their anatomical proximity to the thyroid gland and to the carotid sheath, they are especially vulnerable to injury during thyroidectomy and carotid endarterectomy.

In addition, the normal reflexes of swallowing depend on normal sensation of the pharyngeal mucosa. Thus, postoperative dysphagia can result from direct injury or irritation of the pharyngeal mucosa from endotracheal tubes or esophageal tubes. In cardiac surgical procedures, an esophageal ultrasound probe is frequently used for monitoring. While the overall incidence of postoperative dysphagia following cardiac surgery is 4%, the use of this device is associated with a 7-fold higher incidence. In addition to those discussed above, other procedures that are associated with high risk of

postoperative dysphagia include anterior cervical spine operations, operations for Zencker's diverticulum, and radical lymph node dissection.

Operations for gastroesophageal reflux disease (GERD) have a 6% rate of postoperative esophageal dysphagia, characterized by a sensation of swallowed food sticking in the lower esophagus, and delayed esophageal emptying. While there is no increased risk of aspiration in these patients, postoperative feeding can be significantly impaired, leading to weight loss and malnutrition. Esophageal dysphagia can also complicate other esophageal procedures, such as esophagectomy.

Swallowing is a complex reflex and treatment of postoperative dysphagia requires a thorough understanding of normal swallowing and the surgical complications of swallowing. Dysphagia can be clinically silent, leading to a dangerous condition where no complication is suspected until the patient is orally fed, and develops aspiration pneumonia. When clinical signs do occur, they may be subtle, such as delay in initiation of swallow, regurgitation, or gurgling voice. Clinically apparent aspiration is characterized by cough, regurgitation, and gurgling. Bedside evaluation by a certified speech pathologist is often required for the subtle signs of postoperative dysphagia. Bedside evaluation is up to 90% sensitive for clinical signs and symptoms of aspiration, but cannot detect truly silent aspiration. If silent aspiration is suspected, due to pneumonia or respiratory compromise during feeding, a modified barium swallow is indicated. This examination is conducted in the radiology suite under the supervision of a radiologist and a speech pathologist using real-time fluoroscopy. It enables detection of silent aspiration and allows the speech pathologist to test varying thicknesses of food, in order to determine which thicknesses are safe for the patient, and which are aspirated. The diet can be modified accordingly. If no oral diet can be safely used, then patients with oropharyngeal dysphagia should be fed via an alternate route.

Patients with postoperative esophageal dysphagia should undergo evaluation for structural causes such as anastomotic stricture, and functional causes, such as failed antireflux surgery, or failed procedure for esophageal dysmotility. Barium swallow and upper endoscopy will disclose most anatomic cause of esophageal dysphagia, and some esophageal strictures will respond to endoscopic dilatation. Esophageal dysphagia due to dysmotility requires evaluation with esophageal manometry. Disorders of motility can sometimes be managed medically or with surgical procedures. Alternative routes of feeding can be used in conjunction with these treatments, or can be used as salvage treatment when other treatments fail.

Postoperative Ileus

Ileus is a condition of generalized bowel dysmotility that frequently impairs feeding in the postoperative setting. Ileus typically occurs after abdominal surgery, even if the bowel itself is not altered. It has been shown that laparotomy alone, without intestinal manipulation leads to impaired gastrointestinal motility. The small bowel is typically affected the least, and can maintain organized peristaltic contractions throughout the perioperative period. The stomach usually regains a normal pattern of emptying in 24 h, and the colon is last to regain motility, usually in 48–72 h.

The exact mechanism that causes postoperative ileus is not completely known; however, physiologic studies have demonstrated the significant contribution of both inhibitory neural reflexes and local mediators within the intestinal wall. Inhibitory neural reflexes have been shown to be present within the neural plexes of the intestinal wall itself, and as reflex arcs traveling back and forth from the intestine to the spinal cord. These neural pathways may account for the development of ileus during laparotomy without bowel manipulation. In addition, inflammatory mediators such as nitric oxide are present in manipulated bowel and in peritonitis and may play a role in development of ileus.

Clinically, ileus can be recognized from clinical signs. Abdominal distension, nausea, and the absence of bowel sounds and flatus should prompt the diagnosis. Abdominal X-ray imaging typically shows dilated loops of small bowel and colon. Bowel obstruction must also be considered with these clinical findings, however, and CT or other contrast imaging may be required to confirm or rule out obstruction.

Ileus can also appear following nonabdominal surgery, and can result from effects of medications, most often narcotics, electrolyte abnormalities, especially hypokalemia, and a wide variety of other factors.

Occasionally, the patient sustains a prolonged period of postoperative ileus. This can be due to a large number of contributing factors, such as intra-abdominal infection, hematoma, effects of narcotics and other medications, electrolyte abnormalities, and pain. In addition, there can be prolonged dysmotility from certain bowel operations, such as intestinal bypass.

The role of laparoscopic surgery in prevention of ileus is controversial. In theory, with less handling of the bowel laparoscopically and with smaller incisions, there should be less stimulation of the local mediators and neural reflexes. Animal studies

comparing open and laparoscopic colon surgery indicate earlier resumption of normal motility studies and bowel movements with the laparoscopic approach. Human trials have not been conclusive. Several series demonstrate earlier tolerance of postoperative feeding with the laparoscopic approach to colon resection; however, these have been criticized for selection bias, and such studies are impossible to conduct in a blind fashion.

Early mobilization has long been held to be useful in prevention of postoperative ileus. While standing and walking in the early postoperative period have been proven to have major benefits in pulmonary function and prevention of pneumonia, mobilization has no demonstrable effect on postoperative ileus.

In the expected course of uncomplicated abdominal surgery, the stomach is frequently drained by a nasogastric tube for the first 24 h after surgery, and the patient is not allowed oral intake until there is evidence that colonic motility has returned, usually best evidenced by the passage of flatus. Earlier feeding, and no gastric drainage after bowel surgery can be attempted for healthy patients undergoing elective abdominal surgery, with a high rate of success, provided clinical symptoms of ileus are not present. In such patients, the use of effective preventive strategies is highly effective. These include maintenance of normal serum electrolytes, use of epidural analgesia, and avoidance of complications such as infection and bleeding. The routine use of nasogastric tubes for drainage in the postoperative period after abdominal surgery has come into question since the mid 1990s.

The most effective strategy for management of postoperative ileus following abdominal surgery has been the development of epidural analgesia. Randomized trials have shown that the use of nonnarcotic (local anesthetic based) epidural analgesia at the thoracic level in the postoperative period results in decreased period of postoperative ileus in elective abdominal surgery. Ileus reduction is not seen in lumbar level epidural analgesia, suggesting that inhibitory reflex arcs involving the thoracic spinal cord may play a major role in postoperative ileus.

Narcotic analgesia, while effective for postoperative pain, has been shown to lengthen the duration of postoperative ileus, especially when used as a continuous infusion or as patient-controlled analgesia (PCA). Patients report better control of postoperative pain with continuous infusion or PCA as compared to intermittent parenteral dosing. Many studies have been done comparing various types of opioid analgesics, in attempts to find a type that does not prolong ileus. There has been no clearly superior drug identified; all currently available opioids cause ileus. Opioid antagonists such as naloxone have been used in trials to decrease ileus in chronic narcotic use, and there is evidence that antagonists are effective in that setting; however, in postoperative ileus the antagonists have not been shown to be clinically useful, again suggesting that other mechanisms are contributing to it.

Early Postoperative Bowel Obstruction

Early postoperative bowel obstruction refers to mechanical bowel obstruction, primarily involving the small bowel, which occurs in the first 30 days following abdominal surgery. The clinical picture may frequently be mistaken for ileus, and these clinical conditions can overlap. The clinical presentation of early postoperative bowel obstruction is similar to bowel obstruction arising *de novo*: crampy abdominal pain, vomiting, abdominal distension, and obstipation. The incidence of early postoperative bowel obstruction has been variable in published series, due to difficulty in differentiating ileus from early postoperative bowel obstruction, but the reported range is from 7 to 9.5% of abdominal operations.

Retrospective large series show that about 90% of early postoperative bowel obstruction is caused by inflammatory adhesions. These occur as a result of injury to the surfaces of the bowel and peritoneum during surgical manipulation. The injury prompts the release of inflammatory mediators that lead to formation of fibrinous adhesions between the serosal and peritoneal surfaces. As the inflammatory mediators are cleared, and the injury subsides, these adhesions eventually mature into fibrous, firm, band-like structures. In the early postoperative period, the adhesions are in their inflammatory, fibrinous form and, as such, do not usually cause complete mechanical obstruction.

Internal hernia is the next most common cause of early postoperative bowel obstruction, and can be difficult to diagnose short of repeat laparotomy. Internal hernia occurs when gaps or defects are left in the mesentery or omentum, or blind gutters or sacs are left in place during abdominal surgery. The typical scenario is colon resection involving extensive resection of the mesentery for lymph node clearance. If the resulting gap in the mesentery is not securely closed, small bowel loops may go through the opening and not be able to slide back out. A blind gutter may be constructed inadvertently during the creation of a colostomy. When the colostomy is brought up to the anterior abdominal wall, there is a space between the colon and the lateral abdominal wall, which may also 'trap' the mobile loops of small bowel. Defects in the closure of the fascia

during open or laparoscopic surgery can cause obstruction from incarcerated early postoperative abdominal wall hernia. Fortunately, internal hernia is a rare occurrence in the early postoperative period; however, it must be suspected in cases where bowel anastomoses or colostomies have been constructed. Unlike adhesive obstruction, internal hernia requires operative intervention due to the high potential for complete obstruction and strangulation of the bowel.

Intussusception is a rare cause of early postoperative bowel obstruction in adults, but occurs more frequently in children. Intussusception occurs when peristalsis carries a segment of the bowel (called the lead point) up inside the distal bowel like a rolled up stocking. The lead point is usually abnormal in some way, and typically has some intraluminal mass, such as a tumor or the stump of an appendix after appendectomy. Other rare causes for early postoperative bowel obstruction include: missed causes of primary obstruction at the index laparotomy, peritoneal carcinomatosis, obstructing hematoma, and ischemic stricture.

Management of early postoperative bowel obstruction depends on differentiation of adhesive bowel obstruction (the majority) from internal hernia and the other causes, and from ileus. Clinicians generally rely on radiographic imaging to discern ileus from obstruction. For many years plain X-ray of the abdomen was used: if the abdominal plain film showed air-distended loops of bowel and air/fluid levels on upright views, the diagnosis of obstruction was favored. However, plain radiographs can be misleading in the postoperative setting, and the overlap of ileus and obstruction can be confusing. Upper gastrointestinal (GI) contrast studies using water-soluble agent has better accuracy, and abdominal computed tomography (CT) using oral contrast has been shown to have 100% sensitivity and specificity in differentiating early postoperative bowel obstruction from postoperative ileus.

Once the diagnosis is made, management is tailored to the specific needs of the patient. Decompression via nasogastric tube is usually indicated, and ileus can be treated as discussed. Adhesive bowel obstruction warrants a period of expectant management and supportive care, as the majority of these will resolve spontaneously. Most surgical texts recommend that the waiting period can be extended to 14 days. If the early bowel obstruction lasts longer than 14 days, less than 10% resolve spontaneously, and exploratory laparotomy is indicated. The uncommon causes of early postoperative bowel obstruction, such as internal hernia, require more early surgical correction, and should be suspected in the setting of complete obstipation, or when abdominal CT suggests internal hernia or complete bowel obstruction.

Nutritional Support in the Postoperative Setting

Nutritional support in the postoperative patient, as with any patient, begins with a thorough evaluation of nutritional status. The standard assessment including recent eating habits, changes in weight, physical examination, and relevant laboratory values applies to all patients. Ideally, the surgical patient should undergo this assessment preoperatively, and have postoperative feeding planned appropriately. The majority of patients who undergo elective operation are nutritionally replete, and return to normal oral alimentation soon after surgery. Unfortunately, patients undergoing emergency surgery frequently do not get a nutritional assessment, and their nutritional risk should be assessed promptly postoperatively.

In addition to the patient's nutritional status at time of surgery, the clinician must make some estimate of the time until resumption of normal intake. As discussed above, postoperative complications can make this prediction uncertain, but early recognition of complications should be stressed, to enable prompt initiation of nutritional support and prevent worsening of nutritional status. The type of operation and individual patient risk factors such as age and existing diagnoses of dysmotility or obstruction should also be taken into account when planning postoperative nutritional support.

Postoperative dysphagia can frequently be managed by alteration of the oral diet. Evaluation of swallowing by a speech pathologist will enable the selection of the appropriate diet. The patients that most often benefit from modification of the oral diet are those who have undergone procedures in the neck or in the thoracic cavity. Occasionally, these patients have severe dysphagia and cannot take any oral diet. Enteral feeding via tubes (inserted nasally or surgically) can be provided until the dysphagia resolves. In elective procedures on the esophagus or upper aerodigestive tract, a feeding tube is often placed at the index operation to facilitate postoperative feeding, when the preoperative nutritional assessment indicates a high risk of postoperative dysphagia. A surgically placed feeding tube can remain in place indefinitely; a nasal feeding tube should not remain in place longer than a few weeks, due to the risk of nasal erosion and sinusitis.

The standard enteral formulas can be used in the postoperative setting, and as with all patients, the enteral nutritional prescription must be individualized. The postoperative patient's estimated nutrient goals should take into account the added stress of wound healing, especially if there are large open wounds, such as in the trauma and burn population, and infection, if present. Tolerance of enteral feeding can be a difficult clinical problem in the postoperative setting. These patients may have lingering ileus or unsuspected early postoperative bowel obstruction, and gastrointestinal motility may be impaired by critical illness, preexisting conditions such as diabetes, and pain control medication. Tolerance can be enhanced by advancement of the tip of the feeding catheter into the small bowel, which is resistant to ileus. Advancement of a nasogastric catheter can be accomplished at the bedside, and chances of success are improved by positioning of the patient with the right side dependent and with the use of promotility agents such as metoclopramide and erythromycin. Advancement of a percutaneous or surgically placed feeding catheter usually requires endoscopic or fluoroscopic guidance in the endoscopy suite or in radiology.

Postoperative bowel obstruction can be a difficult clinical problem, as this patient population may have some increased nutritional risk preoperatively, the duration of the bowel obstruction is difficult to predict, and some of these patients may require reoperation to relieve the obstruction. In addition, nutritional support options are limited, as both the oral and enteral routes are not available. This patient group should be considered for postoperative parenteral nutrition. Most nutritional guidelines, including the American Society for Parenteral and Enteral Nutrition (ASPEN), call for initiation of parenteral nutrition if the anticipated duration of obstruction is 7 days or greater. If the patient is already nutritionally compromised at the time of surgery, parenteral nutritional support is indicated sooner.

In the case of prolonged postoperative ileus, parenteral nutrition may be required especially if the patient is nutritionally compromised preoperatively. As with bowel obstruction, an estimate of the duration of impaired intake is required, but unlike obstruction, there is no clear guideline as to how long the ileus may last. As discussed, all measures to avoid or curtail ileus should be undertaken: avoidance of opioid analgesia, correction of electrolyte abnormalities, and treatment of infection. Opioid antagonists and most promotility agents have not been proven clinically effective in treating postoperative ileus. There are two motility agents in clinical use that have shown some promise and are under investigation: cisapride and neostigmine. Cisapride increases motility throughout the gastrointestinal tract by stimulation of acetylcholine release in the intrinsic myenteric nerve plexes and has been shown to shorten the duration of postoperative ileus in randomized trials when administered intravenously. Unfortunately, this drug has severe cardiac side effects and was withdrawn from use in the US and the UK in 2000. Neostigmine inhibits the removal of acetylcholine neurotransmitter from nerve synapses and thereby increases intestinal motility. This drug has been used effectively in cases of dysmotility syndromes and has been used successfully in case reports and in small prospective trials for postoperative ileus. Larger trials await completion to gain acceptance in general use.

In general, the same guidelines for early postoperative bowel obstruction apply to ileus: if resolution within 7 days is not anticipated, or if the patient is nutritionally compromised, parenteral nutrition is indicated.

Summary

The vast majority of postoperative patients resume their preoperative feeding plan without significant delay. Complications of surgery can interfere with or permanently alter the patient's ability to resume normal intake. These complications can take many forms and cause functional or anatomic dysfunction of any portion of the gastrointestinal tract. Specific procedures have well documented rates of complication, enabling the clinician to individualize treatment of postoperative feeding disorders. Nutritional support in the postoperative setting may be required, depending on nutritional risks of the patient and the anticipated length of nutritional impairment.

See also: **Colon**: Structure and Function; Disorders. **Small Intestine**: Structure and Function; Disorders. **Stomach**: Structure and Function; Disorders.

Further Reading

ASPEN Board of Directors (2002) Guidelines for the Use of Parenteral and Enteral Nutrition in Adult and Pediatric Patients. *Journal of Parenteral and Enteral Nutrition* **26S**: 1SA–96SA.

Holte K and Kehlet H (2000) Postoperative ileus: a preventable event. *British Journal of Surgery* **87**: 1480–1493.

Klein S, Kinney J, and Jeejeebhoy K (1997) Nutrition support in clinical practice: review of published data and recommendations for future research directions. Summary of a conference sponsored by the National Institutes of Health, American Society for Parenteral and Enteral Nutrition, and the American Society for Clinical Nutrition. *Journal of Parenteral and Enteral Nutrition* **21**: 133–156.

Rolandelli RH, Gupta D, and Wilomre DW (2002) Nutritional support. In: Wilmore DW, Cheung LY, Harken AH, Holcroft JW, Meakins JL, and Soper NJ (eds.) *ACS Surgery. Principles and Practice*, pp. 1–22. New York: WebMD Corporation.

Rousou JA, Tigh DA, and Garb JL (2000) Risk of dysphagia after transesophageal echocardiography during cardiac operations. *Annals of Thoracic Surgery* **69**: 486–490.

Sajja SBS and Schein M (2004) Early postoperative small bowel obstruction. *British Journal of Surgery* **91**: 683–691.

Long-term Nutritional Management

E Lin and T R Ziegler, Emory University, Atlanta, GA, USA

Long-Term Nutrition Support

The initial goal of nutritional support in the surgical patient is to prevent or reverse the catabolic effects of disease or injury. This article addresses the clinical application of nutritional support once the patient is stabilized and the initial injury contained. Long-term nutritional support usually refers to transition from the acute hospital setting into a long-term chronic care setting or the patient's home. The ability to transition nutrition therapy is the result of innovations in central venous catheters, enteral catheters, nutrient formulations, and infusion pumps and the availability of dedicated home health industry. However, in considering a patient for home nutritional support, one of the most important criteria is the availability of strong family support systems.

The indications for long-term nutrition support, specifically for home, include any patient who is unable to meet nutrient requirements by oral intake to avert the consequences of malnutrition. In the United States, cancer patients represent the most frequent use (>40%) of home parenteral and enteral nutrition support. Inflammatory bowel disease is the second most common reason for home parenteral nutrition, and swallowing disorders are the second most common reason for home enteral nutrition. Other indications for long-term home nutrition support are outlined in **Tables 1** and **2**.

Table 1 Body fuel reserves in a 70-kg man

Component	*Mass (kg)*	*Energy (kcal)*	*Days available*
Water and minerals	49	0	0
Protein	6	24 000	13
Glycogen	0.2	800	0.4
Fat	15	140 000	78
Total	70	164 800	91.4

Energy Requirements

Overall nutritional assessment is undertaken to determine the severity of nutrient deficiencies or excess and to aid in predicting nutritional requirements. Pertinent information is obtained by determining the presence of weight loss, chronic illnesses, or dietary habits that influence the quantity and quality of food intake. Social habits predisposing to malnutrition and the use of medications that may influence food intake or urination should also be investigated. Physical examination seeks to assess loss of muscle and adipose tissues, organ dysfunction, and subtle change in skin, hair, or neuromuscular function reflecting frank or impending nutritional deficiency. Anthropometric data (arm circumference and muscle area) and biochemical determinations (creatinine excretion, albumin, prealbumin, total lymphocyte count, and transferrin) may be used to substantiate the patient's history and physical findings. It is imprecise to rely on any single or fixed combination of the previous findings to accurately assess nutritional status or morbidity. Appreciation for the stresses and natural history of the disease process, in combination with nutritional assessment, remains the basis for identifying patients in acute or anticipated need of nutritional support.

A fundamental goal of nutritional support is to meet the energy requirements for metabolic processes, core temperature maintenance, and tissue repair. Failure to provide adequate nonprotein

Table 2 Energy equivalent of substrate oxidation (per gram)

Substrate	*O_2 consumed (l/g)*	*CO_2 produced (l/g)*	*Respiratory quotient*	*kcal/g*	*Recommended daily need g/kg/day*
Glucose	0.75	0.75	1.0	4.0	7.2
Dextrose	—	—	—	3.4	—
Lipid	2.0	1.4	0.7	9.0	1.0
Protein	1.0	0.8	0.8	4.0	0.8

Table 3 Caloric adjustments above basal energy expenditure (BEE) in hypermetabolic conditions

Condition	*kcal/kg/day*	*Adjustment above BEE*	*g Protein/kg/day*	*Nonprotein calorie:nitrogen*
Normal/moderate malnutrition	25–30	1.1	1.0	150:1
Mild stress	25–30	1.2	1.2	150:1
Moderate stress	30	1.4	1.5	120:1
Severe stress	30–35	1.6	2.0	90–120:1
Burns	35–40	2.0	2.5	90–100:1

energy sources will lead to dissolution of lean tissue stores. The requirement for energy may be measured by indirect calorimetry or estimated from urinary nitrogen excretion, which is proportional to resting energy expenditure. Basal energy expenditure (BEE) may also be estimated by the equations of Harris and Benedict:

$$\text{BEE(men)} = 66.47 + 13.75\ (W) + 5.0\ (H) - 6.76\ (A)\ \text{kcal/day}$$

$$\text{BEE(women)} = 655.1 + 9.56\ (W) + 1.85\ (H) - 4.68\ (A)\ \text{kcal/day}$$

where W is weight (kg), H is height (cm), and A is age (years).

These equations, adjusted for the type of surgical stress, are suitable for estimating energy requirements in more than 80% of patients. It has been demonstrated that the provision of 25–30 kcal/kg/day will adequately meet energy requirements in most postsurgical patients, with low risk of overfeeding. The prescribed nutrient substrate requirements during long-term nutrition are to replace body nutrient stores and facilitate protein synthesis (**Table 3**). An appropriate nonprotein calorie:nitrogen ratio of 150 to 1 (1 g N = 6.25 g protein) should be maintained, which is the basal calorie provided to prevent use of protein as energy source. In the absence of severe renal or hepatic dysfunction precluding the use of standard nutritional regimens, approximately 0.25–0.35 g of nitrogen/kg of body weight should be provided daily.

Vitamins and Minerals

The requirements for vitamins and essential trace minerals usually can be easily met in the average patient with an uncomplicated postoperative course. Therefore, vitamins are usually not given in the absence of preoperative deficiencies (**Table 4**). Patients maintained on elemental diets or parenteral hyperalimentation require complete vitamin and mineral supplementation. Commercial enteral diets contain varying amounts of essential minerals and vitamins. It is necessary to ensure that adequate replacement is available in the diet or by supplementation. Numerous commercial vitamin preparations are available for intravenous or intramuscular use, although most do not contain vitamin K and some do not contain vitamin B_{12} or folic

Table 4 Common manifestations of vitamin and mineral deficiencies in adults

Deficiency	*Major manifestations of deficiencies*
Vitamin	
Vitamin A	Night blindness, corneal/conjunctiva drying
Vitamin D	Rickets, osteomalacia, bone pain
Vitamin E	Chronic cholestasis, spinaocerebellar ataxia, hyporeflexia
Vitamin K	Hemorrhagic disorders
Thiamin (B_1)	Dry beriberi (mental status changes, peripheral neuropathy), wet beriberi (heart failure)
Riboflavin	Sebaceous gland inflammation
Niacin	Dermatosis, pellagra
Pyridoxine (B_6)	Peripheral neuropathy
Biotin	Dry and scaling skin, glossitis
Vitamin B_{12}	Pernicious anemia, neuropathy, myelopathy, glossitis
Folic acid	Similar to B_{12} deficiency
Pantothenic acid	Headache, insomnia, paresthesia
Vitamin C	Scurvy (weakness, listlessness, musculoskeletal pain), perifollicular hemorhage
Essential fatty acids	Hair loss, dry skin, eczematoid dermatosis
Minerals	
Calcium	Dementia, encephalopathy, tetany
Phosphorus	Mental status changes, erythrocyte hemolysis, paresthesia
Potassium	Respiratory failure, paralytic ileus, tetany, arrhythmias
Magnesium	Hypocalcemia, hypokalemia, neuromuscular spasms, gut malabsorption
Iodine	Goiter
Iron	Microcytic anemia, fatigue, dyspnea
Copper	Hypochromic anemia (unresponsive to iron), neutropenia, osteoporosis
Zinc	Dermatosis, photophobia, night blindness, impaired wound healing, alopecia, diarrhea
Fluoride	No acute clinical signs
Selenium	Cardiomyopathy, myalgia, white nail beds
Chromium	Glucose intolerance,
Cobalt	No acute clinical signs known
Molybdenum	Headache, night blindness, lethargy
Manganese	Hair thinning, weight loss, dermatitis

acid. Supplemental trace minerals may be given intravenously by commercial preparations. Essential fatty acid supplementation may also be necessary, especially in patients with depletion of adipose stores.

Overfeeding

Overfeeding usually results from overestimation of caloric needs. Overestimation occurs when actual body weight is used, such as in critically ill patients with significant fluid overload and in obese patients. Indirect calorimetry can be used to quantify energy requirements. Estimated dry weight should be obtained from preinjury records or family members. Adjusted lean body weight can also be calculated. Clinically, increased oxygen consumption, increased carbon dioxide production, fatty liver, suppression of leukocyte function, and increased infectious risks have been documented with overfeeding.

Enteral Nutrition

Rationale for Enteral Nutrition

Enteral nutrition is generally preferred over parenteral nutrition based on reduced cost and associated risks of the intravenous route. Laboratory models have long demonstrated that luminal nutrient contact reduces intestinal mucosal atrophy when compared with parenteral or no nutritional support. Studies comparing postoperative enteral and parenteral nutrition in patients undergoing gastrointestinal surgery have demonstrated reduced infection complications and acute phase protein production when fed by the enteral route. However, prospective randomized studies for patients with adequate nutritional status (albumin $\geq$4 g/dl) undergoing gastrointestinal surgery demonstrate no differences in outcome and complications when administered enteral nutrition compared to maintenance of intravenous fluids alone in the initial days following surgery. Furthermore, intestinal permeability studies in well-nourished patients undergoing upper gastrointestinal cancer surgery have demonstrated normalization of intestinal permeability by postoperative day 5. At the other extreme, a meta-analysis of critically ill patients demonstrated a 44% reduction in infection complications in those receiving enteral nutritional support compared to those receiving parenteral nutrition. In addition, most prospectively randomized studies of severe abdominal and thoracic trauma demonstrate very significant reductions in infection complications for patients given early enteral nutrition compared to those who are unfed or receiving parenteral nutrition. The exception has been in studies of patients with closed-head injury because no significant differences in outcome are demonstrated between early jejunal feeding and other nutritional support modalities. Moreover, early gastric feeding following closed-head injury was frequently associated with underfeeding and calorie deficiency due to difficulties in overcoming gastroparesis and the high risk of aspiration.

The early initiation of enteral feeding in burn patients, although sensible and supported by retrospective analysis, is an empiric practice supported by limited prospective trials.

Recommendations for instituting early enteral nutrition to surgical patients with moderate malnutrition (albumin = 2.9–3.5 g/dl) can only be made by inferences due to a lack of data directly pertaining to this population. For these patients, it is prudent to offer enteral nutrition based on measured energy expenditure of the recovering patient or if complications arise that may alter the anticipated course of recovery (e.g., anastomotic leaks, return to surgery, sepsis, and failure to wean from ventilator). Other clinical scenarios with substantiated benefits from enteral nutritional support include permanent neurologic impairment, oropharyngeal dysfunction, short bowel syndrome, and bone marrow transplantation.

Collectively, the data support the use of early enteral nutritional support following major trauma and in patients who are anticipated to have prolonged recovery after surgery. Healthy patients without malnutrition undergoing uncomplicated surgery can tolerate 10 days of partial starvation (i.e., maintenance intravenous fluids only) before any significant protein catabolism occurs. Earlier intervention is likely indicated in patients with poorer preoperative nutritional status.

Initiation of enteral nutrition should occur immediately after adequate resuscitation, most readily determined by adequate urine output. The presence of bowel sounds and the passage of flatus or stool are not absolute requisites for initiating enteral nutrition, but feedings in the setting of gastroparesis should be administered distal to the pylorus. Gastric residuals $\geq$200 ml in 4–6 h or abdominal distention will require cessation of feeding and adjustment of infusion rate. Concomitant gastric decompression with distal small bowel feedings may be appropriate in certain patients, such as closed-head injury patients with gastroparesis. There is no evidence to support withholding enteric feedings for patients following bowel resection or in those with low-output enterocutaneous fistulas of $<$500 ml/day, but low-residue formulations may be preferred. Enteral feeding should also be offered to patients with short bowel syndrome or clinical malabsorption, but caloric needs, essential minerals,

and vitamins should be supplemented with parenteral modalities.

Enteral Formulas

The functional status of the gastrointestinal tract determines the type of enteral solutions to be used. Patients with an intact gastrointestinal tract will tolerate complex solutions, but patients who have not been fed via the gastrointestinal tract for prolonged periods are less likely to tolerate complex carbohydrates such as lactose. In patients with malabsorption, such as in inflammatory bowel diseases, absorption may be improved by provision of dipeptides, tripeptides, and medium-chain triglycerides (MCTs). However, MCTs are deficient in essential fatty acids, which necessitates supplementation with long-chain triglycerides (LCTs).

In general, factors that influence the choice of enteral formula include the extent of organ dysfunction (e.g., renal, pulmonary, hepatic, and gastrointestinal), the nutrient needs to restore optimal function and healing, and the cost of specific products. There are no conclusive data to recommend one category of product over another, and nutritional support committees typically develop the most cost-efficient set of enteral formulary for the most commonly encountered disease categories within the institution.

Low-residue, isotonic formulas Most low-residue, isotonic formulas provide a caloric density of 1.0 kcal/ml and approximately 1500–1800 ml is necessary to meet daily requirements. These low-osmolarity compositions provide baseline carbohydrates, protein, electrolytes, water, fat, and fat-soluble vitamins (some do not have vitamin K) and typically have a nonprotein calorie:nitrogen ratio of 150:1. These contain no fiber bulk and therefore leave minimum residue. These solutions are usually considered 'standard' or first-line formula for stable patients with an intact gastrointestinal tract.

Isotonic formulas with fiber These formulas contain soluble and insoluble fibers that are most often soy based. Physiologically, fiber-based solutions delay intestinal transit time and may reduce the incidence of diarrhea compared with nonfiber solutions. Fibers stimulate pancreatic lipase activity and are degraded by gut bacteria into short-chain fatty acids, an important fuel for colonocytes. There are no contraindications for using fiber-containing formulas in critically ill patients.

Immune-enhancing formulas These solutions are fortified with special nutrients that are purported to enhance various aspects of immune or solid organ function. Such additives include glutamine, arginine, branched-chain amino acids, omega-3 fatty acids, nucleotides, and β-carotene. Although several trials have proposed that one or more of these additives reduce surgical complications and improve outcome, results have not been uniformly corroborated by other trials. From the addition of amino acids, these formulas generally double the amount of protein (nitrogen) found in standard formula. The severalfold higher cost for these solutions prohibits liberal use of these formulas.

Calorie-dense formulas The primary distinction of these formulas is a larger amount of calories for the same volume. Most commercial products of this variety provide 1.5–2 kcal/ml and therefore are suitable for patients requiring fluid restriction or those unable to tolerate large-volume infusions. As expected, these solutions have higher osmolality than standard formulas and are very suitable for intragastric feedings.

High-protein formulas This variety is available in isotonic and nonisotonic mixtures and is proposed for critically ill or trauma patients with high protein requirements. These formulas comprise nonprotein calorie:nitrogen ratios between 80 and 120:1.

Elemental formulas These formulas contain predigested nutrients and provide proteins in the form of small peptides. Complex carbohydrates are limited and fat content, in the form of MCTs and LCTs, is minimal. The primary advantage of such a formula is ease of absorption, but the inherent scarcity of fat, associated vitamins, and trace elements limits its long-term use as a primary source of nutrients. Due to its high osmolarity, dilution or slow infusion rates is usually necessary, particularly in critically ill patients. These formulas have been used frequently in patients with malabsorption, gut impairment, and pancreatitis, but cost is significantly higher than that of standard formulas.

Renal failure formulas The primary benefits of the renal formula are the lower volume required to meet daily requirements and the lower concentrations of potassium, phosphorus, and magnesium. This formulation almost exclusively contains essential amino acids and has a high nonprotein calorie:nitrogen ratio, but does not contain trace elements or vitamins.

Pulmonary failure formulas In these formulas, fat content is usually increased to 50% of total calories, with a corresponding reduction in carbohydrate content. The goal is to reduce carbon dioxide production and alleviate ventilation burden for failing lungs.

Hepatic failure formulas Approximately 50% of the proteins in this formula are branched-chain amino acids (leucine, isoleucine, and valine). The goal of such a formula is to reduce aromatic amino acid levels and increase branched-chain amino acids, which can potentially reverse encephalopathy in patients with hepatic failure. However, the use of this formula is controversial because no clear benefits have been derived from clinical trials. In fact, protein restriction should be avoided in patients with end-stage liver disease because they have significant protein–energy malnutrition, predisposing them to additional morbidity and mortality.

Access for Enteral Nutritional Support

The available techniques and repertoire for enteral access have provided multiple options for feeding the gut. Currently utilized methods and preferred indications are summarized in **Table 5**.

Nasoenteric tubes Nasogastric feeding should be reserved for those with intact mental status and protective laryngeal reflexes to minimize risks of aspiration. Indeed, even in intubated patients, nasogastric feedings can often be recovered from tracheal suction. Nasojejunal feedings are associated with less pulmonary complications, but access past the pylorus requires greater effort to accomplish. Blind insertion of nasogastric feeding tubes commonly results in misplacement, and air instillation with auscultation is inaccurate for ascertaining proper positioning. Radiographic confirmation is usually required to verify the position of the nasogastric feeding tube.

Several methods have been recommended for the passage of nasoenteric feeding tubes into the small bowel, including prokinetic agents, right lateral decubitus positioning, gastric insufflation, tube angulation, and clockwise torque. However, the successful placement of feeding tubes by these methods is highly variable and operator dependent. Furthermore, it is time-consuming, and the success rate of intubation past the duodenum into the jejunum by these methods is less than 20%. Fluoroscopy-guided intubation past the pylorus has a greater than 90% success rate and more than half of these intubations result in jejunal placement. Similarly, endoscopy-guided placement past the pylorus has a high success rate, but advancing the tube beyond the second portion of the duodenum by a standard gastroduodenoscope is difficult.

Small bowel feeding is more reliable for delivering nutrition than nasogastric feeding. Furthermore, the risks of aspiration pneumonia can be reduced by 25% with small bowel feeding compared with nasogastric feeding. The benefits of nasoenteric feeding tubes are limited by clogging, kinking, inadvertent displacement or removal, and nasopharyngeal complications. If nasoenteric feeding is required for more than 30 days, access should be converted to a percutaneous one.

Table 5 Options for enteral feeding access

Access option	*Comments*
Nasogastric tube	Short-term use only; aspiration risks; nasopharyngeal trauma; frequent dislodgement
Nasoduodenal/ nasojejunal	Short-term use; lower aspiration risks in jejunum; placement challenges (radiographic assistance often necessary)
Percutaneous endoscopic gastrostomy (PEG)	Endoscopy skills required; may be used for gastric decompression or bolus feeds; aspiration risks; can last 12–24 months; slightly higher complication rates with placement and site leaks
Surgical gastrostomy	Requires general anesthesia and small laparotomy; may allow placement of extended duodenal/ jejunal feeding ports; laparoscopic placement possible
Fluoroscopic gastrostomy	Blind placement using needle and T-prongs to anchor stomach; can thread smaller catheter through gastrostomy into duodenum/ jejunum under fluoroscopy
PEG-jejunal tube	Jejunal placement with regular endoscope is operator dependent; jejunal tube often dislodges retrograde; two-staged procedure with PEG placement, followed by fluoroscopic conversion with jejunal feeding tube through PEG
Percutaneous endoscopic jejunostomy	Direct endoscopic placement with enteroscope; placement challenges; greater injury risks
Surgical jejunostomy	Commonly applied during laparotomy; general anesthesia; laparoscopic placement usually requires assistant to thread catheter; laparoscopy offers direct visualization of catheter placement
Fluoroscopic jejunostomy	Difficult approach with injury risks; not commonly done

Percutaneous endoscopic gastrostomy The most common indications for percutaneous endoscopic gastrostomy (PEG) placement include impaired swallowing mechanisms, oropharyngeal or esophageal obstruction, and major facial trauma. It is frequently utilized for debilitated patients requiring caloric supplementation, hydration, or frequent medication dosing. It is also appropriate for patients requiring passive gastric decompression. Relative contraindications for PEG placement include ascites, coagulopathy, gastric varices, gastric neoplasm, and lack of a suitable abdominal site. Most tubes are 18–28 Fr in size and may be used for 12–24 months.

Identification of the PEG site requires endoscopic transillumination of the anterior stomach against the abdominal wall. A 14-gauge angiocatheter is passed through the abdominal wall into the fully insufflated stomach. A guidewire is threaded through the angiocatheter, grasped by snares or forceps, and pulled out through the mouth. The tapered end of the PEG catheter is secured to the guidewire and pulled into position out of the abdominal wall. Once the PEG tube is secured without tension against the abdominal wall, many have reported using the tube within hours of placement. It has been the practice of some to connect the PEG tube to a drainage bag for passive decompression for 24 h prior to use, allowing more time for the stomach to seal against the peritoneum.

If endoscopy is not available or technical obstacles preclude PEG placement, the interventional radiologist can attempt the procedure percutaneously under fluoroscopic guidance by first insufflating the stomach against the abdominal wall with a nasogastric tube. If this is unsuccessful, surgical gastrostomy tube placement can be considered, particularly with minimally invasive methods. When surgery is indicated, it may be wise to consider directly accessing the small bowel for nutrition delivery.

Although PEG tubes enhance nutritional delivery, facilitate nursing care, and are superior to nasogastric tubes, serious complications can occur in approximately 3% of cases, including wound infection, necrotizing fasciitis, peritonitis, aspiration, leaks, dislodgement, bowel perforation, enteric fistulae, bleeding, and aspiration pneumonia. For patients with significant gastroparesis or gastric outlet obstruction, feedings through PEG tubes are obviously hazardous. In this instance, the PEG tube can be used for decompression and allows access for converting the PEG tube to a transpyloric feeding tube.

PEG–jejunostomy and direct percutaneous endoscopic jejunostomy Although gastric bolus feedings are more physiologic, patients who cannot tolerate gastric feedings or have significant aspiration risks should be fed directly past the pylorus. In the PEG–jejunostomy method, a 9- to 12-Fr tube is passes through an existing PEG tube, past the pylorus, into the duodenum. This can be achieved by endoscopic or fluoroscopic guidance. With weighted catheter tips and guidewires, the tube can be further advanced past the ligament of Treitz. However, long-term PEG–jejunostomy malfunction has been reported to be greater than 50% as a result of retrograde tube migration into the stomach, kinking, or clogging.

The same techniques are used for direct percutaneous endoscopic jejunostomy (DPEJ) placement as for PEG tube placement, but DPEJ requires an enteroscope or colonoscope to reach the jejunum. DPEJ malfunctions are probably less frequent than PEG–jejunostomy malfunctions, and kinking or clogging are usually averted by placement of larger caliber catheters. The success rate of DPEJ placement is variable because of the complexity of endoscopic skills required to locate a suitable jejunal site. In such cases, surgical jejunostomy tube placement is more appropriate, especially when minimally invasive techniques are available.

Surgical gastrostomy and jejunostomy In a patient undergoing complex abdominal or trauma surgery, thought should be given during surgery to the possible routes for subsequent nutritional support because laparotomy affords direct access to the stomach or small bowel. The only absolute contraindication to feeding jejunostomy is distal intestinal obstruction. Relative contraindications include severe edema of the intestinal wall, radiation enteritis, inflammatory bowel disease, ascites, severe immunodeficiency, and bowel ischemia. Needle catheter jejunostomies can also be used with a minimal learning curve. The drawback is usually related to clogging and knotting of the 6-Fr catheter.

Abdominal distention and cramps are frequent complications of early enteral nutrition. Some have also reported impaired respiratory mechanics as a result of intolerance to enteral feedings. These are mostly correctable by temporarily discontinuing feeds and resuming at a lower infusion rate.

Pneumatosis intestinalis and small bowel necrosis are infrequent, but significant, problems associated with patients receiving jejunal tube feedings. Several contributing factors have been proposed, including the hyperosmolar consistency of enteral solutions, bacterial overgrowth, fermentation, and metabolic breakdown products. The common pathophysiology is believed to be bowel distention and consequent

reduction in bowel wall perfusion. Risk factors for these complications include cardiogenic and circulatory shock, vasopressor use, diabetes mellitus, and chronic obstructive pulmonary disease. Therefore, enteral feedings in the critically ill patient should be delayed until adequate resuscitation has been achieved. Alternatively, dilution of standard enteral formula, delaying the progression to goal infusion rates, or using monomeric solutions with low osmolality requiring less digestion by the gastrointestinal tract have been successfully employed.

Parenteral Nutrition

Parenteral nutrition involves the continuous infusion of a hyperosmolar solution containing carbohydrates, proteins, fat, and other necessary nutrients through an indwelling catheter inserted into the superior vena cava. In order to obtain the maximum benefit, the ratio of calories to nitrogen must be adequate (at least 100–150 kcal/g nitrogen) and both carbohydrates and proteins must be infused simultaneously. When the sources of calories and nitrogen are given at different times, there is a significant decrease in nitrogen utilization. These nutrients can be given in quantities considerably greater than the basic caloric and nitrogen requirements, and this method has proved to be highly successful in achieving growth and development, positive nitrogen balance, and weight gain in a variety of clinical situations. Clinical trials and meta-analysis of parenteral feeding in the perioperative period have suggested that preoperative nutritional support may benefit some surgical patients, particularly those with extensive malnutrition. Short-term use of parenteral nutrition in critically ill patients (duration <7 days) when enteral nutrition may have been instituted are associated with higher infection complications.

Rationale for Parenteral Nutrition

The principal indication for parenteral nutrition is when use of the gastrointestinal tract for feedings is not possible. In some instances, intravenous nutrition may be used to supplement inadequate oral intake. The safe and successful use of parenteral nutrition requires proper selection of patients with specific nutritional needs, experience with the technique, and an awareness of the associated complications. As with enteral nutrition, the fundamental goals are to provide sufficient calories and nitrogen substrate to promote tissue repair and to maintain the integrity or growth of lean tissue mass. The following are cases in which parenteral nutrition has been used in an effort to achieve these goals:

1. Newborn infants with catastrophic gastrointestinal anomalies, such as tracheoesophageal fistula, gastroschisis, omphalocele, or massive intestinal atresia
2. Infants who fail to thrive from gastrointestinal insufficiency associated with short bowel syndrome, malabsorption, enzyme deficiency, meconium ileus, or idiopathic diarrhea
3. Adult patients with short bowel syndrome secondary to massive small bowel resection (<100 cm without colon or ileocecal valve, or <50 cm with intact ileocecal valve and colon)
4. Enteroenteric, enterocolic, enterovesical, or high-output enterocutaneous fistulas (>500 ml/day)
5. Surgical patients with prolonged paralytic ileus following major operations (>7–10 days), multiple injuries, or blunt or open abdominal trauma, or patients with reflex ileus complicating various medical diseases
6. Patients with normal bowel length but with malabsorption secondary to sprue, hypoproteinemia, enzyme or pancreatic insufficiency, regional enteritis, or ulcerative colitis
7. Adult patients with functional gastrointestinal disorders such as esophageal dysmotility following cerebral vascular accident, idiopathic diarrhea, psychogenic vomiting, or anorexia nervosa
8. Patients with granulomatous colitis, ulcerative colitis, and tuberculous enteritis, in which major portions of the absorptive mucosa are diseased
9. Patients with malignancy, with or without cachexia, in whom malnutrition might jeopardize successful delivery of a therapeutic option
10. Failed attempts to provide adequate calories by enteral tube feedings or high residuals
11. Critically ill patients who are hypermetabolic for more than 5 days or when enteral nutrition is not feasible

Conditions contraindicating hyperalimentation include the following:

1. Lack of a specific goal for patient management, or where instead of extending a meaningful life, inevitable death is prolonged
2. Periods of hemodynamic instability or severe metabolic derangement (severe hyperglycemia, azotemia, encephalopathy, hyperosmolality, and fluid-electrolyte disturbances) requiring control or correction before attempting hypertonic intravenous feeding
3. Feasible gastrointestinal tract feeding (in the vast majority of instances, this is the best route by which to provide nutrition)
4. Patients in good nutritional status

5. Infants with less than 8 cm of small bowel, since virtually all have been unable to adapt sufficiently despite prolonged periods of parenteral nutrition
6. Patients who are irreversibly decerebrate

Total parenteral nutrition Also referred to as central parenteral nutrition, total parenteral nutrition requires access to a large-diameter vein to deliver the entire nutritional requirements of the individual. Dextrose content is high (15–25%) and all other macro- and micronutrients are deliverable by this route.

Peripheral parenteral nutrition The lower osmolarity of this solution secondary to reduced dextrose (5–10%) and proteins (3%) allows administration via peripheral veins. Some nutrients cannot be supplemented due to inability to concentrate them into small volumes. Therefore, peripheral parenteral nutrition is not appropriate for repleting patients with severe malnutrition. It can be considered if central routes are not available or if supplemental nutritional support is required. Typically, peripheral parenteral nutrition is used for less than 2 weeks. Beyond this time, total parenteral nutrition should be instituted.

Initiating Parenteral Nutrition

The basic solution contains a final concentration of 15–25% dextrose and 3–5% crystalline amino acids. The solutions are usually prepared sterilely in the pharmacy from commercially available kits containing the component solutions and transfer apparatus. Preparation in the pharmacy under laminar flow hoods reduces the incidence of bacterial contamination of the solution. Proper preparation with suitable quality control is absolutely essential to avoid septic complications. The proper provision of electrolytes and amino acids is dependent on routes of fluid and electrolyte loss, renal function, metabolic rate, cardiac function, and the underlying disease state.

Intravenous vitamin preparations should also be added to parenteral formulas. Rarely to deficiencies of vitamin occur if such preparations are utilized. In addition, because vitamin K is not part of any commercially prepared vitamin solution, it should be supplemented on a weekly basis. During prolonged fat-free parenteral nutrition, essential fatty acid deficiency may become clinically apparent, manifested by a dry, scaly dermatitis and loss of hair. The syndrome may be prevented by periodic infusion of a fat emulsion at a rate equivalent to 10–15% of total calories. Essential trace minerals may be required after prolonged total parenteral nutrition and may be supplied by direct addition of commercial preparations. The most frequent presentation of trace mineral deficiencies is the eczamatoid rash developing both diffusely and in intertriginous areas in zinc-deficient patients. Other rare trace mineral deficiencies include microcytic anemia associated with copper deficiency and glucose intolerance presumably related to chromium deficiency. The latter complications are seldom seen except in patients receiving parenteral nutrition for extended periods of time. The daily administration of commercially available trace mineral supplements will obviate most such problems.

Depending on fluid and nitrogen tolerance, parenteral nutrition solutions can generally be increased over 2 or 3 days to achieve the desired infusion rate. Insulin may be supplemented as necessary to ensure glucose tolerance. Occasionally, additional intravenous fluids and electrolytes may be necessary with persistently high fluid losses. The patient should be carefully monitored for development of electrolyte, volume, acid–base, and septic complications. Vital signs and urinary output are regularly observed, and the patient should be weighed regularly. Frequent adjustments of the volume and composition of the solutions are necessary during the course of therapy. Electrolytes are drawn daily until stable and every week thereafter. Blood counts, blood urea nitrogen, liver functions, phosphate, and magnesium are determined at least weekly.

Capillary blood sugar level is checked at least once daily during the first few days of the infusion and at frequent intervals thereafter. Relative glucose intolerance may occur following initiation of parenteral nutrition, which often manifests as glycosuria. If blood sugar levels remain elevated or glycosuria persists, the dextrose concentration may be decreased, the infusion rate slowed, or regular insulin added to each bottle. The increase in blood glucose concentration observed after initiating an parenteral nutrition may be temporary because the normal pancreas increases its output of insulin in response to the continuous carbohydrate infusion. In patients with diabetes mellitus, additional insulin may be required.

Potassium is essential to achieve positive nitrogen balance and replace depleted intracellular stores. In addition, a significant shift of potassium ion from the extracellular to the intracellular space may take place because of the large glucose infusion, with resultant hypokalemia, metabolic alkalosis, and poor glucose utilization. In some cases, as much as 240 mEq of potassium ion daily may be required.

Hypokalemia may cause glycosuria, which should be treated with potassium, not insulin. Thus, before giving insulin, the serum potassium level must be checked to avoid exacerbating the hypokalemia.

Patients with insulin-dependent diabetes mellitus may exhibit wide fluctuations in blood glucose during parenteral nutrition. Partial replacement with lipid emulsions for dextrose calories may alleviate these problems in selected patients. Lipid emulsions derived from soybean or safflower oils are widely used as an adjunctive nutrient to prevent the development of essential fatty acid deficiency. There is no evidence of enhanced metabolic benefit when more than 10–15% of calories are provided as lipid emulsions. Although the administration of 500 ml of 20% fat emulsion one to three times per week is sufficient to prevent essential fatty acid deficiency, it is common to provide fat emulsions on a daily basis to provide additional calories. The triple mix of carbohydrate, fat, and amino acids is infused at a constant rate during a 24-h period. The theoretical advantages of a constant fat infusion rate include increased efficiency of lipid utilization and reduced impairment of reticuloendothelial function normally identified with bolus lipid infusions. The addition of lipids to an infusion bag may alter the stability of some micronutrients in a dextrose–amino acid preparation.

There are several advantages to infusing parenteral nutrition in a cyclic manner over 12–16 h during the hours of sleep. First, it permits the patient to pursue a normal lifestyle during the day hours. Second, it minimizes the risks of fatty liver and hepatomegaly that are associated with long-term continuous calorie and protein infusions.

Intravenous access methods Temporary or short-term access can be achieved with a 16-gauge percutaneous catheter inserted into a subclavian or internal jugular vein threaded into the superior vena cava. More permanent access, with the intention of providing long-term or home parenteral nutrition, can be achieved by placement of a catheter with a subcutaneous port for access, by tunneling a catheter with a substantial subcutaneous length, or by threading a long catheter through the basilic or cephalic vein into the superior vena cava. In some patients, implanted vascular access ports are an acceptable option.

Complications of Parenteral Nutrition

Technical complications One of the more common and serious complications associated with long-term parenteral feeding is sepsis secondary to contamination of the central venous catheter. Contamination of solutions should be considered but is rare when proper pharmacy protocols have been followed. This problem occurs more frequently in patients with systemic sepsis and in many cases is due to hematogenous seeding of the catheter with bacteria. One of the earliest signs of systemic sepsis may be the sudden development of glucose intolerance (with or without temperature increase) in a patient who previously has been maintained on parenteral alimentation without difficulty. When this occurs or if high fever (>38.5 °C) develops without obvious cause, a diligent search for a potential septic focus is indicated. Other causes of fever should also be investigated. If fever persists, the infusion catheter should be removed and cultured. If the catheter is the cause of fever, removal of the infectious source is usually followed by rapid defervescence. Some centers are replacing catheters considered at low risk for infection over a guidewire. Should evidence of infection persist for 24–48 h without a definable source, the catheter should be replaced in the opposite subclavian vein or into one of the internal jugular veins and the infusion restarted. It is prudent to delay reinserting the catheter by 12–24 h, especially if bacteremia is present.

Other complications related to catheter placement include the development of pneumothorax, hemothorax, hydrothorax, subclavian artery injury, thoracic duct injury, cardiac arrhythmia, air embolism, catheter embolism, and cardiac perforation with tamponade. All these complications may be avoided by strict adherence to proper techniques. The use of multilumen catheters may be associated with a slightly increased risk of infection. This is most likely associated with greater catheter manipulation and intensive use. Catheter infections are highest when catheters are placed in the femoral vein, lower when placed in the jugular vein, and lowest when placed in the subclavian vein. When catheters are indwelling for less than 3 days, infection risks are negligible. If indwelling time is 3–7 days, the infection risk is 3–5%, and an indwelling time of more than 7 days is associated with a catheter infection risk of 5–10%.

Metabolic complications Hyperglycemia may develop with normal rates of infusion in patients with impaired glucose tolerance or in any patient if the hypertonic solutions are administered too rapidly. This is a particularly common complication in latent diabetics and in patients subjected to severe surgical stress or trauma. Treatment of the condition consists of volume replacement with correction of electrolyte abnormalities and the administration of

insulin. This complication can be avoided with careful attention to daily fluid balance and frequent monitoring of blood sugar levels and serum electrolytes.

Increasing experience has emphasized the importance of not 'overfeeding' the parenterally nourished patient. This is particularly true of the depleted patient in whom excess calorie infusion may result in carbon dioxide retention and respiratory insufficiency. In addition, excess feeding has also been related to the development of hepatic steatosis or marked glycogen deposition in selected patients. Cholestasis and formation of gallstones are common in patients receiving long-term parenteral nutrition. Mild but transient abnormalities of serum transaminase, alkaline phosphatase, and bilirubin may occur in many parenterally nourished patients. Failure of the liver enzymes to plateau or return to normal over 7–14 days should suggest another etiology.

Intestinal atrophy Lack of intestinal stimulation is associated with intestinal mucosal atrophy, diminished villous height, bacterial overgrowth, reduced lymphoid tissue size, reduced IgA production, and impaired gut immunity. The full clinical implications of these changes are not well realized, although bacterial translocation has been demonstrated in animal models. The most efficacious method to prevent these changes is to provide nutrients enterally. In patients requiring full parenteral nutrition, it may be feasible to infuse small amounts of trophic feedings via the gastrointestinal tract.

Special Formulations

Glutamine and arginine Glutamine is the most abundant amino acid in the human body, comprising nearly two-thirds of the free intracellular amino acid pool, of which 75% is found within the skeletal muscles. In healthy individuals, glutamine is considered a nonessential amino acid because it is synthesized within the skeletal muscles and the lungs. Glutamine is a necessary substrate for nucleotide synthesis in most dividing cells and hence provides a major fuel source for enterocytes. It also serves as an important fuel source for immunocytes, such as lymphocytes and macrophages, as well as a precursor for glutathione, a major intracellular antioxidant. During stress states such as sepsis or in tumor-bearing hosts, peripheral glutamine stores are rapidly depleted and the amino acid is preferentially shunted as a fuel source toward the visceral organs and tumors, respectively. These situations create, at least experimentally, a glutamine-depleted environment of which the consequences include enterocyte and immunocyte starvation.

The beneficial effects of glutamine supplementation demonstrated experimentally are multifaceted (**Table 6**). However, glutamine metabolism during stress in humans may be more complex than in previously reported animal data. More advanced methods of detecting glutamine traffic in patients with gastrointestinal cancer have not demonstrated more tumor sequestration of glutamine than in normal intestine. There are data demonstrating decreased dependency on total parenteral nutrition in severe cases of short bowel syndrome when glutamine therapy with modified diets and growth hormones are used. However, in patients with milder forms of short bowel syndrome and better nutritional status, glutamine supplementation did not demonstrate appreciable enhancement in intestinal absorption. In healthy subjects, glutamine-supplemented total parenteral nutrition did not attenuate endotoxin-induced symptoms or proinflammatory cytokine release compared to standard total parenteral nutrition. Although it is hypothesized that provision of glutamine may preserve immune cell and enterocyte function and enhance nitrogen balance during injury or sepsis, the pool of clinical evidence in support of this phenomenon in human subjects remains inconclusive.

Arginine, also a nonessential amino acid in healthy subjects, first attracted attention for its immunoenhancing properties, wound-healing benefits, and improved survival in animal models of sepsis and injury. As with glutamine, the benefits of experimental arginine supplementation during stress states are diverse. Clinical studies in which arginine was administered enterally have demonstrated net nitrogen

Table 6 Experimental benefits of glutamine and arginine supplementation

Glutamine
Enhances bowel absorptive capacity after intestinal resection
Decreases intestinal permeability
Early resolution of experimental pancreatitis
Maintains nitrogen balance
Promotes liver regeneration after hepatectomy
Restores mucosal IgA function
Enhances bacterial clearance in peritonitis
Protects postradiation enterocyte viability
Restores intracellular glutathione levels
Facilitates tumor sensitivity to chemotherapy and radiation therapy
Enhances natural killer (NK) and lymphocyte-activated killer (LAK) cell function
Arginine
Minimizes hepatic ischemia–reperfusion injury
Reduces intestinal bacterial translocation
Enhances NK and LAK cell function
Increases nitrogen retention and protein synthesis

retention and protein synthesis compared to isonitrogenous diets in critically ill and injured patients and following surgery for certain malignancies. Some of these studies are also associated with in vitro evidence of enhanced immunocyte function. The clinical utility of arginine in improving overall patient outcome remains an area of investigation.

Omega-3 fatty acids The provision of omega-3 polyunsaturated fatty acids (canola oil or fish oil) displaces omega-6 fatty acids in cell membranes, which theoretically reduce the proinflammatory response from prostaglandin production.

Nucleotides RNA supplementation in solutions is purported, at least experimentally, to increase cell proliferation, provide building blocks for DNA synthesis, and improve T-helper cell function.

Patient monitoring In the first month of home nutrition support, nutrition and metabolic assessments should be performed weekly. In the stable patient, the frequency of these assessments can be reduced to monthly and then quarterly. Hepatic steatosis, cholestasis, and cholelithiasis are all known sequelae of long-term parenteral nutrition. Regular assessments are necessary because specific nutrient deficiencies, such as selenium, vitamin C, iron, and thiamin, are known in patients on long-term parenteral nutrition support. Patients on a fat-free diet or who receive infrequent lipid infusions are at risk of developing essential fatty acid deficiencies that manifest as dermatitis, scaling, and sparse hair growth. These can be circumvented by the provision of long-chain fatty acids in the diet on a regular basis. Zinc and copper deficiencies are particularly prevalent in patients who have short bowel syndrome or malabsorptive states. In many instances, routine biochemical screening may not adequately reflect the functional level of a particular nutrient but, rather, the concentration in a body compartment such as the intravascular space. In these instances, treatment for any nutrient deficiencies should be initiated based on anticipation or clinical suspicion.

See also: **Colon**: Disorders; Nutritional Management of Disorders. **Energy**: Requirements. **Fatty Acids**: Omega-3 Polyunsaturated. **Gall Bladder Disorders**. **Liver Disorders**. **Nutritional Support**: In the Home Setting; Adults, Enteral; Adults, Parenteral. **Small Intestine**: Disorders. **Stomach**: Disorders. **Supplementation**: Role of Micronutrient Supplementation. **Surgery**: Long-term Nutritional Management.

Further Reading

A.S.P.E.N. (1998) Special report: Safe practices for parenteral nutrition formulations. *Journal of Parenteral and Enteral Nutrition* **22**: 49–66.

A.S.P.E.N. (2002) *Nutritional Considerations in the Intensive Care Unit*. Silver Spring, MD: American Society of Parenteral and Enteral Nutrition.

A.S.P.E.N. board of directors (1998) *Clinical Pathways and Algorithms for Delivery of Parenteral and Enteral Nutrition Support for Adults*. Silver Spring, MD: American Society for Parenteral and Enteral Nutrition.

A.S.P.E.N. board of directors (1999) Standards for home nutrition support. *Nutrition in Clinical Practice* **14**: 151–162.

Jonas CR, Griffiths DP, Bergman GF, Leader LM, and Ziegler TR (2001) Nutrient pharmacotherapy. In: Rolandelli R and Rombeau JL (eds.) *Clinical Nutrition: Parenteral Nutrition,* vol. 3, pp. 562–579. Philadelphia: WB Saunders.

Lin E, Calvano SE, and Lowry SF (1998) Systemic response to injury. In: Schwartz SI (ed.) *Principles of Surgery*, 7th edn., pp. 3–51. New York: McGraw-Hill.

Lin E, Goncalves JA, and Lowry SF (1998) Efficacy of nutritional pharmacology in surgical patients. *Current Opinion in Nutrition and Metabolic Care* **1**: 41–50.

Lin E, Kotani J, and Lowry SF (1998) Nutritional modulation of immunity and inflammatory response. *Nutrition* **14**: 545–550.

Lin E and Lowry SF (2002) Substrate metabolism in surgery. In: Norton JA (ed.) *Surgery: Scientific Basis and Clinical Evidence*, pp. 95–104. New York: Springer-Verlag.

Lord L (1997) Enteral access devices. *Nursing Clinics of North America* **32**(4): 685–704.

Mascioli EA, Lopes SM, Champagne C, and Driscoll DF (1996) Essential fatty acid deficiency and home total parenteral nutrition patients. *Nutrition* **12**: 245–249.

Shils M and Brown RO (1999) Parenteral nutrition. In: Shils ME, Olson JA, Shike M, and Ross AC (eds.) *Modern Nutrition in Health and Disease*, 9th edn., pp. 1657–1688. Baltimore, MD: Williams & Wilkins.

Ziegler TR (2001) Fuel metabolism and nutrient delivery in critical illness. In: Becker KL (ed.) *Principles and Practice of Endocrinology and Metabolism*, pp. 2102–2107. Philadelphia: JB Lippincott.

Ziegler TR, Bazargan N, Leader LM, and Martindale RG (2000) Glutamine and the gastrointestinal tract. *Current Opinion in Clinical Nutrition and Metabolic Care* **3**: 355–362.

Ziegler TR, Puckett AB, Griffiths DP, and Galloway JR (1997) Interactions between nutrients and growth factors in cellular growth and tissue repair. In: Ziegler TR, Pierce GF, and Herndon DN (eds.) *Growth Factors and Wound Healing: Basic Science and Potential Clinical Applications*, pp. 104–150. New York: Springer-Verlag.

T

TEA

D J Baer and S C Chen, US Department of Agriculture, Beltsville, MD, USA

Published by Elsevier Ltd.

Throughout the world, tea is one of the most consumed beverages, second only to water. Based on data from the Tea Association of the United States, the estimated wholesale value of the tea industry increased 273% between 1990 and 2002, and it is now valued in excess of $5 billion. The greatest increase has been in the ready-to-drink market (925% increase). One reason for the popularity in tea consumption is the interest of consumers to improve their health. Although tea is a poor source of most classically defined essential nutrients, it is a good and important source of many phytochemicals that have been presumptively linked to reduction of risk of cancer, cardiovascular disease, and diabetes. Many of these phytochemicals have strong antioxidant properties, and presumably these antioxidant properties are associated with improved health.

Epidemiological studies have provided the evidence suggesting an association between the consumption of plant-based foods and diets with a reduced risk of diseases, especially cardiovascular disease and cancer. These findings, together with experimentally data, have spurred further research activities and encouraged the consumption of fruits, vegetables, and other foods and beverages with high concentrations of nutrient (e.g., vitamin C, vitamin E, and carotenoids) and nonnutrient antioxidants (e.g., polyphenols). Epidemiologic studies that focus on food consumption indicate great potential benefit from consumption of foods high in antioxidant phytonutrients. Experimental studies include *in vitro* and cell culture studies; *ex vivo* studies using biological material from experimental animals and humans (e.g., determination of susceptibility of low-density lipoprotein to oxidation); and investigations using animal models of cancer, cardiovascular disease, diabetes, arthritis, macular degeneration, and the diseases of most other organ systems. Many experimental studies focus on the antioxidant compounds that are found in foods and not on the foods per se as they are typically consumed. Although disease prevention based on inference from these studies seems promising, results from dietary intervention studies are inconclusive or contradictory. Moreover, the number of well-designed, carefully controlled dietary intervention studies, which show cause and effect, is limited, and this represents a critical gap in our understanding of these compounds as they relate to health promotion and disease prevention.

More than 8000 phenolic compounds have been identified in plants. The structure of these compounds ranges from simple monomers to highly complex, heterogeneous polymers with molecular sizes in excess of 30 kDa. Of these phenolic compounds, more than 5000 are structurally defined as polyphenolic 'flavonoids.' Although there is an abundance of these compounds in our food supply, many studies have focused on the activity of flavonoids from tea, wine, and cocoa because these are the most common sources of flavonoids in typical American diets. Compared to most foods, tea contains a much higher concentration of flavonoids (in the form of flavan-3-ols), which are provided with a minimal amount of additional energy. The high concentration of water-soluble polyphenols, limited contribution of other nutrients (especially other antioxidants), ease of product preparation and administration, and subject acceptability make tea an ideal source for these potentially important compounds for human intervention studies.

Tea Composition

The predominant phytochemicals of beverage tea are catechins, theaflavins, and thearubigins. Catechins are one form of a class of polyphenolic compounds termed 'flavonoids,' which are all 2-phenyl benzopyran-based compounds. The flavonoids include six major structurally related subclasses of compounds: flavones, flavonols, flavanones, anthocyanidins, isoflavones, and flavanols. Catechins belong to the

flavanol subclass and are structurally defined as flavan-3-ols. The major catechins of the tea leaf are epicatechin, epicatechin gallate, epigallocatechin, and epigallocatechin gallate. These four catechins comprise approximately 25% of the weight of a dry tea leaf. Flavonols (kaempferol, quercetin, and myricitin glycosides) are minor flavonoid components of the tea leaf (approximately 3% of dry weight). Polysaccharides and cellulose (approximately 20% of dry weight) and protein (approximately 15% of dry weight) are other major components of tea leaf, and other nutrients and compounds comprise less than 5% of dry weight each.

Green, oolong, and black are the predominantly consumed types of tea. They are produced from the leaf of the same plant, *Camellia sinseis*. Green tea, which is consumed largely in Asia, is processed with the intent to minimize leaf damage and fermentation. Catechins constitute 30–40% of the green tea solids. To produce oolong and black tea, the leaves are rolled and fermented. This processing releases polyphenol oxidase that initiates polymerization and oxidation of the catechins to theaflavins and thearubigins. In contrast to the heterogeneous and unknown chemical structure of thearubigins, theaflavins are chemically well defined. There are four primary theaflavins in black tea and they represent between 1 and 6% of the dry weight of the solids (theaflavin, theaflavin 3-gallate, theaflavin 3′-gallate, and theaflavin 3,3′-digallate). Thearubigins represent between 10 and 20% of the dry weight of black tea solids. Oolong tea is processed and fermented to a lesser degree than black tea, and therefore it contains a higher concentration of the theaflavins and lower concentration of thearubigins than black tea. Black tea is popular in Western countries, and oolong tea is sold commercially in the United States and is often served in Chinese restaurants. Theaflavins and thearubigins are responsible for the red-amber color of oolong and black tea and for its astringency.

Another important group of flavan-3-ol compounds are the procyanidin polymers. These compounds are catechin polymers linked by C4 to C6 and C4 to C8 bonds. The most predominant polymers are those ranging from dimers to decamers and are found in a variety of foods, including chocolate, cocoa, cereal grains, nuts, fruits, and vegetables.

In Vitro and *Ex Vivo* Evidence of Antioxidant Properties of Polyphenols

One reason for the interest in polyphenols is their excellent *in vitro* ability to scavenge stable free radicals and reactive oxygen species. In fact, *in vitro*, many polyphenols found in fruits and vegetables are better scavengers than the essential nutrient antioxidants. For example, compared to vitamin E and vitamin C, flavonols (quercetin, kaempferol, and rutin) are 1.3- to 4.7-fold more effective at scavenging free radicals. Flavan-3-ols and theaflavins are 2.5–6.9 times more effective than vitamin C and vitamin E at free radical quenching. Tea beverage has strong antioxidant properties, being 3.5–3.8 times more effective than vitamin C and vitamin E at free radical scavenging. Green tea is more effective than oolong tea, which is more effective than black tea. In comparison, carotenoids (lycopene, α-carotene, and β-carotene) and xanthophylls (β-cryptoxanthin, zeaxanthin, and lutein) are 1.3- to 2.9-fold more effective at *in vitro* free radical scavenging.

In addition to scavenging radicals and oxidants, tea polyphenols have also been demonstrated *in vitro* to be able to chelate metals, inhibit redox-sensitive transcription factors, inhibit prooxidant enzymes, and induce phase II enzymes. These activities may explain, at least in part, the mechanisms for the antioxidant effects of tea.

An *in vitro* and *ex vivo* test of the biological activity of an antioxidant is to assess its ability to protect LDL particles from oxidation that, theoretically, can be critical for the prevention of atherosclerosis since modified LDL particles (e.g., oxidized LDL) may play an important role in the etiology of atherosclerosis. There is considerable but not overwhelming evidence that flavonoids and foods rich in flavonoids protect LDL from oxidation, which may be one mechanism by which these foods decrease risk of cardiovascular disease. Catechins are more effective than vitamins E and C at protecting LDL particles from oxidation. Green tea is more effective than black tea. Notwithstanding these *in vitro* data, results from many studies have demonstrated that there is no effect of tea consumption on *ex vivo* LDL oxidizability.

Similar to tea polyphenols, resveratrol, a major polyphenol in red wine, has also been found to have a protective effect against LDL oxidation in some but not all studies. In contrast to the disparate findings from wine and tea and their flavonoids on LDL protection, studies consistently suggest that cocoa, chocolate, and the procyanins found in cocoa protect LDL from oxidation, both *in vitro* and *ex vivo*. *In vitro* studies of apples, apple juice, and apple extracts (rich in anthocyanins, flavan-3-ols, and flavonols) indicate that these foods also protect LDL.

The effect of antioxidants to protect lipids (fatty acids), protein, and DNA from oxidation is also

important to disease prevention. Of particular interest is the prevention of damage to the vasculature. Maintaining proper endothelial function and reactivity (a nitric oxide-mediated event) is critical since endothelial dysfunction is associated with the risk, etiology, and complications of cardiovascular disease and diabetes. Overproduction of reactive oxygen species, such as peroxynitrite, prevents the NO-facilitated vasodilation and leads to impaired endothelial function. In addition to impaired endothelial function, peroxynitrite induces other detrimental cellular changes and apoptosis. Nutrient antioxidants (vitamins E and C) have been shown to reverse oxidative stress and restore endothelial function in diabetics. Polyphenols from tea, wine, cocoa, and other foods have been found to be effective at *in vitro* scavenging of peroxynitrite. In addition to the effect of peroxynitrite on endothelial function, this and other reactive oxygen intermediates can damage protein, lipids, and DNA.

In summary, flavonoids from tea, wine, cocoa, and other food sources are very effective free radical and reactive oxygen species scavengers *in vitro*. However, *ex vivo* data, especially for the protection of LDL particles, are not consistent. Polyphenols are also effective at scavenging reactive oxygen intermediates. These reactive oxygen species can have a detrimental effect on endothelial function and can damage macromolecules. These events are thought to be important in the etiology of cardiovascular disease, cancer, and diabetes. Confirmation that these effects occur *in vivo* in humans with consumption of foods high in antioxidant phytonutrients is needed.

Epidemiologic Studies of Flavonoids and Tea Intake

Flavonoids

An association of flavonoid intake with a reduction of risk of cardiovascular disease has been suggested from several but not all epidemiologic studies. Among the first studies to demonstrate that flavonoid intake is inversely related to risk of coronary heart disease mortality was the Zutphen Elderly Study. Incidence of first fatal and nonfatal stroke is also lower in this population for individuals with the highest quartile intake of flavonoids compared to those with the lowest quartile. Despite the fact that intake of only a few flavonoids was measured (quercetin, kaempferol, myricetin, apigenin, and luteolin), tea, which is relatively low in these compounds, was identified as the most important dietary source. Onions and apples are also important sources of these flavonoids in this population. In the Seven Countries Study, flavonoid intake is inversely associated with coronary heart disease mortality. However, in Wales, flavonol intake (mainly from tea) is not associated with a decrease in risk of ischemic heart disease. In fact, in this population there is an increase in risk of ischemic heart disease between men consuming more than eight cups of tea per day and those consuming less than two cups per day. In other countries, such as Finland, the association between flavonoid intake and total and coronary mortality is not significant. However, in a Finnish cross-sectional study, quercetin intake was associated with a lower risk of ischemic heart disease and a trend in reduction of risk of type 2 diabetes. In the United States, the Health Professionals Study data suggest that there is no relationship between flavonol and flavone intake and risk of coronary heart disease in men. However, these data suggest that flavonoids may be protective for men with established coronary heart disease. Moreover, moderate and heavy tea drinkers in the year prior to an acute myocardial infarction had lower mortality compared to nondrinkers.

Tea

In epidemiologic studies designed to link tea intake with risk for cardiovascular disease, the data are inconsistent. Studies conducted in Norway and Scotland found no association between tea consumption and coronary heart disease and mortality. In the Netherlands, tea consumption is inversely related to severe aortic atherosclerosis but unrelated to mild or moderate aortic atherosclerosis. Moreover, tea consumption is inversely related to fatal myocardial infarction but not to nonfatal events. In Japan, green tea consumption is inversely related to the degree of stenosis for those consuming more than four cups per day but not for those consuming two or three cups per day compared to those consuming one cup per day. In the United States, tea consumption is associated with lower risk of myocardial infarction. In a meta-analysis of several studies, the incidence of myocardial infarction decreased 11% with an increase in tea consumption of three cups per day. However, data on tea consumption tend to be heterogeneous and biased, and geographic region explained much of the heterogeneity. For example, there was an increase in risk of coronary heart disease in the United Kingdom and an increase in risk of stroke in Australia but a decreased risk in other regions.

Results of epidemiologic data on cancer prevention from tea consumption are also mixed. An

inverse relation between urinary levels of tea polyphenols and gastric cancer was found in Chinese but not Japanese men. In the United States, tea consumption is related to a reduced risk of colon cancer. However, a comparable study was conducted in The Netherlands in which no association was found between black tea consumption and the risk of colorectal, stomach, lung, or breast cancer. A review of 30 papers examining populations in 12 countries found no consistent support for tea being a chemopreventive agent against colon and rectal cancers.

Wine and Chocolate

Epidemiologic studies of wine and alcohol have found a J-shaped curve, with nondrinkers and heavy drinkers at increased risk for all-cause mortality compared to moderate consumers. It is difficult to separate the effects of flavonoid and alcohol in these studies. There are limited epidemiologic data linking chocolate consumption with risk of chronic disease: The available data suggest that there is no association between chocolate consumption and risk of coronary heart disease.

In summary, epidemiologic evidence relating the consumption of flavonoids from tea or other foods to risk of cardiovascular disease and cancer is inconclusive. Geographic region appears to impact the results. This confounding factor may relate to other dependent dietary factors, lifestyle, environment, genetics, type of tea consumed and the quality of the flavonoid databases being used, and the quality of food intake data.

Antioxidant and Biomarker Evidence from Intervention Studies in Humans

There are limited data from diet-controlled randomized crossover studies of humans on tea and other flavonoid-containing foods. Most intervention studies, apart from design considerations, suffer from lack of diet control, making them difficult to interpret. Results from intervention studies that employ dietary recalls, food records, and self-administered diets are notorious for introducing error that can mask treatment effects. Clinical studies in humans have focused on the antioxidant capacity of blood and oxidative damage to protein, lipid, and DNA as well as a number of risk factors associated with cardiovascular disease, including lipids, hemostasis, platelet aggregation, endothelial function, and blood pressure. Interventions have included high- and low-flavonoid diets, tea, chocolate, cocoa, wine, grape extracts, and fruit juices.

Changes in Antioxidant Capacity

Antioxidant capacity of the blood may be one indicator of a food's ability to act as an *in vivo* antioxidant. Two commonly used measures of overall antioxidant capacity are the ORAC assay (oxygen radical absorbance capacity) and the FRAP assay (ferric reducing ability of plasma). Dietary interventions do alter the antioxidant capacity of blood. For example, individuals consuming a high cocoa and chocolate diet for 2 weeks have higher serum ORAC than when they consume a control diet. However, these results are not always consistent. For example, in individuals consuming procyanidins in similar amounts of cocoa powder and chocolate, there is no change in plasma ORAC after 6 or 12 weeks.

Green tea consumption rapidly increases FRAP but the effect appears to be short-lived and is perhaps related to the short half-life of the antioxidant tea phytochemicals in plasma. Moreover, green tea appears to have a greater antioxidant capacity (FRAP) than black tea, and the consumption of black or green tea with or without milk does not appear to effect the rise of the blood's antioxidant capacity. Extracts of food such as green tea extract and grape seed extract also increase postprandial antioxidant status. Although these findings are compelling, a more important outcome than a change in plasma antioxidant capacity is the protection of lipid, protein, and DNA from damage. This is particularly true since some studies show no effect on total antioxidant capacity but do show a change in the degree of protein or lipid oxidation.

DNA Oxidation and Damage

Protecting DNA from oxidative damage is thought to be an important mechanism in cancer prevention. Flavonoids may be important in reducing DNA damage, especially in smokers, who are exposed to significant, self-induced oxidative stress. For example, the rate of sister chromatid exchange in mutagen-stimulated peripheral lymphocytes was higher in smokers than nonsmokers, but the rate in smokers was reduced to that of the nonsmokers after 6 months' consumption of two or three cups of green tea per day. Another marker of DNA oxidation is 8-hydroxy-2′-deoxyguanosine. White blood cell content and urinary excretion of 8-hydroxy-2′-deoxyguanosine are reduced in smokers and nonsmokers following consumption of green tea compared to no tea. Furthermore, the reduction is greater for smokers than for nonsmokers. However, not all results are consistent. For example, in a study of healthy subjects consuming a

high- and low-flavonol diet (primarily from tea and onions), there was no difference in DNA base damage after 2 weeks, despite a significant increase in plasma quercetin after consumption of the high-flavonol diet.

Protein Oxidation and Damage

There are fewer data on the role that flavonoids may play on protecting protein from oxidative damage, and as with protection of DNA from oxidation, results of studies on their effect are not consistent. Quercetin provided from black current and apple juice increases the concentration of plasma 2-amnio-adipic semialdehyde residues (a marker of protein oxidation), and the increase in concentration of these semialdehyde residues increases with increasing intake of quercetin. These results suggest a prooxidant effect of the diet on protein (there was no effect on total antioxidant capacity and a decrease in lipid oxidation). However, consumption of grape seed extract (three times per day for 1 week) does not affect plasma 2-amnio-adipic semialdehyde concentration. Moreover, consumption of a green tea extract does not alter hemoglobin protein oxidation.

Lipid Oxidation and Damage

Whereas intake of quercetin from black current and apple juice appears to increase protein oxidation, it decreases plasma malondialdehyde concentration, a marker of lipid oxidation. Isoprostanes, which are specific and sensitive markers of lipid peroxidation, are not different after consumption of a high-flavonoid or low-flavonoid diet, after consumption of green or black tea, or after consumption of black tea and hot water. Similarly, there is little change in urinary isoprostane excretion after supplementation of dark chocolate and cocoa or after consumption of red and white wine (although dealcoholized wine does decrease urinary isoprostane).

Biomarkers of Diseases

Perhaps the most important outcomes in intervention studies are changes in risk factors or biomarkers of disease, especially for cardiovascular disease. Lipoprotein cholesterol (HDL and LDL) concentrations are considered important biomarkers for risk of cardiovascular disease. In a large ($N = 65$) study of men and women consuming six cups of black tea per day, there was no change in plasma LDL-cholesterol, HDL-cholesterol, or triglycerides compared to a control beverage. However, in a controlled diet study of 12 individuals with slightly elevated LDL-cholesterol, consumption of five cups of black tea per day compared to a control, caffeine-containing beverage resulted in a 6.5% decrease in total cholesterol and an 11.1% decrease in LDL-cholesterol after 3 weeks of consumption. There were no concomitant changes in HDL-cholesterol or triglycerides. A cholesterol-lowering effect of a theaflavin-enriched green tea extract has also been observed in a large ($N = 240$), double-blind, randomized outpatient study of individuals with mild to moderate hypercholesterolemia. In this study, individuals consumed either a capsule containing 375 mg of theaflavin-enriched green tea extract or a placebo capsule. After 12 weeks, the individuals taking the theaflavin-enriched green tea extract capsule had an 11.3% decrease in plasma total cholesterol and a 16.4% decrease in LDL-cholesterol from their pretreatment concentrations. The were no significant changes in total cholesterol or LDL-cholesterol in the placebo group and no change in HDL-cholesterol or triglyceride concentration in either group. In an average US diet supplemented with cocoa powder and dark chocolate, no change in LDL-cholesterol or the ratio of total cholesterol to HDL-cholesterol was found but HDL-cholesterol increased.

Although black tea was found to decrease platelet aggregation *in vitro*, similar to water, it had no acute or chronic effect on platelet aggregation in patients with coronary artery disease or in healthy individuals. Chocolate procyanidins increase plasma prostacyclin and decrease plasma leukotrienes, a possible mechanism by which chocolate flavonoids can decrease platelet activation. These changes occur rapidly (within 2 h) after consumption of chocolate and quickly disappear. However, in a well-controlled intervention, there was no effect of chocolate and cocoa powder on thromboxane and 6-keto-prostaglandin $F_{1\alpha}$ urinary excretion.

In addition to traditional biomarkers of cardiovascular disease, there are newer markers associated with cardiovascular disease that are related to endothelial function. One measure is brachial artery reactivity. In response to acute and chronic tea consumption, black tea consumption improves arterial reactivity in individuals with coronary artery disease. Tea consumption also improves arterial reactivity in mildly hypercholesterolemic individuals. Similar improvement in endothelial function occurred in individuals consuming purple grape juice.

Changes in whole body concentration of cytokines (especially those that control synthesis of acute phase proteins such as IL-6 and TNF-α), acute phase proteins (especially C-reactive protein), and soluble adhesion molecules (especially eSelectin, ICAM, and VCAM) are important indicators of

changes in cell signaling pathways, subclinical inflammation, and interactions of circulating cells with the endothelium. Some of these proteins have been identified as independent risk factors of cardiovascular disease and they remain important, emerging markers of disease. Despite the improvement in vascular reactivity associated with tea consumption, cell adhesion molecules do not appear to be altered by tea consumption in smokers and nonsmokers. These adhesion molecules are also related to endothelial function and are involved in the early etiology of atherosclerosis. Similar to tea, cocoa powder and chocolate consumption also do not alter proinflammatory cytokines, adhesion molecules, and acute phase proteins.

In summary, the observation that total circulating antioxidant capacity increases with consumption of foods high in flavonoids appears to be consistent in most studies, and tea is an important source of dietary flavonoids. However, improved circulating antioxidant status is not always consistent with a decrease in oxidative damage to DNA, protein, or lipid. Endothelial function, an important risk factor for cardiovascular disease, also appears to improve in response to catechins. A potential limitation of many studies is the lack of a controlled diet, which may account for the variability in results observed.

See also: **Antioxidants**: Diet and Antioxidant Defense; Observational Studies; Intervention Studies. **Cholesterol**: Sources, Absorption, Function and Metabolism. **Coronary Heart Disease**: Prevention. **Lipoproteins**.

Further Reading

Arab L and Il'yasova D (2003) The epidemiology of tea consumption and colorectal cancer incidence. *Journal of Nutrition* **133**: 3310S–3318S.

Balentine DA, Wiseman SA, and Bouwens LC (1997) The chemistry of tea flavonoids. *Critical Reviews in Food Science and Nutrition* **37**: 693–704.

Benzie IF, Szeto YT, Strain JJ *et al.* (1999) Consumption of green tea causes rapid increase in plasma antioxidant power in humans. *Nutrition and Cancer* **34**: 83–87.

Bingham SA, Vorster H, Jerling JC *et al.* (1997) Effect of black tea drinking on blood lipids, blood pressure and aspects of bowel habit. *British Journal of Nutrition* **78**: 41–55.

Davies MJ, Judd JT, Baer DJ *et al.* (2003) Black tea consumption reduces total and LDL cholesterol in mildly hypercholesterolemic adults. *Journal of Nutrition* **133**: 3298S–3302S.

Dreosti IE (2000) Antioxidant polyphenols in tea, cocoa, and wine. *Nutrition* **16**: 692–694.

Hara Y (2000) *Green Tea: Health Benefits and Applications*. New York: Marcel Dekker.

Hodgson JM, Croft KD, Mori TA *et al.* (2002) Regular ingestion of tea does not inhibit *in vivo* lipid peroxidation in humans. *Journal of Nutrition* **132**: 55–58.

Knekt P, Kumpulainen J, Jarvinen R *et al.* (2002) Flavonoid intake and risk of chronic diseases. *American Journal of Clinical Nutrition* **76**: 560–568.

Kris-Etherton PM, Lefevre M, Beecher GR *et al.* (2004) Bioactive compounds in nutrition and health—Research methodologies for establishing biological function: The antioxidant and anti-inflammatory effects of flavonoids on atherosclerosis. *Annual Review of Nutrition* **24**: 511–538.

Kris-Etherton PM, Lichtenstein AH, Howard BV *et al.* for the Nutrition Committee of the American Heart Association Council on Nutrition, Physical Activity, and Metabolism (2004) Antioxidant vitamin supplements and cardiovascular disease. *Circulation* **110**: 637–641.

Lambert JD and Yang CS (2003) Mechanisms of cancer prevention by tea constituents. *Journal of Nutrition* **133**: 3262S–3267S.

Maron DJ, Lu GP, Cai NS *et al.* (2003) Cholesterol-lowering effect of a theaflavin-enriched green tea extract: A randomized controlled trial. *Archives of Internal Medicine* **163**: 1448–1453.

Peters U, Poole C, and Arab L (2001) Does tea affect cardiovascular disease? A meta-analysis. *American Journal of Epidemiology* **154**: 495–503.

Princen HM, van Duyvenvoorde W, Buytenhek R *et al.* (1998) No effect of consumption of green and black tea on plasma lipid and antioxidant levels and on LDL oxidation in smokers. *Arteriosclerosis, Thrombosis, and Vascular Biology* **18**: 833–841.

Stensvold I, Tverdal A, Solvoll K *et al.* (1992) Tea consumption. Relationship to cholesterol, blood pressure, and coronary and total mortality. *Preventive Medicine* **21**: 546–553.

Wiseman SA, Balentine DA, and Frei B (1997) Antioxidants in tea. *Critical Reviews in Food Science and Nutrition* **37**: 705–718.

Teeth *see* **Dental Disease**

THIAMIN

Contents
Physiology
Beriberi

Physiology

D I Thurnham, University of Ulster, Coleraine, UK

Thiamin is a water-soluble vitamin and the structure comprises a pyrimidine and a thiazole ring linked by a methylene bridge (**Figure 1**). In its metabolically active forms, the hydroxyl group on the thiazole moiety is replaced by one, two, or three phosphate groups to form three phosphorylated coenzymes. A well-nourished human adult body contains approximately 30 mg of thiamin—approximately 80–90% as thiamin diphosphate (TDP), 10% as thiamin triphosphate (TTP), and a small amount of thiamin monophosphate (TMP) and thiamin. Like most water-soluble vitamins, there is no definable store in the body; the only reserves are thiamin coenzymes that are present in most cells in combination with appropriate thiamin-requiring enzymes. The predominant need for thiamin is linked to energy production but there is increasing evidence that thiamin is also needed for additional neurological functions. Thiamin is found in the aleuron layer of cereal grains as well as in animal food products such as liver. Man's desire for high-extraction cereal products in situations in which the diets contained little more than the cereal was a main contributory factor to the scourge of beriberi throughout much of Southeast Asia at the end of nineteenth and beginning of the twentieth century. Thiamin is relatively unstable and destroyed by poor cooking habits, and it is susceptible to degradation in foods that are not stored properly. Thiamin turnover is also quite rapid, and the absence of stores means that a continuous supply of thiamin is required. So thiamin status can be fairly rapidly impaired by factors affecting intake (e.g., vomiting and alcohol abuse) or excessive excretion (e.g., induced by diuretics). Thus, thiamin deficiency is sometimes a problem in pregnancy, in alcohol abuse, and in the elderly. Seasonal outbreaks can also occur in poor developing countries when energy output is high and cereals may have been stored for many months and food supplies are restricted.

Dietary Sources of Thiamin

Thiamin is present in most foods but cereal products provide most thiamin for most people in the world, although the source is fundamentally different in developing and more industrialized countries. In the developing world, unrefined cereal grains and/or starchy roots and tubers provide 60–85% of dietary thiamin, whereas most dietary thiamin in industrialized countries is obtained from fortified cereal products. In the United Kingdom, for example, wheat flour is fortified with 2.4 mg thiamin per kilogram and many breakfast cereals contain 30% or more of the daily thiamin requirement per portion. Thiamin is present in greatest amounts in brewers yeast, the germ and aleuron layers of fresh wheat, egg yolk, and mammalian liver. It is also present in meat flesh, particularly pork, and vegetables, nuts, and legumes (**Table 1**). Milk from both humans (0.49–0.79 μmol/l; 0.23 μg/4.2 MJ (1000 kcal)) and cows (1.18–1.48 μmol/l) is a poor source of thiamin. Thiamin is actively secreted into milk by the lactating mother, and it is of interest that the amount of thiamin in human milk is not increased by supplements, but the concentration

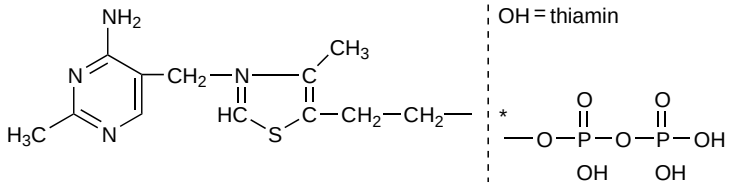


Figure 1 Thiamin and thiamin diphosphate (asterisk). Thiamin monophosphate and triphosphate are formed by the similar addition of one or three phosphate groups at the asterisk.

Table 1 Thiamin content of common foods

Food group	*Food item*	*Thiamin content (mg/100 g)*
Bread	Wholemeal	0.26
	White	0.18
	Hovis	0.52
Breakfast cereals	Cornflakes (fortified)	1.8
	Rice Krispies	2.3
	Weetabix	1.0
Flour	Wholemeal (100%[a])	0.46
	Brown (85%)	0.42
	White (fortified) (70%)	0.28–0.33
Milk, cheeses		0.03–0.06
Eggs	Cooked (various)	0.07–0.09
	Yolk raw	0.30
Vegetables (cooked)	Various leaf and root types	0.02–0.07
	Dahl, chick peas, green, beans, etc.	0.05–0.14
Pork products	Gammon rashers (lean)	1.0
	Bacon (various)	0.36–0.55
	Pork meat	0.5–0.88
	Liver (stewed)	0.21
Other meats	Beef (various)	0.03–0.09
	Lamb (various)	0.04–0.14
	Lamb liver	0.56
	Chicken (various)	0.04–0.10
	Game	~0.30
Yeast (dried)		2.33

[a]Percentages indicate the level of extraction in flour preparation.
Source: Paul AA, Southgate DAT (1978) *McCance & Widdowson's* The composition of food, 4th edn. London: HMO.

and of course the volume consumed increase during the first 6 weeks of lactation.

Refined foods in general, such as fat, sugar, and alcohol, are poor sources of thiamin. Polished rice is particularly low in thiamin (80 μg/100 g) and is especially important because of its widespread consumption and importance as a source of calories. Cereal grains lose thiamin during refining, but the process of parboiling rice before milling enables most of the thiamin to be retained (190 μg/100 g) since it migrates into the starchy endosperm during the procedure. Proper storage of cereal grains is also important to maintain thiamin activity. Studies in The Gambia, West Africa, found that old season millet, which had been stored under thatch and in high humidity, when consumed in the middle of the rainy season had thiamin concentrations (11 μg/100 g) that were 6–12 times lower than cooked samples obtained immediately postharvest. Imported rice used in the village likewise only contained 10 μg/100 g at the time of consumption.

Because of the water-soluble properties of thiamin, it can be leached from food during cooking. Thiamin is stable in slightly acid water up to boiling point but is unstable in alkaline solution that oxidizes it

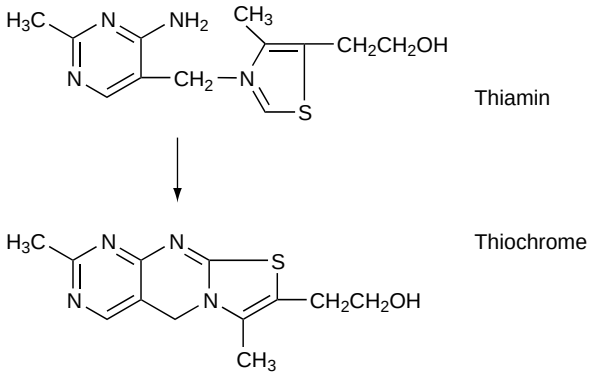


Figure 2 Structures of thiamin and thiochrome.

quantitatively to thiochrome (**Figure 2**). In addition, anti-thiamin factors in food can accelerate thiamin losses. Paralysis in foxes fed raw carp led to the discovery of the thiaminase enzymes. Two thiaminases are found in food. Thiaminase I is found in fish, shellfish, ferns, and some bacteria and catalyzes a base exchange reaction between thiazole and another base. Thiaminase II is a hydrolytic enzyme that cleaves the vitamin at the methylene bridge and is found mainly in bacteria. The thiaminases are heat labile, so only food that is eaten raw or fermented may loose thiamin during its preparation or in the gastrointestinal tract. There are also heat-stable anti-thiamin factors that are found in ferns, tea, betel nuts, large numbers of plants and vegetables, and some animal tissues. Anti-thiamin factors bind with varying degrees of attachment to thiamin and may or may not interfere with the bioavailability of thiamin. Diphenols, especially those with the hydroxyl groups in the *ortho* position, tend to react to give products that are both thiochrome negative and microbiologically inactive (i.e., thiamin is deactivated). Thus, in areas of northern and northeastern Thailand where tea drinking, chewing fermented tea leaves, chewing betel nuts, and consuming raw/fermented fish are common practices, thiamin deficiency still occurs despite thiamin intakes of 0.44–0.50 mg/4.2 MJ.

Absorption and Ethyl Alcohol

In food, thiamin occurs mainly as phosphate coenzymes and the predominant form is TDP (also called thiamin pyrophosphate and cocarboxylase). The phosphate coenzymes are broken down in the gut by phosphatases to give free thiamin for absorption. Thiamin is absorbed mainly from the upper intestine, and less thiamin is absorbed on an empty stomach than when taken with a meal. The latter could be due to the alkaline conditions in the duodenum, which are prevented by the presence of food.

Absorption of up to 2 mg per meal occurs by an active saturable process involving a sodium-dependent adenosine triphosphatase and against a concentration gradient. During absorption, thiamin is phosphorylated to the monophosphate ester (TMP). Thiamin is absorbed via the portal venous system. Further phosphorylation to TDP occurs on entry into all tissues. TDP can cross the blood–brain barrier, where a portion is converted to TTP, although even in the brain, TDP is the predominant form of thiamin. A second passive absorption process operates when intakes of thiamin are >5 mg but the maximum that can be absorbed from an oral dose is 2–5 mg.

The active process of absorption is impaired by ethyl alcohol. For example, 55% of a 5 mg dose of orally administered, labeled thiamin was recovered over 72 h in healthy adults, but this was reduced by 25–40% if they were previously given 1.5–2 g alcohol/kg. In people with fatty livers who had previously been abusing alcohol, mean thiamin absorption was reduced by 60%. However, the passive absorption of thiamin is not inhibited by alcohol, nor does it block entry of thiamin into the liver or interfere with thiamin metabolism in the tissues. Absorption of thiamin may also be reduced by gastrointestinal disturbances, such as vomiting and diarrhea, ulcerative colitis, and neoplasia, and in patients with hepatic disease and achlorhydria.

Transport, Storage, and Excretion

Thiamin with some TMP (19–75 nmol/l) circulates in the blood bound to albumin. When the binding capacity of plasma albumin is exceeded, or thiamin is in excess of tissue needs, it is rapidly excreted in the urine. Most thiamin in erythrocytes is present as TDP principally bound to the enzyme transketolase. Likewise, in most other tissues, there is very little free thiamin and it is mostly present as TDP (90%) in coenzymes bound to respective enzymes and a smaller amount of TTP (10%) in nervous tissues. The concentration of thiamin in specific tissues is on the order of 2–3 μg/g for heart muscle; 1 μg/g for brain, liver, and kidney; and 0.5 μg/g in skeletal muscle. Thiamin supplements can increase these concentrations slightly and prolonged febrile illnesses are likely to reduce them. Thiamin is mainly excreted intact in the urine but there are small amounts of thiochrome (**Figure 2**) and other thiazole and pyrimidine metabolites. A linear relationship exists between intake and excretion of thiamin until intake falls to an amount approaching minimum requirements when excretion decreases rapidly indicating a renal conservation mechanism.

There is concern that the long-term use of diuretics in the management of chronic congestive heart failure (CHF) may impair thiamin status and, as a consequence, impair myocardial function. The diuretic drug furosemide has been the subject of much attention. In healthy volunteers, a dose-dependent increase in urine flow accompanied by an increase in the urinary thiamin excretion rate have been demonstrated. In furosemide-treated patients, the concomitant presence of thiamin in the urine and biochemical deficiency of thiamin from measurements in blood has been shown. These results suggest that furosemide treatment can override the renal conservation mechanism. In one study, 23 patients with chronic CHF receiving 80–240 mg furosemide daily for 3–14 months were studied along with 16 age-matched controls without heart failure and not taking diuretics. No subjects in either group were identified as consuming inadequate thiamin intake or having increased thiamin requirements. However, biochemically, 21 of the 23 CHF patients and 2 of the controls were thiamin deficient. Furthermore, 5 of the CHF patients were treated with intravenous thiamin (100 mg thiamin HCl twice daily for 7 days). Biochemical thiamin status normalized and echocardiographic assessment of left ventricular ejection fraction increased in 4 of the 5 patients. Because no other changes were made in the patients' therapeutic regimen, the results suggest that the improvement in cardiac contractility was due to the correction of the thiamin deficiency.

Biological Functions

Thiamin functions as the coenzyme TDP in the metabolism of carbohydrates and branched-chain amino acids (α-keto-isocaproic, α-keto-β-methyl valeric, and α-keto-isovaleric acids). In association with Mg^{2+} ions, TDP is important (1) in various dehydrogenase complexes for the oxidation of α-keto acids (pyruvate, α-ketoglutarate, and the branched-chain α-keto acids) and (2) in the formation of α-ketols among the hexose and pentose phosphates catalyzed by transketolase (EC 2.2.1.1). Thus, a deficiency of thiamin has severe consequences for energy generation and amino acid interconnections, and these have important links with lipid metabolism, cell replication, and neural activity.

Two principal dehydrogenase complexes that require the participation of TDP are pyruvate dehydrogenase, which generates acetyl-CoA, and the oxidative decarboxylation of α-ketoglutarate to succinyl-CoA (**Figure 3**). Pyruvate dehydrogenase is situated at the junction of the glycolysis pathway,

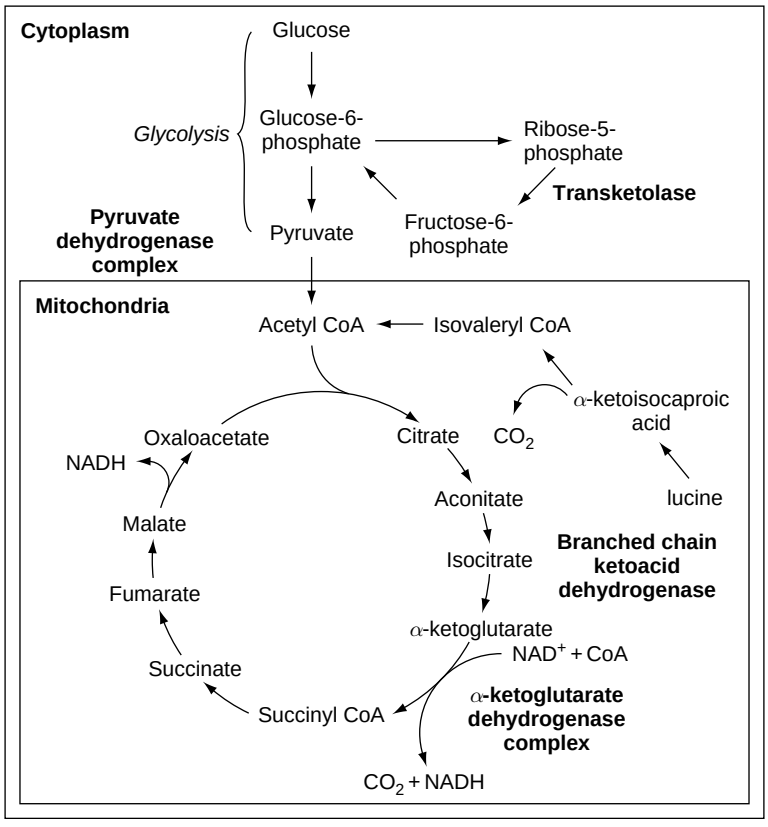


Figure 3 The four principal sites of action of thiamin diphosphate coenzyme in carbohydrate metabolism.

where it enters the tricarboxylic acid cycle. Acetyl-CoA is a key source of energy for mitochondrial oxidation and the production of adenosine triphosphate (ATP) as well as an important precursor in lipid metabolism. The impaired functioning of pyruvate dehydrogenase leads to a lactic acidosis, with increased concentrations of serum pyruvate and/or lactate especially as a result of exercise. The lactate acidosis can be explained by the fact that ATP depletion stimulates glycolysis, thus generating more pyruvate. As pyruvate concentrations increase, lactate dehydrogenase converts some of the pyruvate to lactate, producing the lactic acidosis. The increases in these compounds formed the basis of the earliest biochemical test for thiamin deficiency, which was later made more reproducible by taking the blood soon after moderate exercise (e.g., climbing a few steps).

Many features of beriberi indicate that thiamin plays an important role in neural tissues. TTP is specifically found in nervous tissues, but although this triphosphorylated metabolite of thiamin has been known for approximately 30 years, its precise role is still in doubt. TDP in the dehydrogenase complexes is undoubtedly also required for normal function. Some of the earliest biochemical studies on the brain documented abnormalities in the oxidative metabolism of glucose and a disruption in energy supply may underlie many of the neurochemical changes and structural lesions associated with thiamin deficiency. For example, acetyl-CoA produced by pyruvate dehydrogenase is a precursor of the parasympathetic transmitter molecule acetylcholine, but the obligatory requirement of glucose as an energy source for nervous tissue indicates the essentiality of TDP. Likewise, the cytosolic enzyme transketolase is also present in nervous tissue, and as a key enzyme in the HMS it may be important in minimizing oxidant stress. The HMS generates NADPH, which is required to maintain glutathione in the reduced state.

The cellular and subcellular localization of the enzymes responsible for metabolism of thiamin phosphates in nervous tissues may indicate possible sites of action of the specific metabolites. Thiamin that enters the brain is phosphorylated by thiamin pyrophosphokinase to form TDP. The concentration of thiamin phosphates is 3 or 4 times higher in neurons than in neuroglia, and the activity of thiamin diphosphatase (TDPase), which converts TDP to TMP, is 20 times higher in neurons than neuroglia. Thiamin monophosphatase is only detected in neuroglia. Within the neuron, TDPase is mostly

localised in the microsomal fraction. Thiamin triphosphatase (TTPase), which converts TTP to TDP, is particularly enriched in presynaptic terminals. Stimulation of nerves or treatment with certain neuroactive drugs result in decreases in TDP and particularly TTP in the nerve, with an increase in free TMP in the surrounding fluid. It is postulated that TTP plays an essential role in nerve transmission involving a gating mechanism for sodium and potassium ion transport via the specific ATPase. Some evidence for this comes from patients with Leigh's disease (pathologically similar to Wernicke–Korsakoff syndrome), in whom severe neurological disease is accompanied by a deficiency in TTP but normal TDP concentrations.

The well-documented role of mitochondria in programmed cell death and the importance of thiamin for oxidative metabolism have stimulated investigators to examine brain thiamin homeostasis in neurodegenerative diseases. Diminished thiamin-dependent processes, abnormal metabolism, and oxidative stress accompany the neurodegeneration of Alzheimer's disease (AD), Huntington's disease, Wernicke–Korsakoff syndrome, progressive supranuclear palsy, and the adult-onset neurodegenerative diseases that are caused by genes containing variable numbers of CAG repeats within their coding regions. Abnormalities in the thiamin-dependent processes have also been linked with thiamin-responsive maple syrup urine disease, Leigh's disease (a subacute necrotizing encephalomyelopathy), sudden infant death syndrome, cerebellar degeneration, thiamin-responsive anemia, ataxia, and disorders of energy metabolism including pyruvate dehydrogenase deficiency. The extent to which disturbances in thiamin metabolism are a cause or a consequence of the disease process is still under examination.

Assessment of Thiamin Status

Thiamin status can be assessed using methods that measure thiamin or its metabolites in plasma, erythrocytes and urine (**Table 2**). Samples are acidified to stabilize the thiamin and precipitate any protein. Usually, thiamin is oxidized to thiochrome (**Figure 2**) using cyanogen bromide in alkaline solution and measured by fluorescence with or without chromatography. Concentrations of thiamin in urine and plasma tend to reflect dietary intake, being high when intake is adequate and low when dietary sources are poor. Erythrocyte thiamin is mainly in the form of the coenzyme TDP, which can be extracted from washed erythrocytes, derivatized as described previously, and quantified by high-performance liquid chromatography. The most popular test, however, is the erythrocyte transketolase (ETKL) stimulation test, which measures enzyme activity with and without added TDP. The reference range for ETKL activity in well-nourished, thiamin-adequate people is reported to be 570–830 mU/g hemoglobin. The stimulation test measures the proportion of the apoenzyme in red cell homogenate (i.e., the proportion that is not bound to TDP and represents the degree of thiamin deficiency). Studies have shown that results from the urinary assay for thiamin agree reasonably well with those obtained by the ETKL stimulation test.

Table 2 Biochemical assessment of thiamin status

Test	*Acceptable*	*Marginal risk*	*High risk*
Urinary thiamin (μmol/mol creatinine)[a]			
1–3 years	>66	45–66	<45
4–6 years	>45	32–45	<32
Adults	>25	10–25	<10
Erythrocyte transketolase activity			
Activity coefficient	<1.11	1.11–1.25	>1.25
TDP effect (%)	<11	11–25	>25
Red cell thiamin concentrations (nmol/l)	749 ± 196		~560[b]
Whole blood thiamin concentrations (nmol/l)	166–266		<133

[a]Converted from μg/g creatinine using the factor (×0.376).
[b]Based on a decrease of 25% in red cell thiamine diphosphate (TDP).

One of the reasons for the popularity of the ETKL stimulation test is that sensitivity is still good even in the presence of thiamin deficiency. In all other measurements of thiamin status, as deficiency approaches, the quantity of thiamin or its metabolites diminishes in the biological fluid. Low concentrations of a product are usually more difficult to measure and precision deteriorates, or the amount of sample has to be increased to provide sufficient material to detect. In contrast with the ETKL stimulation test, in an acute thiamin deficiency, ETKL activity is maintained and only the amount of TDP decreases, so the test becomes more sensitive. However, in chronic thiamin-deficient states, the apoenzyme of ETKL is reported to be unstable *in vivo*, and in the absence of the coenzyme, concentrations of the apoenzyme decrease, with the result that *in vitro* stimulation may show normal thiamin status. Thus, in situations in which chronic thiamin deficiency is suspected as a result of a long-term marginal thiamin intake, alcohol abuse, or use of diuretics for many months, one or more of the concentration tests may be useful as an adjunct to the stimulation test.

Certain precautions should be taken in handling samples for thiamin analysis. Urine should be acidified to avoid degradation and stored below −20 °C. Heparinized whole blood should be collected and immediately put on ice. For total erythrocyte TDP measurements, cells are separated from plasma within 2 h when possible, washed in saline, and diluted 1:1 with saline prior to acidification. Centrifugation of the acidified mixture provides a clear extract that can be stored for no more than 5 days at 4 °C or longer at ≤−20 °C. Washed red cells are also used for the ETKL assay. Duplicate tubes of the red cells in saline suspension with and without added TDP are mixed and can be stored at −70 °C prior to enzymatic analysis of ETKL activity. Even at −70 °C, however, storage should be for no more than a few weeks. The ETKL apoenzyme is unstable, and even in the tubes to which TDP has been added, if mixing did not thoroughly expose all apoenzyme to the added coenzyme, deterioration will occur and results will be unreliable.

Recommended Dietary Allowances

Quantifying thiamin requirements is based on a variety of biochemical data. Early results indicated that a thiamin intake of 0.4 mg/day on a low-energy intake was close to the absolute minimum requirement. Epidemiological evidence suggested that beriberi occurred when the intake of thiamin was <0.2 mg thiamin per 4.2 MJ (1000 kcal); however, when 0.188 mg/4.2 MJ was fed to sedentary elderly men for 2 years, no indisputable alteration in clinical state occurred. Thiamin requirements are strongly influenced by physical activity and at higher energy intakes with liquid formula diets containing 11.76 and 15.12 MJ (2800 and 3600 kcal), there was good agreement between thiamin excretion and ETKL stimulation to interpret thiamin status at different levels of thiamin intake. Increasing intake from 0.2 to 0.23 mg/4.2 MJ moved first the urinary excretion and then ETKL activation out of the deficient range. Both measurements were normalized at intakes of 0.3 mg/4.2 MJ, and to allow for variance the recommended nutrient intake adopted by the Department of Health in the United Kingdom was 0.4 mg/4.2 MJ. This amount is recommended for all groups of the population since additional needs in pregnancy and lactation are met by increased energy intakes. It was recommended that formula feed should contain not less than 0.3 mg/4.2 MJ.

Women are less affected by beriberi than are men even when they are consuming the same diet, but there is no consistent indication that men have greater needs than women. Differences between the sexes that may affect susceptibility to beriberi need further investigation (e.g., the amount of food eaten by the sexes when supplies are short or of poor quality, metabolic responses to infection during illness, and differences in energy requirements). The close association between thiamin metabolism and carbohydrate metabolism means that thiamin requirements are determined by basal metabolic rate (BMR) and physical activity. BMR of men is slightly higher than that of women of the same weight, but total energy expenditure can vary 1.4 to 2.5 times BMR depending on physical activity.

Drug–Nutrient Interactions

Mention has already been made of the influence of alcohol and diuretics on thiamin status. Oral contraceptives are reported to have no effects on thiamin status.

Toxicity

High intakes of thiamin administered orally are nontoxic. The rapidly saturable thiamin absorption mechanism limits the amount taken up from a single dose to ~2.5 mg, and thiamin present in excess of protein binding capacity is excreted. However, there are reports of toxicity from chronic intakes in excess of 50 mg/kg or >3 g/day with a wide variety of clinical signs, including headache, irritability, insomnia, rapid pulse, weakness, rapid pulse contact dermatitis, pruritus, and, in one case, death.

See also: **Alcohol**: Disease Risk and Beneficial Effects. **Carbohydrates**: Regulation of Metabolism. **Cereal Grains**. **Drug–Nutrient Interactions**. **Thiamin**: Beriberi.

Further Reading

Bender DA (1984) B vitamins in the nervous system. *Neurochemistry International* **6**: 297–321.

Department of Health (1991) *Dietary Reference Values for Food Energy and Nutrients for the United Kingdom*, Report on Health and Social Subjects No. 41. London: HMSO.

Gibson GE and Zhang H (2002) Interactions of oxidative stress with thiamine homeostasis promote neurodegeneration. *Neurochemistry International* **40**: 493–504.

Hoyumpa AMJ, Nichols SG, Wilson FA, and Schenker S (1977) Effect of ethanol on intestinal (Na,K) ATPase and intestinal thiamin transport in rats. *Journal of Laboratory & Clinical Medicine* **90**: 1086–1095.

Sauberlich HE, Herman YF, Stevens CO, and Herman RH (1979) Thiamin requirements of the adult human. *American Journal of Clinical Nutrition* **32**: 2337–2248.

Seligmann H, Halkin H, Rauchfleisch S *et al.* (1991) Thiamine deficiency in patients with congestive heart failure receiving long-term furosemide therapy: A pilot study. *American Journal of Medicine* **91**: 151–155.

Sinclair HM (1982) Thiamin. In: Barker BM and Bender DA (eds.) *Vitamins in Medicine*, 4th edn., pp. 114–167. London: Heinemann Medical Books.

Suter PM and Vetter W (2000) Diuretics and vitamin B_1: Are diuretics a risk factor for thiamin malnutrition? *Nutrition Reviews* **58**: 319–323.

Thurnham DI (2004) An overview of interactions between micronutrients and between micronutrients with drugs, genes and immune mechanisms. *Nutrition Research Reviews* **17**: 211–240.

WHO/FAO (2002) *Human Vitamin and Mineral Requirements; Report of a Joint FAO/WHO Expert Consultation.* Rome: WHO/FAO.

Beriberi

D I Thurnham, University of Ulster, Coleraine, UK

Beriberi is caused by a deficiency of thiamin (also called thiamine, aneurin(e), and vitamin B_1). Classic overt thiamin deficiency causes cardiovascular, cerebral, and peripheral neurological impairment and lactic acidosis. The disease emerged in epidemic proportions at the end of the nineteenth century in Asian and Southeast Asian countries. Its appearance coincided with the introduction of the roller mills that enabled white rice to be produced at a price that poor people could afford. Unfortunately, milled rice is particularly poor in thiamin; thus, for people for whom food was almost entirely rice, there was a high risk of deficiency and mortality from beriberi. Outbreaks of acute cardiac beriberi still occur, but usually among people who live under restricted conditions. The major concern today is subclinical deficiencies in patients with trauma or among the elderly. There is also a particular form of clinical beriberi that occurs in patients who abuse alcohol, known as the Wernicke–Korsakoff syndrome. Subclinical deficiency may be revealed by reduced blood and urinary thiamin levels, elevated blood pyruvate/lactate concentrations and α-ketoglutarate activity, and decreased erythrocyte transketolase (ETKL) activity. Currently, the *in vitro* stimulation of ETKL activity by thiamin diphosphate (TDP) is the most useful functional test of thiamin status where an acute deficiency state may have occurred. The stimulation is measured as the TDP effect.

Epidemiology

Beriberi presents in several different clinical forms (**Table 1**). Beriberi became endemic following the introduction of steam-powered rice mills, which enabled milled rice to be produced cheap enough so that almost everybody could afford it and consume it. It was particularly serious at the end of nineteenth and the beginning of the twentieth centuries when seasonal epidemics of wet beriberi occurred with many deaths. The disease affected mainly the Chinese and Japanese populations, although outbreaks were reported in India and among settlers in the New World during the long cold winters, and the disease was not necessarily confined to rice-eating populations. Where acute cardiac beriberi occurred, dry beriberi was also present but usually in the older members of the community.

Milled rice has a thiamin concentration that is particularly poor (80 μg/100 g), but social conditions

Table 1 Forms of beriberi in man

Subclinical beriberi	Identified by transketolase activity or other biochemical tests of thiamin status. May be associated with early subjective symptoms such as anorexia, weakness, dysthesie, and depression. Responds rapidly to treatment with thiamin.
Wet beriberi	Subacute or cardiac beriberi frequently having muscular pains, oedema of feet and legs, enlarged heart, and tachycardia. Responds rapidly to treatment with thiamin. Major form and was typically seasonal in endemic areas. Acute fulminant type of beriberi in which the main feature is dominated by insufficiency of the heart and blood vessels. Responds rapidly to treatment with thiamin.
Dry beriberi	Chronic, atrophic type of polyneuropathy in which the main features are of a weak wasted person, with painful musculature making walking difficult, impaired sensory nerves and tendon reflexes, and flaccid paralysis of the motor nerves. Poor or no response to treatment with thiamin.
Infantile beriberi	Usually acute wet beriberi. Responds rapidly to treatment with thiamin.
Wernicke–Korsakoff syndrome	Predominantly neurological, affecting walking and vision in most and memory and cardiac function in over 50% of patients. Wernicke or ocular component responds rapidly to treatment but the Korsakoff psychosis responds slowly or not at all.

Modified from Thurnham DI (1978) Thiamin. In: Rechcigl M Jr (ed.) *Nutrition Disorders*, pp. 3–14. West Palm Beach, FL: CRC Press.

at the time of the large epidemics contributed to the problems. Bonded labor was common, with workers living on the work premises most of the time and paid mainly in the form of rice. In addition, reports at the time suggest that the rice was of uncertain freshness and quality, and that it could be so mouldy, matted, and lumpy that it had to be remilled and washed, with a further loss of thiamin. The social conditions prevented natural eating practices because workers had little money to purchase additional food and they were dependent on what they were given. Likewise, badly stored cereals can lose up to 90% of the thiamin content, and toxins associated with mould growth have been implicated in causing sickness that may well precipitate clinical beriberi.

Reports suggest that the acuteness of the outbreak of beriberi and the interrelationship of thiamin deficiency with deficiencies of other nutrients probably had a major role in determining the nature of the pathological changes and lesions produced. For example, it is reported that protein energy malnutrition almost always accompanied subacute beriberi, reflecting the link between impoverishment and the disease. In contrast, it is also suggested that severe beriberi more often affected the more active, stronger, or supposedly better nourished members of the community. The younger, stronger rickshaw puller was most likely to suffer severe beriberi. Likewise, infantile beriberi appeared to affect the male infant who 'tended to be overfed.' This enigma may be due to thiamin intakes from a diet containing a high proportion of rice being insufficient to meet the thiamin requirement posed by the higher calorie intakes of the more active community members. In the case of infantile beriberi, slightly deficient mothers probably produced milk that was only marginally adequate in thiamin and/or infants were given supplements of thiamin-poor rice. It is a common habit even today for rural mothers to give very young infants, even beginning at 1 week of age, a bolus of masticated rice to supplement the milk intake. The inability to match thiamin intake to energy needs may also explain why nonspecific pyrexia was a precipitating factor for beriberi. A 1 °C rise in body temperature is associated with a 10% increase in basal metabolic rate. It has been suggested that more than half the mild cases of beriberi were associated with a nonspecific bout of fever, and such cases responded less readily to treatment with thiamin.

Parboiled rice is partially cooked before milling, and this prevents beriberi because the thiamin is dispersed through the grain (190 μg/100 g). The advantages of this were clearly seen in Malaya, where at the end of nineteenth century there were large-scale immigrations of young, able-bodied Chinese to work in the tin mines and Indians to work on the rubber estates. In both cases, immigrants often lived in remote regions where there was little opportunity to purchase local food and they were dependent on imported rice. It was the Chinese who, because of their dietary preference for milled rice, died in enormous numbers.

Although ways of avoiding the disease were known to the Japanese navy at the end of the nineteenth century, since the director general of the medical department had demonstrated that the disease was almost eradicated if the traditional rice diet was supplemented with fish, vegetables, meat, and barley, this information was not widely available, and supplementation was not feasible by the vast majority of people. It was widely believed that the cause of beriberi was an infection or toxin resulting from bad food. In particular, Pasteur's work on the microbiological cause of infections led many to search for an infectious agent, but none could be consistently identified. The scale of the problem for the colonial powers in Southeast Asia in the latter part of the nineteenth and early twentieth centuries should not be underestimated. Labour was cheap but the death toll posed enormous problems. Extracts from reports at the time are illuminating: In 1887, there were 690 deaths out of 1931 native government officers in Sumatra, infant mortality was 445 per 1000 live births in the Philippines in 1910, and one report stated that there were so many deaths that "there was insufficient earth to bury the corpses."

The Dutch government sought to resolve the situation by appointing a medical bacteriologist, Christiaan Eijkman, to travel to Indonesia to investigate the problem. Working in Java, he showed within 6 years that beriberi was a nutritional problem and that a paralytic condition closely resembling the polyneuritic symptoms of beriberi could be produced in chickens by feeding them both stale and freshly cooked polished rice. However, it was Funk in 1911 who first reported the isolation of a 'vital amine' from rice polishings that had anti-beriberi properties. Funk was the first person to coin the word 'vitamine' as a substance essential for life. The structure and synthesis of thiamin were reported in 1936.

Currently, clinical beriberi no longer occurs with the devastating effects of former years. Considerable improvements have occurred in nutrition worldwide, the diversity of foods available, the quality of food due to improved storage methods, and social and economic structures in many countries, especially in Southeast Asia. However, sporadic

outbreaks do occur, which are usually of the acute, fulminating type of beriberi, and deaths still, occur often in young men aged 20–40 years. Usually, a combination of factors is responsible, but once the cause is identified, treatment is cheap and readily available and, if given rapidly, tragic circumstances can be averted.

Two iatrogenic causes of subclinical beriberi are known, namely that associated with diuretic treatment and one resulting from alcohol abuse. Both are of concern because the use of diuretics is introduced to manage cardiovascular disease, a condition that will deteriorate if thiamin status is impaired, and alcohol abuse can lead to Wernicke–Korsakoff syndrome, which can have many of the features of both wet and dry beriberi.

Severe multisystem trauma, endotoxemia, or situations in which there is a raised metabolic demand for thiamin, such as pregnancy, thyrotoxicosis, and intercurrent illness or impaired absorption (e.g., alcohol abuse or gastrointestinal disease or resection), can produce subclinical evidence of thiamin deficiency or more severe life-threatening aspects of beriberi, such as renal and/or cardiovascular failure. The elderly may be particularly at risk of subclinical thiamin deficiency. One Belgian study on patients with a mean age of 83 years reported that 40% had a raised TDP effect (>15%), in whom there was a high proportion of Alzheimer's disease, depression, cardiac failure, and falls. The diuretic furosemide was also more frequently taken by the thiamin-deficient patients.

Table 2 Aetiological factors contributing to thiamin deficiency

Dietary thiamin deficiency	Commonly milled rice
High dietary carbohydrate to fat ratio	Metabolism of carbohydrate requires thiamin, whereas metabolism of fat spares thiamin requirements
Heavy physical activity	Predisposes to beriberi when accompanied by low intake of thiamin
Protein energy malnutrition	Older literature reports sometimes accompanies subacute beriberi indicating importance of impoverished diet
Poor storage conditions for food	Fall 6- to 10-fold in thiamin content of cereals. Moulds may accelerate decay as well as increase risk of toxins
Thiaminases	Two known, but only of importance when uncooked foods are consumed
Anti-thiamin factors	Factors in food that chelate with thiamin and potentially reduce bioavailability
Alcohol abuse	Alcohol impairs the active absorption mechanism for thiamin
Infection and trauma	Increase requirements for thiamin to support increased carbohydrate metabolism and energy production
Diuretics, long-term use	Accelerate thiamin excretion and appear to block thiamin control mechanism
Seasonal factors	Combination of heavy work load, impoverished diet, and last season's (badly stored) cereals
Male sex	Some evidence that men have higher thiamin requirements than women but more likely to be a combination of the first three factors listed here

Etiology

The factors associated with the various forms of beriberi are listed in **Table 2**. Beriberi is caused by a lack of thiamin in the diet, but the onset of the disease and the symptoms associated with the disease are influenced by one or more of the other etiological factors. Wet beriberi (i.e., cardiac beriberi) and Wernicke's encephalopathy are conventionally described as acute manifestations of the disease and respond most rapidly to treatment. In contrast, dry beriberi is described as due to a chronic deficiency of thiamin and does not respond well to treatment. However, experimental acute deficiency studies, which very rapidly produced subjective feelings of malaise and weakness at the slightest exertion, very rarely produced evidence of oedema and peripheral pain. These observations suggest that all forms of beriberi are probably preceded by an indeterminate period of chronic thiamin deficiency during which pathophysiological adaptations to the marginal nutritional state occur. Thus, physiological adaptations to the vascular system may well have occurred particularly in those who did heavy physical work and needed to overcome the weakness and malaise imposed by a low thiamin diet. The factor(s) that precipitated the clinical disease may not be thiamin at all. Platt, in his descriptions of beriberi in China, recounts how humid weather and infections such as malaria increased the number of cases of wet beriberi. The extra energy needed to cool the body in hot conditions or fuel the rise in temperature during infection may have imposed a critical burden on energy production that the system could not meet, and beriberi ensued.

However, the increased number of cases associated with heat, humidity, and malaria may also be due to a seasonal decline in the quality of food. A 6- to 12-fold decline in thiamin content is reported for millet when stored under traditional thatched storage houses in The Gambia, and reports suggest that much of the rice consumed late in the

season was not in the best condition. Some of the products introduced by mould growth may possess anti-thiamin properties that impair thiamin bioavailability. Thus, the ratio of thiamin to calories is likely to fall during the agricultural year and to be at its worst when calorie requirements are at their highest for land preparation and weeding. Land preparation also takes place just prior to or at the beginning of the rainy season, when the prevalence of malaria and diarrhoeal diseases increases.

Thiaminases are inactivated by cooking; thus, the enzymes are only a problem where certain foods are eaten raw. It has been suggested that in northern Thailand, where consumption of fermented raw fish products is widely practised and raw molluscs are eaten, thiamin status may be impaired by these food habits. Even as recently as 2001, marginal thiamin status was reported in more than 50% of women 3 months postpartum despite thiamin supplements of 100 mg/day during pregnancy. The deficiencies were found in Karen refugee women living on the Thai–Burmese border and whose diet contained fermented fish, tea leaves, and betel nuts—substances suspected of containing thiaminases. Polyphenol compounds in tea and many vegetables may also posses anti-thiamin properties and impair bioavailability, but their etiological importance in causing thiamin deficiency is difficult to assess.

Alcohol is an important factor in causing thiamin deficiency because it inhibits the active transport of thiamin across the gut and when abused it impairs the quality of the diet consumed. Diuretics accelerate the excretion of thiamin and appear to override the renal conservation mechanism. Their use is of potential concern in elderly people whose diet may be poor for other medical reasons and their physicians may be unaware of their need for supplemental nutrient.

Both sexes are vulnerable to the effects of thiamin deficiency, but in many of the sporadic outbreaks that have been reported, there appears to have been a male excess. This may be due to higher thiamin requirements in men than women because of their higher lean body mass or to hormonally driven sex differences. However, it is also possible that the cause is due to a higher risk of a thiamin:calorie imbalance in men compared to women. In many rural communities, men traditionally eat first and may satisfy their calorie requirements, whereas their womenfolk make do with the leftovers. Because of their greater physical strength, men frequently do heavier work than women, requiring more energy (i.e., more food to meet their requirements). Thus, men may consume more of the thiamin-depleted cereals in the diet to satisfy calorie needs and in so doing achieve a poorer thiamin:calorie ratio than women.

Experimental Thiamin Deficiency in Man and Measurement of Thiamin Status

In young and healthy nonalcoholic subjects, subjective symptoms appear after 2 or 3 weeks of deficient diet but urinary thiamin will already be falling (**Table 3**). Characteristic early symptoms include anorexia, weakness, dysthesiae, and depression. At this stage, urinary thiamin will be almost zero, ETKL activity depressed, and the TDP effect approximately 15–30%. After 6–8 weeks the only objective signs at rest may be a slight fall in blood pressure and moderate weight loss, although urinary thiamin will now be negligible and the TDP

Table 3 Effects of thiamin deficiency on urinary thiamin, the erythrocyte transketolase TDP effect, and early clinical symptoms of thiamin deficiency in human volunteers

Days of deficiency	*Urinary thiamin (μg/day)*[a]	*TDP effect (%)*[a]	*Clinical signs of deficiency following diets containing 150–350 μg thiamin/day*[b]
5	50	0–10	Mostly studies report no signs but one study (360 μg/day) found within 1 week chest
10	25	~15	pains, extreme lassitude, anorexia, palpitation, and burning feet
21–28	<25	~30	Loss of body weight, anorexia, general malaise, insomnia, increased irritability, fatigue on slightest exertion
30–40	Negligible	≥40	Increased malaise, loss of body weight, intermittent claudication and polyneuritis, bradycardia, peripheral oedema,[a] cardiac enlargement,[a] ophthalmoplegia
>45	10–20	>40	Additional signs of nausea and dizziness appeared
~75	10–20		Additional signs of vomiting, low blood pressure, and tenderness of calves

[a]Biochemical data and report of oedema and cardiac enlargement from Brin (1964), in which healthy male medical students were fed 200 μg thiamin per day for 6 weeks. TDP effect is a measure of thiamin status obtained by measuring the activity of erythrocyte transketolase in the presence and absence of added thiamin diphosphate.

[b]Clinical signs adapted from several studies. Investigators were impressed by the rapid degree of debility induced by the specific withdrawal of thiamin from the diet. In one group (150 μg/day for 75 days, four female mental patients), the authors reported that the condition more closely resembled 'neurasthenia' than beriberi and noted that oedema, cardiac dilation, and peripheral pain characteristic of classic beriberi were all absent (reported by Carpenter, 2002).

effect ≥35%. After 2 or 3 months, apathy and weakness become extreme, calf muscle tenderness develops, and there is loss of recent memory, confusion, ataxia, and sometimes persistent vomiting. Urinary thiamin will be negligible and the TDP effect may be normal (because apo-ETKL is unstable even *in vivo*), but ETKL activity should be considerably depressed.

The clinical symptoms resulting from experimental thiamin deficiency in man have usually responded rapidly to treatment with thiamin. In one feeding study, however, two mental patients were kept for 110 days on a diet providing 200 μg thiamin daily and 1 mg of thiamin by injection 1 day each week; thus, their overall weekly average was 350 μg/day. They developed a polyneuropathy characterised by defects in the sensory nervous pathways, loss of tendon reflexes, and paralysis of the legs, which took many weeks to respond to large doses of thiamin, and in one case response was still incomplete after 4 months of treatment. The slow cure suggested that degeneration of peripheral nerves had occurred, as is indicated in the dry form of beriberi, in which the neurological lesions are irreversible.

Clinical Features of Beriberi

Depletion and repletion studies suggest that intakes >300 μg/4.2 MJ are compatible with normal biochemistry and good health, and clinical signs of thiamin deficiency occur at intakes of thiamin below 200 μg/4.2 MJ (1000 kcal). The disease as studied from the 1880s onward in Asians subsisting on white rice began typically with weakness, 'wandering pains' in the legs, and lack of feeling in the feet. Some patients then developed oedema (the presence of excessive amounts of fluid in the intercellular tissue spaces of the body) of the legs, trunk, and face. In severe cases, sufferers found it increasingly difficult to catch their breath and would die of heart failure. The clinical features of subacute and acute wet beriberi are summarized in **Table 4**. The main form was subacute beriberi, which was typically seasonal in endemic areas. There are reports that the peripheral muscles most severely affected were those most frequently used; thus, in male laborers it was the legs. Aching pain, tightness, and cramps in the calf and associated muscles were usually a first cause of complaint, and pain on squeezing the calves was one of the most useful diagnostic tests for beriberi. In women who performed repetitive tasks involving hands and arms, a loss of sensation in the fingers was frequently a first cause of complaint.

Dry beriberi is essentially a chronic condition showing muscular atrophy and polyneuritis and frequently occurring in older adults. Walking is usually difficult because of the weak wasted and painful musculature, and in the later stages feeding and dressing may also become impossible. When bed-ridden and cachetic (extreme state of malnutrition and wasting), patients become very susceptible to infections. Sensory nervous function is impaired (hypoesthesia) almost to the point of anesthesia. Hypoesthesia is particularly evident in the extremities and progressively extends over the outer aspects of the legs, thighs, and forearms. Motor nerve disturbances also begin in the extremities and ascend progressively. Flaccid paralysis of the extensor muscles precedes that affecting the flexors and results in 'wrist drop' and 'foot drop' (**Figure 1**). Loss of the Achilles tendon reflex usually precedes an impaired patellar reflex.

Mortality from infantile beriberi mainly affected breast-fed infants between the second and fifth months of life, when solid foods were often first introduced. The introduction of white rice porridges, poor in thiamin, to a rapidly growing child and/or the increased exposure to infections when solids are introduced may both have contributed to infantile beriberi. The onset of the disease was rare in the first month and early signs could be mild and somewhat subjective (e.g., vomiting, restlessness, anorexia, and insomnia). Early signs could progress to subacute infantile beriberi, the acute and usually fatal condition, or a chronic form. Features of acute infantile beriberi are presented in **Table 5**. The subacute form was characterized by slight oedema in the form of puffiness, vomiting, abdominal pain, oliguria, dysphagia, and convulsions. In addition, aphonia (soundless cry) was often a feature of subacute infantile beriberi and may have been due to nerve paralysis or oedema of vocal cords. Vomiting was also a feature of chronic infantile beriberi and could be accompanied by inanition, anemia, aphonia, neck retraction, opisthotonus, oedema, oliguria constipation, and meteorismus (swelling of the abdominal cavity from gas in the intestine). Opisthotonus is a characteristic of acute thiamin deficiency in birds and is described as due to a tetanic spasm in which the spine and extremities are bent backwards.

In alcoholic and other malnourished subjects, one of the early signs of thiamin deficiency is anorexia. In alcohol abuse, the overwhelming desire for alcohol may outweigh all other interest in food, leading to generalized malnutrition. Alcohol specifically blocks the active absorption of thiamin and alcohol abuse can progress to the potentially fatal condition known as Wernicke–Korsakoff syndrome. The typical clinical features of Wernicke's encephalopathy comprise ophthalmoplegia, nystagmus (usually

Table 4 Common features of wet beriberi

	Subacute beriberi	*Acute fulminating beriberi*[a]
Digestive system	Anorexia is common; constipation more frequent than diarrhea	Vomiting is common, often with intense thirst Liver enlarged and tender and the epigastric region spontaneously painful
Neurological	Aching pain, stiffness, tightness, or cramps in calf or associated muscles Increasing muscular tenderness and weakness with fatigue pains resembling muscular ischemia, especially at night Pain on squeezing calves Inability to rise from squatting position without use of hands Diminished reflexes of ankle and knees usually bilaterally Hypoesthesia or paraesthesia presenting as 'pins and needles,' numbness particularly over the tibia, formication (like ants running on the skin) or itching	Pupils dilated with anxious expression on face Aphonia frequently present and patient moans with cries of a special kind as a result of hoarseness produced by paralysis of laryngeal muscles Reflexes of ankle or knee lost or diminished
Cardiac	Oedema of feet and legs often appearing first on dorsa of feet and extending up legs but may also appear on back of hands and as puffiness in face Heart enlarged with tachycardia and bounding pulse Raised venous pressure (see **Figure 2**) with percussion sometimes revealing dilation of right auricle and ventricle Heart murmurs if present are usually systolic Apex beat is downward and outwardly displaced Neck vein possibly distended showing visible pulsations Dyspnea upon exertion Palpations, dizziness, and giddiness Extremities possibly cold and pale with peripheral cyanosis but where circulation is maintained, skin warm due to vasodilatation Electrocardiograms often undisturbed but QRS complex may show low voltage and inversion of T waves indicating disturbed conduction	Patients severely dyspneic, have violent palpitations of the heart, are extremely restless, experience intense precordial agony but accessory muscles of respiration on slightly brought into action Widespread and powerful undulating pulsations visible in the region of the heart, epigastrium, and neck due to a tumultuous heart action Facial cyanosis more marked during inspiration Pulse is moderately full, regular, even with frequency of 120-150/min A wavelike motion may be felt over the heart On percussion, the heart is enlarged both to the left and right but mainly the latter, and the apex beat may reach the axilla Raised systolic pressure and low diastolic pressure give the 'pistol shot' sound on auscultation over the large arteries Rapidly increasing oedema may extend from legs to trunk and face with associated pericardial, pleural, and other serous effusions
Urine	Nocturia; no albuminuria	Oliguria or anuria; no albuminuria or glycosuria

[a] The whole picture of acute fulminating beriberi is dominated by insufficiency of heart and blood vessels and this tends to mask all other features of the subacute form, although these are often present and accentuated. Death is accompanied by a systolic pressure falling to 70–80 mm, the pulse becomes thinner, and the veins dilate. The rough whistling respiration deteriorates and rales appear. The patient dies intensely dyspneic but usually fully conscious.

horizontal), ataxic gait, and an abnormal mental state that can range from mild delirium to global confusion. Liver disease and tachycardia occur in more than 50% of cases. Korsakoff's psychosis is characterized by a profound amnesia, disorientation, and often confabulation. The clinical features of Wernicke–Korsakoff syndrome are listed in **Table 6**.

Management/Treatment

Patients in whom cardiac and renal signs of thiamin deficiency are identified usually respond well to treatment. The dose given and route used will vary with the seriousness of the deficiency. Intravenous doses as high as 250 mg/day for 14 days and intramuscular doses of 25 mg followed by thrice daily oral doses of 10 mg have been reported for wet beriberi and are followed by a marked increase in urinary output and improvement in cardiac function. Peripheral neuropathy (dry beriberi) is more resistant to treatment. Patients with the ocular signs of Wernicke's disease usually respond to two or three daily injections of 50 mg thiamin. Long-term oral treatment of other manifestations of Wernicke–Korsakoff syndrome with doses up to 50 mg/day is reported, although benefit is variable and considerably influenced by patients' ability to avoid further alcohol consumption. It is unlikely that patients receiving oral thiamin will absorb more than 5–7 mg/day, but in patients likely to abuse alcohol, absorption by passive diffusion of high thiamin doses

Figure 1 Patient with dry beriberi showing evidence of motor nerve disturbances resulting in a flaccid paralysis of the extensor muscles and 'wrist drop' and 'foot drop'.

is the only way to ensured that the patient will receive any thiamin. In addition, as in all patients who show evidence of nutritional deficiency, the likelihood of other coexisting deficiencies should not be overlooked and multinutrient treatment is probably desirable.

Table 5 Features of acute infantile beriberi and frequency of occurrence

	Features	*Frequency (%)*
Appearance	Pale and cyanotic appearance, oedematous, ill-tempered with abdominal distension	40
Voice	Hoarseness	80
	Sometimes groaning	50
Digestive system	Vomiting	80
	Dyspepsia	46
Cardiac	Tachycardia, <200 beats/min	83
	Heart dilated	31
	Femoral sound on ascultation	5
Lungs	Rapid breathing	83
	Accentuation of the 2nd pulmonary sound	
Neurological	Tendon reflex usually increased	74
	Less frequently decreased	26
	Convulsions	17
Urinary	Oliguria	65
Other	Slight fever	50
	Uneasiness	50

Modified from Thurnham DI (1978) Thiamin. In: Rechcigl M Jr (ed.) *Nutrition Disorders*, pp. 3–14. West Palm Beach, FL: CRC Press.

Finally, it is important to realise that untreated thiamin deficiency can result in sudden death.

Lipid-Soluble Thiamin Derivatives

In recent years, several lipid-soluble derivatives of thiamin have been introduced, of which the best known is benfotiamine. Advantages of these compounds appear to be increased absorption, but by the diffusion mechanism only, and greatly increased transketolase activity. Transketolase is the rate-limiting enzyme of the nonoxidative branch of the pentose phosphate pathway. Benfotiamine has been shown to be useful for the management of rare genetic disorders in thiamin transport and may also prove useful to prevent damage from diabetic hyperglycemia. One study demonstrated that benfotiamine prevented experimental retinopathy. Diabetic hyperglycemia is accompanied by an increase in the potentially pathogenic glycolytic metabolites glyceraldehyde-3-phosphate and fructose-6-phosphate. Benfotiamine, by increasing transketolase activity, stimulates the pentose phosphate pathway to metabolise these glycolytic intermediates into pentose-5-phosphates and prevent the intracellular increase of potentially toxic products.

Case Study

A good example of the specific effect of thiamin in the treatment of beriberi is illustrated by the response of a 29-year-old male who was admitted with an unexplained acute renal failure and had been anuric for 24 h. The physicians' report on his symptoms should be compared with the common clinical features of wet beriberi shown in **Table 4**. The patient's physical state and voice were extremely weak but speech was copious and confused. He complained intermittently of severe central chest and epigastric pain. A central cyanosis was present and he had a respiratory rate of 36 beats per minute. His temperature was normal and peripheries were lukewarm. He had gross generalised oedema. The jugular venous pressure became grossly elevated (**Figure 2**). Pulse rate was 100 beats per minute, regular, and weak at the wrist, although the carotid pulses were visibly bounding. Blood pressure was 80/60. There was a marked parasternal heave present, with a loud pulmonary second heart sound. The chest was clear; the abdomen was obese.

The father reported that the patient's usual beer intake was 6–12 pints daily and his one regular meal was usually no more than a sausage roll or a pie. Prior to admission, for 6 weeks he had felt too

Table 6 Clinical features of Wernicke–Korsakoff syndrome and frequency of occurrence

	Features	*Frequency (%)*[a]
Ocular disorders	Nystagmus (ocular ataxia — rhythmical oscillation of the eyeballs), almost always horizontal and in 50% of cases associated with vertical nystagmus on upward gaze	85
	Paralysis of one or more of the ocular muscles	50
	Sluggish reaction by pupils to light	19
Ataxia (inability to coordinate muscles)	Gait	87
	Legs	20
	Arms	12
	Speech	87
Polyneuropathy	Limbs only affected, mainly the legs only	82
	Of arms and legs	18
	Common symptoms include weakness, paresthesia, pain, loss of tendon reflexes and of sensation and motor power	
	Some cases of foot drop or wrist drop or both	
Cerebral function	Global confusional state, profound disorientation, apathy, deranged perception and of memory, drowsiness, inattentiveness, indifference	56
	Disorder of memory: both retrograde and ante-retrograde amnesia, confabulation	57
	Alcohol abstinence syndrome	16
Cardiac	Tachycardia	51
General medical abnormalities	Disorders of skin and mucous membranes	36
	Redness and/or papilliary atrophy of the tongue	29
	Liver disease	60

[a]Percentages based on 188–245 cases.
Modified from Thurnham DI (1978) Thiamin. In: Rechcigl M Jr (ed.) *Nutrition Disorders*, pp. 3–14. West Palm Beach, FL: CRC Press.

tired to go out in the evening, and for 2 weeks he had suffered epigastric discomfort and had eaten nothing. Eight days before admission, he developed painful calf stiffness and he became too weak to go to work. He had a painful dry cough and dyspnoe on the slightest exertion. Finally, confusion, cyanosis, and intermittent vomiting led to admission.

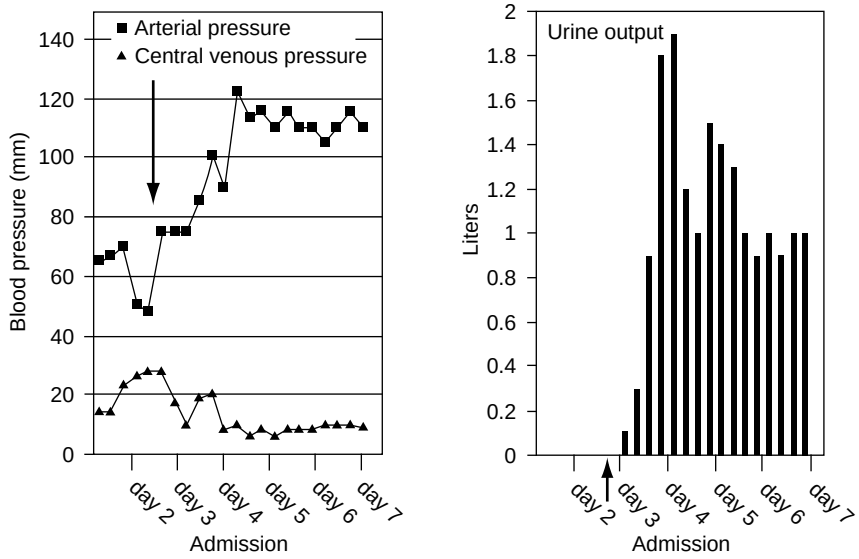


Figure 2 (Left) Arterial and central venous blood pressure and (right) urine output of a patient who was admitted with unexplained acute renal failure in a very weak physical state and whose speech (although very weak) was copious and confused. The patient was discovered to be a regular beer drinker consuming 6–12 pints daily, and his usual food intake amounted to no more than a sausage roll or pie. He had become progressively weaker over the past 8 weeks and had eaten nothing at all in the past 2 weeks. After excluding other diagnoses, it was suspected that the patient had fulminant beriberi and he was treated with thiamin after 36 h. The figures display the rapidly increasing arterial pressure, fall in venous pressure, and a rapid resumption in renal function following thiamin treatment. The patient lost ~20 l of urine during the first 7 days in the hospital. (Modified from Anderson SH, Charles TJ and Nicol AD (1985) Thiamine deficiency at a district general hospital: Report of five cases. *Quarterly Journal of Medicine* **55**: 15–32.)

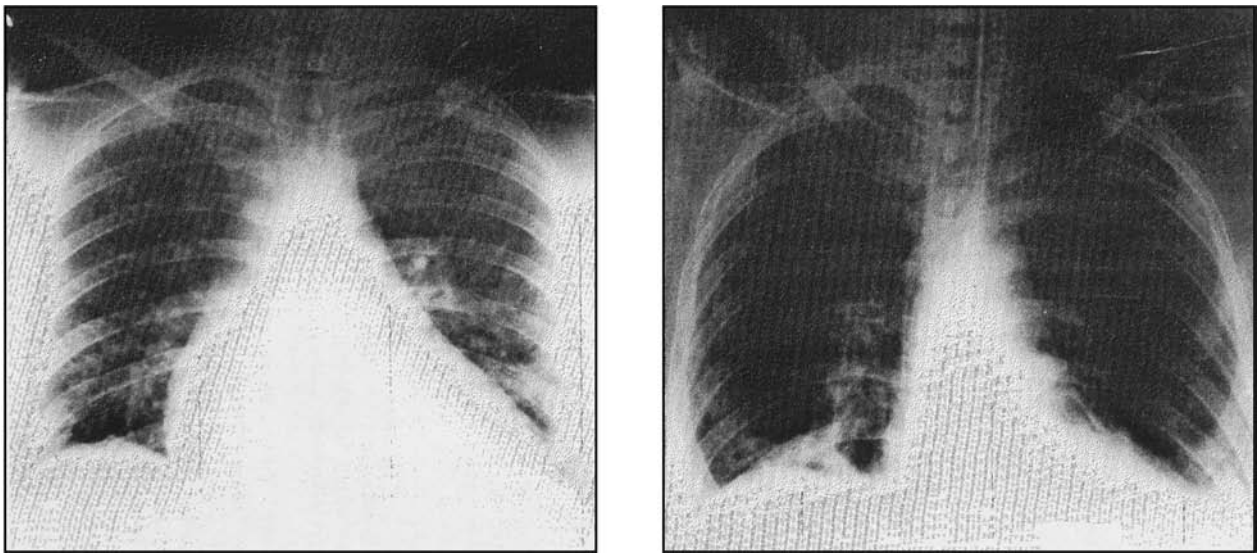


Figure 3 Chest radiographs of the patient described in **Figure 2** obtained on admission (left) and 14 days after high-dose, parenteral thiamin treatment (right). On admission, the heart was grossly enlarged, extending downward and to the right, with a cardiothoracic ratio of 0.63. After treatment for 14 days, the cardiothoracic ratio was 0.44. (Modified from Anderson SH, Charles TJ and Nicol AD (1985) Thiamine deficiency at a district general hospital: Report of five cases. *Quarterly Journal of Medicine* **55**: 15–32.)

The first diagnoses considered were myocardial infarction, pulmonary embolism, and overwhelming septicaemia, and he was placed on dialysis and received appropriate treatments. His lack of response at 36 h, continuing low systolic pressure (70 beats per minute), increasingly gross hyperdynamic precordial signs, and moribund appearance led to a diagnosis of beriberi. Treatment with intravenous thiamin (250 mg for 14 days) brought about a dramatic response (**Figure 2**). Within 6 h peripheral pulses were strong, blood pressure had risen to 105 systolic, and central venous pressure had fallen by half. By 12 h the parasternal heave was less marked and diuresis of up to 6 l per day ensued. After 24 h, plasma urea concentration peaked at 50.4 mmol/l and creatinine at 832 μmol/l, and thereafter there was a steady fall over the next 2 weeks during which thiamin treatment continued and dialysis stopped. He lost a net 20 l of fluid over the first 7 days in the hospital and creatinine clearance 3 weeks after admission was 178 ml/min, indicating a return to normal kidney function. Other biochemical abnormalities resolved over the 2 weeks on high-dose thiamin, including the chest radiograph (**Figure 3**). It is interesting to note, however, that when he was discharged 3 months after admission, he was walking with a calliper because of a right-sided foot drop (**Figure 1**). The persistence of the foot drop is a further indication of the greater difficulty in reversing neurological consequences, in contrast to the cardiac effects, of thiamin deficiency.

Toxicity

Chronic intakes in excess of 50 mg/kg, or more than 3 g/day, are toxic to adults with a wide variety of clinical signs, including headache, irritability, insomnia, rapid pulse, weakness, contact dermatitis, pruritis, and, in one case, death. Early researchers also indicated that regular administration or contact with thiamin occasionally led to allergic response, contact dermatitis, or hypersensitivity.

See also: **Alcohol**: Effects of Consumption on Diet and Nutritional Status. **Cereal Grains**. **Fish**. **Thiamin**: Physiology.

Further Reading

Anderson SH, Charles TJ, and Nicol AD (1985) Thiamine deficiency at a district general hospital: Report of five cases. *Quarterly Journal of Medicine* **55**: 15–32.

Brin M (1964) Erythrocyte as a biopsy tissue for functional evaluation of thiamine adequacy. *Journal of the American Medical Association* **187**: 762–766.

Burgess RC (1958) VI Special problems concerning beriberi. B. Infantile beriberi. Proceedings of a conference on beriberi, endemic goitre and hypervitaminosis A entitled 'Nutritional Disease.' *Proceedings of the Federation of Association of Societies of Experimental Biology* **17**(supplement 2): 39–46.

Carpenter KJ (2002) Acute versus marginal deficiencies of nutrients. *Nutrition Reviews* **60**: 277–280.

Fehily L (1940) Infantile beriberi in Hong Kong. *Caduceus* **19**: 78–93.

Friedemann TE, Kmieciak TC, Keegan PK, and Blum Sheft B (1948) The absorption, excretion and destruction of orally-administered thiamin by human subjects. *Gastroenterology* **11**: 101–114.

Hammes H-P *et al.* (2003) Benfotiamine blocks three major pathways of hyperglycemic damage and prevents experimental diabetic retinopathy. *Nature Medicine* **9**: 294–299.
Ibner FL, Blass JP, and Brin M (1982) Thiamin in the elderly, relation to alcoholism and to neurological degenerative disease. *American Journal of Clinical Nutrition* **36**: 1067–1082.
Luyken R (ed.) (1990) *Polyneuritis in Chickens, or the Origin of Vitamin Research* [First English edition of papers by Christian Eijkman published 1890–1896]. Basle: Roche.
McConachie I and Haskew A (1988) Thiamine status after major trauma. *Intensive Care Medicine* **14**: 628–631.
Pepersack T, Garbusinski J, Robberecht J *et al.* (1999) Clinical relevance of thiamine status amongst hospitalised elderly patients. *Gerontology* **45**: 96–101.
Tang CM, Rolfe M, Wells JC, and Cham K (1989) Outbreak of beri-beri in the Gambia. *Lancet* **2**: 206–207.
Thurnham DI (1978) Thiamin. In: Rechcigl M Jr (ed.) *Nutrition Disorders*, pp. 3–14. West Palm Beach, FL: CRC Press.
Thurnham DI (2004) An overview of interactions between micronutrients and of micronutrients with drugs, genes and immune mechanisms. *Nutrition Research Reviews* **17**: 211–240.
Victor M, Adams RD, and Collins GH (1971) The clinical findings. In: F Plum and FH McDowell (eds.) *The Wernicke–Korsakoff Syndrome*, pp. 16–34. Oxford: Blackwell.
Williams RR (1961) *Towards the Conquest of Beriberi* Cambridge, MA: Harvard University Press.

THIRST

J Leiper, University of Aberdeen, Aberdeen, UK

Role of Thirst in Water Balance

Approximately 70% of the lean body mass of an individual is composed of water, with approximately two-thirds of the total body water (TBW) volume being held within the cells of the body (intracellular pool) and the remaining one-third (extracellular pool) is divided between the circulating blood plasma (intravascular pool) and the fluid-filled spaces between the cells (interstitial pool). The volume and distribution of the body fluids are mainly determined by the amounts of body water and sodium. In man, TBW content is regulated daily to within approximately 0.2% of lean body mass in normal, temperate conditions by factors that control input and output. The kidneys regulate water excretion in excess of the evaporative loss and the fecal and obligatory urine losses. Water intake occurs in the form of food and drink, with the sensation of thirst underpinning drinking behavior.

The mechanisms that monitor the body's hydration status also interact with the thirst control centers in the brain to regulate the desire to drink. This article is concerned with the physiological factors that govern the perception of thirst and how this is altered by drinking.

Perception of Thirst

Thirst is a sensation that is best described as the desire to drink. The reason for drinking may not be directly involved with a physiological need for water intake, but it can be prompted by habit, ritual, taste, nutrients, craving for alcohol, caffeine, or other drug in a beverage, or a desire to consume a fluid that will give a warming or cooling sensation. Much of the perception of thirst is a learned or conditioned process, with signals such as dryness of the mouth or throat initiating drinking, whereas a feeling of fullness of the stomach can stop ingestion before a fluid deficit has been restored.

The thirst response is thought to be regulated by neural modulators that operate as a reward mechanism, integrating the effective requirement for water intake with the sensations of taste and pleasantness of the fluid ingested. Thus, when the individual is hypohydrated multiple areas of the brain are activated, promoting the intensity of the thirst sensation. As the water deficit is restored the feeling of thirst diminishes and this subjective sensation correlates well with a reduction in neural activation. However, areas of the brain associated with taste that are activated by water when thirsty remain as active following drinking to satiety when water is ingested.

Although it is true that thirst in man is a poor indicator of acute hydration status and that daily fluid intake is normally in excess of obligatory water loss, the preservation of TBW volume under a variety of environmental and nutritional stresses is remarkably robust and is mainly due to the drive to drink, which the sensation of thirst chronically provokes.

Assessment of Thirst

In humans two main techniques have been used to identify the perception of thirst and its alleviation by drinking. The first method is to monitor the volume

of drink voluntarily ingested by an individual within an allotted time period and to compare the amount drunk with the volume of fluid required to restore a given water deficit or other imbalance of the body water pools. The other method is to assess the individual's perceived rating of thirst by asking him or her to record on a visual analog scale his or her responses to a series of questions that are thought to relate to the sensation of thirst (**Figure 1**). The questionnaire technique has the advantages that it allows a series of measurements to be made before, during, and following the period of drinking, and it appears to give an indication of the relative strength of a given stimulus. Although in many studies both methods are used to gauge the sensation of thirst and the responses have been correlated to a number of physiological parameters that are known to influence the drive to drink, it is widely recognized that there is no consistently reliable measure of the thirst sensation.

A more recently introduced technique has been the use of noninvasive methods of imaging to identify the specific regions of the brain that are activated during the genesis and satiation of the thirst response. Different techniques are used to image brain activation. Positron emission tomography and functional magnetic resonance imaging are both being used to visualize brain activation by detecting either temporal changes in blood flow or changes in the chemical composition of regions in the brain that occur when individuals are exposed to specific stimuli. The number of brain regions, their

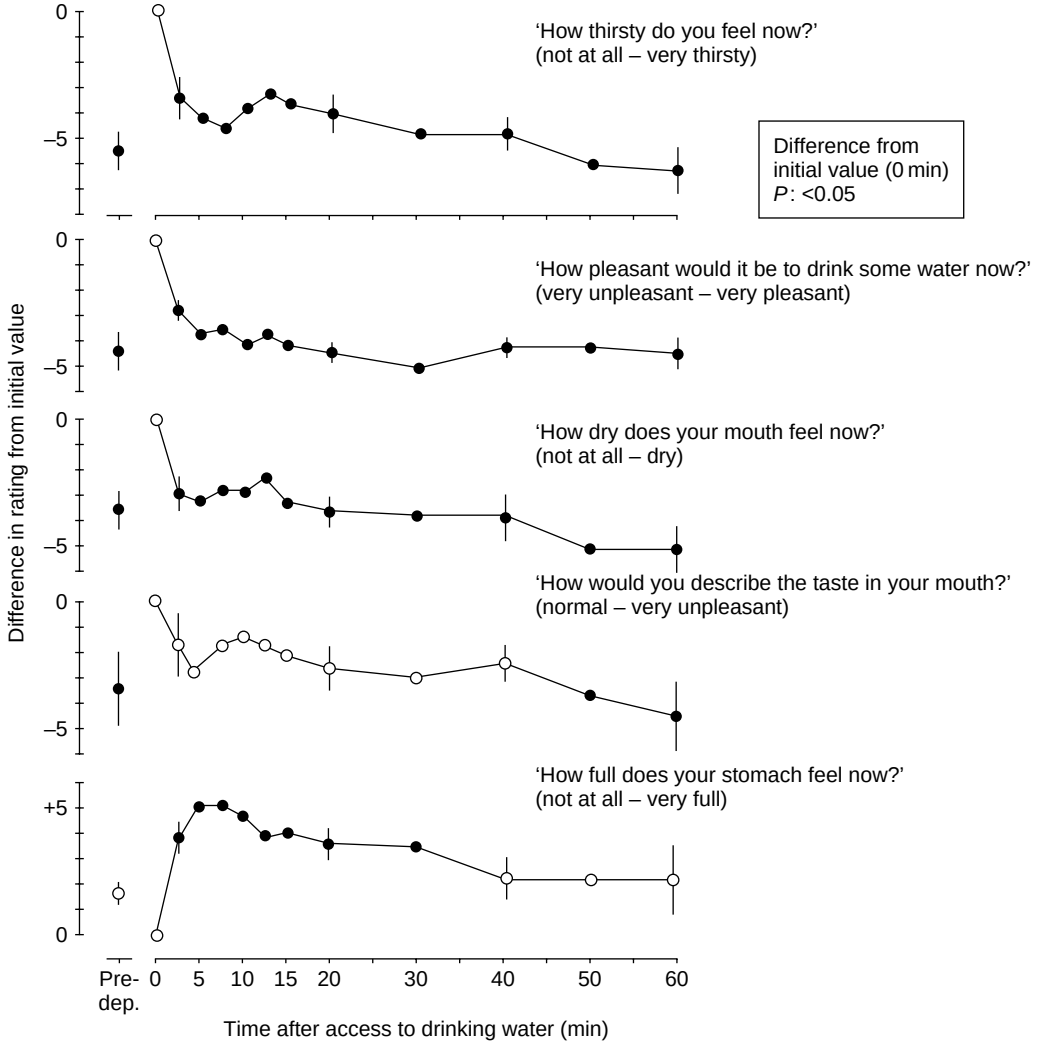


Figure 1 Subjective responses to a series of five questions assessed by visual analog scale ratings from individuals deprived of water for 24 h. Questions were asked before and during a 60-min rehydration period. The data are shown as differences (in cm) from initial values and significant differences are indicated by filled symbols. (Redrawn with permission from Rolls BJ, Wood RJ, Rolls ET *et al.* (1980) Thirst following water deprivation in humans. *American Journal of Physiology* **239**: R476–R482.)

specificity, and the intensity of activation has also been correlated with the subjective perception of thirst.

Physiological Regulation of Thirst

Because thirst is the major factor controlling water intake, the physiological regulation of thirst is associated with the need to maintain a relatively stable volume of TBW. Although water is lost from the body continually, albeit usually in relatively small amounts, and hence the body is almost always developing a water deficit, water intake is intermittent. The amount of fluid usually ingested is in excess of that required to replace the losses incurred since the last water intake. The factors that initiate, maintain, and end the drinking response are various and are not fully understood. However, because the regulation of the volume and composition of the various water pools of the body play an essential role in controlling the perception of thirst, an understanding of the homeostatic mechanisms involved has given us the best insight we have into the complexities of the perception of thirst.

The total volume, distribution, and composition of body fluids must be regulated within narrow limits for normal cellular function to be maintained. Body water is passively distributed between the extracellular and intracellular pools according to osmotic, oncotic, and hydrostatic forces as shown in **Figure 2**. The sodium and chloride contents of the extracellular fluid constitute the two greatest osmotically active components of this fluid and are therefore important in maintaining its volume. Potassium, phosphate, and protein fulfill a similar role in regulating the intracellular fluid volume. The distribution of water between the intravascular and extravascular pools is dependent on the balance of hydrostatic and oncotic pressures across the capillaries and postcapillary venules.

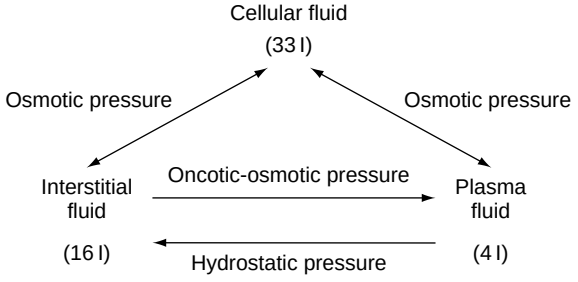


Figure 2 Diagramatical representation of the forces that regulate the distribution of the body water pools. The volumes given are those determined in a single male subject with a lean body mass of 75.8 kg.

Variation in the water-to-solute ratio of a body fluid pool results in changes in the tonicity and hence effective osmolality of the fluid. Because the various body water pools are in dynamic equilibrium with each other (**Figure 2**) there is a tendency for adjustments to occur throughout the body as water moves from regions of low solute concentration to those of higher solute concentration. Changes in plasma osmolality are relatively easy to monitor; therefore, there is a tendency to equate changes in the circulation as the effector of fluid balance control. However, it is important to remember that any alteration in one body pool will affect the others and that receptors that initiate responses affecting water balance may reside at sites far removed from the circulation.

Loss of water from the body or an increase in the circulating solute concentration cause an increase in the osmolality of primarily the extracellular fluid; water then moves into the extracellular space from the cells producing a reduction in cell volume. Changes in plasma osmolality are therefore thought to be signaled to the effector mechanisms by changes in the cell volume of specific specialized cells, collectively termed osmoreceptors. Because the main solute determining the tonicity of the extracellular fluid is sodium, there has been debate about whether the receptor cells detect changes in osmolality or changes in sodium ion content. The evidence suggests that at least the majority of the receptors respond to osmolality rather than to sodium concentration. These osmoreceptors have a regulatory role not only in the perception of thirst but also in the maintenance of the circulating levels of hormones that regulate the excretion of water and solute by the kidneys (**Figure 3**). Because increases in the extracellular osmolality effectively decrease the volume of the cells in the body, this form of dehydration is termed cellular dehydration.

Alteration in the volume of the extracellular fluid pool without changes in its osmolality also affects the fluid balance hormone concentrations and the sensation of thirst. Changes in the volume of blood in the circulation affect the blood and capillary pressures and atrial filling pressure. The effect on capillary pressure will tend to redistribute body water and help to adjust the circulating fluid volume, and the change in venous return to the heart will alter the cardiopulmonary and arterial stretch receptor (baroreceptor) activity. The level of afferent activity from these baroreceptors directly affects both the sensation of thirst and the secretion of some fluid balance hormones. Additionally, modifications to the arterial blood pressure can directly affect renal perfusion, which together with

baroreceptor activity to the kidneys regulates the renin–angiotensin system (**Figure 4**). Although the effect on the kidneys can influence the perception of thirst, the main renal response is to regulate urine water and solute excretion. A decrease in the volume of the extracellular pool with no concomitant change in plasma osmolality is termed extracellular dehydration.

When humans are given access to fluids after the development of a water deficit, their drinking response usually follows a pattern of rapid ingestion of more than 50% of the total intake followed by intermittent consumption of relatively small volumes of drink over a longer period. Although initiation of the response to drink is due mainly to osmotic or blood volume (volemic) changes, there appear to be

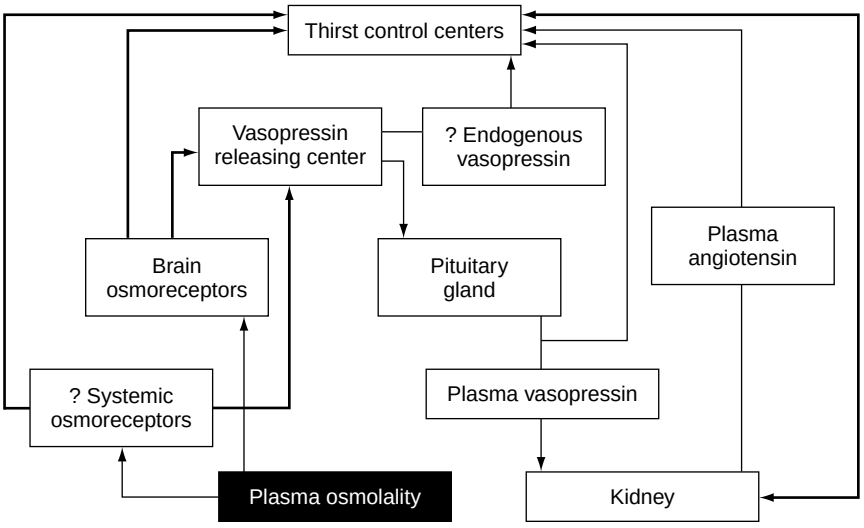


Figure 3 Schematic representation of the main factors proposed in osmotically induced regulation of the sensation of thirst and their interaction with the control of diuresis. A rise in plasma osmolality will tend to stimulate greater excitatory activity, whereas a decrease in osmolality will activate more inhibitory inputs. Neural pathways are indicated by thick arrows and hormonal input by thin arrows.

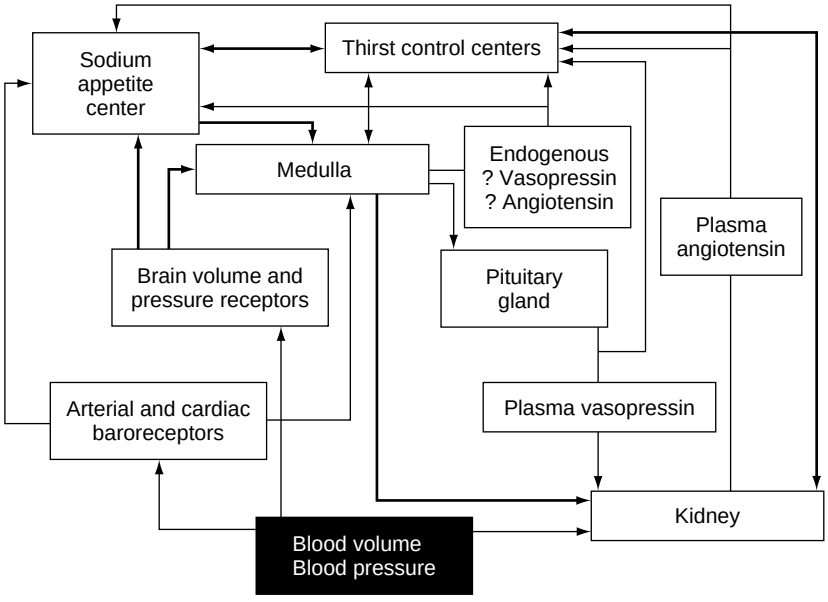


Figure 4 Schematic representation of the main factors proposed in volemic-induced regulation of the sensation of thirst and their interaction with the control of diuresis and sodium appetite. A fall in circulating blood volume will decrease baroreceptor activity, which will increase excitatory activity, whereas a rise in volume will have the opposite effect. Reduction in blood pressure will decrease renal perfusion, which will activate the renal renin–angiotensin system. Neural pathways are indicated by thick arrows and hormonal input by thin arrows.

other mechanisms involved in the control of the continuation and satiety responses. Receptors in the mouth, oesophagus, and gastrointestinal tract appear to be major factors in the acute regulation of thirst satiation, with the effects that the volume and solute content of the ingested drink have on restoring the fluid deficits controlling the chronic regulation of thirst (**Figure 5**). There is a close relationship between eating and drinking, with approximately 70% of daily fluid intake normally being associated with meals (**Figure 6**). The desire to drink while eating is probably produced by a series of responses, including the mechanochemical composition of the food before absorption, the neuroendocrine response to digestion, the movement of water into the intestine during digestion, and the osmotic solute load that occurs following absorption. The intake of minerals is essential to replace those lost from the body and for growth. The majority of mineral intake is supplied by the food ingested, and indications of a desire or appetite for ingesting specific minerals have been shown in animals and man. Although sodium appetite has been linked to the sensation of thirst, anatomically and functionally the controlling mechanisms are distinct and separate.

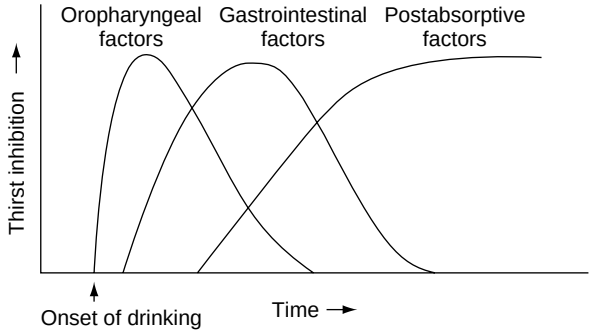


Figure 5 Schematic diagram depicting the proposed onset, duration, and overlap of various inhibitory signals to continue fluid ingestion following initiation of drinking in response to a fluid deficit. (Redrawn with permission from Verbalis JG (1990) Inhibitory controls of drinking: satiation of thirst. In Ramsay DJ and Booth DA (eds.) *Thirst: Physiological and Psychological Aspects*, pp. 313–330. London: Springer-Verlag.)

Mechanisms of Thirst Regulation

The sensation of thirst is regulated separately by both the osmotic pressure and the volume of the body fluids and as such is closely related to the control mechanisms that are responsible for the secretion of the fluid balance hormones, which affect water and solute reabsorption in the kidneys and play a role in blood pressure control. These hormones—arginine vasopressin, atrial natriuretic peptide, oxytocin, and the renin–angiotensin–aldosterone system—are central to the regulation of thirst. The hypothalamus and forebrain appear to be the main areas involved in the control of thirst and antidiuresis, and collectively these parts of the brain have been termed the

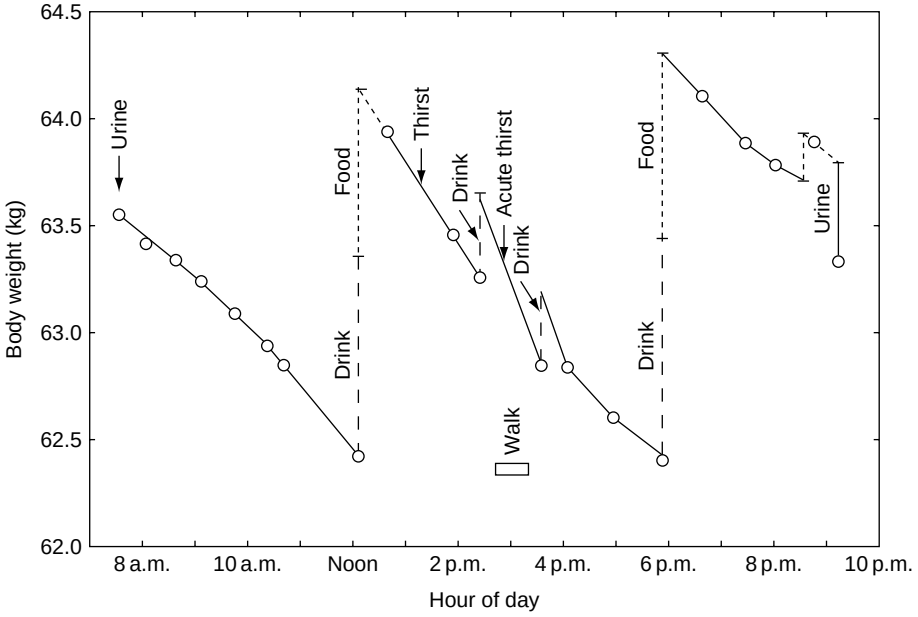


Figure 6 Changes in body weight during 13 h in the desert. The majority of the volume of drink ingested was associated with food intake. Sweat loss varied from 150 to 700 ml/h, and total fluid intake was 3.05 l. At the end of the period body mass was essentially the same at the beginning and end of the day; therefore, water intake and output were equal. (From Adolph ED, *Physiology of Man in the Desert*. Copyright © 1947. Reprinted by permission of John Wiley & Sons Inc.)

thirst control centres. Neurons that are responsive to changes in osmolality, intravascular volume (volemia), and blood pressure are found within these areas of the brain, as are other receptors that are responsive to many of the fluid balance hormones. Neural pathways from the thirst control centers and the kidneys may allow some direct integration between the control of thirst and excretion, whereas within the brain all of the major fluid balance hormones are present as neurohormones. Afferent input from systemic receptors monitoring osmolality, circulating sodium concentration, and changes in intravascular volume and pressure also have roles in controlling the feeling of thirst. Therefore, there appears to be a complex integrated system for both monitoring the status of the body water pools and controlling intake and excretion (**Figures 3** and **4**). Many of the regulatory mechanisms controlling water balance appear to overlap, with several stimuli appearing to subserve the same response; however, it is assumed that this effect is required in order to ensure that the blocking of one type of stimulus will not prevent homeostatic control.

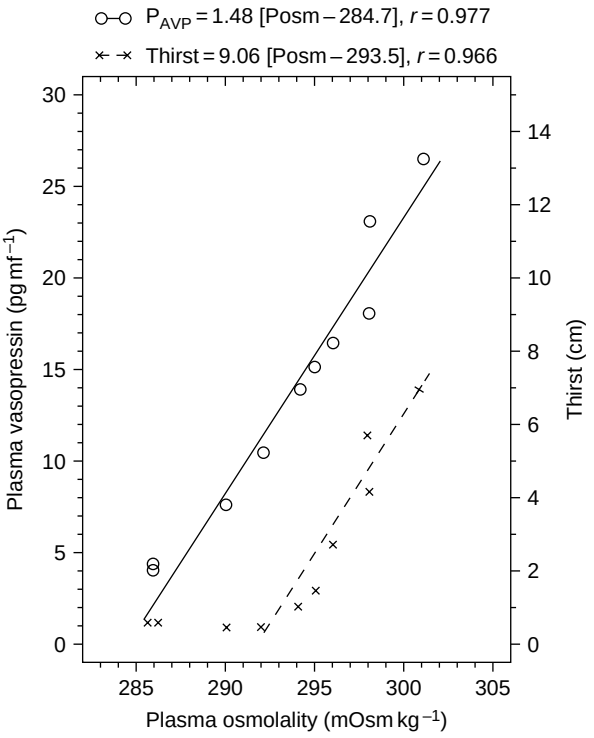


Figure 7 The relationship of plasma vasopressin (o) and thirst (×) to plasma osmolality in a volunteer during infusion of 5% saline. (Redrawn with permission from Robertson GL (1984) Abnormalities of thirst regulation (Nephrology forum). *Kidney International* **25**: 460–469.)

Osmotic Regulation of Thirst

The osmolality of circulating plasma is normally maintained within a very narrow limit between 270 and 295 mosmol/kg, with the circulating levels of the antidiuretic hormone arginine vasopressin playing a major role in its homeostatic regulation. An increase of as little as 2 or 3% in plasma osmolality is sufficient to produce a strong sensation of thirst and a significant increase in circulating arginine vasopressin concentration (**Figure 7**). The osmoreceptors that monitor the tonicity of the body pools appear to reside mainly in an area of the brain that lacks a blood–brain barrier; therefore, they appear to respond mainly to changes that occur in the osmolality of the blood rather than in the cerebral interstitium. Although the changes in the circulating levels of arginine vasopressin and the perception of thirst appear to parallel one another, it is unlikely that the same receptors are responsible for both responses. It is likely that there are different neurons that react to the same stimulus. However, there may be some neurohormonal interaction between the osmotically activated thirst centers and the 'vasopressin-releasing center' in the brain, and arginine vasopressin-responsive neurons have been detected within the thirst centers (**Figure 3**).

The current theory of the osmotic control of thirst suggests that there is constant output of both inhibitory and excitatory neural activity from the respective osmoreceptors to the thirst centers and the arginine vasopressin-releasing centers. Within the normal range for plasma osmolality, the inhibitory and excitatory activities in the thirst centers effectively cancel out one another and there is neither a sensation of thirst or satiety. However, at this level of activity there is release of a basal level of arginine vasopressin that is sufficient to maintain a state of half-maximum antidiuresis. A rise in plasma osmolality above the normal level stimulates greater excitatory output causing an increase in the feeling of thirst and higher levels of circulating arginine vasopressin. Raised levels of arginine vasopressin increase the concentrating ability of the kidneys. A decrease in plasma osmolality below the normal range increases the inhibitory output producing a feeling of satiety, and arginine vasopressin release is suppressed allowing urinary dilution (**Figure 3**).

Cells and fibers within the brain have been shown to contain several hormones, including angiotensin and vasopressin, within the same cell. Although neurons associated with the thirst centers can be activated *in vitro* by vasopressin, it is not clear whether peripheral- or neural-generated arginine vasopressin levels influence the perception of thirst.

Volemic Regulation of Thirst

The receptors that initiate hypovolemic thirst are generally thought to be the cardiovascular baroreceptors, which respond to underfilling of the circulation by reducing their inhibitory nerve impulse activity to the thirst centers. However, in areas of the brain associated with the thirst centers there are neurons that are separately responsive to volemic, pressure, and osmotic changes. This suggests that at least part of the response to changes in blood volume originates in the brain. It is thought that changes in blood pressure and osmolality are monitored mainly within the brain, whereas variations in blood volume are principally sensed by the peripheral baroreceptors, with a degree of overlap between the different receptors. The mechanisms that respond to changes in intravascular volume and pressure appear to be not as sensitive as those responsive to osmotic changes; for example, a decrease of approximately 10% of the plasma volume is required to initiate hypovolemic thirst. Because fairly large variations in blood volume and pressure occur during normal daily activity, such as postural changes and physical activity, this apparent lack of sensitivity presumably prevents overactivity of the volemic control mechanisms. As with osmotic thirst, the control of volemic thirst is thought to be a balance between continuous inhibitory and excitatory neural activity, although in this system the basal level appears to be essentially inhibitory. Another difference in the basic control mechanism between the two systems is due to the requirement for both solute, mainly sodium, and water to restore the extracellular volume. Therefore, extracellular dehydration causes an initial thirst and a delayed increase in sodium appetite.

Reduction in the intravascular volume sufficient to lower cardiac output and arterial blood pressure decreases afferent activity from the low- and high-pressure cardiovascular baroreceptors to the thirst centers and increases sympathetic activity to the kidneys. Because afferent input from the baroreceptors to the thirst centers is inhibitory, a decrease in activity produces a reflex increase in the perception of thirst and also appears to directly stimulate arginine vasopressin release. The increase in sympathetic activity to the kidneys directly promotes greater renal renin release. In addition, reduction in blood pressure lowers the renal perfusion pressure, which stimulates renin release both as a direct pressure effect and by decreasing the delivery of sodium to the kidneys.

Increased activation of the renin–angiotensin–aldosterone system also helps regulate hypovolemic thirst. While circulating levels of both vasopressin and aldosterone affect water and sodium reabsorption in the kidneys and thereby control water and solute loss, angiotensin acts directly on the thirst and sodium appetite centers to stimulate their respective responses. Neurons that are stimulated by angiotensin are found in areas of the brain that lack a blood–brain barrier; therefore, circulating angiotensin has direct access to both centres. In addition, the release of neurally generated angiotensin is promoted by suitable neuron activity responding to sensory stimuli (**Figure 4**).

There are a variety of neural and hormonal responses that interact to modulate and control both thirst and urine excretion. A number of other hormones, including oxytocin, atrial natriuretic peptide, tachykinins, neuropeptide Y, thyroid hormones, corticotrophin-releasing factor, and steroid hormones, have also been shown experimentally to affect the drinking response. Under normal conditions of water and solute loss, both osmotic and volemic dehydration occur; therefore, stimuli from receptors for both systems are usually involved in the sensation of thirst. Increases in extracellular osmolality appear to be more effective than hypovolemia in promoting the thirst and hence drinking response. More than 70% of the stimulus to drink appears to be generated by increased osmolality.

Sensory Regulation of Thirst

The sensations of a dry mouth or desire for a specific taste or effect also generate the desire to drink when there may be no physiological requirement to drink. A dry mouth promotes changes in neural activity in the parahippocampus, amygdala, thalamus, cingulate, insula, allocortex, and transitional cortex of the brain. This finding has strengthened the hypothesis that the perception of thirst is a primitive vegetative function that appeared long before vertebrates evolved.

Drinking water activates areas of the anterior insular and frontal opercular cortex that are also involved in the perception of taste. Areas of the orbitofrontal cortex are also activated by the ingestion of water or sweet or salty tastes, but activation is greatest when subjects are thirsty and it diminishes when subjects have drunk water to satiety. This has been interpreted as functionally separate areas of the brain, one of which responds to taste stimuli that are not diminished following drinking to satiety, whereas the other is highly active during drinking when water is physiologically required but reduces as the need for water is met.

Mechanisms for Terminating the Sensation of Thirst

Although undoubtedly decreasing osmolality and increasing extracellular volume promote a reduction in the perception of thirst by reactivating inhibitory neuron activity, usually there is a decrease in the perception of thirst and termination of drinking before circulating osmolality, volume, and hormonal levels have returned to predehydration levels. Although it could be argued that receptors in the brain may be responsible for the early cessation of the perception of thirst, the majority of the evidence suggests that it is receptors in the upper gastrointestinal tract that promote the early termination of drinking. Although the nature and neural connections of these proposed receptors have not been fully characterized, most appear to have an inhibitory response. It has been suggested that because much of the thirst and drinking response is behavioral, an individual learns what volume of drink is required to restore a given water deficit. Termination of drinking therefore could be a learned response that anticipates a future fluid deficit or matches a known current level of dehydration. The stimuli for gauging the current level of dehydration may be the same as that which initiates the sensation of thirst.

The mere presence of liquid, particularly cold liquid, in the mouth reduces the perception of thirst. Receptors in the mouth and oesophagus responding to different chemical, tactile, pressure, and temperature stimuli are thought to be part of the mechanism that influences the relative intensity of the perception of thirst. The neural activity involved in swallowing and perhaps oropharyngeal and gastric receptors are thought to be effective in sensing or metering the volume of liquid ingested. Distension of the stomach tends to inhibit drinking due to increased gastric stretch receptor activity, although this response does not always reduce the perception of thirst. Taste and other psychological factors can have a stimulatory effect on consumption of a drink that is considered to be palatable.

The continuation and termination of the acute sensation of thirst are regulated by a series of stimuli that operate before all of the drink consumed has been absorbed and before disturbances in the body water pools have been corrected. A variety of receptors located from the mouth to the upper part of the small intestine, and probably neural control from the higher centers of the brain, appear to monitor and regulate the initial volume consumed. After absorption, if restoration of body water pools does not occur the sensation of thirst is once again initiated, presumably by the same homeostatic stimuli that initially evoked the feeling of thirst, and drinking restarts. The integration of the pre- and postabsorptive stimuli modulates the sensation of thirst and finally the volume of drink consumed.

Fluid Requirements

Renal reabsorption can reduce the volume of water and solute loss and hence slow the rate of progress of a fluid deficit, but it cannot stop its development. Intake of fluid either as food or as drink is required to restore a fluid deficit. Daily fluid intake is highly variable between individuals and the rate of loss is dependent on factors such as environmental temperature, physical activity, sweating rate, antidiuretic function, and dietary solute load. A representative normal daily water turnover in a sedentary individual living in a temperate climate and eating a typical Western diet is approximately 2 or 3 l, and a minimum daily fluid intake of approximately 1.7 l is necessary to conserve fluid balance. The water content of the typical Western diet approximates to about 1 l and metabolically derived water produces in the order of approximately 300 ml, which together almost offset the daily obligatory water loss. Therefore, in many situations the requirement for fluid intake can actually be very low.

There are conditions in which water loss is greater than that indicated previously and replacement obviously requires a compensatory increase in daily fluid intake (**Figure 6**). Urine volume is related to the solute content of the diet, with a minimum volume of approximately 500 ml being necessary to eliminate the daily solute load. Diets rich in protein or foods with high sodium content require a greater obligatory urine output for excretion. The renal concentrating ability at maximum antidiuresis determines the minimum urinary water loss for a given dietary solute load. Normally, there is a wide range of urinary osmolality such that the same solute load can be excreted in 500 ml of urine with an osmolality of 1400 mosmol/kg or in 23 l of urine with an osmolality of 30 mosmol/kg. This feature of renal excretion allows body water balance to be maintained while fluid intake volume is varied.

Prolonged relatively intense muscular activity, elevated ambient temperature, and fever all increase the rate of evaporative sweat loss. Individual sweat rates are highly variable, but daily losses of between 10 and 15 l have been recorded. Daily faecal losses associated with a Western diet are usually between 100 and 200 ml; however diarrhea, particularly infectious diarrhea, can produce prodigious fecal water losses that are potentially fatal.

Inappropriate fluid intake can be produced following lesions or development of tumors in regions of the brain associated with the thirst centers. Diabetes insipidus promotes an increase in the volume of fluid ingested, which is caused by a lowering of the basal threshold set point for osmotic thirst. A similar, although less pronounced, lowering of the osmotic thirst threshold occurs during pregnancy. In both the young and the elderly, the thirst response can be blunted and inappropriate drinking habits may occur. Psychogenic disturbances in the sensation of thirst and hence fluid intake have also been reported for a variety of clinical conditions.

See also: **Appetite**: Physiological and Neurobiological Aspects; Psychobiological and Behavioral Aspects. **Behavior**. **Body Composition**. **Brain and Nervous System**. **Caffeine**. **Dehydration**. **Diarrheal Diseases**. **Electrolytes**: Water–Electrolyte Balance. **Infants**: Nutritional Requirements. **Older People**: Nutritional Requirements; Nutrition-Related Problems; Nutritional Management of Geriatric Patients. **Pregnancy**: Energy Requirements and Metabolic Adaptations; Safe Diet for Pregnancy. **Sodium**: Physiology; Salt Intake and Health.

Further Reading

Adolph ED Associates (1947) *Physiology of Man in the Desert*. New York: Interscience.

de Araujo IET, Kringelbach ML, Rolls ET, and McGlone F (2003) Human cortical responses to water in the mouth, and the effects of thirst. *Journal of Neurophysiology* **90**: 1865–1876.

Denton D, Shade R, Zamarippa F *et al.* (1999) Correlation of regional cerebral blood flow and change of plasma sodium concentration during genesis and satiation of thirst. *Proceedings of the National Academy of the United States of America* **96**: 2532–2537.

Engell DB, Maller O, Sawka MN *et al.* (1987) Thirst and fluid intake following graded hypohydration levels in humans. *Physiology and Behaviour* **40**: 229–236.

Johnson AK and Edwards GL (1990) Central projections of osmotic and hypovolaemic signals in homeostatic thirst. In: Ramsay DJ and Booth DA (eds.) *Thirst: Physiological and Psychological Aspects*, ILSI Human Nutrition Reviews, pp. 149–175. London: Springer-Verlag.

Maughan RJ (1994) Fluid and electrolyte loss and replacement in exercise. In: Harries BJ, Williams C, Stanish CW, and Micheli A (eds.) *Oxford Textbook of Sports Medicine*, pp. 82–93. New York: Oxford University Press.

Phillips PA, Rolls BJ, Ledingham JGG, and Morton JJ (1984) Body fluid changes, thirst and drinking in man during free access to water. *Physiology and Behaviour* **33**: 357–363.

Ramsay DJ (1989) The importance of thirst in the maintenance of fluid balance. *Clinical Endocrinology and Metabolism* **3**(2): 371–391.

Ramsay DJ (1990) Water: Distribution between compartments and its relationship to thirst. In: Ramsay DJ and Booth DA (eds.) *Thirst: Physiological and Psychological Aspects*, ILSI Human Nutrition Reviews, pp. 23–34. London: Springer-Verlag.

Ramsay DJ and Thrasher TN (1986) Hyperosmotic and hypovolemic thirst. In: de Caro G, Epstein AN, and Massi M (eds.) *The Physiology of Thirst and Sodium Appetite*, pp. 83–96. New York: Plenum Press.

Robertson GL (1984) Abnormalities of thirst regulation (Nephrology forum). *Kidney International* **25**: 460–469.

Rolls BJ and Rolls ET (1982) *Thirst*. Cambridge, UK: Cambridge University Press.

Rolls BJ, Wood RJ, and Rolls ET (1980) Thirst: The initiation, maintenance, and termination of drinking. *Progress in Psychobiology and Physiological Psychology* **9**: 263–321.

Rolls BJ, Wood RJ, Rolls ET *et al.* (1980) Thirst following water deprivation in humans. *American Journal of Physiology* **239**: R476–R482.

Verbalis JG (1990) Inhibitory controls of drinking: Satiation of thirst. In: Ramsay DJ and Booth DA (eds.) *Thirst: Physiological and Psychological Aspects*, ILSI Human Nutrition Reviews, pp. 313–330. ILSI Human Nutrition Reviews, London: Springer-Verlag.

Tocopherol *see* **Vitamin E**: Metabolism and Requirements; Physiology and Health Effects

Trace Elements *see* **Chromium**. **Copper**. **Immunity**: Effects of Iron and Zinc. **Iodine**: Physiology, Dietary Sources and Requirements. **Iron**. **Manganese**. **Selenium**. **Zinc**: Physiology

***Trans* Fatty Acids** *see* **Fatty Acids**: *Trans* Fatty Acids

TUBERCULOSIS

Contents
Nutrition and Susceptibility
Nutritional Management

Nutrition and Susceptibility

J P Cegielski, Centers for Disease Control and Prevention, Atlanta, GA, USA
D N McMurray, Texas A&M University, College Station, TX, USA

Published by Elsevier Ltd.

The oral traditions of medicine and public health state that malnutrition is an important risk factor for the development of tuberculosis (TB). Malnutrition profoundly affects cell-mediated immunity (CMI), and CMI is the principal host defense against TB. It makes biological sense. Although most health professionals readily accept this principle, much of this belief is based on uncontrolled observations such as disaster situations or on backwards logic from the cachexia common among TB patients. In fact, the evidence in humans is surprisingly thin from the perspective of scientific rigor. Also, few data, if any, quantify the extent of the relative or attributable risk of TB due to malnutrition. Moreover, data from experimental animals, until recently, were based on animal models that largely were not relevant to human TB infection and disease. Malnutrition may account for a greater population-attributable risk of TB than HIV infection, and certainly a much more correctable one.

Malnourished individuals have an increased likelihood of primary or latent infection progressing to active disease. In populations with substantial latent TB infection, the occurrence of malnutrition may be an important determinant of the incidence of TB. The potential public health impact of malnutrition on the global incidence of TB was summarized in the US Surgeon General's 1988 *Report on Nutrition and Health*, which emphasized that malnutrition was the leading cause of acquired, correctable immune system dysfunction throughout the world. The United Nations Food and Agriculture Organization (FAO) estimated that 841 million people in developing countries, or 20% of the world's 1990–1992 population, were undernourished. Modest decreases in resistance affecting such large numbers of people may result in substantial increases in TB incidence at a population level. Population groups at highest risk for poor nutrition are also at high risk for TB, with poverty being the common denominator.

In general, there are three streams of evidence relating the risk of TB to malnutrition: observations in humans, experimental work in animal models, and inferences from related work in microbiology and immunology. In humans, direct evidence for the risk of TB due to malnutrition is sparse, and the data have not been reviewed critically in more than three decades. *In vitro* studies have generated a substantial body of evidence documenting the negative effects of malnutrition on cell-mediated immune function and on the immunology of TB. Although one can reason from the *in vitro* evidence, it cannot replace *in vivo* data. This article summarizes the evidence from observations in human populations and from experimental animal models with relevance to human TB.

TB risk is of two kinds: the risk of becoming infected with *Mycobacterium tuberculosis* and the risk of the infection progressing to TB disease. This article focuses on the risk of infection progressing to disease because that is where CMI comes into play. There is no evidence of a direct relationship between malnutrition and the risk of initial infection. Although both TB and malnutrition are linked with poverty, the data reviewed here suggest no independent association between malnutrition and primary or latent TB infection.

Human Data

Ecological Studies

Ecological studies present fairly convincing evidence that nutrition, isolated to some extent from other historical circumstances, has played a direct role in TB morbidity and mortality. In 1938, Faber reported on TB epidemiology in Denmark during World War I. For most of the war, neutral Denmark exported the bulk of its meat, fish, poultry, and dairy products to the extent that the local diet lacked these protein-, vitamin-, and mineral-rich foods. TB rates increased

as in the warring countries. The German blockade of Denmark in 1918, however, created a surfeit of these foods and TB rates plummeted. In contrast, TB rates in the neighboring warring countries continued to increase unabated.

The second study involves the Trondheim, Norway, Naval Training School, where the high rate of TB among recruits in the early twentieth century was ascribed to crowded, poor housing and unhygienic conditions. In 1954, Leitch reported that TB rates did not decrease after improved housing and hygiene were provided. Subsequently, the diet was fortified with milk, margarine, cod liver oil, whole wheat bread, and fresh fruits and vegetables. TB morbidity promptly declined to the prevailing level for young adults of that area.

The third important contribution to the early literature was Leyton's study of TB morbidity among British and Russian prisoners of war (POW) held in German POW camps during World War II. The prisoners shared the same diet, but the British received Red Cross food supplements amounting to 30 g protein and 1000 kcal per day. In a subsequent radiographic survey, the TB rate among the British was only 1.2% and their plasma proteins were higher than those of the Russians, who had a TB rate of 15–19%. Both groups shared the same living and working conditions and had the same chance for infection. In the malnourished prisoners, TB was more severe, the onset was more rapid, and patients died rapidly with large pulmonary cavities and massive tissue breakdown. Granuloma formation was poor in the malnourished prisoners, supporting the idea that there was a deficit of CMI in this group.

Lastly, McKeown's seminal work in the 1950s and 1960s elaborated on the concept that the decline in TB mortality in England and Wales from 1770 to 1900 was most likely due to improving standards of living in general and to the nutritional status of the population in particular. Through extensive critical reasoning, McKeown excluded the alternative explanations: (i) advances in medicine and medical effort or (ii) natural selection. What remained, according to McKeown, was that the environment and thereby the resistance of the host improved. The death rate from TB at the beginning of the nineteenth century was approximately 40/1000 person-years. It declined to 14/1000 person-years by the end of the century and accounted for nearly half of the overall decline in mortality in the nineteenth century. McKeown concludes, "Having confidently excluded therapy and genetic selection, with reservations, we are left with changes in the environment as the most acceptable reason for the trend of mortality from tuberculosis." Of the features of the environment to be considered, "the evidence in respect of diet seems ... highly suggestive."

Case Series

Nutrition, immune function, and infection interact in complex and dynamic patterns. Protein–energy malnutrition compromises CMI and increases susceptibility to or severity of infections. Conversely, infection can rapidly lead to nutritional stress and weight loss, thereby worsening nutritional status and immunologic function. Therefore, understanding the temporal relationship between the onset of malnutrition and the development of the infectious disease is crucial in order to correctly assess any possible cause–effect relationship.

Since 1968, several case series of postsurgical patients undergoing intestinal bypass surgery for morbid obesity have provided observational data in which nutritional status and incident cases of TB were observed in the same individuals in the correct temporal sequence. These patients experience rapid weight loss and malabsorption due to their short-circuited bowels. In several series, the incidence rate of TB was 1–4% among postoperative patients throughout various durations of follow-up. This range was much higher than expected based on historical or population comparisons. Similarly, partial gastrectomy for ulcer disease was shown to predispose men to TB, but the association was 14 times more likely for men whose weight was $<85\%$ of ideal than for men whose weight was normal for their height. Although the patients in these series do not represent persons at risk for TB in general, and there are no contemporaneous controls, the observations are worth noting due to the sharply increased incidence of TB following nutritional insult.

Cross-Sectional and Case–Control Studies

Cross-sectional and case–control studies generally suffer from the same inherent fatal flaw: Patients with and without TB disease are compared in terms of their concurrent nutritional status. However, TB causes wasting, depression of the immune system, and other changes resembling malnutrition. Therefore, the intrinsic uncertainty regarding the sequence of cause and effect in case–control and cross-sectional study designs becomes intractable. Although these studies demonstrate substantial macro- and micronutrient deficits in TB patients, one cannot infer a causal role for nutritional deficiency in the development of disease from these data because the chronological sequence is unclear and TB plays a role in the development of the nutritional deficits.

Two cross-sectional studies on vitamin D metabolism in relation to TB focused on the molecular and cellular mechanisms of the interaction rather than on the direction of causality. These studies examined the dynamics of $1,25(OH)_2$ vitamin D_3 in lymphocytes and macrophages from patients with TB compared to patients without TB. Investigators determined that lymphocytes obtained by bronchoalveolar lavage from patients with TB expressed specific receptors for $1,25(OH)_2$ vitamin D_3 but not 25(OH)D3. These were primarily $CD4^+$ T lymphocytes. Peripheral blood lymphocytes did not express these receptors. Furthermore, uncultured cells recovered by bronchoalveolar lavage and blood mononuclear cells from normocalcemic patients with TB both produced $1,25(OH)_2$D3. The amount correlated with the number of $CD8^+$ T lymphocytes present but not other cell types. Purified T lymphocytes from all patients with TB produced $1,25(OH)_2$D3, which correlated closely with that produced by unseparated lavage cells. Since $1,25(OH)_2$D3 can improve the capacity of macrophages to kill mycobacteria, these results support the conclusion that cellular interactions mediated in part by $1,25(OH)_2$D3 may be important in the antituberculosis immune response. These data are consistent with those for experimental animals noted later.

Cohort Studies

Very few follow-up studies have been performed with the explicit purpose of understanding the relationship between nutrition and the incidence of TB. The unique strength of cohort studies is that nutritional status is measured prior to the onset of TB.

Only two cohort studies have examined the relationship between micronutrients and TB incidence; both of these included vitamin C. In the 1940s, Getz and coworkers followed 1100 men who were free of TB at baseline, by clinical and radiographic criteria, for up to 5 years with serial clinical, radiographic, and laboratory examinations. Among 16 men who developed TB, blood levels of vitamins A and C were consistently lower than those of men who remained free of TB. Plasma vitamin A levels were low in 13 of 16 men who developed TB compared to 30% (318/1058) of those who did not. Similarly, plasma vitamin C levels were low in all of the subjects who developed active TB compared to only 11% (117/1013) of those who did not. Exposure to TB did not differ between the men who developed TB and those who did not.

Investigators in Finland randomized 26 975 healthy male smokers aged 50–69 years to supplementation with tocopherol, β-carotene, both, or neither. The subjects were followed for a mean of 6.7 years for diagnoses of cancer identified through a registry of all hospital discharges and the associated diagnoses in the region. Hemilä and coworkers analyzed the dietary data in this cohort for vitamin C and vitamin C-rich foods and the discharge registry for diagnoses of TB. In more than 173 000 person-years of follow-up, 167 cases of TB were detected. Higher intake of fruits and vegetables was associated with lower risk of TB. Among those with increased intakes of vitamin C and fruits and vegetables, the adjusted relative risk of TB decreased to 0.4 (95% confidence interval, 0.24–0.69). This study is noteworthy for its size and quality of data. However, detecting TB through hospital discharges selects TB patients who were sick enough to require hospital admission. Lower intakes of fruits and vegetables and of vitamin C may be associated with higher rates of hospitalization rather than higher rates of TB.

Regarding macro indicators of nutritional status, as part of the long-term follow-up of participants in the large-scale BCG vaccine trials in Georgia and Alabama, Comstock and Palmer reported that the incidence of TB was 2.2 times higher in children with 0–4 mm subcutaneous fat than in those with 10 mm subcutaneous fat. Cegielski examined the relationship between undernutrition and the incidence of TB based on data from the first National Health and Nutrition Examination Survey (NHANES-1) and the NHANES-1 Epidemiological Follow-Up Study (NHEFS). NHANES-1 was a cross-sectional survey of a representative sample of the US population from 1971 to 1975. In the NHEFS, the adult subjects of NHANES-1, aged 25–74 years at baseline, were followed up longitudinally with serial waves of questionnaires and examinations. Follow-up exceeded 95%. Through 1987, 64 cases of TB were detected. In proportional hazards analysis, having body mass index, average skin-fold thickness, or upper arm muscle area in the lowest decile of the population increased the adjusted hazard of TB from 6- to 10-fold, controlling for other known risk factors for TB.

Palmer's group studied the relationship of TB incidence to naturally acquired delayed-type tuberculin sensitivity among US Navy recruits. Nearly all Navy recruits from 1949 to 1951 were skin tested and followed longitudinally. Of 68 754 subjects with follow-up data, tuberculin sensitivity was recorded as >0 mm for 8704 (12.7%). During 4 years of follow-up, 109 developed TB: $28/10^5$ among those with 0 mm skin test reactions, $29/10^5$ among those with 1–9 mm reactions, and $157/10^5$ among those with 10 mm or greater reactions. Later, these investigators related the risk of

TB to 'body build' by obtaining height and weight data from the entrance medical examination on a stratified random sample of 1138 subjects. A weight–height index was constructed based on deviation of weight from the median weight-for-height of the study sample. There were no significant differences in tuberculin sensitivity by weight, height, or weight–height index. In contrast, the weight–height index was strongly associated with TB incidence: TB incidence was $75/10^5$ for those 15% or more below the median weight for their height and decreased to $19/10^5$ for those at least 5% overweight for their height ($p < 0.01$ for both purified protein derivative (PPD) groups). The trends were the same regardless of the degree of tuberculin sensitivity, although incidence rates were higher among those with 10 mm PPD reactions. Edwards extended Palmer's study to more than 823 000 Navy recruits and found that TB developed three times more often in young men 10% or more below their ideal body weight than in those 10% or more above it.

Curiously, these authors avoid any reference to inadequate nutrition in the recruits. Instead, they conclude that the results demonstrate an association between body build and TB disease, that some unknown factor associated with body build is an important determinant of host susceptibility to TB disease but not to primary infection. A small but consistent body of literature has accumulated on the relationship of body build to TB and was reviewed by Snider in 1987. One study stands out. The country of Norway attempted to screen all people older than age 14 years for TB with compulsory mass miniature radiography from 1963 to 1975. This screening program covered 42–85% of the population, varying by age group. Height and weight were measured accurately for approximately 80% of those screened. From these data, Tverdal reported results from more than 1.7 million Norwegians, with follow-up via the national notification system through 1982 (i.e., 8–19 years of follow-up; mean, 12.1 years). A total of 2531 incident cases of TB were identified. The incidence of pulmonary TB, both sputum smear positive and smear negative, declined logarithmically with increasing body mass index (BMI) for both sexes, all age groups, and at all durations of follow-up: The age-adjusted incidence of new pulmonary TB was five times higher in the lowest BMI category than in the highest. This relationship was not observed for extrapulmonary TB. Tverdal argued that the association could not be explained by preexisting nutritional status or TB. As with the US Navy studies, he argues that this relationship is a function of body build and, aside from this single mention, does not discuss nutrition. Comstock suggested that body build may influence susceptibility to TB because of differences in pulmonary mechanics, but no studies have attempted to address this hypothesis.

Interpreting the findings of these large studies in terms of body build rather than nutritional status disregards the well-established concept that body weight is a function of the balance over time between caloric intake and energy expenditure. Clearly, increased or decreased intake can transform a thin individual into an overweight one or an obese person into an underweight one. With physical training and appropriate intake, a person with either body type may become muscular and fit. Therefore, the concept of body build as a fixed phenotype that, by itself, predisposes to or protects against TB may be inadequate. Reinterpreting the findings of these studies in terms of nutritional status would be useful. A more inclusive view may be that body habitus as a function of genetic and early environmental influences and nutritional status as a function of ongoing nutrient intake and physical activity each affect the incidence of TB. Sorting out the mechanisms and relative contribution of each remains a challenge for future research.

Intervention Trial

Micronutrient deficiencies in relation to TB are difficult to study in isolation in human beings. In this respect, a unique study of the effect of micronutrient supplementation on TB incidence was reported by Downes in 1949. In a controlled trial among the families of black TB patients in Harlem, New York City, 194 of 218 families under public health supervision in 1941 were examined and divided into two groups matched for family size. The families were allocated alternately to receive vitamin and mineral supplements versus no supplements, along with the health department's standard health education program. The education program included intensive nutrition education. The two groups were similar in terms of prior attack rates and mortality from TB, prevalence of primary and reinfection TB at the start of the study, sputum smear positivity among the index cases, and relation of the index case to the rest of the family. In addition, the groups were similar in terms of their income, the proportion receiving welfare, the degree of crowding within the home, and their food habits. After 5 years of follow-up, using an intention-to-treat analysis, the risk of TB in the control group was 2.8 times the risk of TB in the vitamin group. However, there was substantial noncompliance with the supplements. The relative risk of TB among the controls (1096 person-years of follow-up) compared to those who actually took the vitamin

supplements for the entire follow-up period (644 person-years of follow-up) was 5.9. The relative risk among contacts compared to those who did not take the supplements despite being allocated to that group (27% of the supplement group) or who only took them for some of the follow-up period (33% of the supplement group) (total of 598 person-years of follow-up) was only 1.82. Therefore, vitamin supplementation substantially reduced the risk of TB among family contacts of active TB cases.

This study probably underestimates the efficacy of micronutrient supplements for two reasons related to an underlying secular trend: The economic status and food habits of both groups improved substantially during the period of the study. Both these secular trends would diminish the apparent effect of vitamin and mineral supplements. A third consideration, unrelated to secular trends, was that noncompliance with the supplements by more than half the individuals in the supplemented group would further bias the measure of effect toward the null. With these three factors working against the experimental intervention, it is likely that the effect may be greater than the findings demonstrated.

Nutrition and the Immune Response to BCG Vaccine in Humans

One study design permits prospective evaluation of the effects of malnutrition on the immune response to mycobacterial proteins closely related to *M. tuberculosis*, namely delayed-type hypersensitivity (DTH) responses following BCG vaccination. Satyanarayana *et al*. showed that milder grades of malnutrition did not affect the skin test response to PPD 6 months after immunization with BCG, but that children with kwashiorkor were skin test negative. Chandra and Newberne demonstrated that the DTH skin test response to tuberculin and to numerous other antigens is reduced in protein–energy malnutrition in children and adults. Among TB patients, PPD skin test reactivity was directly proportional to serum transferrin level, a sensitive indicator of protein malnutrition. Similarly, malnourished individuals do not develop skin test responses to tuberculin as often or as large after BCG vaccination as do well-nourished children. Importantly, this effect has been demonstrated even in modest protein–energy malnutrition.

Nutrition–Immunity–Tuberculosis: Limitations of Human Data

Although the Surgeon General's report summarizes data on the interaction of nutrition and infection, it cautions that few reliable data have been obtained in human subjects on the influence of individual essential nutrients or of protein–energy nutrition on specific immune system functions and their interactions. Nutrition, immune function, and infection interact in complex and dynamic patterns. Many intervening and unknown variables affect the relationship. Protein–energy malnutrition (PEM) impairs CMI and worsens infections. Conversely, infection can lead rapidly to weight loss, malnutrition, and immunologic dysfunction. Indeed, PEM is only partly due to food deprivation. Common infectious diseases, such as diarrheal diseases and respiratory and parasitic infections, are major contributing and precipitating factors in PEM. However, in patients with TB it is nearly impossible to determine accurately their nutritional status prior to the onset of TB and, therefore, determine whether malnutrition led to TB or whether TB led to malnutrition. This problem naturally leads to the use of experimental animal models to elucidate the causal links between nutritional deficiencies, immune system function, and TB.

Experimental Animal Data

Guinea Pig Model of Pulmonary Tuberculosis

In the past 30 years, a modest literature has accumulated regarding the link between diet, antimycobacterial immunity, and disease resistance in TB. The vast majority of this work has been conducted in a highly relevant guinea pig model of low-dose pulmonary TB. The pathogenesis of TB in this model mimics essentially all of the important aspects of TB in humans.

Early studies established that moderate, chronic deficiencies of protein and other nutrients (e.g., zinc) could be induced in guinea pigs, and that the resulting nutritional states had many of the metabolic hallmarks of human dietary deficiencies. In general, the design of these experiments called for giving BCG vaccine to one group of guinea pigs among a larger group receiving different diet treatments and measuring a number of antigen-specific immune responses *in vitro* and *in vivo* several weeks later. Groups of vaccinated and nonvaccinated animals from each diet group were then challenged with an aerosol containing a low dose of virulent *M. tuberculosis*. The ability of the guinea pigs to control the infection was assessed quantitatively by culture of viable mycobacteria from the lungs and spleens.

Protein

Most of the work involving this model has been carried out with moderate, chronic protein deficiency.

Feeding a 10% ovalbumin-based diet over several weeks resulted in a dramatic loss of T cell functions in BCG-vaccinated guinea pigs. Thus, protein-deprived animals had much smaller PPD skin tests and their lymphocytes proliferated poorly to mitogenic and antigenic stimuli *in vitro*. PPD-stimulated T cells from low-protein animals produced significantly less interleukin (IL)-2 and interferon (IFN), and macrophage–lymphocyte cocultures from malnourished animals produced less tumor necrosis factor-α (TNF-α) in response to infection of the macrophages with virulent *M. tuberculosis*. Following virulent pulmonary challenge, protein-deficient guinea pigs were unable to form mature, well-circumscribed granulomas in the lung, and they expressed significantly less BCG-induced resistance in the lung and spleen. Not only was BCG-induced protection diminished by protein deficiency but also the response to exogenous reinfection was impaired. Furthermore, although immune cells from normally nourished guinea pigs adoptively protected syngeneic, protein-deficient guinea pigs against aerosol infection, the reverse was not true. That is, immune lymphocytes from low-protein animals did not protect naive, normally nourished recipients.

Protein malnutrition altered the absolute and relative numbers of total T lymphocytes and various subpopulations, including $CD2^+$, $CD4^+$, $CD8^+$, and Fc receptor-bearing T cell subsets in the circulation and lymphoid organs (e.g., spleen and bronchotracheal lymph nodes draining the infected lung). Taken together, these results imply that protein deficiency is accompanied by alterations in the ability of guinea pigs to regulate the normal recirculation and trafficking of T lymphocytes that would be required for the formation of protective granulomas. These phenomena could be explained by diet-induced changes in the production or function of chemokines, which have been observed to be altered in TB, or by perturbations in the expression of adhesion molecules on T cells or endothelial cells.

Finally, macrophages from TB patients are known to produce suppressive factors for T cells, including transforming growth factor-β (TGF-β). Alveolar macrophages are particularly effective at downregulating T cell proliferation in many species, including humans. Although protein deficiency was not associated with loss of some macrophage functions in the guinea pig model, alveolar macrophages suppress the mitogen-induced proliferation of autologous splenic lymphocytes at macrophage-to-lymphocyte ratios of 1:4 or greater. More important, the intrinsic suppression of alveolar macrophages was enhanced 10-fold in this system when the cells were derived from protein-deficient guinea pigs. In a separate series of experiments, it was demonstrated that TGF-β was produced in higher amounts by cells from protein-deprived guinea pigs, and that recombinant human TGF-β injected daily into guinea pigs infected with virulent *M. tuberculosis* suppressed T cell functions and impaired bacillary control in the lungs and spleens of treated animals. Thus, macrophages from protein-deprived guinea pigs appear to be more suppressive for T lymphocyte functions, and this suppression may be mediated, in part, by overproduction of TGF-β.

It should be noted that the profound loss of T cell-mediated resistance that accompanies chronic dietary protein deprivation in this model is substantially and rapidly reversible. BCG-vaccinated guinea pigs maintained on a low-protein diet during the entire 6-week period postvaccination, but given a normal diet beginning on the day of virulent pulmonary challenge, displayed PPD skin test reactivity and vaccine-induced control of bacillary loads in the lungs and spleens 2–4 weeks later. These reactions were indistinguishable from those in BCG-vaccinated animals that had never been protein deficient. One possible interpretation of these observations is that protein deficiency interferes with the expression, but not the development, of T cell-mediated protective mechanisms in TB.

These basic observations were confirmed and extended by studies performed in protein-malnourished mice. Using a high-dose, intravenous challenge model, Chan and colleagues observed many of the same T cell defects that have been reported in low-protein guinea pigs, including loss of control of the virulent infection and impaired granuloma formation and also recovery of resistance following refeeding with an adequate diet. They concluded that loss of resistance to TB in their model was a result of diminished nitric oxide (NO) production by activated macrophages, which occurred secondary to an IFN-γ defect in malnourished animals. These are important studies because they confirm the fundamental nature of the effects of protein deprivation in TB even when such crucial variables as host species and infection dose and route are altered.

Micronutrients

Zinc Chronic dietary zinc deficiency was found to exert a profound suppressive effect on T lymphocyte functions in BCG-vaccinated guinea pigs. Thus, there was significant anergy in response to PPD skin tests in zinc-deficient animals and dramatic loss of PPD-induced lymphoproliferation *in vitro*. The activity of a cytokine, macrophage migration

inhibitory factor, was also impaired by zinc deficiency. Taken together, these data implied that zinc deficiency interfered with the ability of BCG vaccine to induce protection against virulent, pulmonary challenge. However, no differences were observed between the bacillary loads of zinc-deficient and normally nourished guinea pigs 4 weeks following aerosol infection, and BCG exerted the same protective effect regardless of zinc status in this model.

Vitamin D Calcitriol [1,25(OH)$_2$ vitamin D$_3$] is a potent coactivator of macrophages. Several *in vitro* studies demonstrated that the addition of calcitriol to cultured human macrophages enhanced the ability of the cells to control the intracellular replication of virulent *M. tuberculosis* over several days in culture. The role of dietary vitamin D deficiency was examined in the guinea pig model of pulmonary TB. Feeding a diet completely devoid of vitamin D for several weeks resulted in marked depletion of serum levels of the calcitriol precursor, 25(OH) vitamin D$_3$, and resulted in significant loss of some T cell functions in BCG-vaccinated guinea pigs. However, vitamin D deficiency did not alter the course of TB disease in nonvaccinated guinea pigs, nor did it impair the protective efficacy of BCG vaccination in this model.

Conclusions from Experimental Animal Studies

The previously discussed studies confirm that protein deficiency, in particular, can have devastating consequences on both innate and vaccine-induced resistance against TB in animal models. Certain micronutrient deficiencies, although not as well studied, also appear to affect the immune response to *M. tuberculosis*, but the effect on the course of the disease is less clear. The precise mechanisms by which diet exerts these effects remain to be elucidated. However, the results of the experiments summarized previously point to defects in T cell trafficking and antigen-induced proliferation, the inability to form mature granulomas, diminished production of 'protective' cytokines (e.g., IL-2, IFN-γ, and TNF-α) and antimycobacterial effector molecules (e.g., NO in mice), and increased suppression by adherent cells, perhaps secondary to increased TGF-β production.

Summary

This article critiques known studies in human populations and in relevant animal models to cover the *in vivo* evidence concerning the risk of TB due to malnutrition. Although TB is clearly related to malnutrition, the risk relative to specific levels and types of PEM and micronutrient deficiencies remain to be defined. Analysis of the NHANES-1 Epidemiological Follow-Up Study provides plausible estimates of the relative risk in a representative nationwide sample of adults. The 6- to 10-fold increase in relative risk includes mild to moderate as well as severe undernutrition. Severe PEM may increase the relative risk more than mild or moderate malnutrition, but severe malnutrition occurs in a small fraction of the population, even in low-income countries, except, for example, in famine, war, or disaster situations. Mild to moderate PEM or micronutrient deficiencies affect larger fractions of the population at risk for TB, so prevention efforts may not be successful if they target only severely undernourished groups.

The questions we would like answered include not only how much of the TB burden in a population is due to malnutrition but how is TB due to malnutrition? As suggested by work in the aerosol-infected guinea pig model, protein undernutrition in particular impairs host defense against TB and the impairment is rapidly reversed with nutritional rehabilitation. Changes in the movement and proliferation of T lymphocyte subpopulations in response to specific antigens, in the production of key cytokines, in the formation of organized granulomas, and in macrophage activation have been identified as important components of the process.

Acknowledgements

This article is adapted with Permission from Cegielski JP, Mc Murray DN. "The Relationship between Malnutrition and Tuberculosis." Int J Tuberc Lung Dis 2004; 8(3): 286–298.

See also: **Ascorbic Acid**: Deficiency States. **Cytokines**. **Immunity**: Effects of Iron and Zinc. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **Tuberculosis**: Nutritional Management. **Vitamin D**: Physiology, Dietary Sources and Requirements. **Zinc**: Physiology; Deficiency in Developing Countries, Intervention Studies.

Further Reading

Dai G, Phalen SW, and McMurray DN (1998, July) Nutritional modulation of host responses to mycobacteria. *Frontiers in Bioscience:* 110–122.

Ellner JJ (1997) Review: The immune response in human tuberculosis—Implications for tuberculosis control. *Journal of Infectious Diseases* **176**: 1351–1359.

Harries AD, Nkhoma WA, Thompson PJ, Nyangulu DS, and Wirima JJ (1988) Nutritional status in Malawian patients with pulmonary tuberculosis and response to chemotherapy. *European Journal of Clinical Nutrition* **42**: 445–450.

Kielmann AA, Uberol IS, Chandra RK, and Mehra VL (1976) The effect of nutritional status on immune capacity and immune

responses in preschool children in a rural community in India. *Bulletin of the World Health Organization* **54**: 477–483.
McMurray DN (1994) Guinea pig model of tuberculosis. In: Bloom BR (ed.) *Tuberculosis: Pathogenesis, Protection and Control*, pp. 135–147. Washington, DC: American Society for Microbiology.
McMurray DN (1998) Impact of nutritional deficiencies on resistance to experimental pulmonary tuberculosis. *Nutrition Review* **56**: S147–S152.
McMurray DN and Bartow RA (1992) Immunosuppression and alteration of resistance to pulmonary tuberculosis in guinea pigs by protein undernutrition. *Journal of Nutrition* **122**: 738–743.
McMurray DN, Bartow RA, Mintzer CL, and Hernandez-Frontera E (1990) Micronutrient status and immune function in tuberculosis. *Annals of the New York Academy of Sciences* **587**: 59–69.
Roland CG (1992) *Courage under Siege: Starvation, Disease, and Death in the Warsaw Ghetto*, pp. 154–174. New York: Oxford University Press.
Satyanarayana K, Bhaskaran P, Seshu VC, and Reddy V (1980) Influence of nutrition on post-vaccinial tuberculin sensitivity. *American Journal of Clinical Nutrition* **33**: 2334–2337.
Schwenk A and Macallan DC (2000) Tuberculosis, malnutrition, and wasting. *Current Opinion in Clinical Nutrition and Metabolic Care* **3**: 285–291.
Scrimshaw NS, Taylor CE, and Gordon JE (1968) Effect of malnutrition on resistance to infection. In: Scrimshaw NS, Taylor CE, and Gordon JE (eds.) *Interactions of Nutrition and Infection*, pp. 60–142. Geneva: World Health Organization.
Snider DE (1982) Jejunoileal bypass for obesity: A risk factor for tuberculosis. *Chest* **81**: 531–532.
Snider DE (1987) Tuberculosis and body build. *Journal of the American Medical Association* **258**: 3299.
Tverdal A (1986) Body mass index and incidence of tuberculosis. *European Journal of Respiratory Diseases* **69**: 355–362.

Nutritional Management

J P Cegielski, Centers for Disease Control and Prevention, Atlanta, GA, USA
L Demeshlaira, Emory University, Atlanta, GA, USA

Published by Elsevier Ltd.

Nutritional Status at the Time of Diagnosis

Compared to people without tuberculosis (TB), TB patients have significantly lower body mass index (BMI), skinfold thicknesses, mid-upper arm circumference, and overall proportion of body fat. In the United States, weight loss was present at diagnosis in 45% of patients. In Tanzania, among 200 consecutive adult patients with sputum smear-positive pulmonary TB, 77% of males and 58% of females had a BMI $<18.5\,kg/m^2$, whereas slightly more than 20% had a BMI $<16\,kg/m^2$. In Malawi, TB patients' muscle strength was substantially weaker than the control group as measured by hand-grip dynamometry (mean, 29.9 kg versus 36.5 kg in males; mean, 23.3 kg versus 32.1 kg in females), suggesting less somatic protein. TB patients in Malawi had 35% lower fat mass and 19% lower lean body mass than controls. BMI in both countries was lower in patients with a long symptom history and with extensive disease than in the complementary groups. In countries with a relatively high prevalence of both TB and HIV, one-third to two-thirds of TB patients have HIV infection. Patients with concurrent HIV infection tend to be even more malnourished. In the United Kingdom, TB patients had significantly lower serum albumin levels than controls (mean, 37 g/l versus 46 g/l), suggesting the possibility of protein undernutrition. Among Asians living in the United Kingdom, BMI among TB patients ($19.3\,kg/m^2$) was 13% lower than that of controls ($22.2\,kg/m^2$), skinfold thickness was 13% lower, and arm muscle circumference was 20% lower.

The relationship between concurrent TB and malnutrition is complex. Even though malnutrition is an important risk factor for TB, TB in turn results in cachexia, anorexia, and asthenia. As in other infectious diseases, resting energy expenditure is increased and intake is decreased. Weight loss and muscle wasting may in part be related to tumor necrosis factor-α (TNF-α) and other proinflammatory cytokines that participate in the immune response against TB. On the one hand, TNF-α plays a critical protective role against *Mycobacterium tuberculosis*. On the other hand, TNF-α results in wasting and anorexia. One of the first research teams to identify TNF called it 'cachectin' because the team identified it as the chemical mediator of weight loss in cancer and chronic inflammatory diseases. In experimental mycobacterial infections, TNF-α is required for the control of bacillary growth and the protective granulomatous response. Patients receiving anti-TNF-α monoclonal antibodies for the treatment of rheumatoid arthritis developed reactivation TB at much higher rates than rheumatoid arthritis patients not receiving TNF blockers. One school of thought believes that TNF and other cytokines released from the population of activated macrophages responding to *M. tuberculosis* account for much of the weight loss and tissue damage that characterize this disease. Other experts disagree, however, and consider the matter unresolved. Protein utilization may be altered by the cytokine milieu. Anabolic pathways may be functionally blocked due to preferential oxidation of ingested amino acids for energy rather than for protein synthesis.

Protein–energy malnutrition (PEM) rarely occurs without micronutrient deficiencies as well. TB patients have been found to be deficient in vitamins A, B_6, and D as well as zinc, copper, and iron. Of these, vitamins A

and D, zinc, and copper play roles in cell-mediated immune responses. The D vitamins, for example, are important macrophage activators. One study found that in addition to serum albumin, blood hemoglobin, plasma retinol, and plasma zinc were significantly lower in malnourished TB patients than in well-nourished TB patients, well-nourished healthy controls, and undernourished but otherwise healthy controls.

Comorbidity and Nutrition in Tuberculosis Patients

TB often occurs in association with other diseases or conditions that have nutritional implications. HIV infection is a strong risk factor for the progression of primary or latent infection with *M. tuberculosis* to active TB disease. Tuberculin skin test (TST)-positive, HIV-infected people develop active TB at the rate of 7–10% per year. TST-positive individuals without HIV infection develop active TB at the rate of 5–10% over their lifetime. HIV infection, of course, has profound implications for nutritional status and nutrient requirements. Diabetes mellitus also increases the risk of progressive primary or reactivation TB. Nutrition is a central consideration in the pathogenesis and management of diabetes mellitus. TB is more frequent in alcoholic individuals. Alcoholism is associated with profound alterations in nutrition that affect production of red blood cells, intermediary metabolism, and mucous epithelial tissue such as the lining of the gastrointestinal system. These include deficiencies of thiamine, all fat-soluble vitamins (A, D, E, and K), folic acid, pyridoxine, ascorbic acid, cobalamin, and zinc. Therefore, nutritional management of TB patients frequently must take these and other comorbidities into account.

Effect of Malnutrition on the Course of Tuberculosis

To the extent that cellular immune function affects recovery from TB, one might expect malnutrition to retard or impair the response to treatment. The adverse consequences of malnutrition on the course and outcome of TB disease have been well documented. Malnutrition increases the risk of death in TB patients nearly twofold. Among approximately 1200 patients with TB followed prospectively, 10.9% of patients with moderate to severe malnutrition died in the first 4 weeks compared to 6.5% of patients with normal nutritional status or mild malnutrition. Another study found that TB patients with a BMI $<17.0\,kg/m^2$ were at increased risk of early death. These findings cannot be interpreted to mean malnutrition *per se* causes the increased mortality risk. For example, the duration or extent of disease or drug resistance may be the primary cause(s) of death. At the same time, these disease characteristics would lead to even worse malnutrition. In children, weight for age is an important indicator of the prognosis. In general, the severity of malnutrition is an important indicator of the progress of the disease, and normalization of body weight in response to treatment is a positive sign.

In addition, anemia and hypoalbuminemia are associated with severity of the clinical course of TB. Severe anemia has been negatively correlated with the response to treatment and is associated with early mortality from TB. Anemia in TB patients may be normocytic and normochromic, or it may be microcytic and hypochromic, suggesting a relationship to iron homeostasis in some cases. Iron metabolism is disrupted in active TB. *Mycobacterium tuberculosis* requires iron to grow and avidly scavenges iron from its environment. One aspect of the host response to TB is intense sequestration of iron that restricts microbial growth. Whether iron supplementation or iron chelation may play a role in the treatment of TB remains speculative.

Nutritional Management of Tuberculosis Patients

The extent to which optimal nutrient intake improves the body's ability to heal during (and after) TB treatment is not well established. Fortunately, the immunological deficits associated with PEM and various micronutrient deficiencies are reversed rapidly with nutritional rehabilitation. One would expect this to improve recovery. The scientific literature is limited, however, with regard to the effect of nutrition on the outcomes of treatment. Only one randomized, double-blind trial is known, comparing vitamin A and zinc supplementation with placebo in 80 patients with pulmonary TB in Jakarta, Indonesia. BMI was $<18.5\,kg/m^2$ in 64% of TB patients, plasma retinol was $<0.70\,\mu mol/l$ in 32%, and plasma zinc was $<10.7\,\mu mol/l$ in 30%. After treatment, plasma zinc was similar in the two groups, and plasma retinol was significantly higher in the supplement group. Sputum conversion and resolution of the radiographic extent of lung disease were slightly more rapid in the supplement group. One other prospective study compared 6 weeks of high-energy nutritional supplements plus education to education alone among TB patients starting treatment, but the outcome was nutritional status, not TB disease status. At 6 weeks, the supplement group had gained substantially more body

weight (2.6 versus 0.8 kg) and total lean body mass (1.2 versus 0.0 kg) significantly more in the trunk than limbs. By 24 weeks, however, the relative difference in weight gain between the two groups was smaller (4.4 versus 2.7 kg), and the difference was mostly in fat mass, especially in the limbs. The patients' treatment outcomes were not compared, but at both time points, the supplement group performed better in tests of muscle strength or endurance.

It is clear that TB patients gain weight during treatment, and weight gain is one of the traditional metrics for treatment effectiveness. The extent to which diet influences the course and outcome of treatment is less clear. A large study in Madras in the 1950s compared the nutritional aspects of sanatorium versus home treatment for pulmonary TB and found no difference in the outcome after 12 months of treatment. In the sanatorium patients, however, sputum was converted to negative faster, and fewer patients relapsed. Both diet and food intake were substantially better in the sanatorium patients, and these patients gained more weight. However, adherence to treatment may have been better also. In addition, strict bed rest in the sanatorium may have affected energy balance and nutrient utilization. In the Malawian study noted previously, weight, arm muscle area, arm fat area, and hand grip strength increased significantly by 4 weeks and continued to increase at 8 weeks. However, among the Asian population in the United Kingdom described previously, after 12 months of treatment arm muscle circumference and serum albumin were still lower than those of controls, suggesting that full recovery may take even longer. The dynamics of nutritional status in relation to TB treatment in a very low-income population were described in the Tanzanian study of 200 patients noted previously. Patients gained weight irrespective of their clinical response, and weight gain continued for up to 6 months—the duration of treatment. The mean weight gain was 10 kg. Weight gain was most closely correlated with the duration of hospitalization. After treatment was completed and patients were released from the hospital, the mean weight of cured patients declined by 2.5 kg, but it declined 7.3 kg among patients who relapsed. Apparently, the hospital diet was substantially better than what was normally available to these patients.

In the treatment of TB patients, the nutritional goals are to supply adequate nutrients (i) to fuel the elevated resting energy expenditure and catabolic state and (ii) to support the extensive cellular proliferation, metabolic pathways, and protein synthesis required to repair damaged tissues and replenish somatic reserves. A deficit of 10 kg, evenly divided between lean and fat mass, represents 60 000 kcal that needs to be replaced in addition to daily nutrient requirements. A rough guideline includes at least 2 g of high-quality protein and 50 kcal of energy per kilogram body weight per day, with caloric intake distributed as 40–50% from carbohydrates, 30–40% protein, and not more than 20% to 30% fat, including essential fatty acids. Micronutrients should be supplemented to at least the recommended daily allowance, although some experts may increase the intake of vitamins and minerals needed for cellular immune function and tissue regeneration, especially vitamins A, B_6, and D and the minerals copper, iron, and zinc. Calcitriol, for example, plays an important role in macrophage activation and has been shown to contribute to the control of intracellular growth of *M. tuberculosis*.

Nutritional Issues with Antituberculosis Drug Toxicity

The standard treatment for all newly diagnosed patients with active TB lasts at least 6 months and includes isoniazid, rifampicin, pyrazinamide, and ethambutol for 2 months followed by isoniazid and rifampicin for 4 months. Isoniazid interferes with vitamin B_6 metabolism. The term vitamin B_6 refers to a group of vitamins including pyridoxine, pyridoxal, and pyridoxamine. Isoniazid combines with pyridoxal or pyridoxal phosphate to form hydrazones, which are potent inhibitors of pyridoxal kinase. Thus, isoniazid blocks the formation of the coenzyme form of the vitamin. Normally, pyridoxal phosphate participates in metabolic transformations of several amino acids.

In the absence of vitamin B_6 supplements, approximately 2% of patients treated with 5 mg/kg will develop peripheral neuritis, increasing up to 20% at higher doses or in high-risk patients such as diabetics or alcoholics. Daily administration of 25 mg of vitamin B_6 prevents peripheral neuritis and nearly all nervous system side effects of isoniazid. A similar situation exists for the second-line anti-TB drug, D-cycloserine, which has useful antibiotic properties but inhibits a wide range of pyridoxal phosphate-requiring enzymes. Cycloserine is used only for patients with drug-resistant TB, but in such patients doses of 200–300 mg of vitamin B_6 have been recommended.

Although vitamin B_6 occurs in many foods, if intake is limited to types of food with low B_6 content, deficiency can occur, especially in TB patients

whose intake is compromised by anorexia and metabolism accelerated by chronic inflammation. Foods such as peas, beans, and cereals contain adequate amounts. Meat and eggs are not good sources, however, and high protein intake increases vitamin B_6 requirements. Deficiencies of vitamin B_6, like other nutrients, can increase because of significant intestinal malabsorption or loss by gastrointestinal dysfunction.

Treatment of TB may induce other problems that affect nutritional status and nutrient intake. Three of the first-line drugs—isoniazid, rifampicin, and pyrazinamide—all carry a small risk of chemical hepatitis, ranging from the asymptomatic elevation of hepatic transaminases to severe and potentially fatal hepatitis. Although nutritional factors do not contribute to the cause, hepatitis has many important consequences affecting nutritional status and nutrient intake. Other anti-TB drugs that affect nutrition include *para*-aminosalicylic acid and ethionamide, which commonly cause moderate to severe gastrointestinal disturbances such as abdominal pain, nausea, vomiting, and anorexia.

Barriers to Nutritional Rehabilitation

Many barriers to successful nutritional rehabilitation do not include the disturbed metabolism or pathological processes of the disease. The following cognitive and behavioral factors may play important roles in promoting or impeding nutritional rehabilitation:

- *Education for health care workers:* It is necessary to promote education related to the nutrition of TB patients to health care workers in virtually every country, but especially in regions with a high TB incidence. These workers will be responsible for planning and implementing the nutritional program, and their participation is critical.
- *Information/education for patients:* Information and education for patients about the principles of good nutrition and dietary strategies is also lacking. We found no published data on TB patients' knowledge about nutrition/nutritional status or the role of nutrition in treating TB.
- *Resources:* TB mainly occurs among the poorest people in both industrialized and developing countries. A primary problem is the lack of resources for treatment other than the essential anti-TB drugs. The resource deficiency exceeds the capacity of international donor groups. In addition, the resource deficiency includes the unavailability of food, a common problem for people in low- and middle-income countries.
- *Cultural and individual food preferences:* Specific food preferences in different regions and among different people strongly affect willingness to eat and drink. Traditions regarding food are among the strongest ties with one's culture. This is particularly relevant to immigrants, refugees, displaced people, and other migrating populations that comprise important risk groups for TB. In addition to primary factors (e.g., availability of food), secondary factors may be responsible for the apparent lack or misuse of food. One may determine which secondary factors in the person's environment have been responsible for inadequate nutrient consumption and rectify them to the extent possible.

Nutritional support is an important component of the comprehensive care of people with TB. In theory, a nutritionist–dietician can tailor a specific dietary prescription for each TB patient based on a careful nutritional assessment and the patient's clinical status, including any comorbidities. In low- and middle-income countries, however, this degree of effort may not be possible and it may not be necessary, given the sparse data reviewed in this article. Virtually all TB patients should be given nutritional support and recommendations, but a generally healthful standardized diet with adequate protein, energy, essential fats, and micronutrients may suffice for the large majority of patients to achieve the potential benefits. A deeper understanding of the essential role of nutrition in TB pathogenesis and resolution may help improve treatment practices and improve outcomes of TB patients.

The extent of nutritional deficits among TB patients is reason enough for intensified nutritional support. The association of malnutrition with worse clinical outcomes and the possibility of favorably influencing the course of treatment add to the impetus for further work to be carried out to identify the optimal strategies for nutritional intervention.

See also: **Anemia**: Iron-Deficiency Anemia. **Lung Diseases**. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **Tuberculosis**: Nutrition and Susceptibility. **Vitamin B_6**. **Weight Management**: Approaches.

Further Reading

Edwards LB, Livesay VT, Acquaviva FA, and Palmer CE (1971) Height, weight, tuberculous infection, and tuberculous disease. *Archives of Environmental Health* **22**: 106–112.

Harries AD, Nkhoma WA, Thompson PJ, Nyangulu DS, and Wirima JJ (1988) Nutritional status in Malawian patients with pulmonary tuberculosis and response to chemotherapy. *European Journal of Clinical Nutrition* **42**: 445–450.

Karyadi E, West CE, Schultnik W *et al.* (2002) A double-blind, placebo-controlled study of vitamin A and zinc supplementation in persons with tuberculosis in Indonesia: Effects on clinical response and nutritional status. *American Journal of Clinical Nutrition* **75**: 720–727.

Kennedy N, Ramsay A, Uiso L *et al.* (1996) Nutritional status and weight gain in patients with pulmonary tuberculosis in Tanzania. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **90**: 162–166.

McMurray DN (1998) Impact of nutritional deficiencies on resistance to experimental pulmonary tuberculosis. *Nutrition Reviews* **56**: S147–S152.

Onwubalili JK (1988) Malnutrition among tuberculosis patients in Harrow, England. *European Journal of Clinical Nutrition* **42**: 363–366.

Palmer CE, Jablon S, and Edwards PQ (1957) Tuberculosis morbidity of young men in relation to tuberculin sensitivity and body build. *American Review of Tuberculosis and Pulmonary Disease* **76**: 517–539.

Schwenk A and Macallan DC (2000) Tuberculosis, malnutrition, and wasting. *Current Opinion in Clinical Nutrition and Metabolic Care* **3**: 285–291.

Snider DE (1987) Tuberculosis and body build. *Journal of the American Medical Association* **258**: 3299.

Tumor *see* **Cancer**: Epidemiology and Associations Between Diet and Cancer; Epidemiology of Gastrointestinal Cancers Other Than Colorectal Cancers; Epidemiology of Lung Cancer

U

ULTRATRACE ELEMENTS

F Nielsen, Grand Forks Human Nutrition Research Center, Grand Forks, ND, USA

This article is reproduced from the first edition, pp. 1884–1897, © 1999, Elsevier Ltd., with revisions made by the Editor.

Definition

In the earlier part of this century, scientists could qualitatively detect small amounts of several mineral elements in living organisms. In reports, these elements were described as being present in 'traces' or 'trace amounts.' It is not surprising that these elements soon became known as trace elements. Today, sophisticated analytical techniques have permitted the accurate measurement of the amount of many mineral elements, some at very low concentrations, in biological material. The trace elements found in living organisms may be essential, that is, indispensable for growth and health, or they may be nonessential, fortuitous reminders of our geochemical origins or indicators of environmental exposure. Some of the nonessential trace elements can be beneficial to health through pharmacological action. All of the trace elements are toxic when intake is excessive.

Trace elements are those elements of the periodic table that occur in animals or humans in amounts measured in mg per kg of body weight or less. The trace elements essential for health are usually required by humans in amounts measured in mg per day; these elements include copper, iron, manganese, and zinc. The individual trace elements are discussed elsewhere in the encyclopedia. Since 1980, the term 'ultratrace element' has appeared in the nutritional literature. Ultratrace elements have been defined as those elements with estimated dietary requirements usually less than $1\,mg\,kg^{-1}$, and often less than $50\,\mu g\,kg^{-1}$ of diet for laboratory animals. For humans, the term often is used to indicate an element with an established, estimated, or suspected requirement of less than 1 mg per day or generally indicated by μg per day. At least 18 elements could be considered ultratrace elements: aluminum, arsenic, boron, bromine, cadmium, chromium, fluorine, germanium, iodine, lead, lithium, molybdenum, nickel, rubidium, selenium, silicon, tin, and vanadium. Emerging evidence indicates that silicon should be categorized as a trace element instead of an ultratrace element. However, knowledge about the practical importance or beneficial actions of silicon is in a state similar to that for most of the ultratrace elements; thus, it is considered as one of them here. Cobalt perhaps also belongs in the ultratrace category; however, it is required only in the form of vitamin B_{12} and thus is usually discussed as a vitamin.

The quality of the experimental evidence for nutritional essentiality varies widely for the ultratrace elements. The evidence for the essentiality of three elements, iodine, molybdenum and selenium, is substantial and noncontroversial; specific biochemical functions have been defined for these elements. The nutritional importance of iodine and selenium are such that they have separate entries in this encyclopedia. Molybdenum, however, is given very little nutritional attention, apparently because a deficiency of this element has not been unequivocally identified in humans other than individuals nourished by total parenteral nutrition or with genetic defects causing disturbances in metabolic pathways involving this element. Specific biochemical functions have not been defined for the other 15 ultratrace elements listed above. Thus, their essentiality is based on circumstantial evidence, which most often is that a dietary deprivation in an animal model results in a suboptimal biological function that is preventable or reversible by an intake of physiological amounts of the element in question. Often the circumstantial evidence includes an identified essential function in a lower form of life, and biochemical actions consistent with a biological role or beneficial action in humans. The circumstantial evidence for essentiality is substantial for arsenic, boron, chromium, nickel, silicon, and vanadium. The evidence for essentiality for the

other elements is generally limited to a few gross observations in one or two species by one or two research groups. However, it should be noted that two of these ultratrace elements have beneficial actions when ingested in high (pharmacological) amounts: they are fluorine, which prevents tooth caries, and lithium, which is used to treat manic-depressive disorders.

Although aluminum has a separate article, and the elements cadmium, lead, and nickel are discussed in the entry the focus in those entries is toxicity; thus, these elements will be included in the following discussion. Chromium, however, which also has a separate entry, will not be included.

Absorption, Transport, and Storage

Homeostasis (maintenance of a steady optimal concentration of an element in the body) regulation involves the processes of absorption, storage, and excretion. The relative importance of these three processes varies among the ultratrace elements. The amount absorbed from the gastrointestinal tract is often the controlling mechanism for positively charged ultratrace elements such as aluminum, nickel, and tin. With these trace elements, if the body content is low, or if intake is low, the percentage of the element absorbed from the gastrointestinal tract is increased, and vice versa. Elements that exist mainly as negatively charged ions or oxyanions, such as arsenic, boron, and fluoride, are usually absorbed quite freely and completely from the gastrointestinal tract. Excretion through the urine, bile, swear, and breath is, therefore, the major mechanism for controlling the amount of these ultratrace elements in an organism. By being stored at inactive sites or in an inactive form, some ultratrace elements are prevented from causing adverse reactions when present in high quantities. An example of this homeostatic process is the binding of cadmium by the cysteine-rich protein called metallothionein. Release of an ultratrace element from storage forms also can be important in preventing deficiency.

Absorption of ultratrace elements from the intestinal lumen can occur in three ways. These are described below.

1. Passive diffusion – passive transport driven by a difference in concentration of the element between the two sides of the luminal membrane and the mucosa. Transmembrane movement of ions occurs through pores or channels within the membrane and is an energy-independent process. A significant amount of passive transport across the intestinal mucosa may occur through a paracellular pathway, or the transport between cells across intercellular right junctions.
2. Facilitated diffusion – the transfer of an element across the membrane by carrier proteins embedded in the membrane. Facilitated transport resembles simple diffusion because it is not energy dependent and is driven by a difference in the ion concentration between two sides of a membrane. Facilitated transport occurs much more rapidly than simple diffusion and is saturable because of a finite number of carrier proteins.
3. Active transport – the accumulation within, or the extrusion from, a cell of an element in opposition to a concentration gradient. Active transport is saturable, is energy dependent and involves a carrier protein that usually is quite specific for an element. The mechanisms of absorption for the various ultratrace elements are given in **Table 1**; this table also lists the known transport and storage vehicles for these elements.

Metabolism and Excretion

Knowledge about chemical changes that must occur before excretion for most of the ultratrace elements is quite limited. Perhaps the best characterized is inorganic arsenic, which is methylated into monomethylarsonic acid and dimethylarsinic acid, and organic arsenic, which is converted into, or remains mostly as, arsenobetaine before being excreted in the urine. Other ultratrace elements that are known to be incorporated into biochemical metabolites for transport and/or excretion include aluminum bound to transferrin, cadmium incorporated into metallo-thionein, nickel as the α-2-macroglobulin nickeloplasmin or bound to albumin and L-histidine, and vanadium converted into vanadyl-transferrin and vanadyl-ferritin (see **Table 1**). A known important metabolite of molybdenum is a small nonprotein cofactor containing a pterin nucleus that is present at the active site of molybdoenzymes. More than 40% of molybdenum not attached to an enzyme in liver also exists as this cofactor bound to the mitochondrial outer membrane. This form can be transferred to an apoenzyme of xanthine oxidase or sulfite oxidase, which transforms it into an active enzyme molecule. Molecules of biological importance for the ultratrace elements are shown in **Table 2**. The ultratrace elements are excreted from the body mainly via the feces and urine. Fecal excretion of absorbed ultratrace elements

Table 1 Absorption, transport, and storage characteristics of the ultratrace elements

Element	*Major machanism(s) for homeostasis*	*Means of absorption*	*Percentage of ingested absorbed*	*Transport and storage vehicles*
Aluminum	Absorption	Uncertain; some evidence far passive diffusion through the paracellular pathway; also, evidence for active absorption through processes shared with active processes of calcium; probably occurs in proximal duodenum; citrate combined with aluminum enhances absorption	Less than 1%	Transferrin carries aluminum in plasma; bone a possible storage site
Arsenic	Urinary excretion: Inorganic arsenic as mostly dimethylarsinic acid and organic arsenic as mostly arsenobetaine	Inorganic arsenate becomes sequestered in or on mucosal tissue, then absorption involves a simple movement down a concentration gradient; organic arsenic absorbed mainly by simple diffusion through lipid regions of the intestinal boundary	Soluble inorganic forms, >90%; slightly soluble inorganic forms, 20–30%; inorganic forms with foods, 60–75%; methylated forms, 45–90%	Before excretion inorganic arsenic is converted into monomethyl arsonic acid and dimethylarsinic acid; arsenobetaine not biotransformed; arsenocholine transformed to arsenobetaine
Boron	Urinary excretion	Ingested boron is converted into $B(OH)_3$ and absorbed in this form, probably by passive diffusion	Greater than 90%	Boron transported through the body as undissociated $B(OH)_3$; bone a possible storage site
Bromine	Urinary excretion	Probably passive diffusion because no apparent saturable component	75–90%	None identified
Cadmium	Absorption	May share a common absorption mechanism with other metals (e.g., zinc) but mechanism is less efficient for cadmium	5%	Incorporated into metallothionein, which probably is both a storage and transport vehicle
Fluorine	50% daily intake excreted in urine; about 50% daily intake stored in bone and developing teeth	Absorption by passive diffusion and inversely related to pH. Significant portion absorbed as hydrogen fluoride from stomach; absorption of fluoride also occurs throughout the small intestine	76–90%	Exists as fluoride ion in plasma; hydrogen fluoride is the form in diffusion equilibration across cell membranes. Stored in bone
Germanium	Urinary excretion	Has not been conclusively determined but most likely is by passive diffusion	Greater than 90%	None identified

Continued

Table 1 Continued

Element	*Major machanism(s) for homeostasis*	*Means of absorption*	*Percentage of ingested absorbed*	*Transport and storage vehicles*
Lead	Absorption	Uncertain; thought to be by passive diffusion in small intestine, but evidence has been presented for an active transport, perhaps involving the system for calcium	Adults 5–15% Children 40–50%	Bone is a repository for lead
Lithium	Urinary excretion	Passive diffusion by paracellular transport via the tight junctions and pericellular spaces	Lithium chloride highly absorbed – greater than 90%	Bone can serve as a store for lithium
Molybdenum	Urinary and biliary excretion	Uncertain, possible that molybdate is moved both by diffusion and by active transport, but at high concentrations active transport is a small portion of flux; absorption occurs rapidly in stomach and continues throughout the small intestine	50–93%	Molybdate in blood loosely attached to erythrocytes and tends to bind α_2-macroglobulin. Liver and kidney retain highest amount of molybdate
Nickel	Both absorption and urinary excretion	Uncertain, evidence both for passive diffusion (perhaps as an amino acid or other low molecular weight complex) and for energy driven transport; occurs in the small intestine	$<$10% with food	Transported in blood principally bound to serum albumin with small amounts bound to L-histidine and α_2-macroglobulin; no organ accumulates physiological amounts of nickel
Rubidium	Excretion through kidney and intestine	Resembles potassium in its pattern of absorption; rubidium and potassium thought to share a transport system	Highly absorbed	None identified
Silicon	Both absorption and urinary excretion	Mechanisms involved in intestinal absorption have not been described	Food silicon near 50%; insoluble or poorly soluble silicates $=1\%$	Silicon in plasma believed to exist as undisassociated monomeric silicic acid
Tin	Absorption	Mechanisms involved in intestinal absorption have not been described	About 3%. Percentage increases when very low amounts are ingested	None identified. Bone might be a repository
Vanadium	Absorption	Vanadate has been suggested to be absorbed through phosphate or other anion transport systems; vanadyl has been suggested to use iron transport systems. Absorption occurs in the duodenum	$<$10%	Converted into vanadyl-transferrin and vanadyl-ferritin; whether transferrin is the transport vehicle and ferritin is the storage vehicle for vanadium remains to be determined. Bone is a respository for excess vanadium

Table 2 Excretion, retention, and possible biological roles of the ultratrace elements

Element	*Organs of high content (typical concentration)*	*Major excretory route after Ingestion*	*Molecules of biological importance*	*Possible biological role*
Aluminum	Bone (1–12 $\mu g\,g^{-1}$) Lung (35 $\mu g\,g^{-1}$)	Urine; also significant amounts in bile	Aluminum binds to proteins, nucleotides, and phospholipids; aluminum-bound transferrin apparently is a transport molecule	Enzyme activator
Arsenic	Hair (0.65 $\mu g\,g^{-1}$) Nails (0.35 $\mu g\,g^{-1}$) Skin (0.10 $\mu g\,g^{-1}$)	Urine	Methylation of Inorganic oxyarsenic anions occurs in organisms ranging from microbial to mammalian; methylated and products include arsenocholine, arsenobetaine, dimethylarsinic acid, and methylarsonic acid; arsenite methyltransferase and monomethylarsonic acid methyltransferase use *S*-adenosylmethionine for the methyl donor	Metabolism of methionine, or involved in labile methyl metabolism; regulation of gene expression
Boron	Bone (1.6 $\mu g\,g^{-1}$) Fingernails (15 $\mu g\,g^{-1}$) Hair (1 $\mu g\,g^{-1}$) Teeth (5 $\mu g\,g^{-1}$)	Urine	Boron biochemistry essentially that of boric acid, which forms ester complexes with hydroxyl groups, preferably those adjacent and *cis*, in organic compounds. Five naturally occurring boron esters (all antibiotics) synthesized by various bacteria have been characterized	Cell membrane function or stability such that it influences the response to hormone action, transmembrane signaling or transmembrane movement of regulatory cations or anlons
Bromine	Hair (3.0 $\mu g\,g^{-1}$) Liver (4.0 $\mu g\,g^{-1}$) Lung (6.0 $\mu g\,g^{-1}$) Testis (5.0 $\mu g\,g^{-1}$)	Urine	Exists as Br Ion *in vivo*, binds to proteins and amino acids	Electrolyte balance
Cadmium	Kidney (14 $\mu g\,g^{-1}$) Liver (4 $\mu g\,g^{-1}$)	Urine and gastrointestinal tract	Metallothionein, a high sulfhydryl-containing protein involved in regulating cadmium distribution	Involved in metallathionein metabolism and utilization
Fluorine	Bones (1–5 $mg\,g^{-1}$) Teeth (500 $\mu g\,g^{-1}$)	Urine	Exists as fluoride ion or hydrogen fluoride in body fluids; about 99% of body fluorine found in mineralized tissues as fluoroapatite	Role in biological mineralization
Germanium	Bone (9 $\mu g\,g^{-1}$) Liver (0.3 $\mu g\,g^{-1}$) Pancreas (0.2 $\mu g\,g^{-1}$) Testis (0.5 $\mu g\,g^{-1}$)	Urine	None identified	Role in immune function
Lead	Aorta (1–2 $\mu g\,g^{-1}$) Bone (25 $\mu g\,g^{-1}$) Kidney (1–2 $\mu g\,g^{-1}$) Liver (1–2 $\mu g\,g^{-1}$)	Urine; also significant amounts in bile	Plasma lead mostly bound to albumin; blood lead binds mostly to hemoglobin but some binds a low molecular weight protein in arythrocytes	Facilitates iron absorption and/or utilization

Continued

Table 2 Continued

Element	*Organs of high content (typical concentration)*	*Major excretory route after Ingestion*	*Molecules of biological importance*	*Possible biological role*
Lithium	Adrenal gland ($60\,ng\,g^{-1}$) Bone ($100\,ng\,g^{-1}$) Lymph nodes ($200\,ng\,g^{-1}$) Pituitary gland ($135\,ng\,g^{-1}$)	Urine	None Identified	Regulation of some endocrine function
Molybdenum	Kidney ($0.4\,\mu g\,g^{-1}$) Liver ($0.6\,\mu g\,g^{-1}$)	Urine; also significant amounts in bile	Molybdoenzymes of aldehyde oxidase, xanthine oxidase/ dehydrogenase and sulfite oxidase in which molyldenum exists as a small nonprotein factor containing a pterin nucleus; molybdate ion (MoO_4^{2+}), the form that exists in blood and urine	Molybdoenzymes oxidize and detoxify various pyrimidines, purines, and pteridines; catalyze the transformation of hypoxanthine to xanthine and xanthine to uric acid; and catalyze the conversion of sulfite to sulfate
Nickel	Adrenal glands ($25\,ng\,g^{-1}$) Bone ($33\,ng\,g^{-1}$) Kidney ($10\,ng\,g^{-1}$) Thyroid ($30\,ng\,g^{-1}$)	Urine as low molecular weight complexes	Binding of Ni^{2-} by various ligands including amino acids (especially histidine and cysteine), proteins (especially albumin), and a macroglobulin called nickeloplasmin important in transport and excretion. Ni^{2+} component of urease; Ni^{3-} essential for enzymatic hydrogenation, desulfurization, and carboxylation reactions in mostly anaerobic microorganisms	Cofactor or structural component in specific metalloenzymes; role in a metabolic pathway involving vitamin B_{12} and folic acid. Role similar to potassium; neurophysiological function
Rubidium	Brain ($4\,\mu g\,g^{-1}$) Kidney ($5\,\mu g\,g^{-1}$) Liver ($6.5\,\mu g\,g^{-1}$) Testis ($20\,\mu g\,g^{-1}$)	Urine: also significant amounts excreted through intestinal tract	None identified	Role similar to potassium; neurophysiological function
Silicon	Aorta ($16\,\mu g\,g^{-1}$) Bone ($18\,\mu g\,g^{-1}$) Skin ($4\,\mu g\,g^{-1}$) Tendon ($12\,\mu g\,g^{-1}$)	Urine	Silicic acid ($SiOH_4$) is the form believed to exist in plasma; magnesium orthosilicate is probably the form in urine. The bound form of silicon has never been rigorously identified	Structural role in some mucopolysaccharide or collagen; role in the initiation of calcification and in collagen formation
Tin	Bone ($0.8\,\mu g\,g^{-1}$) Kidney ($0.2\,\mu g\,g^{-1}$) Liver ($0.4\,\mu g\,g^{-1}$)	Urine; also significant amounts in bile	Sn^{2+} is absorbed and excreted more readily than Sn^{4+}	Role in some redox reaction
Vanadium	Bone ($120\,ng\,g^{-1}$) Kidney ($120\,ng\,g^{-1}$) Liver ($120\,ng\,g^{-1}$) Spleen ($120\,ng\,g^{-1}$) Testis ($200\,ng\,g^{-1}$)	Urine; also significant amount in bile	Vanadyl (VO^{2+}), vanadate ($H_2VO_4^-$ or VO_3^-) and peroxovanadyl [V–OO]; VO^{2+} complexes with proteins, especially those associated with iron (e.g., transferrin, hemoglobin)	Lower forms of life have haloperoxidases that require vanadium for activity; a similar role might exist in higher forms of life

None of the suggested biological functions or roles of any of the ultratrace elements have been conclusively or unequivocally identified in higher forms of life except for those of molybdenum.

usually results from biliary excretion, but may be of nonbiliary origin (e.g., through the pancreas or intestine). Ultratrace elements may also be excreted through sweat and breath. Ultratrace elements also are removed from the body through the loss of blood (e.g., menses), skin, hair, semen, saliva, and nails. **Table 2** gives the major routes of excretion for the ultratrace elements.

Requirements and High Intakes

As already mentioned, the ultratrace elements other than selenium and iodine are a disparate group in terms of their possible requirement or nutritional importance for human health and well-being. Although molybdenum has known essential functions, it has no unequivocally identified practical nutritional importance. The other 14 ultratrace elements discussed here have been suggested to be essential based on circumstantial evidence. This evidence is presented below along with some indication of possible requirement (extrapolated from the deficient animal intakes shown in **Table 3**), and some indication as to what constitutes a high intake.

Aluminum

A dietary deficiency of aluminum in goats reportedly results in increased abortions, depressed growth, incoordination and weakness in hind legs, and decreased life expectancy. Aluminum deficiency has also been reported to depress growth in chicks. Other biochemical actions that suggest aluminum could possibly act in an essential role include the *in vitro* findings that it activates the enzyme adenylate cyclase, enhances calmodulin activity, stimulates DNA synthesis in cell cultures, and stimulates osteoblasts to form bone through activating a purative G, protein-coupled cation sensing system.

If humans have a requirement for aluminum, for which there is currently no evidence, it probably is much less than $1.0\,mg\,day^{-1}$. Aluminum toxicity apparently is not a concern for healthy individuals. Cooking foods in aluminum cook-ware does not lead to detrimental intakes of aluminum. High dietary ingestion of aluminum probably is not a cause of Alzheimer's disease. However, high intakes of aluminum through such sources as buffered analgesics and antacids by susceptible individuals (e.g., those with impaired kidney function including the elderly and low-birthweight infants) may lead to pathological consequences and thus should be avoided. For most healthy individuals, an aluminum intake of $125\,mg\,day^{-1}$ should not lead to toxicological consequences.

Arsenic

Arsenic deprivation has been induced in chickens, hamsters, goats, pigs, and rats. In the goat, pig, and rat, the most consistent signs of deprivation were depressed growth and abnormal reproduction characterized by impaired fertility and elevated perinatal mortality. Other notable signs of deprivation in goats were depressed serum triacylglycerol concentrations and death during lactation. Myocardial damage was also present in lactating goats. Other signs of arsenic deprivation have been reported, including changes in mineral concentrations in various organs. However, listing all signs reported to be caused by arsenic deficiency may be misleading because studies with chicks, rats, and hamsters have revealed that the nature and severity of the signs are affected by a number of dietary and other factors. For example, female rats fed a diet that is conducive to kidney calcification have more severe calcification when dietary arsenic is low; kidney iron was also elevated. Male rats fed the same diet do not show these changes.

Other factors that affect the response to arsenic deprivation include methionine, arginine, choline, taurine, and guanidoacetic acid. In other words, the signs of arsenic deprivation were changed and generally enhanced by nutritional stressors that affected sulfur amino acid or labile methyl-group metabolism; this suggests that arsenic has a biochemical function that affects these substances. Further evidence for this suggestion is the finding that arsenic deprivation slightly increases liver S-adenosylhomo-cysteine (SAH) and decreases liver S-adenosyl-methionine (SAM) concentrations in animal models, thus resulting in a decreased SAM/SAH ratio; SAM and SAH are involved in methyl transfer. Additionally, arsenite can induce the isolated cell production of certain proteins known as heat shock proteins. The control of production of these proteins in response to arsenite apparently is at the transcriptional level, and involves changes in the methylation of core histones. It also has been shown that arsenic can increase the methylation of the *p53* promoter, or DNA, in human lung cells.

It has been suggested, based upon animal data, that a possible arsenic requirement for humans eating 8.37 MJ (2000 kcal) would be $12–25\,\mu g\,day^{-1}$; this is near the typical daily intake shown in **Table 3**. Because of mechanisms for the homeostatic regulation of arsenic (including methylation,

Table 3 Human body content, and deficient, typical, and rich sources of intakes of ultratrace elements

Element	*Apparent deficient intake (species)*	*Human body content*	*Typical human daily dietary intake*	*Rich sources*
Aluminum	160 μg kg^{-1} (goat)	30–50 mg	2–10 mg	Baked goods prepared with chemical leavening agents (e.g., baking powder), processed cheese, grains, vegetables, herbs, tea, antacids, buffered analgesics
Arsenic	<25 μg kg^{-1} (chicks) <35 μg kg^{-1} (goat) <15 μg kg^{-1} (hamster) <30 μg kg^{-1} (rat)	1–2 mg	12–60 μg	Shellfish, fish, grain, cereal products
Boron	<0.3 mg kg^{-1} (chick) 0.25–0.35 mg per day (human) <0.3 mg kg^{-1} (rat)	10–20 mg	0.5–3.5 mg	Food and drink of plant origin, especially noncitrus fruits, leafy vegetables, nuts, pulses, avocados, legumes, wine, cider, beer, peanut butter
Bromine	0.8 mg kg^{-1} (goat)	200–350 mg	2–8 mg	Grain, nuts, fish
Cadmium	<5 μg kg^{-1} (goat) <4 μg kg^{-1} (rat)	5–20 mg	10–20 μg	Shellfish, grains – especially those grown on high-cadmium soils, leafy vegetables
Fluorine	<0.3 mg kg^{-1} (goat) <0.45 mg kg^{-1} (rat)	3 g	Fluoridated areas, 1–3 mg Nonfluoridated areas, 0.3–0.6 mg	Fish, tea, fluoridated water
Germanium	0.7 mg kg^{-1} (rat)	3 mg	0.4–3.4 mg	Wheat bran, vegetables, leguminous seeds
Lead	<32 μg kg^{-1} (pig) <45 μg kg^{-1} (rat)	Children less than age 10 years, 2 mg Adults, 120 mg	15–100 μg	Seafood, plant foodstuffs grown under high-lead conditions
Lithium	<1.5 mg kg^{-1} (goat) <15 μg kg^{-1} (rat)	350 μg	200–600 μg	Eggs, meat, processed meat, fish, milk, milk products, potatoes, vegetables (content varies with geological origin)
Molybdenum	< 25 μg kg^{-1} (goat) <25 μg day^{-1} (human) <30 μg kg^{-1} (rat)	10 mg	50–100 μg	Milk and milk products, dried legumes, pulses, organ meats (liver and kidney), cereals, and baked goods
Nickel	<100 μg kg^{-1} (goat) <20 μg kg^{-1} (rat)	1–2 mg	70–260 μg	Chocolate, nuts, dried beans and peas, grains
Rubidium	180 μg kg^{-1} (goat)	360 mg	1–5 mg	Coffee, black tea, fruits and vegetables (especially asparague), poultry, fish
Silicon	<2.0 mg kg^{-1} (chick) <4.5 mg kg^{-1} (rat)	2–3 g	20–50 mg	Unrefined grains of high fiber content, cereal products, beer, coffee
Tin	<20 μg kg^{-1} (rat)	7–14 mg	1–40 mg	Canned foods
Vanadium	<10 μg kg^{-1} (goat)	100 μg	10–30 μg	Shellfish, mushrooms, parsley, dill seed, black pepper, some prepared foods, grains, beer, wine

then excretion in urine), its toxicity through oral intake is relatively low; it is actually less toxic than selenium, an ultratrace element with a well-established nutritional value. Toxic quantities of inorganic arsenic generally are reported in milligrams. For example, reported estimated fatal acute doses of arsenic for humans range from 70 to 300 mg or about 1.0 to 4.0 mg per kg body weight.

Some forms of organic arsenic are virtually nontoxic; a 10 g per kg body weight dose of arsenobetaine depressed spontaneous motility and respiration in male mice, but these symptoms disappeared within 1 h. Results of numerous epidemiological studies have suggested an association between chronic overexposure to arsenic and the incidence of some forms of cancer; however, the role of arsenic in carcinogenesis remains controversial. Arsenic does not seem to act as a primary carcinogen, and is either an inactive or extremely weak mitogen. In the USA, a standard known as a reference dose (RfD; lifetime exposure that is unlikely to cause adverse health effects) of 0.3 μg per kg body weight per day, or 21 μg day^{-1} for a 70 kg human, has been suggested for inorganic arsenic. Because of safety factors in the determination, the RfD for arsenic conflicts with the possible arsenic requirement; this conflict is similar to that for some other mineral elements including zinc. These conflicts are currently being addressed by nutritionists and toxicologists.

Boron

Listing the signs of boron deficiency for animal models is difficult because most studies have used stressors to enhance the response to changes in dietary boron. Thus, the response to boron deprivation varied as the diet changed in its content of nutrients such as calcium, phosphorus, magnesium, potassium, and vitamin D. Although the nature and severity of the changes varied with dietary composition, many of the findings indicated that boron deprivation impairs calcium metabolism, brain function, and energy metabolism. Studies also suggest that boron deprivation impairs immune function and exacerbates adjuvant-induced arthritis in rats. Feeding low boron to humans (<0.3 mg day^{-1}) altered the metabolism of macrominerals, electrolytes, and nitrogen, as well as oxidative metabolism, and produces changes in erythropoiesis and hematopoiesis. Boron deprivation also altered electroencephalograms to suggest depressed behavioral activation and mental alertness, depressed psychomotor skills and cognitive processes of attention and memory, and enhanced some effects of estrogen therapy such as increases in concentrations of serum 17β-estradiol and plasma copper. Other findings suggest that boron may have an essential function. *In vitro* it competitively inhibits oxidoreductase enzymes, which require pyridine or flavin nucleotides, and enzymes such as serine proteases, which form transition state analogs with boronic acid or borate derivatives. Boron has an essential function in plants, in which it influences redox actions involved in cellular membrane transport. This latter finding supports the hypothesis that boron has a role in cell membrane function or stability such that it influences the response to hormone action, transmembrane signaling, or transmembrane movement of regulatory cations or anions. Another finding in support of this hypothesis is that boron influences the transport of extracellular calcium into and the release of intracellular calcium in rat platelets activated by thrombin.

An analysis of both human and animal data has resulted in the suggestion by a World Health Organization (WHO) publication that an acceptable safe range of population mean intakes of boron for adults could well be 1.0–13 mg day^{-1}. In other words, 1.0 mg probably covers any requirement and 13 mg will not lead to any toxicological consequences. However, the US and Canada concluded in 2002 that there was still insufficient evidence to establish a clear biological function for boron in humans, so no recommended dietary intake was set for those countries. Boron has a low order of toxicity when administered orally. Toxicity signs in animals generally occur only after dietary boron exceeds 100 μg g^{-1}. The low order of toxicity of boron for humans is shown by the use of boron as a food preservative between 1870 and 1920 without apparent harm. It was reported in 1904 that when doses equivalent to more than 0.5 g of boric acid were consumed daily, disturbances in appetite, digestion, and health occurred. It was concluded in this report that this quantity of boron per day was too much for an average person to receive regularly. The upper limit (UL) for the US and Canada has been set at 20 mg day^{-1} based on extrapolation from animal studies.

Bromine

It has been reported that a dietary deficiency of bromide results in depression of growth, fertility, hematocrit, hemoglobin, and life expectancy, and increases in milk fat and spontaneous abortions in goats. Other biological actions that suggest bromine could possibly act in an essential role include the findings that bromide alleviates growth retardation caused by hyperthyroidism in mice and chicks, and insomnia exhibited by many hemodialysis patients has been associated with bromide deficit.

If humans have a requirement for bromide, which has not yet been shown to be the case, based on deficient intakes for animals it is probably no more

than 1.0 mg day^{-1}. Bromine ingested as the bromide ion has a low order of toxicity; thus bromine is not of toxicological concern in nutrition.

Cadmium

Deficiency of cadmium reportedly depresses growth of rats and goats. Other *in vitro* biochemical actions that suggest cadmium could possibly act as an essential element include the finding that it has transforming growth factor activity and stimulates growth of cells in soft agar.

If humans have a requirement for cadmium, which is still uncertain, based on deficient intakes for animals it is probably less than 5 μg day^{-1}. Although cadmium may be an essential element at these extremely low amounts, it is of more concern because of its toxicological properties. Cadmium has a long half-life in the body and thus high intakes can lead to accumulation, resulting in damage to some organs, especially the kidney. The toxicological aspects of cadmium have been discussed earlier (*See:* xx).

Fluorine

Reported unequivocal or specific signs of fluoride deficiency are almost nonexistent. A study with goats indicated that a fluoride deficiency decreases life expectancy and caused pathological hisrology in the kidney and endocrine organs. Most of the evidence accepted as showing a need for fluoride comes from studies in which it was orally administered in pharmacological doses. Pharmacological doses of fluoride have been shown to prevent tooth caries, improve fertility, hematopoiesis and growth in iron-deficient mice and rats, prevent phosphorus-induced nephro-calcinosis, and perhaps prevent bone loss leading to osteoporosis.

Although fluoride is not generally considered an essential element in the classical sense for humans, it still is considered a beneficial element. Because of this, in the US-Canada, the AI has been set, on the basis of reducing dental caries without adverse effects, at: 0.01 mg day^{-1} for infants 0–6 months; 0.5 mg for 6–12 months; 0.7 mg for 1–3 years; 1 mg for 4–8 years; 2 mg for 9–13 years; 3 mg for 14–18 years; 3 mg for women and 4 mg for men. These intakes provide amounts of fluoride that will give protection against dental caries and generally not result in any consequential mottling of teeth; they should not be considered intakes that are needed to prevent a nutritional deficiency of fluoride. Chronic fluoride toxicity through excessive intake mainly through water supplies and industrial exposure has been reported in many parts of the world. Chronic toxicity resulting from the ingestion of water and food providing in excess of 2.0 mg day^{-1} is manifested by dental fluorosis or mottled enamel ranging from barely discernible with intakes not much above 2.0 mg day^{-1} to stained and pitted enamel with much higher amounts. Crippling skeletal fluorosis apparently occurs in people who ingest 10–25 mg day^{-1} for 7–20 years. The UL (mg per day) is 0.7 mg for 0–6 months, 0.9 mg for 7–12 months, 1.3 mg for 1–3 years, and 2.2 mg for 4–8 years, and 10 mg for all older age groups including pregnant and lactating women.

Germanium

A low germanium intake has been found to alter bone and liver mineral composition and decrease tibial DNA in rats. Germanium also reverses changes in rats caused by silicon deprivation, and is touted as having anticancer properties because some organic complexes of germanium can inhibit tumor formation in animal models.

If humans have a requirement for germanium, based on animal deprivation studies, it is probably less than 0.5 mg day^{-1}. The toxicity of germanium depends upon its form. Some organic forms of germanium are less toxic than inorganic forms. Inorganic germanium toxicity results in kidney damage. Some individuals consuming high amounts of organic germanium supplements contaminated with inorganic germanium have died from kidney failure. Although germanium has long been believed to have a low order of toxicity because of its diffusible state and rapid elimination from the body, until more knowledge is obtained about the intakes at which germanium becomes toxic, they probably should not greatly exceed those found in a typical diet. An intake of no more than 5.0 mg day^{-1} would meet any possible need for germanium and most likely will be below the level found to have toxicological consequences.

Lead

A large number of findings have come from one source that suggests that a low dietary intake of lead is disadvantageous to pigs and rats. Apparent deficiency signs found include: depressed growth; anemia; elevated serum cholesterol, phospholipids and bile acids; disturbed iron metabolism; decreased liver glucose, triacylglycerols, LDL-cholesterol and phospholipids; increased liver cholesterol; and altered blood and liver enzymes. A beneficial action of lead (2 μg g^{-1} versus 30 ng g^{-1} diet) is that it alleviates iron deficiency signs in young rats.

If humans have a requirement for lead, which has not yet been demonstrated to be the case, it is probably less than 30 μg day^{-1} based on animal deprivation studies. Although lead may have beneficial effects at low intakes, lead toxicity is of more concern than lead deficiency. Lead is considered one of the major environmental pollutants because of the past use of lead-based paints and the combustion of fuels containing lead additives. The toxicological aspects of lead are discussed elsewhere (*See:* xx).

Lithium

Lithium deficiency reportedly results in depressed fertility, birthweight, and life span, and altered activity of liver and blood enzymes in goats. In rats, lithium deficiency apparently depresses fertility, birthweight, litter size, and weaning weight. Other *in vitro* biochemical actions suggesting that lithium could possibly act as an essential element include the stimulation of growth of some cultured cells, and having insulinomimetic action. Lithium is best known for its pharmacological properties; it is used to treat manic-depressive psychosis. Its ability to affect mental function perhaps explains the report that incidence of violent crimes is lower in areas with high-lithium drinking water.

If humans have a requirement for lithium, based on animal deprivation studies it is probably less than 25 μg day^{-1}, which is much less than the usual dietary intake (see **Table 3**). Lithium is not a particularly toxic element, but the principal disadvantage in the use of lithium for psychiatric disorders is the narrow safety margin between therapeutic and toxic doses. About 500 mg lithium per day is needed to raise serum concentrations to be effective in these disorders; this is close to the concentration where mild toxicity signs of gastrointestinal disturbances, muscular weakness, tremor, drowsiness, and a dazed feeling begin to appear. Severe toxicity results in coma, muscle tremor, convulsions, and even death.

Molybdenum

The evidence for the essentiality of molybdenum is substantial and conclusive. Molybdenum functions as a cofactor in enzymes that catalyze the hydroxylation of various substrates. Aldehyde oxidase oxidizes and detoxifies various pyrimidines, purines, pteridines, and related compounds. Xanthine oxidase/dehydrogenase catalyzes the transformation of hypoxanthine to xanthine, and xanthine to uric acid. Sulfite oxidase catalyzes the transformation of sulfite to sulfate. Attempts to produce molybdenum deficiency signs in rats, chickens, and humans have resulted in only limited success, and no success in healthy humans. Deficiency signs in animals are best obtained when the diet is supplemented with massive amounts of tungsten, an antagonist of molybdenum metabolism. Nonetheless, reported deficiency signs for goats and pigs are depressed food consumption and growth, impaired reproduction characterized by increased mortality in both mothers and offspring, and elevated copper concentrations in liver and brain. A molybdenum-responsive syndrome found in hatching chicks is characterized by a high incidence of late embryonic mortality, mandibular distortion, anophthalmia, and defects in leg bone and feather development. The incidence of this syndrome was particularly high in commercial flocks reared on diets containing high concentrations of copper, another molybdenum metabolism antagonist.

Examples of nutritional standards that have been set for molybdenum are the current US-Canada recommendations, which are the following: Adequate Intake for infants aged 0–0.5 years, 2 μg and aged 0.5–1 years, 3 μg; RDA for children 1–3 years, 17 μg; 4–8 years, 22 μg; 9–13 years, 34 μg; 14–18 years, 43 μg; women from 19–>70 years, 34 μg; and men aged 19–>70 years, 45 μg. The recommended intake is 50 μg day^{-1} in pregnancy and lactation. These values were set using balance data in adults with extrapolation to the other groups. Usual dietary intakes are substantially higher than these recommendations. Large oral doses are necessary to overcome the homeostatic control of molybdenum; thus, it is a relatively nontoxic nutrient. The UL for children 1–3 years is 300 μg, for 4–8 years, 600 μg, and 9–13 years, 1100 μg. For adolescents the UL is 1700 μg, and for adults, 2000 μg, including pregnant and lactating women, based on doses that caused reproductive damage in animals.

Nickel

Based on recent studies with rats and goats, nickel deprivation depresses growth, reproductive performance and plasma glucose, and alters the distribution of other elements in the body, including calcium, iron, and zinc. As with other ultratrace elements, the nature and severity of signs of nickel deprivation are affected by diet composition. For example, vitamin B_{12} status affects signs of nickel deprivation in rats, and the effects suggest that vitamin B_{12} must be present for optimal nickel function. The nickel function also may involve folic acid because an interaction between these two affected the vitamin B_{12} and folic acid-dependent pathway of methionine synthesis from homocysteine. Nickel might function as a cofactor or structural component in specific metalloenzymes in higher organisms

because such enzymes have been identified in bacteria, fungi, plants, and invertebrates. These nickel-containing enzymes include urease, hydrogenase, methylcoenzyme M reductase, and carbon monoxide dehydrogenase. Moreover, nickel can activate numerous enzymes *in vitro*.

Based on a lack of human studies, no recommended intake levels have been set for humans. Life-threatening toxicity of nickel through oral intake is unlikely. Because of excellent homeostatic regulation, nickel salts exert their toxic action mainly by gastrointestinal irritation and not by inherent toxicity. Based on extrapolation from animal studies, the UL has been set for the US and Canada at the following doses of soluble nickel salts: 1–3 years, 0.2 mg; 4–8 years, 0.3 mg; 9–13 years, 0.6 mg; and all adolescents and adults, 1 mg.

Rubidium

Rubidium deficiency in goats reportedly results in depressed food intake and life expectancy, and increased spontaneous abortions. If rubidium is required by humans, the requirement probably would be no more than a few hundred micrograms per day, based on animal data. Rubidium is a relatively nontoxic element and thus is not of toxicological concern from the nutritional point of view.

Silicon

Most of the signs of silicon deficiency in chickens and rats indicate aberrant metabolism of connective tissue and bone. For example, chicks fed a silicon-deficient diet exhibit structural abnormalities of the skull, depressed collagen content in bone, and long-bone abnormalities characterized by small, poorly formed joints and defective endochondral bone growth. Silicon deprivation can affect the response to other dietary manipulations. For example, rats fed a diet low in calcium and high in aluminum accumulated high amounts of aluminum in the brain; silicon supplements prevented the accumulation. Also, high dietary aluminum depressed brain zinc concentrations in thyroidectomized rats fed low dietary silicon; silicon supplements prevented the depression. This effect was not seen in nonthyroidectomized rats. Other biochemical actions suggest that silicon is an essential element. Silicon is consistently found in collagen, and in bone tissue culture has been found to be needed for maximal bone prolylhydroxylase activity. Silicon deficiency decreases ornithine aminotransferase, an enzyme in the collagen formation pathway, in rats. Finally, silicon is essential for some lower forms of life in which silica serves a structural role and possibly affects gene expression.

Much of the silicon found in most diets probably occurs as aluminosilicates and silica from which silicon is not readily available. Owing to lack of evidence for a biological role for silicon in humans, no recommended intakes have been set. Silicon is essentially nontoxic when taken orally. Magnesium trisilicate, an over-the-counter antacid, has been used by humans for more than 40 years without obvious deleterious effects. Other silicates are food additives used as anticaking or antifoaming agents.

Tin

A dietary deficiency of tin has been reported to depress growth, response to sound, and feed efficiency, alter the mineral composition of several organs, and cause hair loss in rats. Additionally, tin has been shown to influence heme oxygenase activity and has been associated with thymus immune and homeostatic functions.

Owing to lack of data no recommended intakes have been set for tin. Inorganic tin is relatively nontoxic. However, the routine consumption of foods packed in unlacquered tin-plated cans may result in excessive exposure to tin, which could adversely affect the metabolism of other essential trace elements including zinc and copper. Because $50\,\mathrm{mg\,day^{-1}}$ of tin was found to affect zinc and copper metabolism, routine intakes near this amount probably should be avoided.

Vanadium

Vanadium-deprived goats were found to exhibit an increased abortion rate and depressed milk production. About 40% of kids from vanadium-deprived goats died between days 7 and 91 of life with some deaths preceded by convulsions; only 8% of kids from vanadium-supplemented goats died during the same time. Also, skeletal deformations were seen in the forelegs, and forefoot tarsal joints were thickened. In rats, vanadium deprivation increases thyroid weight and decreases growth. Other biochemical actions support the suggestion that vanadium could possibly act in an essential role. *In vitro* studies with cells and pharmacological studies with animals have shown that vanadium has: insulin-mimetic properties; numerous stimulatory effects on cell proliferation and differentiation; effects on cell phosphorylation-dephosphorylation; effects on glucose and ion transport across the plasma membrane: and effects on oxidation-reduction processes. Some algae, lichens, fungi, and bacteria contain enzymes that require vanadium for activity. The enzymes include nitrogenase in bacteria, and bromoperoxidase, iodoperoxidase, and chloroperoxidase in algae, lichens, and fungi, respectively. The

haloperoxidases, catalyze the oxidation of halide ions by hydrogen peroxide, thus facilitating the formation of a carbon–halogen bond. The best known haloperoxidase in animals is thyroid peroxidase. Vanadium deprivation in rats affects the response of thyroid peroxidase to changing dietary iodine concentrations. Since a functional role for vanadium has not been determined in humans no recommended intakes have been set.

Vanadium can be a relatively toxic element. Green tongue, cramps and diarrhea, and neurological effects have occurred in humans ingesting vanadium salts. Based on renal damage in animals, the UL for adults is 1.8 mg vanadium salts per day, with insufficient data to set a UL for other age groups.

Dietary Sources

The requirements for the ultratrace elements will be met if a person consumes a diet based on the dietary guidelines recommended by. For some areas of the world, especially in developing countries where traditional, monotonous diets are based primarily on a cereal (particularly rice) or tuber staple, the intake of several ultratrace elements (e.g., boron, molybdenum) could possibly be low. Reported typical dietary intakes (mostly for industrialized countries) and rich sources of the ultratrace elements are shown in **Table 3**.

See also: **Aluminum**. **Chromium**. **Cobalamins**. **Cofactors**: Organic. **Dental Disease**. **Food Safety**: Heavy Metals. **Iodine**: Physiology, Dietary Sources and Requirements. **Selenium**.

Further Reading

Chappell WR, Abernathy CO, and Cothern CR (eds.) (1994) *Arsenic Exposure and Health*. Northwood: Science and Technology Letters.

Ciba Foundation (1986) *Silicon Biochemistry*. Chichester: John Wiley & Sons.

Editorial (1994) Health effects of boron. *Environmental Health Perspectives Supplement* **102**(supplement 7).

FAO/WHO (1996) *Trace Elements in Human Nutrition and Health*. Geneva: World Health Organization.

Frieden E (ed.) (1984) *Biochemistry of the Essential Ultratrace Elements*. New York: Plenum.

Institute of Medicine (2002) *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc*. Washington DC: National Academy Press.

Mertz W (ed.) (1986, 1987) *Trace Elements in Human and Animal Nutrition*, vols 1 and 2. Orlando and San Diego: Academic Press.

Nielsen FH (1986) Other elements: Sb, Bs, B, Br, Cs, Ge, Rb, Ag, Sr, Sn, Ti, Zr, Be, Bi, Ga, Au, In, Nb, Sc, Te, Tl, W. In: Mertz W (ed.) *Trace Elements in Human and Animal Nutrition*, vol. 2, pp. 415–463. Orlando: Academic Press.

Nielsen FH (1994) Ultratrace minerals. In: Shils ME, Olson JA, and Shike M (eds.) *Modern Nutrition in Health and Disease*, 8th edn, vol. 1, pp. 269–286. Philadelphia: Lea & Febiger. (9th edition with updated chapter on ultratrace minerals in progress.)

Nielsen FH (1996) Other trace elements. In: Ziegler EE and Filer LJ Jr (eds.) *Present Knowledge in Nutrition*, 7th edn, pp. 353–376. Washington DC: ILSI Press.

Sigel H and Sigel A (eds.) (1995) *Metal Ions in Biological Systems*, vol. 23, Nickel and its Role in Biology. New York: Marcel Dekker.

Sigel H and Sigel A (eds.) (1995) *Metal Ions in Biological Systems*, vol. 31, Vanadium and its Role in Life. New York: Marcel Dekker.

UNITED NATIONS CHILDREN'S FUND

J P Greaves, London, UK
R Shrimpton, Institute of Child Health, London, UK

The United Nations Children's Fund (UNICEF) is that part of the United Nations (UN) system that has a special mandate to improve the welfare of children and women. Beginning with a restricted mandate for a limited period, the organization now has a permanent concern for the whole child (rights to health, education, protection, etc.) throughout the world. This article describes the history of the organization, with special reference to how its approach to nutrition has evolved over the years. The article also describes its role, structure, and funding.

History

UNICEF was established for a limited period by the General Assembly of the UN in December 1946 as the United Nations International Children's Emergency Fund to provide assistance to children in Europe and China suffering from the devastations of World War II. In 1950, its life was extended for an additional 3 years, with a mandate covering all developing countries, and in 1953 the General

Assembly gave it permanent status in recognition of the chronic and continuing needs of children. Its name was changed to that by which it is known today: 'International' was dropped as being redundant and 'Emergency' as too restrictive, but the familiar acronym was retained.

Early Days

In its early days, UNICEF gave priority to relief and rehabilitation and the provision of material assistance, laying the foundations for its reputation as an efficient supply agency. UNICEF needed people 'in the field' to ensure supplies reached those for whom they were intended; this was the origin of UNICEF's network of country offices. The 'specialized agencies'—the Food and Agriculture Organization of the United Nation (FAO) and the World Health Organization (WHO)—provided technical advice from a regional or headquarters base.

In the year after it was created, UNICEF requested FAO and the WHO Interim Commission (WHO itself not formally having been established) for technical advice on child nutrition. A joint committee of specialists in nutrition and pediatrics was established, and its advisory report provided basic information on such matters as the energy and protein needs of different age groups, the importance of breast feeding, the negative effects of infections on nutritional status, and the consequences of deficiencies of micronutrients such as iodine, iron, and vitamin A. These remain major issues in public nutrition, although understanding has increased of how to address them successfully. UNICEF has developed its own technical capacity and has become recognized as primarily a developmental rather than a humanitarian organization, dealing in ideas as well as supplies.

During its early years, UNICEF's work on nutrition concentrated on support to direct child feeding through schools, especially through provision of dried skim milk, which was in plentiful supply. This was sometimes coupled with the development of local dairy industries. Fortification of milk powder with vitamin A was an issue—and sometimes still is today. Later, following the common scientific perception of the day, and guided by the Protein Advisory Group (PAG) of the UN system, UNICEF shifted attention from school to preschool feeding and to commercial production of low-cost, high-protein supplementary foods for children. However, costs could not be kept so low that those for whom they were chiefly designed could afford them, and although a frequent condition for UNICEF provision of plant was that the government should subsidize the product for the poor, this commitment could not be maintained.

Applied Nutrition and Nutrition Planning

During the 1960s, UNICEF provided support, with FAO technical inputs, to so-called applied nutrition programs (ANPs), which were essentially educational programs at the local level encouraging the production and consumption of 'nutritious foods.' Emphasis was placed on horticulture and raising of small animals; kitchen, school, and community gardens; and use of appropriate technology to store, preserve, and prepare foods and conserve fuel. Nutrition education was always a component and such programs were often linked to provision of health services, such as immunization, potable water, and environmental sanitation. Although some of these programs attempted to be responsive to local needs and to mobilize local resources, insufficient attention was given in practice to these critical matters, and ANPs tended to be regarded as of peripheral significance. Nevertheless, in some countries they were acknowledged to have laid the groundwork for national nutrition policies, to have provided the first practical experience of intersectoral cooperation at various levels, and to have increased recognition of the need to involve local people in community programs.

The 1970s saw attempts to introduce nutrition strategies into regional and national development planning. In 1971, the UNICEF executive board declared that "the best action was through the establishment of national food and nutrition policies." Governments were asked to consider specific measures designed to improve nutritional conditions of mothers and children of low-income families. However, these were primarily food based, and the insight of the Mixed Committee of the League of Nations, with its call in 1937 for multisectoral approaches to problems of hunger and malnutrition with a "marriage of agriculture and medicine," seemed to have been lost. Many found it difficult to accept that malnutrition could result either from infection or from an inadequate diet, or often from some combination of the two. Meanwhile, UNICEF continued to try to deal with poor child nutrition through interventions of various complementary kinds, including public health, small-scale agriculture, and the promotion of women's groups.

During this period, experience of program delivery and coverage and better understanding of the development process led to the formulation of concepts that have had a profound effect on the operational approach to improving nutrition. One of these

was the basic services concept, first presented by UNICEF to the World Food Conference in 1974 but accepted as policy by the executive board and endorsed by the General Assembly of the UN in 1976. The essence of this holistic approach lies in promoting and responding to community initiatives, the involvement of local community or village-level workers, appropriate technology, and effective support, technical supervision, and referral services. The objective is to foster self-reliance. Similar principles are behind the primary health care (PHC) approach, endorsed by the WHO/UNICEF jointly sponsored international conference on PHC in Alma Ata in 1978. The PHC strategy involves shifting the focus of attention to the primary level of health care, involving the community in its establishment and management, and recognizing the many different sectoral activities that contribute to improved health, among which the Declaration of Alma Ata included the "promotion of food supply and proper nutrition." Experience of these approaches has demonstrated the difficulty of generating community participation and the generally inadequate attention given to training and support of village-level workers. More fundamentally, problems have been encountered in the comprehension and acceptance of the philosophy of the approaches, indicating the importance of effective communication. Often, nongovernmental organizations (NGOs) are able to overcome these handicaps: The challenge then is to expand their local successes into the national scene.

Child Survival and Development

In 1982, UNICEF launched a new initiative, known as the Child Survival and Development Revolution, which focused on the young child and emphasized low-cost actions appropriate for the family but suitable for national application. This was characterized by some as selective PHC. Key components (all reflecting the synergism between nutrition and infection) were the protection and promotion of breastfeeding, immunization against the six EPI (WHO's Expanded Programme of Immunization) diseases, oral rehydration therapy in the control of diarrhea, and regular growth monitoring of the child primarily to help the mother promote optimal growth. Sometimes referred to as the GOBI strategy, it relied heavily on mass communication and social mobilization. The component that received most support was immunization, and in 1991 the UN recognized the attainment of the 1990 goal toward universal child immunization (UCI): 80% of all infants immunized against tuberculosis, polio, diphtheria, pertussis, tetanus, and measles.

UNICEF had become concerned by the erosion of breast-feeding consequent on urbanization and encouraged by aggressive marketing of breast milk substitutes, and it was actively involved with WHO and NGOs in steps that led to the adoption in 1981 by the World Health Assembly of the International Code of Marketing of Breast-Milk Substitutes. In 1990, UNICEF convened with WHO, with the cosponsorship of the US Agency for International Development and the Swedish International Development Authority, an international conference on the protection, promotion, and support of breast-feeding. This conference issued the Innocenti Declaration (named for its venue at the historic Spedale degli Innocenti in Florence, home of UNICEF's International Child Development Centre), which called for the reinforcement of a 'breast-feeding culture' and its vigorous defence against incursions of a 'bottle-feeding culture.'

Growth monitoring was perceived as a useful tool for promoting satisfactory growth of children, which itself represented the outcome of influences of diet and disease. Furthermore, child growth was held to be a sensitive and reliable indicator of overall development, and UNICEF advocated that nutritional status should be considered along with more conventional economic indicators in assessing situations and determining policy, in the context of its advocacy of 'structural adjustment with a human face.' Growth charts and weighing scales were widely distributed, but the approach came into some disrepute when it was recognized that too often it was perceived as an end in itself, a sort of technological fix, with little attention to how monitoring should be done and how the results should be used.

Aware of the importance of communication and concerned about identifying simple messages about survival and development of universal validity, although often requiring local adaptation or addition of specificity, UNICEF in 1989 published with WHO and the United Nations Educational, Scientific and Cultural Organization *Facts for Life: What Every Family and Community Has a Right to Know*. More than 15 million copies are in use in 215 languages. The third edition, published in 2002 by UNICEF and seven other UN agencies including the World Bank, covers 13 topics, including safe motherhood, breast feeding, nutrition and growth, diarrhea, and HIV/AIDS.

Nutrition Strategy and the World Summit for Children Goals

In 1990, the UNICEF executive board approved a new strategy for improving nutrition of children

and women in developing countries. The strategy stemmed from the Convention on the Rights of the Child: freedom from hunger and malnutrition are recognized as basic human rights, and continued malnutrition is unacceptable. The strategy proposed a methodology for the identification of appropriate actions through situation assessment and analysis rather than through a predetermined set of technical interventions. This so-called triple A cycle (assessment, analysis, and action, followed by reassessment, etc.) is applicable at household, district, and national levels.

Nutrition status is seen as an outcome. Immediate determinants are dietary intake and infectious disease. Underlying influences can be grouped into three major clusters: household food security, health services coupled with a healthy environment, and care for children and women. The degree to which the three conditions necessary for good nutrition are fulfilled depends on the availability and control of human, economic, and organizational resources at different levels of society: household, community, national, and international. Education has an important role, against a background of political, economic, cultural, and ideological factors. The conceptual framework advocated by the strategy for analyzing the nutrition situation is shown in **Figure 1**.

The strategy proposed a number of nutrition goals for the 1990s, shared with WHO. As endorsed by heads of government at the World Summit for Children in 1990, the goals were

- reduction of severe and moderate malnutrition among under-fives by half of 1990 levels;
- reduction of the rate of low birth weight (less than 2.5 kg) to less than 10%;
- virtual elimination of iodine deficiency disorders (IDD);
- virtual elimination of vitamin A deficiency (VAD) and its consequences, including blindness;
- reduction of iron deficiency anemia (IDA) in women by one-third of 1990 levels; and
- empowerment of all women to exclusively breastfeed their children for 4–6 months (later changed to 6 months) and to continue breast feeding with complementary food for up to 2 years of age or beyond.

The summit endorsed a number of other goals related to women's health and education, child health and sanitation, and basic education, all of which are relevant to the attainment of the goals for nutrition. Governments committed themselves to prepare and execute national plans of action to implement the summit goals. In 1992, the FAO/WHO International Conference on Nutrition in Rome included in its World Declaration and Plan of Action for Nutrition a commitment to the nutritional goals of the World Summit for Children.

In 1991, the first meeting on a global scale to pursue summit goals was held in Montreal,

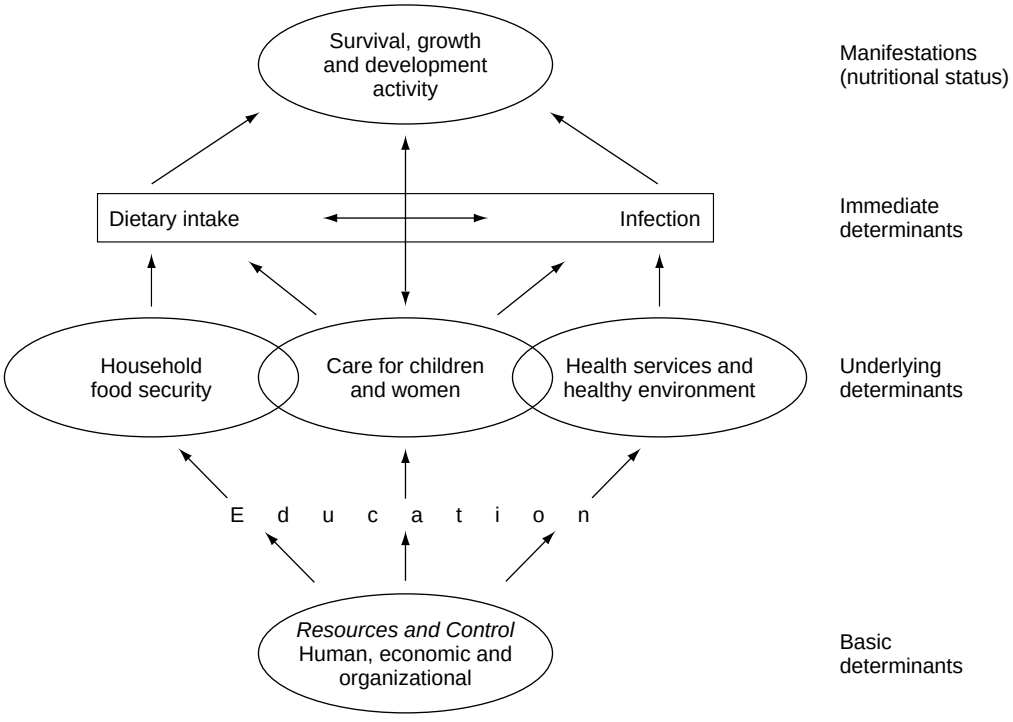


Figure 1 Determinants of child survival and development and nutritional status.

Canada: it was a policy conference on overcoming micronutrient malnutrition titled Ending Hidden Hunger. The strategy to control the three micronutrient deficiencies was to involve a country-specific combination of short-term and long-term measures, including dietary diversification, food fortification, and nutrient supplementation. In 1991, UNICEF also launched with WHO the Baby-Friendly Hospital Initiative, which formally recognizes maternity facilities that follow specified practices that enable mothers to make an informed choice about how to feed their babies and that helps them to establish and maintain lactation.

By 1993, UNICEF and WHO adopted mid-decade goals, which included goals for IDD, VAD, and breast-feeding, as a way to ensure early progress toward achieving the end-decade goals. The mid-decade goals were the more doable 'top down' goals that could be verified by coverage surveys. For IDD, it was Universal Salt Iodisation (USI), to be verified by the coverage of iodized salt at the household level. For VAD, it was the coverage of children aged 6–59 months consuming high-dose vitamin A capsules in the previous 6 months, and for breast-feeding it was the extent of exclusive breast-feeding at 4 months and the coverage of baby-friendly hospitals. The mid-decade push and prioritization ensured real progress toward these goals by the end of the decade.

By 2000, most children in more than 40 countries were receiving at least one high-dose vitamin A capsule yearly. Between 1998 and 2000, an estimated 1 million child deaths were prevented by vitamin A supplementation. USI was also a great success. Whereas in 1990 less than 20% of households were consuming iodized salt, by 2000 this was 90% or more in 24 countries, with a further 21 achieving 70–90%. The exclusive breast-feeding rates also increased by 10% during the decade, and by 2001 more than 15,000 hospitals in 136 countries had been certified as baby-friendly. These large increases in coverage of iodized salt, and also of vitamin A capsule distribution, and exclusive breast-feeding, can be translated into millions of child lives saved, and child disabilities prevented, and improvement in children's development improved.

Nutrition in the World Fit for Children

At the Special Session on Children held at the UN General Assembly in May 2002, the nations of the world committed to building a world fit for children and adopted a resolution that includes the World Fit for Children (WFFC) goals for 2010. Among the WFFC goals are the reduction of child malnutrition among children younger than 5 years of age by at least one-third, with special attention to children younger than 2 years of age, and reduction of the rate of low birth weight by at least one-third.

UNICEF has adopted in its medium-term strategic plan for 2002–2005 five priorities that derive from emphasis on the rights of the child: integrated early childhood development (IECD); immunization 'plus' (signifying services that can utilize the same delivery system, for example, vitamin A supplementation); girls' education; improved protection of children from violence, abuse, exploitation and discrimination; and fighting HIV/AIDS.

The IECD approach is a holistic one that looks at the whole child and starts from before birth to preschool age, with the emphasis on children younger than 3 years of age. IECD is where nutrition now resides in terms of UNICEF programs. Reduction of protein energy malnutrition (PEM) through implementation of the nutrition strategy requires what have been called diagonal approaches—those that respect both vertical goal orientation and horizontal community process capacity building. IECD should facilitate the tackling of PEM since it is about understanding how to deliver a set of integrated services to children in their community. UNICEF's capacity to do this alone is limited, and partnerships are sought with others, including the World Bank.

With regard to emergencies, UNICEF continues to be responsive to natural and man-made disasters within the constraints of its mandate and resources, and in close cooperation with other parts of the UN system. Thus, UNICEF would not normally provide food aid, this being the province of the World Food Programme (WFP), but may provide supplementary or therapeutic foods for children, health supplies, blankets and tents, fuel for heating, water purifiers, cash assistance, and logistical support.

Role

UNICEF's role can be appropriately summarized by the following quotations from its mission statement, adopted by its executive board in January 1996:

> UNICEF is mandated by the United Nations General Assembly to advocate for the protection of children's rights, to help meet their basic needs and to expand their opportunities to reach their full potential. ... UNICEF aims, through its country programmes, to promote the equal rights of women and girls and to support their full participation in the political, social and economic development of their communities.

UNICEF pursues its advocacy and educational role at all levels and in all possible fora. The World

Summit for Children at the UN in New York in September 1990 attracted 71 heads of state and government and also ministers or other senior officials from 88 other countries—at the time the largest gathering of world leaders in history. In May 2002, the UN General Assembly held a special session on children, which was notable for the unique and active participation of approximately 400 children from more than 150 countries. UNICEF had facilitated the collection in these countries of key information about children's rights and well-being to assess progress made for children since the 1990 World Summit. Nutrition (and other) indicators for each country in the world are published in the annual *State of the World's Children* reports.

UNICEF's programming role is essentially a country activity, conducted by the UNICEF representatives and their staff in close collaboration with officials of concerned government departments (e.g., of planning, health, education, and social affairs ministries) and in consultation with representatives of other UN agencies and NGOs. The country programming process, which begins with a situation analysis (describing the situation of women and children in the country and reviewing past programs of cooperation and potential areas of future cooperation), is intended to ensure that the UNICEF country program relates to the needs of children and women in the country, is fully integrated into the government program, reflects government policy and programs as well as policies established for UNICEF by its board, and is complementary to assistance provided by other multilateral and bilateral agencies. The process involves balancing a global ethic with country priorities, but there need be no inconsistency between setting international and national goals and targets and at the same time working to empower local partners to participate in progress toward improved nutrition. 'Bottom-up' advance need not be an alternative to 'top-down' advocacy and policy development.

UNICEF operates one of the largest supply networks in the UN system. In 2001, it procured $596 million worth of supplies relating to health, education, water and sanitation, and nutrition products, including 40% of global vaccine doses for children.

UNICEF works closely with other UN agencies concerned with nutrition, particularly WHO, FAO, WFP, and the World Bank. A long-standing mechanism for harmonization of policies with WHO exists in the form of a Joint Committee on Health Policy (now including UNFPA). UNICEF is a major supporter of the United Nations System Standing Committee on Nutrition (SCN). The SCN (which evolved from the earlier PAG) currently has 19 UN members and also representatives of other agencies, bilateral donors, and NGOs; it is the focal point for promoting harmonized nutrition policies and strategies throughout the UN system and for strengthening collaboration with other partners.

Structure

Within the UN system, UNICEF's comparative advantage relates to its field-based structure. At the end of 2002, 86% of UNICEF's 8083 staff members were located in its 199 field offices, supported by 7 regional offices. UNICEF's headquarters is in New York. In 2002, professional staff numbered 3458, representing 165 different nationalities: 68% were from developing countries, and 45% were female.

UNICEF is governed by an executive board consisting of representatives of 36 countries who serve on a rotational basis. The board is responsible for overall policy and for authorizing receipt and expenditure of funds. The executive director is appointed by the secretary-general of the UN in consultation with the board. The executive director appoints all other staff.

UNICEF has established a unique structure of supporting bodies in most industrialized countries. Known as National Committees for UNICEF, in 2001 there were 37 largely autonomous organizations with an important role of advocacy and fund-raising. They each have a formal relationship with UNICEF and are responsible for fund-raising operations in their own countries, public information and development education, and promotion of UNICEF concerns, including the Convention on the Rights of the Child and follow-up of the Children's Summit. Much of their work is supported through voluntary action.

Funding

UNICEF is funded entirely through voluntary donations, chiefly from governments but also from private institutions and individuals. In 2001, the total income was $1225 million, of which $990 million was for regular programs (unrestricted plus earmarked resources) and $235 million raised for emergencies (**Table 1**). Of the total, governments and

Table 1 UNICEF Income, 1999–2001 (in millions of US$)

	1999	*2000*	*2001*
Regular resources	589	563	551
Other (regular)	332	377	439
Other (emergencies)	197	199	235
TOTAL	1118	1139	1225

intergovernmental organizations contributed 64%, and the National Committees contributed nearly 30%.

Although 152 countries made some contribution to UNICEF in 2001, only 9 contributed more than $40 million from both the governmental and private sectors. The three governments that contributed most in total funds were the United States, Japan, and the United Kingdom. In terms of total funds per head of population, Norway, Sweden, and Denmark contributed most.

UNICEF has developed supportive relationships with international service organizations such as Rotary International, which has contributed more than $462 million for the global eradication of polio (with $167 million directly through UNICEF), and Kiwanis International, which raises funds for the elimination of IDD. Canada and Sweden are also major contributors for such activities.

Regular resources are available for cooperation in country programs approved by the executive board, as well as for programme support and administrative expenditures. Allocations are made to country programs according to three criteria: annual number of deaths of children younger than 5 years per 1000 live births, income level (GNP per caput), and the size of the child population. Thus, most money is allocated to the larger, and poorer, countries. Representatives have the responsibility to negotiate with governments programs of cooperation within these allocations. Such programs may be expanded if other resources become available.

Total expenditure in 2001 was $1246 million—93% for country programs of cooperation and 6% for management and administration of the organization.

See also: **Breast Feeding**. **Children**: Nutritional Requirements; Nutritional Problems. **Infants**: Nutritional Requirements. **Vitamin A**: Deficiency and Interventions. **World Health Organization**.

Further Reading

Annan KA (2001) *We the Children: Meeting the Promises of the World Summit for Children*. New York: United Nations Children's Fund.

Bellamy C (1996) *The State of the World's Children: 1996*, [Contains social goals for 1995 and 2000]. Oxford: Oxford University Press.

Bellamy C (1998) *The State of the World's Children: 1998* [Special focus on nutrition]. Oxford: Oxford University Press.

Black M (1986) *The Children and the Nations. The Story of UNICEF*. New York: United Nations Children's Fund.

Black M (1996) *Children First. The Story of UNICEF, Past and Present*. Oxford: Oxford University Press.

Grant JP (1991) *The State of the World's Children*: 1991 [Contains full texts of the Convention on the Rights of the Child and the World Declaration and Plan of Action from the World Summit for Children]. Oxford: Oxford University Press.

Shrimpton R *et al.* (2002) *UNICEF Nutrition Portfolio Review (1980–1999)*. New York: United Nations Children's Fund.

Shrimpton R and Schultink W (2002) Can supplements help meet the micronutrient needs of the developing world? *Proceedings of the Nutrition Society* **61**: 223–229.

UNICEF (1990) *Strategy for Improved Nutrition of Children and Women in Developing Countries. A UNICEF Policy Review*. New York: United Nations Children's Fund.

UNICEF (2001) *Progress Since the World Summit for Children: A Statistical Review*. New York: United Nations Children's Fund.

UNICEF (2002) *Annual Report, 2002*. New York: United Nations Children's Fund.

URBAN NUTRITION

N Solomons, Center for Studies of Sensory Impairment, Aging and Metabolism (CeSSIAM), Guatemala City, Guatemala

Introduction

For most readers of this encyclopedia, the notion of urban living and life style would be considered the demographic norm or a 'ground state.' In fact, the history of human biological and cultural evolution was written in rural areas. Currently, 53% of the population of developing countries is still living in the rural countryside as agrarian peasants or hunters; this proportion has fallen rapidly from 83% in 1950. The context for the topic of urban nutrition includes: the manner in which wealth and income are distributed among the urban populations; the agricultural efforts within and without the cities to supply food to the urban masses; trends and fashions in marketing and consumption of foods; and the ecological and environmental influences generated in the agrarian countryside and in the urbs themselves.

Urbanization

For 90% of the duration of *Homo sapiens*' evolution, all human endeavor was related to obtaining

food, reproducing and rearing young, or defense and survival. During that period, humans were part of nomadic tribal groups roving over the hunting and foraging ranges that provided their food, the pelts or fiber for their clothing, and other necessities. More recently, some 40 000 years ago, mankind witnessed the advent of the pastoralist lifstyle, with pursuit of grazing livestock across available pasture areas constituting a new form of livelihood.

Only with the domestication of plants and the emergence of the agrarian life style, some 10 000 years ago, did settling of specific terrains become a human attribute. For the first time in human history, a rural family's effort could produce enough food to feed more than just a single family. Agriculture was the precondition for human 'civilization,' as it freed some of the time, for some of the population, from an obligation to gather food as the sole human pursuit. Thus, classes of artisans, traders, clergy, politicians, and warriors could enter the society, supported and fed by the livestock, crops, and produce supplied by farmers. These classes did not live in dispersed terrains but rather created communities that were larger and more complex than a tribal village. The first towns represent either the seats of rulers or the temple complexes of religious elites. As trade expanded, towns and cities developed along the routes of commerce. These were often situated along waterways and at coastal harbors, or scattered along caravan routes. With the rise of warrior classes, protective walls and fortifications began to surround such communities. The Mediterranean and Asia Minor were the sites of the cities of antiquity such as Babylon, Athens and Sparta, Alexandria, Rome, and Carthage.

The evolution of mercantile trade in Europe became a motive for constructing even larger cities, as exemplified by the city-states of Renaissance Italy. Finally, the Industrial Age gave a whole new charge and dimension to cities. The change of production from artisan guilds to workers in factories altered the character of urban life. Crowded teeming cities with tenements, as described by Dickens, came into being in Europe at the beginning of the nineteenth century. These early modern metropolises were smoky, grimy, smelly, unsanitary places, ripe for the occurrence of occupational mishaps and the transmission of contagious diseases. Although the caloric supply might have been sufficient, only the most stably preserved of foodstuffs could be used to feed urban populations before the era of refrigeration and food technology. Fresh fruits and vegetables were available largely to those still living on countryside lands.

It is important to realize that until very recently in world history, the proportion of humanity living in cities remained very small. Most were living in tribal settings, as nomadic pastoralists, or as rural peasants. The United Nation defines an 'urban center' as a concentrated population of at least 20 000 inhabitants. By this definition, only 5% of the world's population was urbanized at the turn of the nineteenth century, 13% at the turn of the twentieth century, and 47% by the year 2000. In 1950, the urban population of developing countries was 17%; by the year 2000, it had reached 44%. The degree of urbanization varies from region to region. In the Americas, 75% of the population is now urban, whereas on the African continent, urbanization hovers around 40%.

The massive megapolises, cities of over 10 million inhabitants, are a recent phenomenon. Initially, the biggest metropolitan areas, such a New York and Tokyo-Yokohama, were in industrialized countries. In 1975, there were only five cities with more than 10 million inhabitants; this rose to 19 in 2000, and is projected to reach 23 by the year 2015. The top-ten of megapolises 10 years hence are projected to be: (1) Tokyo-Yokohama; (2) Bombay; (3) Lagos; (4) Dhaka; (5) Sao Paulo; (6) Karachi; (7) Mexico City; (8) New York; (9) Jakarta; and (10) Calcutta. Only two of these are in currently industrialized nations.

Urbanization and Nutrition Transition

Whatever the era in the evolution of cities, from antiquity to the present, it is likely that the cuisine and dietary fare of urban populations has always differed from that in tribal, nomadic, or agrarian settings of the same countries. On the one hand, if the town were a coastal port or fishing village, consumption of fish and seafood would be higher than in the interior countryside, while populations of trading centers would have more access to exotic, imported items such as Oriental spices and teas or Caribbean rums and molasses. On the other hand, food is produced in the countryside. The agrarian producers have first access to the crops and foodstuffs produced, whereas the cities can only gain that which is transported to the urbs and offered for sale. Generally, then, the basic staples in the cities are those that are traditional to the farmers that serve them.

Despite a history of poorly mechanized industry in urban factories that required heavy physical labor, an even greater daily energy output was required for rural agricultural pursuits. Hence, the amount of dietary energy needed for child growth

or to maintain energy balance is generally lower for urban populations. Until recently, the variety of foods, especially produce,- was more limited in urban markets than on farms and plantations. The meat in urban centers was more likely to be dried and salted than fresh, and one or two staples, grains or tubers, would provide the bulk of dietary energy.

Electricity, food technology, and concentration of wealth in urban areas have changed this panorama in recent years as a process of 'nutrition transition,' a term introduced by Barry Popkin, has come to dominate particularly the urban populations of today's developing nations. More, rather than less, variety of food is the reality of modern cities. Sweetened and flavored processed foods have higher appeal than coarse staple roots and cereals. The 20th century saw a meteoric rise in the demand for and production of cooking oils and fat-based spreads, while refrigeration meant that fluid milk did not have to come from the udder of a dairy animal on one's doorstep to be safe and available. Processed foods – bottled, canned, and frozen – entered the market with the rise of the food sciences and food technology, and sweets and desserts became a larger component of daily fare. Inexpensive vegetable oils from corn, soy, safflower, cottonseed, etc., entered international commerce for cooking, or were hydrogenated and solidified into margarines and shortening. These dietary changes are not restricted to the affluent elite of developed countries, but are part of the change affecting the urban middle and lower classes in developing countries to an ever-increasing extent. However, at the same time, those living in abject poverty in both urban unemployment and rural landlessness could be unable to participate in any of this varied and energy-dense fare. This is the setting for 'nutrition transition' defined by Popkin as 'the rapid shift in the structure of diet in low-income countries and the coexisting problems of under- and overnutrition.' With respect to the overnutrition aspect, developing country residents are participating in the dietary patterns associated with increased risk of chronic diseases.

Contemporary Food Supplies and Diets for the Cities

In terms of availability, factors such as transport, refrigeration, sewage treatment and disposal, etc., have allowed the modern city the option to overcome some of the barriers to having the food supply safe for the urban populace. Westernization of the dietary patterns of cities is an interesting consequence and a reinforcing cycle of supply and demand. Not only domestic production, but also the importation of commodities, becomes an alternative for supplying food to urban populations. The transport infrastructure to get foods to cities from regional farms or across routes for foreign imports has improved with the advent of refrigeration and high-speed transportation. More perishable foods can reach urban tables, not only from the interior of countries but from overseas. Currently, issues of food supplies are generally more an issue of accessibility at the household level than availability in the urban marketplace.

Electrification and refrigeration are factors in household diets, as well. Although most cities have electricity, not all urban homes have their own refrigerators; but the food consumed in these homes will have been refrigerated in its processing, storage, and while on sale in the store. The factor of access to refrigeration has been associated with a major decrease in the use of traditional conservation methods such as salting, pickling, and smoke-curing of food for preservation, as well as improved microbiological quality. In this respect, urbanization may be a factor in reducing the incidence of carcinogenesis associated with mycotoxins, salt, and organic byproducts, as well as the incidence of episodes of some food-borne diseases.

Accessibility (the ability to obtain items actually on the market) is more important in determining what gets onto the tables of urban households. Inequality of resources and wealth within the urban populations gives rise to food insecurity within households in the urban area. Often a pattern of geographical distribution demarcates the zones with the highest risk of insecure and inadequate household food supplies. Generally, the greater diversity of items that one is able to include in one's diet, the lower is one's risk of experiencing food insecurity.

Urban Agriculture

Urban agriculture, using marginal lands within the confines of the metropolitan area, can save up to 20% on outlays of cash for food for poor families at the expense of 1–2 days of labor per week. Urban gardening is seen as a means to improve public health not only through improving economic and food security, but also in providing exercise, psychological, and community well-being, and environmental stewardship. The type of plants that are grown in urban settings are largely foods that contribute micronutrients, and are much less likely to be the

staples that provide the bulk of energy and protein. For reasons of sanitation and zoonotic diseases, as well as waste disposal, domestic livestock in the cities are a much more remote option, although aquaculture with treated waste waters could provide for fish, crustaceans, and mollusks toward meeting the protein needs of urban populations.

Street Foods

Roads and highways may be a phenomenon of rural communication but the street is an urban entity. Street foods are defined by the Food and Agricultural Organization as "ready to eat foods and beverages prepared and/or sold by vendors, especially in streets and others similar public places." With urban life has come both the opportunity, and at times the necessity, to consume one or more daily meals away from one's home. For the affluent and middle-class, cafeterias and restaurants serve the clientele looking for a meal outside of the household. For low-income populations, street foods fill this dietary niche. One-quarter or more of daily calories are often consumed outside of poor urban homes by adult workers and high school children.

The main concerns about street foods have been issues of sanitation and microbiology. The dietetics, anthropology, and economics have remained more obscure. As Sarah Atkinson has said with respect to street foods: "safety has to be assessed not by middle class standards but in relation to other food sources and the environmental conditions in the homes of those buying these foods." Positive aspects are the accessibility of these foods to the poor and the economic empowerment that street foods represent for the vendors. Scrutiny of street food from municipal authorities may come from the fact that it is an unregulated informal activity from which taxes or licensure revenues cannot be collected. Street food is a convenient scapegoat for politicians and bureaucrats when explaining local epidemics that may relate to other failings of the municipal investment in infrastructure.

Nutrition and Health in the Urban World

The way in which people eat is a major determinant of their nutritional stores and status, but issues of life style, health and pollution modify and influence the nutriture of individuals and populations. Each of these factors contributes to the distribution of the simultaneous under- and overnutrition states that characterize nutrition transition.

Labor market shocks impact on maternal pursuits, child work, and schooling. Rural children have traditionally worked on both household and farm chores as part of an integrated family-production pattern. In urban areas, mere economic survival often obligates families to send children to work in factories at low wages, exposing them to occupational risk and interfering with their formal education. To the extent that urban mothers seek income-generating activities outside of the home, monetary resources may be bolstered by child-rearing and child-caring and meal provision can be disrupted. Breast-feeding is one form of meal provision, i.e., that for the infant, that can be influenced adversely by maternal work obligations. In general, less exclusive breast-feeding and shorter total lactation are seen in urban mothers as compared to their counterparts in the countryside.

Urban poverty and undernutrition are growing, in part because of the influx of migrants. Spatial distribution of the pockets of poverty can often be identified on a map of the city. The unempowered and abjectly poor tend to congregate in the least desirable and most precarious areas of the urban landscape. It has been established that such groups have scarce municipal services and low incomes. The relative social power of women in such households is vastly inferior to that of the men. True food insecurity is a legitimate concern in these zones of urban poverty.

Finally, pollution, some of which has its negative effect on the population by contaminating the food supply, cannot be ignored. In Jakarta, where leaded gasoline is still the norm, individuals spending time on the streets became contaminated. Poor air quality (smog) can provoke respiratory distress with a cascade effect on appetite, regular eating, and nutrient retention.

Nutritional Deficiencies of Urban Populations

At a physiological level, nutritional status is a function of the intake, absorption, and retention of nutrients. Any of the micronutrient deficiencies common to human populations can be seen in residents of towns and cities. Household food insecurity or poor caring practices can explain low intakes of total macronutrients or diets with poor micronutrient density and bioavailability. To the extent that sanitation may be better in the urban setting, disease-related wastage of nutrients may contribute less to the process of general undernutrition reflected in poorer growth and higher prevalences of stunting and wasting.

Iron deficiency is prevalent in infants, young children, fertile women, and pregnant and lactating women, independent of social class or setting. If

lead exposure is present in the cities, this can aggravate iron status, especially of the young. In the broadest generality, micronutrient deficiencies are less frequent among urban populations, but a residual prevalence of poor status with respect to riboflavin, zinc, folic acid, and vitamin B_{12} is common in low-income segments of urban populations. Intervention programs are logistically more accessible to urban populations, but relief agencies often direct and target their assistance to the countryside, bypassing the problem that may exist around their headquarters in Third World capitals.

Social determinants related to poverty in urban communities and households constitute the underlying reasons for impaired growth and poorer anthropometric indices. Discrepancies in women's status, that is their social and economic power relative to men in the same households and societies, has a profound influence on the status of their children and the risk of the latter to suffer stunting and wasting. Greater equality between the sexes generally results in overall improved prenatal, obstetrical, and child care, better complementary feeding and treatment of illness, and higher immunization rates. Urban settings would generally tend to foster less traditional social views and provide women with more options for self-realization. Moreover, campaigns and programs for empowering women through education and entry into the work-force would be logistically easier to maintain in towns and cities. Hence, both the better general nutritional status of children in cities may be a consequence of trends toward a greater equality of women and improved decision-making power in favor of their children's nutritional evolution.

Survival through and beyond infancy and preschool years with generally more favorable nutritional status than one's rural counterparts is likely to be reflected in continued better nutriture into childhood and adolescence. A phenomenon of family disintegration, more commonly seen in urban communities, has led to the proliferation of street children who live in the street or of the street. Several aspects of the stereotypical portrait of their life style, including abuse of glue and illicit drugs and participation in child prostitution, would point to increased risks of organic damage and sexually transmitted disease, including HIV/AIDS. On the other hand, through the various modalities of obtaining income from legitimate sales and tasks, to extorsion, theft, and prostitution on the illegitimate side, street children have more disposable income than other urban children from equally humble origins but dependent on meager household means. Little research into comparative nutritional status of street children in the urban setting has been conducted, but fragmentary findings suggest that their income and ingenuity provides them with a more diverse diet than their homebound peers.

The urban elderly are another generational group of concern in terms of their risk to suffer nutritional deficiencies. To the extent that their age, *per se*, makes them more susceptible to undernutrition and social isolation, with the weakening of extended family traditions, and afflicts more urban than rural elders, their situation in the urban context can be more nutritionally precarious.

HIV/AIDS can be more of an urban or a rural health problem depending upon the country in question, its patterns of transmission, and the specific differences of urban and rural culture. Wherever it occurs, for HIV-infected individuals their disease represents a double burden for their nutrition. On the one hand, the ravages of the infection interfere with appetite, disrupt metabolism and deplete nutrient reserves; on the other hand, lost income and the cost of treating the illness jeopardizes individual food budgets and poses the risk of household food security.

Nutritional Excess in Urban Populations

The most important nutritional excesses to consider are those of overweight. This is officially defined as a body mass index (BMI) of $>25\,kg/m^2$. For children, overweight begins at the 85th percentile of normative curves and obesity at the 95th percentile, as defined on the 2000 BMI charts of the US Center for Disease Control and Prevention.

Physical activity patterns are altered from the traditional rural focus role of food production. Even construction and manufacturing work requires increasingly less physical effort. Household chores are favored by electrical labor-saving devices and home delivery of goods and services such as fuel and water. For both adults and children, sports and active recreational pursuits are being replaced by television watching and computer entertainment, including internet, in urban settings. Much of the increase in the number of overweight individuals in cities is ascribed to sedentariness. A new and troubling association of stunted child-overweight mother pairs has been identified, but it is not restricted to urban families. However, low stature seems to be a risk factor for overnutrition as confirmed in China, Singapore, Brazil, and Mexico, especially in their urban areas.

The tendency not only to eat away from home but also omit specific meals is an innovation of modern urban life; family integrity at mealtimes is more

traditionally observed in rural areas Not consuming breakfast is a habit surging in urban households and with it comes proven deficits in attention span and school and work performance. Paradoxically, breakfast-skipping behavior is found to be associated with generally less physical activity and more sedentary life styles and is also linked to excess weight gain. Moreover, higher usage of alcohol and tobacco are found in those who regularly skip the breakfast meal.

Vitamin or mineral excess as a public health problem is rare. It generally occurs as the result of excessive consumption of micronutrient supplements, and this would be more likely to take place in the urban setting.

Diet, Nutrition, and Quality of Life

It is probably safe to conclude that suffering nutritional deficiency and dying a premature death from infectious illness is more likely to occur among the rural agrarian peasantry than among the urban masses. But for rural individuals who survive accidental and infectious deaths, their plant-based diets and rigorous lifelong physical activity patterns makes the goal of 'dying healthy in old age' a stronger possibility. With extended life expectancy among urban populations, ensuring the quality of life in later life is important as both a humanistic and economic consideration. Cardiovascular and malignant diseases produce lingering debility and dependency, robbing individuals of well-being and placing a burden of healthcare on relatives and governmental resources. Global and sustainable interventions to reduce sedentariness and pathogenic dietary practices in the growing segment of the world's population living in cities are an imperative for social and economic stability going forward.

Conclusions

Through the progressive evolution from rural residence in nomadic tribes and pastoral pursuits to settling of farms and plantations, the conditions for creation of towns, cities, and metropolises have emerged. Public health nutrition and the epidemiology of nutrition must focus more on what has been termed 'urban nutrition' as the world's population shifts from rural to urban residency. The essential complexity of providing for human needs in densely populated settings has repercussions for the resulting nutriture of the urban populations and for the nature of their environments. Street foods, street children, urban pollution, and diet and life styles can foster or be associated with undernutrition and reduced productive capacity in some circumstances while contributing to overnutrition and excess noncommunicable disease in others. Rather than use the 'lessons' from the rural experience in developing countries, it is important to direct one's reading, one's research, or both to the study of the urban milieu in order to gain insights for addressing the challenges of urban nutrition. **Table 1** outlines a framework of opportunities for urban nutrition research that derive from the considerations in this review.

Table 1 Opportunities in urban nutrition research

- Basic description of diet and nutrition in selected townships and metropoli
- Rural-urban comparisons
- Role of family and household relationships
- Influence of rural-to-urban migration on nutritional status and dietary status
- Universality and generalizability of dietary factors for degenerative diseases
- Contributions of urban agriculture to the urban food supply
- Influence of urban pollution on the food supply and nutritional status
- Nutritional and health status of street children
- Containing the epidemic of obesity and metabolic syndrome in urban populations

See also: **Anemia**: Iron-Deficiency Anemia. **Breast Feeding**. **Iron**. **Obesity**: Definition, Etiology and Assessment.

Further Reading

Atkinson SJ (1995) Approaches and actors in urban food security in developing countries. *Habitat International* **19**: 151–163.

Caballero B and Popkin BM (2002) *The Nutrition Transition: Diet and Disease in the Developing World*. London: Academic Press.

Delisle H (1989) *Urban Food Consumption Patterns in Developing Countries: Some Issues and Challenges*. Rome: Food and Agricultural Organization.

Food and Agriculture Organization (1989) *Street Foods*. Food and Nutrition Paper No. 46: Rome: FAO.

Gross R (2003) Beyond food and nutrition: How can cities be made healthy. *Asia Pacific Journal of Clinical Nutrition* **11**(supplement): S763–S766.

Gross R, Landfried B, and Herman S (1996) Height and weight as a reflection of the nutritional situation of school-aged children working and living in the streets of Jakarta. *Social Sciences in Medicine* **43**: 453–458.

Krause VM, Tucker KL, Kuhnlein HV *et al.* (1992) Rural-urban variation in limed maize use and tortilla consumption by women in Guatemala. *Ecology of Food and Nutrition* **28**: 279–288.

Phillips DR (1993) Urbanization and human health. *Parasitology* **103**: S97–S107.

Popkin BM (2001) The nutrition transition and obesity in developing countries. *Nutrition Reviews* **52**: 285–298.

Popkin BM and Bisgrove EZ (1988) Urbanization and nutrition in low-income countries. *Food and Nutrition Bulletin* **10**: 3–23.

Solomons NW and Gross R (1995) Urban nutrition in developing countries. *Nutrition Reviews* **53**: 90–95.

V

Vegan Diets *see* **Vegetarian Diets**

Vegetables *see* **Fruits and Vegetables**

VEGETARIAN DIETS

J Dwyer, Tufts University, Boston, MA, USA

Published by Elsevier Ltd.

Introduction

Vegetarian eating patterns are the norm in many parts of the world, while in Western countries they are the exception. Currently, about 2.5% of adults in the US and 4% of adults in Canada follow some sort of self-described vegetarian diet. This article examines the nutritional adequacy of vegetarian diets and some eating patterns and practices that affect it, focusing on Western countries.

Definitions are important in discussing vegetarian diets and nutritional adequacy since both of these terms cover a multitude of disparate characteristics.

Vegetarianism

For most vegetarians today, their investment in vegetarian eating as an all-encompassing philosophical system for organizing all of life is minimal. Food preferences, habit, and flavor take precedence over philosophy in dictating their eating choices. That is, these individuals have vegetarian eating styles but they do not subscribe to vegetarianism.

There is a smaller group of vegetarians whose eating patterns serve as badges of honor for a deeper and more encompassing set of beliefs that includes and is reflected in but is not limited to their eating patterns. This belief system is referred to as vegetarianism. Vegetarianism is a philosophy and belief system rather than simply an eating pattern. Those who subscribe to vegetarianism hold strong convictions about the moral, metaphysical, ethical, or political appropriateness of their eating choices. Often more restrictive animal food and other dietary avoidances are accompanied by the most deeply held views. In fact, it is a minority of all of those who consume vegetarian diets who subscribe to vegetarianism. Advocates of vegetarianism have become more militant in recent years. With some vegans, it remains an ethical, philosophical, or predominantly religious conviction, or a deeply rooted health concern that is privately held and the diet is practiced with great conviction. Other vegans are strongly committed to enlisting additional followers and appear to have become more militant in recent years. Some have adopted broader agendas, including that of animal rights, while others do not. In the US, animal rights organizations such as People for the Ethical Treatment of Animals (PETA), the PETA foundation, the Animal Liberation Front, and the Physicians' Committee for Responsible Medicine (PCRM) in the US are all very vocal. They finance advertizing, journal articles, and educational efforts to encourage adoption of their views. Extensive efforts are directed toward youth. Similar groups are also active in other countries.

Vegetarian Eating Patterns

The term vegetarian diet does not fully describe the variety in nutrient intakes and health status of those

Table 1 Common types of vegetarian dietary patterns categorized by animal food use

Pattern	*Comments*
Meat avoiders	Limit or avoid red meat and other flesh foods; may also restrict poultry, fish, and seafood. Diets are similar in most respects to nonvegetarian diets
Lacto-ovo vegetarians	Avoidances include all meat, poultry, and often fish, but consume milk products and eggs. Iron may be limiting and it can be obtained from iron-fortified cereals. Low-fat dairy products are preferred to keep intakes of saturated fat and total fat moderate
Lacto vegetarians	Avoid all meat, fish, poultry, and eggs. Nutrient considerations same as above
Macrobiotics	Numerous restrictions generally including avoidance of all meat, poultry, milk and eggs, but may consume fish in small amounts. Also avoid sugar and other refined sweeteners, foods that are members of the nightshade family (peppers, egg plant, tomatoes, and potatoes) and tropical fruits. Current variations of the diet are less restrictive than the versions of 30 years ago, but deficiencies of energy, iron, calcium, vitamin B_{12}, vitamin D, and other nutrients may still arise in weanlings, pregnant women, and young children if diets are nutritionally unplanned
Vegans	Avoidances include all animal products including meat, fish, poultry, eggs, and dairy products. Some vegans may also refuse to use any animal products in daily life. Without careful planning, energy, vitamins B_{12} and D, and bioavailable sources of iron may be low. Concentrated sources of energy-dense foods such as sugars and fats are helpful in increasing energy intakes. Vitamins B_{12} and D and calcium can be supplied from fortified soy milk, fortified cereals, and/or dietary supplements of these nutrients. Usually protein is adequate if a variety of protein sources is consumed
Other patterns	Raw food eaters and 'living food' eaters avoid animal foods and eat raw plant foods, including fruits, vegetables and cereals with special health foods such as wheatgrass or carrot juice. Fruitarians consume diets mostly of fruits, nuts, honey, and olive oil. Rastafarians eat a near-vegan diet and avoid alcohol, salt-preserved foods and additives. Yogic groups vary in their eating patterns but are often lacto vegetarian

who follow such eating patterns. There are many different types of vegetarian eating patterns (see **Table 1**). The impact of these patterns on nutritional status and health requires more complete characterization of diet and other aspects of lifestyle than such a simple descriptor.

The terms vegetarian, lactovegetarian, and vegan focus on foods that are left after others have been omitted from the diet. From the nutritional standpoint the animal food groups (e.g., meat fish, fowl, eggs, milk, and milk products) are nutrient-dense foods. In traditional diets of usual foods, they were often rich sources of certain nutrients. Depending on the particular animal food group under consideration, these nutrients may include protein of high biological value, highly bioavailable iron, zinc, calcium, vitamins A, D, B_{12} and B_6, riboflavin, omega 3 fatty acids, and iodine. When these food groups are eliminated entirely from the diet, intakes of the nutrients these groups are rich in may fall short.

Although dietary diversity in the number of food groups consumed is less among vegetarians, diversity within food groups is often considerable and is sufficient to provide adequate amounts of nutrients. Variety within food groups may even be increased on vegetarian diets. For example, among those consuming vegan diets the amount and type of legumes as well as other vegetables and fruits is often increased.

Eating patterns appear to be more closely associated with health outcomes than are nutrient intakes alone. Therefore, it is important to also consider other phytochemicals and zoochemicals in vegetarian diets that may have health effects.

Vegetarians' intakes of some of the bioactive constituents that are rich in fruits and vegetables such as the flavonoids, antioxidants, and dietary fiber may be much higher than those of omnivores. If these substances prove to have beneficial effects on health, such increased intakes may be important.

From the nutritional standpoint there are beneficial trade-offs from vegetarians' limited animal food intakes. Animal foods are major sources of dietary constituents that are in excess in Western diets, such as high amounts of calories, fat, saturated fat, cholesterol, and low amounts of dietary fiber. Consumption of fewer animal foods, especially if it leads to lesser intakes of these constituents may have positive effects on overall nutritional status. Thus, there are both nutritional benefits and risks to limiting animal foods. The exact effects on nutritional status will vary depending on the food group avoided, the degree of limitation, substitutions of other rich food sources, use of fortified foods and dietary supplements, and other changes in dietary intake that occur at the same time. The nutritional goal is to maximize the benefits and minimize the risks by a judicious choice of the type and amount of animal foods or other sources of the nutrients they contain.

Nutritional Adequacy

In English-speaking North America, dietary reference intakes (DRI) have been issued by the Food and Nutrition Board, Institute of Medicine,

National Academy of Sciences. Nutritional adequacy is defined as meeting nutrient needs such as the recommended dietary allowances (RDAs) or the adequate intakes (AI), while avoiding excess and staying below the tolerable upper levels of intakes (UL) and keeping within the macronutrient ranges specified by expert groups. Similar reference standards are available in other countries.

It is useful to screen out those vegetarians who are likely to be at high risk of dietary inadequacy by using some simple characteristics of their diets and lifestyles to determine if further dietary assessment is likely to be needed. The entire pattern of intake (including avoidances, substitutions and additions of foods, and use of dietary supplements) describes the individual vegetarian's profile of nutrient adequacy or inadequacy.

Because dietary practices among vegetarians are so variable, individual assessment of their dietary intakes is recommended. Those at special risk are those in the nutritionally vulnerable groups due to age, life stage (pregnancy, lactation) or illness, especially if they eschew many animal food groups (vegans), have numerous other food avoidances, or hold beliefs that otherwise limit their dietary intakes.

Adequate Vegetarian Diet Patterns

There are many individuals who have little or no risk of dietary inadequacy from their vegetarian eating patterns; for example, an adult male who regards himself as a vegetarian (also referred to as a meat avoider, or semi-vegetarian) but has a dietary pattern that consists solely of occasionally avoiding red meat about half of the time with no other dietary alterations. Such an individual is unlikely to need further dietary assessment.

Some characteristics of sound, adequate vegetarian diet patterns include the following:

- Use of a nutritionally sound food guide of diet planning; vegan and vegetarian food guides that conform to the latest recommendations of expert groups may help to ensure that nutrient needs will be met with balance and without excessive intakes.
- If diet alone does not meet the RDAs, regular use of appropriate vitamin mineral supplements plus use of a nutritionally sound food guide.
- Vegans and some other vegetarians sometimes have multiple food avoidances; intakes of nutrients likely to be deficient can be increased by use of rich sources of whole foods, foods fortified with the nutrients falling short, and/or vitamin or mineral supplements.
- Consumption of a wide variety of food groups and foods within each group.
- Membership of a family with a long history of adherence to healthy vegetarian eating styles.
- Avoidances are limited to a few foods or are sporadic in nature.

Signs of Possibly Inadequate Vegetarian Diet Patterns

The more of the following characteristics that apply, the higher the potential risk of inadequacy and the greater the need for further assessment.

Diet First it is important to examine what food groups, foods, or products are avoided or de-emphasized on vegetarian diets, and then some additional characteristics of dietary patterns, personal characteristics, and belief systems that further increase risk of dietary inadequacy and other health problems:

- **Many types and extensive avoidance of animal food groups**. Assessment of the nutritional adequacy of vegetarian eating patterns begins by examining animal food groups (red meat, poultry, fish and seafood, milk and milk products, eggs) and specific foods within these that are avoided entirely or eaten only in minimal amounts. Unless other foods or food groups rich in these nutrients or nutrient-containing dietary supplements are used, problems may arise.
- **Many types and extensive avoidance of fortified foods, nutrient-containing dietary supplements, and processed foods**. Some vegetarians believe that fortified foods (highly fortified cereals, calcium fortified soy milk and/or juices, B_{12} fortified yeast), processed foods (frozen, canned, and in extreme cases cooked foods for raw food eaters), and nutrient-containing dietary supplements (vitamins, minerals, fatty acids) should be avoided for various philosophical, ideological, or religious reasons, and refuse to use them. Usage needs to be assessed since such avoidances limit options for nutrition intervention strategies.
- **Few acceptable foods and supplements**. Foods and groups that are stressed and emphasized on the vegetarian diet should be noted. If very few foods or food groups are acceptable for one reason or another, or if only special foods are permitted (organic, nonprocessed, etc.) this may further limit intakes. Some vegetarians are willing to use both fortified foods and nutrient-containing dietary supplements. Use of these may have implications for health and should be recorded. Nutrient intervention strategies for increasing intakes of nutrients falling short in diets may be limited

since such individuals may refuse to use fortified foods and/or dietary supplements.
- **Many practices such as fasting, altered diet during illness, and use of special foods for medicinal purposes**. These practices may further increase risks of nutritional inadequacy. If medical care or treatment is avoided, additional risks may accrue.

Other practices Other practices must also be considered:

- **Other lifestyle practices with beneficial potential health impacts**. Vegetarians have other health habits and lifestyles that alter risks for chronic diseases for the better, such as nonsmoking, abstinence from alcohol, and high levels of physical activity. Therefore, differences in their health outcomes are probably due to a range of factors, and not solely to differences in their diets.
- **Lack of ongoing health surveillance by a physician**. Lack of medical supervision increases the chances that preventable or treatable health problems will be dealt with expeditiously.

Personal characteristics Among these are the following:

- **Nutritionally vulnerable because of age or physiological condition**. The very young, the old, the rapidly growing, pregnant and lactating women, pubertal children, the elderly, and the ill and frail all fall into this group. Individuals at especially high risk are those with chronic diseases and conditions that alter nutritional needs who also have inadequate dietary intakes.
- **Low weight for height**. If weight for height, as measured by body mass index, is below 18.5 or if unintentional weight loss has totaled more than about 5–7 kg (10–15 pounds), there is reason for concern.
- **Rapid weight loss**. Unintentional loss of more than 5% of weight in a month is a cause for concern.

Beliefs Ideology is also important:

- **Deeply held beliefs in alternative philosophical or religious systems that govern food choice**. Some individuals feel bound to make their diets conform to their ethical, philosophical, or religious systems. This can further constrain choice and nutrient intakes.
- **Membership of a quasi-philosophical or religious group that includes vegetarian diets that are not planned in line with nutritional recommendations by experts**. Some groups, e.g., the Seventh-Day Adventists or certain other lacto-ovo vegetarian groups, make a conscious effort to incorporate the recommendations of expert groups, such as those of the Food and Nutrition Board/Institute of Medicine and Health Canada in English-speaking North America into the regimens they recommend. In such cases, the group support provided may be of positive benefit and help to ensure nutritional adequacy. However, at times in the past other groups have insisted on regimens that did not incorporate such recommendations. For those who are active in such groups, the group's support may reinforce negative attitudes toward meeting such expert recommendations.

Using the characteristics above, it is usually possible to sort out those consuming vegetarian diets who are at low or no risk of inadequacy from those who may potentially have problems and need further assessment and counseling.

Key Nutritional Concerns for Vegetarians

Of particular concern with respect to risk of inadequacy for vegetarians are energy, vitamins B_{12} and D, riboflavin, omega 3 fatty acids, calcium, iron, zinc, and iodine. Intakes of these tend to be lower on vegan diets, and may also be low on other diets that include extensive avoidances. Vegetarians of all types can easily meet current recommendations for all nutrients if they are willing to use fortified foods and supplements. However, some vegetarians are unwilling to use these options, increasing risks of deficiency and making dietary planning more difficult.

Vegetarian and particularly vegan diets tend to be low in energy, total fat, saturated fat, cholesterol, dietary fiber, and sodium. If processed foods are avoided added sugars are also low. Current recommendations for acceptable macronutrient distribution ranges (AMDR) in the US are for fat 20–55% of calories and for protein 10–35% of calories, with added sugars no more than 25% of calories, and the remainder from other carbohydrates.

Key Nutrients for Vegetarians over the Life Cycle

Well-planned vegan and vegetarian diets can meet nutritional needs at all stages of the life cycle including pregnancy, lactation, infancy, childhood, and adolescence. There are very few longitudinal studies of those on the more restrictive vegetarian diets; an exception is a Dutch cohort of macrobiotic vegetarians who have been followed from birth to adolescence, and who continue to have some health problems. More studies are needed so that long-lasting effects of diet early in life can be better ascertained.

Some vegetarian parents feed their children diets that are inadequate. The problem is not that nutritionally adequate diets cannot be planned, but that the eater's or cook's ideologies and concerns may get in the way. Thus, actual diets, as eaten, may not conform to the recommendations of expert nutritional bodies. Under such circumstances health problems have arisen and continue to do so, especially among infants and children on vegan diets that are limited in other foods as well.

Vegan diets present more problems of micronutrient adequacy than do other vegetarian diets across the life cycle and particularly in infants and children because more food groups are eliminated, sources of vitamins B_{12} and D may be lacking and the caloric density of the diet is lower and bulk higher than that of vegetarians or omnivorous infants. They may also have diets that are limited in other foods, and thus in nutrients such as calcium and iron.

Vegetarian infants are usually breast-fed. They generally thrive until 4–6 months of age, and continue to do so if they are weaned to diets containing cows' milk-based or soy formulas and sufficient food energy. In countries where infant formulas do not provide adequate amounts of micronutrients, dietary supplements may be needed. Today, many more fortified vegan foods are widely available than in the past. They are helpful in meeting the nutrient needs of weanlings and toddlers. Also, dietary supplements are available, and for some vegans these are acceptable alternatives. Good sources of vitamin B_{12} need to be identified and also of vitamin D if exposures to sunlight are not adequate. Sources of other nutrients such as linolenic acid, the omega 3 fatty acid, should be included so that docosahexanenoic acid (DHA) intakes are satisfactory. Vegan diets may also be low in calcium, iron, and zinc, and the forms in which iron and zinc are present may not be highly bioavailable. A source of riboflavin also needs to be identified.

Vegetarian diets need to be carefully monitored when they are fed to young children. Soy milk is not appropriate under 1 year of age. When soy milk is used later in childhood, especially if the child is a vegan, it should be fortified with vitamins D and B_{12} and calcium.

For pubertal children, energy, calcium, iron, vitamins D and B_{12}, as well as iron are of particular concern with respect to dietary adequacy, but these can usually be dealt with by dietary planning.

Current Vegetarian Eating Patterns and Practices

Until about 40 years ago, in Western countries virtually all of the common vegetarian eating patterns involved avoidance of animal flesh (meat and poultry); categorization of vegetarian patterns was relatively straightforward and consisted simply of differentiating between those who ate no animal foods at all (vegan vegetarians), those who also consumed milk and milk products (lacto vegetarians), and those who ate eggs as well (lacto-ovo vegetarians). This simple categorization scheme broke down in the 1960s and 1970s as a result of greater exposure to the cuisines of other cultures, new Eastern religions and philosophical systems with a vegetarian tradition, and other influences, which led to the emergence of new patterns of vegetarianism.

Today, myriad vegetarian eating patterns exist, and they cannot be easily described by focusing on a single dimension, such as animal food intake.

Meatless and vegetarian eating patterns and life styles are growing in popularity today. They continue to be fostered by a greater availability and variety of meat alternatives and analogs for animal products. There is also a good deal of favorable publicity about phytochemicals with supposedly beneficial health effects. At the same time, concerns about the healthfulness of animal foods have been triggered by publicity on the bovine spongiform encephalopathy (BSE) epidemic in the UK, a later epidemic of hoof and mouth disease in cattle, and most recently an epidemic of SARS spread from animals to people. Worries about saturated fat/trans fat coronary artery disease links, dietary fat and cancers, food safety, and other factors probably also contributed to the increased prevalence of vegetarian eating.

At the same time, vegetarian eating patterns are much more heterogeneous today than in the past. The availability and variety of plant foods, as well as commercially available and tasty meat analogs has greatly increased. Fortified foods today include soy milks fortified with vitamins B_{12} and D and a highly bioavailable form of calcium, and highly fortified breakfast cereals. These foods and nutrient-containing dietary supplements make it easier for vegans and vegetarians to obtain nutrients that would otherwise be low or lacking.

Conformity to Nutritional Recommendations

Well-planned vegetarian diets have nutritional profiles that are in line with recent expert recommendations. A well-planned vegetarian diet pattern, if sustained throughout adulthood, may reduce risks of coronary artery and other chronic degenerative diseases associated with excessive weight. Generally, vegetarian diets tend to be low in saturated fat and

cholesterol and high in complex carbohydrates, dietary fiber, magnesium, potassium, folic acid, and antioxidant nutrients such as vitamins C and E. They also tend to be relatively low in energy. Thus, the diet-related risks for a number of chronic degenerative diseases associated with intakes of these nutrients may be decreased on vegetarian diets. Some risks are clearly lower; for example, vegetarians generally tend to have lower weight for height than do nonvegetarians. Constipation tends to be less of a problem in this group, perhaps due in part to the higher intake of dietary fiber.

Conclusions

Vegetarian diets should be planned in accordance with expert nutritional recommendations. When this is followed, such diets are healthful and nutritionally adequate. When they are not planned, the nutrients that are likely to fall short usually differ somewhat from those on nonvegetarian diets. In some cases these deficits can be remedied by dietary counseling. In others differences between ideologies about diet and nutrient needs are such that acceptable dietary strategies cannot be found.

Nutrition scientists and practitioners can help vegetarians who seek their advice by monitoring the nutritional status of high-risk individuals, by identifying food sources of specific nutrients, and by suggesting dietary modifications that may be necessary to meet individual needs when intakes fall short.

See also: **Adolescents**: Nutritional Problems. **Anemia**: Iron-Deficiency Anemia. **Cancer**: Epidemiology and Associations Between Diet and Cancer. **Cobalamins**. **Meat, Poultry and Meat Products**. **Phytochemicals**: Classification and Occurrence; Epidemiological Factors. **Religious Customs, Influence on Diet**. **Supplementation**: Dietary Supplements.

Further Reading

Abrams HL (2000) Vegetarianism: another view. In: Kiple F and Connee K (eds.) *The Cambridge World History of Food*, pp. 1564–1573. Cambridge: Cambridge University Press.

American Dietetic Association and Dietitians of Canada (2003) Position of the American Dietetic Association and Dietitians of Canada: Vegetarian diets. *Canadian Journal of Diet Practice Research* **64**: 62–81.

Draper A, Lewis J, Malhotra N, and Wheeler E (1993) The energy and nutrient intakes of different types of vegetarian: a case for supplements? *British Journal of Nutrition* **70**: 812

Dwyer JT and Jacobs C (1988) Vegetarian children: appropriate and inappropriate diets. *American Journal of Clinical Nutrition* **48**: 811S–818S.

Engs R (2000) *Clean Living Movements: American Cycles of Health Reform*, pp. 1–20. Westport, CN: Praeger Publishers.

Fernandez-Armesto F (2002) *Near a Thousand Tables* New York: Free Press.

Haddad E (1994) Development of a vegetarian food guide. *American Journal of Clinical Nutrition* **59**: 1248S–1254S.

Havala S and Dwyer J (1994) Position of the American Dietetic Association: vegetarian diets. *Journal of the American Dietetic Association* **93**: 1317–1319.

Institute of Medicine (2000) *Dietary Reference Intakes: Applications in Dietary Assessment*. Washington DC: National Academy Press.

Key T, Davey G, and Appleby PN (1999) Health benefits of a vegetarian diet. *Proceedings of the Nutrition Society* **58**: 271–275.

Key T, Fraser GE, Thorogood M, Appleby PN, Beral V, Reeves G *et al.* (1998) Mortality in vegetarians and nonvegetarians: detailed findings from a collaborative analysis of 5 prospective studies. *American Journal of Clinical Nutrition* **70**: S516–S524.

Levenstein H (1993) *Paradox of Plenty: A Social History of Eating in Modern America* Oxford: Oxford University Press.

Mangels AR and Messina V (2001) Considerations in planning vegan diets: infants. *Journal of the American Dietetic Association* **101**: 670–677.

Messina V and Mangels AR (2001) Considerations in planning vegan diets: children. *Journal of the American Dietetic Association* **212**: 661–669.

Messina V and Mangels AR (2001) Considerations in planning vegan diets: children. *Journal of the American Dietetic Association* **101**: 661–669.

Messina V, Melina V, and Mangels AR (2003) A new food guide for North American Vegetarians. *Canadian Journal of Diet in Practice and Research* **64**: 82–86.

Obarzanek E, Sacks FM, Vollmer WM, Bray GA, Miller ER, Lin PH, Karanja NM, Most-Windhouser MM, Moore TJ, Swain JF, Bales CW, and Proschan MA on behalf of the DASH Research Group (2001) Effects on blood lipids of a blood pressure lowering diet: the Dietary Approaches to Stop Hypertension (DASH) Trial. *American Journal of Clinical Nutrition* **74**: 80–89.

Sanders T (1999) Essential fatty acid requirements of vegetarians in pregnancy, lactation and infancy. *American Journal of Clinical Nutrition* **70**: 555S–559S.

Van Dusseldorp M, Scheede J, Refsum H, Ueland PM, Thomas CMG, de Boer E, and Van Staveren WA (1999) Risk of persistent cobalamin deficiency in adolescents fed a macrobiotic diet in early life. *American Journal of Clinical Nutrition* **609**: 664–671.

Van Staveren W, Dhuyvetter JHM, Bons A, Zeelen M, and Hautvast JGAJ (1985) Food consumption and height/weight of Dutch preschool children on alternative diets. *Journal of the American Dietetic Association* **85**: 1579–1584.

Whorton J (2000) Vegetarianism. In: Kiple F and Connee K (eds.) *The Cambridge World History of Food*, pp. 1553–1565. Cambridge: Cambridge University Press.

VITAMIN A

Contents
Physiology
Biochemistry and Physiological Role
Deficiency and Interventions

Physiology

A C Ross, The Pennsylvania State University, University Park, PA, USA

Introduction

Vitamin A is a fat-soluble micronutrient that is required by all vertebrates to maintain vision, epithelial tissues, immune functions, reproduction, and for life itself. It was discovered in 1913 as a minor component in eggs, butter, whole milk, and fish liver oils. It soon became apparent that vitamin A exists in two chemically distinct yet structurally related forms. The first form to be characterized was retinol, a lipid alcohol that is present only in foods of animal origin. Retinol is also known as 'preformed vitamin A' because it can be metabolized directly into compounds that exert the biological effects of vitamin A. A second form of vitamin A, present in deep-yellow vegetables, was characterized as β-carotene, which is synthesized only by plants but can be converted to retinol during absorption in the small intestines. These carotenoids are sometimes referred to as 'provitamin A.' The nutritional requirement for vitamin A can be met by preformed retinol, provitamin A carotenoids, or a mixture, and therefore it is possible to obtain a sufficient intake of vitamin A from carnivorous, herbivorous, or omnivorous diets.

Neither retinol nor the provitamin A carotenoids are directly bioactive. Retinol must be activated in a series of oxidative reactions, while the provitamin A carotenoids must first be cleaved to produce retinol. Of numerous metabolites of vitamin A, two are well recognized as crucial to its physiological functions. 11-*cis*-retinaldehyde (retinal) is a component of the visual pigment required for vision, rhodopsin. Retinoic acid, an acidic derivative, is required for the regulation of gene expression in essentially all tissues.

Besides the natural forms of vitamin A, a large number of structurally related analogs of vitamin A have been synthesized as potential therapeutic agents. The term 'retinoid' applies to both natural dietary forms of vitamin A, its metabolites, and those synthetic analogs that possess some, but usually not all, of the biological activities of vitamin A.

Chemistry

Vitamin A and its metabolites comprise a group of more than a dozen molecules that differ in isomeric form, oxidation state, and whether they are unesterified (free), esterified with a fatty acid, or conjugated.

All-*trans*-Retinol (Vitamin A Alcohol)

Retinol, the parent molecule of the vitamin A family, is a fat-soluble lipid alcohol ($C_{20}H_{30}0$, molecular mass 286.4) composed of a methyl substituted cyclohexenyl (β-ionone) ring, an 11-carbon conjugated tetraene side chain, and a terminal hydroxyl group (**Figure 1A**, R_1). Most of the double bonds can exist in either *trans* or *cis* conformation. All-*trans*-retinol is the most stable and most prevalent form in foods and tissues, but small amounts of other geometric isomers such as 9-*cis*- and 13-*cis*-retinol are found in some cells. The terminal hydroxyl group of retinol can be free or esterified with a fatty acid. Esterification reduces the susceptibility of retinol to oxidation and changes its physical state from a crystalline lipid to an oil. Fatty acid esters of retinol (**Figure 1A**, R_2) are the predominant form of vitamin A in most tissues. In some pharmaceutical products, retinol is present as retinyl acetate. Variant forms of vitamin A are present in some foods and human tissues. For example, vitamin A_2, (3,4-didehydroretinol) is present in freshwater fish, and is also a product of retinol metabolism in human skin.

Oxidized Metabolites of Retinol

Figure 2 illustrates key steps in the metabolism of vitamin A. Retinol is oxidized within cells to generate retinal (**Figure 1A**, R_3) and retinoic acid (**Figure 1A**, R_4). 11-*cis*-Retinal, the isomer of retinal

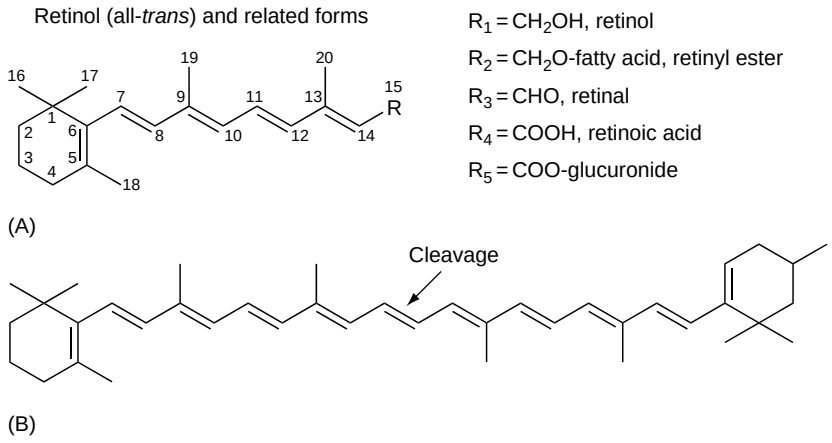


Figure 1 (A) Structure of all-*trans*-retinol and several related forms. (B) Beta-carotene (all-*trans*) showing the position of 15,15′ double bond that through cleavage yields retinal, which can be reduced to form retinal, giving rise to all of the structures indicated in **Figure 1A**.

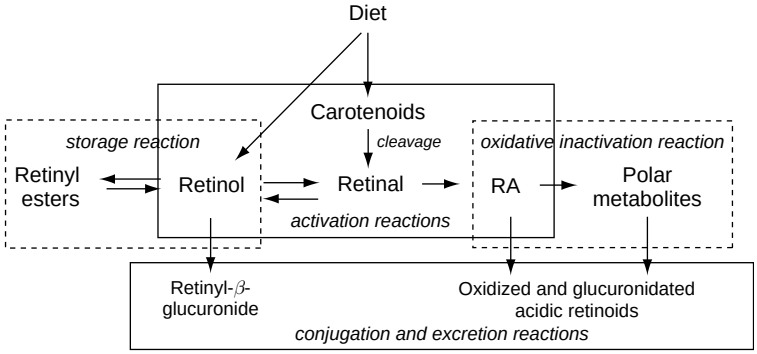


Figure 2 Schematic of principal reactions of vitamin A metabolism.

critical for vision, absorbs light maximally at ~365 nm when in organic solvent, but when coupled with a protein, such as opsin, its peak absorptivity is shifted into the visible range of the electromagnetic spectrum (see 'Vision'). In its all-*trans* isomeric form, retinal is a transient intermediate in the bioconversion of retinol to retinoic acid. Retinoic acid exists in several isomeric forms, two of which (all-*trans*-retinoic acid and 9-*cis*-retinoic acid) interact specifically with nuclear receptor proteins.

Numerous metabolites of retinol or retinoic acid are more polar than retinol or retinoic acid due to additional oxidation of the cyclohexenyl ring, often on carbon 4. Some retinoids, particularly retinoic acid and 4-keto-retinoic acid, may be conjugated with glucuronic acid, forming retinyl- or retinoyl-β-glucuronide (R_5); these metabolites are water-soluble and therefore readily excreted. While some polar and water-soluble retinoids possess bioactivity, most show reduced, or no, activity compared to their precursors.

Carotenoids

Carotenoids are produced only by plants and a few microalgae. In plants, they function as accessory light-gathering pigments that enhance the efficiency of photosynthesis. Of the 600 or so carotenoids found in nature, only β-carotene, α-carotene, and β-cryptoxanthin have the structural features necessary for vitamin A activity. Beta-carotene is a hydrocarbon ($C_{40}H_{56}$, molecular mass 536) with two β-ionone rings, a polyene chain, and structural symmetry around the central 15,15′ double bond (**Figure 1B**). The oxidative cleavage of this bond releases two molecules of retinal, which can be reduced to form vitamin A (retinol). Other isomers of β-carotene with potential nutritional activity include 9-*cis*-β-carotene produced by certain microalgae. Other common carotenoids found in fruits and vegetables, such as lycopene, lutein, and zeaxanthin, are absorbable but they lack structural features essential for vitamin A activity.

Dietary Sources and Nutritional Equivalency

Preformed vitamin A is present at highest concentration in liver and fish oils, and at lower concentrations in nonorgan meats. Food sources of preformed vitamin A and provitamin A are provided in **Table 1**. In 1990, 39% of the vitamin A (including carotenes) in the diets of Americans came from fruits and vegetables. Meats and dairy products each supplied about 20% of the vitamin A consumed. Foods that contain small amounts of vitamin A can still contribute significant amounts of vitamin A to an individual's diet if they are consumed often or in large amounts.

Table 1 Food sources of vitamin A

Food	*%DV**
Animal sources of preformed vitamin A	
Liver, beef, cooked, 3 oz	610
Liver, chicken, cooked, 3 oz	280
Fat-free milk, fortified with vitamin A, 1 cup	10
Cheese pizza, 1/8 of a 12-inch diameter pie	8
Milk, whole (3.25% fat;), 1 cup	6
Cheddar cheese, 1 oz	6
Whole egg, 1 medium	6
Plant sources of β -carotene and other provitamin A carotenoids	
Carrot, 1 raw (7.5 inches long)	410
Carrots, boiled, 1/2 cup, slices	380
Carrot juice, canned, 1/2 cup	260
Sweet potatoes, canned, drained solids, 1/2 cup	140
Spinach, frozen, boiled, 1/2 cup	150
Mango, raw, 1 cup, sliced	130
Vegetable soup, canned, chunky, ready-to-serve, 1 cup	115
Cantaloupe, raw, 1 cup	100
Kale, frozen, boiled, 1/2 cup	80
Spinach, raw, 1 cup	40
Apricot nectar, canned, 1/2 cup	35
Oatmeal, instant, fortified, plain, water, 1 packet	30
Tomato juice, canned, 6 oz	20
Apricots, with skin, juice pack, 2 halves	10
Pepper, sweet, red, raw, 1 ring, 3-inch diam. 1/4-inch thick	10
Peas, frozen, boiled, 1/2 cup	10
Peach, raw, 1 medium	10
Peaches, canned, water pack, 1/2 cup halves or sliced	10
Papaya, raw, 1 cup, cubes	8

*% DV = Daily Value. %DVs are reference numbers based on the Recommended Dietary Allowance (RDA), Percent DVs are based on a 2000 calorie diet. They were developed to help consumers determine if a food contains a lot or a little of a specific nutrient. The DV for vitamin A is 5 000 IU (1 500 micrograms retinol which is 1500 μg RAE). Most food labels do not list a food's vitamin A content. The %DV listed in **Table 1** refer to the vitamin A provided in one serving. Data from Clinical Nutritional Service (2003) *Facts about Dietry Supplements*. Maryland, USA: Warren Grant Magnuson Clinical Center.

Units of Nutritional Activity

Owing to the multiple forms of vitamin A in most diets and the lower efficiency of utilization of carotenoids compared to preformed vitamin A, the total amount of vitamin A (bioactivity) in foods or in the total diet must be expressed in equivalents. Over the years, several equivalency units and conversion factors have been adopted. Most recently, the retinol activity equivalent was adopted by the Institute of Medicine (IOM) in 2001 to replace older units of bioactivity because new information indicated that the conversion of carotenoids is less efficient than previously thought. One microgram of retinol activity equivalent (RAE) is equivalent in terms of activity to 1 μg of all-*trans*-retinol or 2 μg of β-carotene in oily solution. One microgram of RAE is also equivalent to higher amounts of other provitamin A carotenoids in foods because they are less bioavailable than β-carotene in oil. On average, carotenoids must be ingested in the following amounts to provide the equivalent nutritional value of 1 μg of all-*trans*-retinol:

- 2 μg of supplemental β-carotene (in an oily, easily absorbed solution);
- 12 μg of β-carotene in fruits and vegetables (due to association with food matrices); and
- 24 μg of α-carotene or β-cryptoxanthin (due to food matrices and structure of compounds).

Prior to 2001, the retinol equivalent (RE) was used and this unit is still found in most food composition tables. While similar in theory to the RAE, the RE is based on older conversion factors for carotenoids in foods. Using RAE, the vitamin A activity of the provitamin A carotenoids in foods is half that using RE. An older unit, the international unit (IU or USP), which should eventually be replaced by these newer units, is still used in food tables and on some supplement labels. One IU is equal to 0.3 μg of all-*trans*-retinol. Finally, another indicator of nutritional value, % daily value (%DV), is a less quantitative but more convenient means for consumers to compare foods and select those with a substantial portion of a given nutrient. The %DV does not require extensive knowledge of nutritional units; this value appears on food package labels in the US. Besides its application in food labeling, the %DV is a useful value for quickly comparing the vitamin A contents of various common foods.

Transport and Metabolism

Few retinoids are appreciably soluble in water or aqueous body fluids. They gain solubility through association with specific proteins.

Retinol-Binding Protein

Plasma retinol is transported by a specific 21-kDa transport protein, retinal binding protein (RBP). Most RBP is produced in the liver, but some extrahepatic organs also synthesize it. Each molecule of RBP binds one molecule of all-*trans*-retinol noncovalently. In plasma, the retinol-RBP complex (holo-RBP) forms a larger complex with a cotransport protein, transthyretin (TTR), which also binds the hormone thyroxine.

Cellular Retinoid-Binding Proteins

Cellular retinoid-binding proteins (CRBP) are present in the cytoplasm of many types of cells. These proteins are similar in structure and size (~14.6 kDa), and each contains a single binding site that preferentially binds a particular form of retinoid (retinol, retinal, or retinoic acid), often preferring a specific isomer. Four cellular retinol-binding proteins (CRBP-I, -II, -III, and -IV) and two cellular retinoic acid-binding proteins (CRABP-I and -II) are expressed in many cells, yet each has a different tissue distribution. These proteins function as chaperones that confer aqueous solubility on otherwise insoluble retinoids while directing them to specific enzymes that then catalyze their further metabolism. These binding proteins may also play a role, although it is not yet well defined, in the delivery of retinoids (RA) to the nucleus for binding to nuclear retinoid receptors, which are discussed later.

Intestinal Metabolism

Dietary retinyl esters must be hydrolyzed in the lumen of the small intestine before retinol is absorbed, while carotenoids must be absorbed into the intestinal mucosa before being cleaved intracellularly. Several enzymes with retinyl ester hydrolase (REH) activity are present in pancreatic juice or on the brush border of duodenal and jujenal enterocytes (**Figure 3**). Retinol and carotenoids must be solubilized in the lumen in mixed micelles composed of bile acids and products of lipid digestion prior to their uptake into enterocytes. These processes require the release of an adequate amount of bile salts and a minimal quantity of dietary fat (approximately 5%), which must be consumed concomitantly. The retinol thus liberated diffuses into the enterocyte, is bound by CRBP-II, and is then esterified. The newly formed retinyl esters are incorporated into the lipid core of chylomicra, lipoproteins that transport dietary fat into the lymphatic system for absorption. The overall efficiency of retinol absorption is quite high, about 70–90%, and is not significantly downregulated as vitamin A consumption increases.

The efficiency of absorption of β-carotene is considerably lower (9–22%) and more variable than that of retinol. In fact, in controlled studies some subjects have absorbed little, if any, of a test dose of β-carotene. In individuals who do absorb dietary carotenoids, the efficiency of absorption tends to fall as intake increases. The type of carotenoid and its physical form in the ingested foodstuff also affect the efficiency of carotene absorption. Pure β-carotene in an oily solution or supplements is absorbed more efficiently than an equivalent amount of β-carotene in foods. Much of the carotenoid present in foods is bound within a matrix of

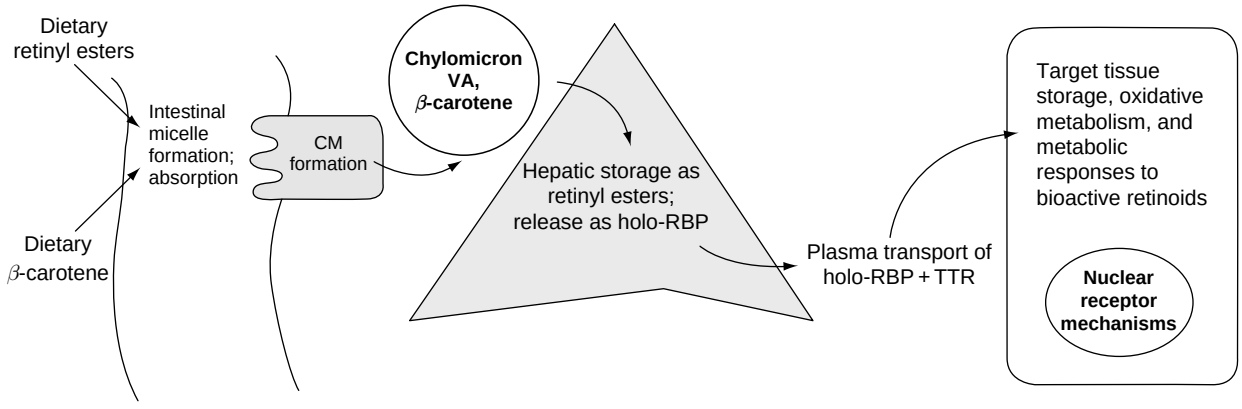


Figure 3 Absorption of dietary vitamin A (VA) via chylomicrons (CM), vitamin A storage in liver, and the release of retinal to plasma as holo-retinol-binding protein (RBP), which combines with transthyretin (TTR), to deliver retinal to organs that produce retinal (eyes) or retinoic acid (essentially all tissues) for the biological functions attributed to vitamin A.

polysaccharides, fibers, and phenolic compounds that is incompletely digested. Although the absorption of provitamin A carotenoids from fruits is generally better than from fibrous vegetables, it is still low as compared to β-carotene in oil (see section on Units).

Once in the enterocyte, provitamin A carotenoids are cleaved by one or more carotene monooxygenases, and the product (initially retinal) is metabolized to form retinol and, subsequently, retinyl esters (see **Figure 2**). In humans about one-third of ingested β-carotene escapes cleavage and, instead, is incorporated intact into chylomicrons.

A small fraction of intestinal β-carotene is oxidized to retinoic acid and absorbed into portal blood. It is speculated that the cleavage of dietary 9-*cis*-β-carotene and its subsequent oxidative metabolism may be a significant source of 9-*cis*-retinoic acid.

Hepatic Vitamin A Uptake, Storage and Release

Once chylomicra enter lymph and plasma, chylomicron remnants are formed rapidly by lipolysis. The majority of chylomicron remnants, still containing vitamin A, are quickly cleared into liver parenchymal cells (hepatocytes) by receptor-mediated endocytosis. Adipose and other tissues, including the mammary glands during lactation, also take up small amounts of newly absorbed vitamin A during the lipolysis of chylomicra. Within a few hours of chylomicron remnant clearance by hepatocytes, most of these newly absorbed retinyl esters are hydrolyzed and the retinol component is re-esterified, forming new retinyl esters. Newly formed retinyl palmitate and stearate are deposited in lipid droplets in vitamin A-storing stellate cells. In the vitamin A adequate state, more than 90% of total body vitamin A is stored in liver stellate cells. Small numbers of similar stellate cells have been described in extrahepatic tissues, suggesting the presence of a network of vitamin A-storing cells throughout the body.

As retinol is needed, stellate cell retinyl esters are hydrolyzed by one or more REHs and retinol is transferred back to hepatocytes for combination with newly synthesized RBP. The holo-RBP complex then passes through the Golgi secretory apparatus and binds noncovalently with a tetramer of TTR. The larger size of this transport complex (~75 kDa) compared to holo-RBP alone (~21 kDa) helps to prevent the rapid loss of retinol and RBP during renal glomerular filtration.

Plasma Concentrations

In normal plasma in the fasting state, more than 95% of retinol is bound to RBP. Retinyl esters also are present during the absorption of vitamin A-rich meals, but they are bound to plasma lipoproteins.

Although there is a significant relationship between plasma retinol and liver vitamin A storage (considered a 'gold standard' for assessing vitamin A status), the relationship is by no means linear. In fact, plasma retinol is nearly constant over a rather wide range of liver vitamin A concentrations, all consistent with vitamin A adequacy. The constancy of plasma retinol reflects its close homeostatic regulation. Plasma retinol levels in normal adults show only minor day-to-day variations, staying close to ~2 and 1.7 $\mu mol\,l^{-1}$ in males and females, respectively. The molar concentration of RBP (1.9–2.4 $\mu mol\,l^{-1}$) is slightly higher and thus RBP is normally 80–90% saturated. A small proportion of circulating RBP is apo-RBP (RBP without retinol). Owing to its reduced affinity for TTR, apo-RBP is readily filtered in the kidneys and catabolized.

When liver vitamin A reserves fall below about 20–30 μg retinol g liver, the secretion of holo-RBP is compromised due to inadequate retinol. Plasma retinol levels begin to fall and, if liver vitamin A continues to decline, plasma levels will fall into the deficient range and will be inadequate to supply retinol to tissues. Essentially all of the vitamin A in liver can be mobilized when it is needed to meet the needs of peripheral tissues. But ultimately, vitamin A intake must increase to bring plasma retinol levels back to the normal range.

Conversely, when vitamin A is consumed in excess of needs, its concentration in liver can increase markedly. When the concentration rises above about 300 $\mu g\,g^{-1}$, as occurs in hypervitaminosis A (see later section), the levels of plasma retinol and RBP remain almost normal but total vitamin A increases due to retinyl esters bound to plasma lipoproteins.

Plasma Vitamin A Kinetics

Both RBP and TTR have a relatively short half-life (~0.5 and 2–3 days, respectively) and, therefore, they must be synthesized continuously to maintain normal plasma levels. Plasma retinol, RBP, and TTR are reduced in states of impaired protein synthesis, which may be due to an inadequate intake of protein or energy or to impairments in metabolism. Plasma RBP and TTR are sometimes used as clinical indicators of visceral protein synthesis. During infection and/or inflammation, plasma retinol, RBP, and TTR fall transiently, even though liver vitamin A reserves

are adequate, due to altered protein synthesis during the acute-phase response. Because multiple nutritional and metabolic disturbances can lead to a similar decrease in plasma retinol, RBP, and TTR, laboratory values must be interpreted with caution.

Studies using computer-based compartmental modeling to analyze plasma retinol kinetics have shown that each molecule of retinol circulates through the plasma compartment several times before it is irreversibly degraded (see 'Tissue Retinoid Metabolism'). In a young man who consumed 105 μmol of retinyl palmitate in a test meal, 50 μmol of retinol passed through his plasma per day, while only 4 μmol day^{-1} was degraded. Unlike retinol, RBP is not recycled, implying that RBP is synthesized in extrahepatic tissues for the release and continued recycling of retinol. Some extrahepatic tissues, such as kidney and adipose, contain RBP mRNA at a level ~5–10% of that in liver. The kidneys evidently play a very significant role in the recycling and conservation of retinol after the glomerular filtration of holo-RBP. Cell culture studies have shown that holo-RBP can bind to renal epithelial cells and cross the epithelium by transcytosis, suggesting a mechanism for the recovery of retinol lost by filtration.

Overall, the body is efficient at conserving retinol, but relatively inefficient in degrading and eliminating excess retinoids. These differences seem to explain the propensity for retinyl esters to accumulate in tissues when vitamin A is consumed in amounts that substantially exceed requirements.

Carotenoids

Carotenoids circulate in plasma in association with low-density and high-density lipoproteins. The level of β-carotene reflects its recent intake, but it also is higher when plasma lipoprotein levels are elevated. Beta-carotene is stored at relatively low concentrations in liver and fatty tissues. A prolonged slow rate of postabsorptive conversion to retinol has been observed in volunteers in isotope kinetic studies.

Tissue Retinoid Metabolism

Tissues obtain retinol from holo-RBP throughout the day, and retinyl esters from chylomicrons and chylomicron remnants after consumption of vitamin A-containing meals. Although the majority of the body's vitamin A is stored in the liver, many organs contain small reserves of retinyl esters. These small local supplies are believed to be critical for the generation of bioactive retinoids, formed through oxidative metabolism (**Figure 2**). Retinol that is liberated by the hydrolysis of retinyl esters is oxidized to retinoic acid in a two-step process in which retinal is an obligate intermediate. The oxidation of retinol to retinal has been attributed to several enzymes of the alcohol dehydrogenase and the short-chain dehydrogenase/reductase gene families. Both of these types of enzymes also oxidize other substrates and the specifics of retinol oxidation in various tissues have been difficult to define. It is likely that the CRBP proteins function as chaperones for retinol during its oxidative metabolism. Since retinal can be reduced to form retinol, the retinal that is generated from the metabolism of carotenoids can give rise to retinyl esters and all of the other metabolites of retinol (see **Figure 2**). In the second oxidative step, which is irreversible, retinal is converted to retinoic acid. This step also may be catalyzed by several enzymes. Retinoic acid is present, sometimes in several isomeric forms, at nanomolar concentrations in many tissues. Its half-life is very short, a few hours or less, which implies that it is produced continuously to maintain tissue retinoic acid levels.

The β-ionone ring of retinol and retinoic acid also can be oxidized, usually at carbon 4 to form 4-hydroxy and 4-oxo metabolites. The metabolism of retinoic acid is, in part, autoregulated due to the ability of retinoic acid to induce the expression of cytochrome P450 enzymes that form 4-oxo derivatives of retinoic acid. At this time, although limited evidence suggests that ring-oxidized retinoids still possess bioactivity, most evidence supports the thinking that they are metabolites in a catabolic pathway, destined for excretion. Ring-oxidized metabolites, which are normally present in plasma in low concentrations, are readily removed by the liver and conjugated with glucuronic acid, which makes them soluble in water. Glucuronides comprise a substantial fraction of the total retinoid excreted in bile and eliminated by the fecal route.

Production and oxidation both serve to maintain normal tissue retinoid levels. Nonetheless, these physiological processes can be overwhelmed by an excess of dietary vitamin A, resulting in excessive vitamin A in tissues and plasma, or by the use of synthetic retinoids that lead to substantial elevations in tissue bioactive retinoids (see the article section 'Hypervitaminosis A and Vitamin A Toxicity').

Physiological Actions

Vision

The retinal pigment epithelium cells (RPE) of the retina form an epithelial cell layer that takes up retinol from choroid capillaries and stores it as

retinyl esters, to be used as substrate for the generation of 11-*cis*-retinal. In the layer of rod and cone photoreceptor cells adjacent to the RPE, 11-*cis*-retinal combines covalently with the protein opsin to generate the visual pigment rhodopsin in rods and, similarly, iodopsin in cones. Each rod outer segment is densely packed with some 10^8 molecules of rhodopsin per cell. The small quantity of vitamin A stored in the retina would be inadequate to maintain vision were it not for the visual cycle, a process in which 11-*cis*-retinal is regenerated after photobleaching. The absorption of light by rhodopsin catalyzes the photoisomerization of the 11-*cis*-retinal moiety of rhodopsin to all-*trans*-retinal (resulting in bleaching), which induces the release of all-*trans*-retinal from opsin. The change in retinal's isomeric configuration is crucial for initiating a signal transduction cascade from the rods to nearby retinal ganglion cells, and thereafter to the optic nerve for transmission to the brain's visual cortex. For vision to continue, 11-*cis*-retinal must be regenerated. This is accomplished in a series of biochemical reactions constituting the visual cycle, some of which take place in the rod cell outer segment and others in the RPE. The regeneration of 11-*cis*-retinal (dark adaptation) is slow (on the order of minutes) compared to the photoisomerization (fractions of a second). However, normal vision continues without a period of blindness as long as retinol can be drawn from retinyl esters stored in the RPE, rapidly isomerized to 11-*cis*-retinol, re-oxidized to 11-*cis*-retinal, and passed to the rod cell outer membrane where rhodopsin is regenerated. When the supply of retinyl esters in the RPE is not adequate, there is significant slowing of the visual cycle, resulting in the condition known as night blindness, a loss of the ability to quickly dark adapt after exposure to bright light. Night blindness is often the first-detected clinical sign of vitamin A deficiency (see the article section 'Hypervitaminosis A and Vitamin A Toxicity').

Cornea The cornea, an avascular tissue, requires retinoic acid for the normal differentiation of the corneal and conjunctival epithelium. Holo-RBP, which is present in the lacrimal glands and tears, is likely to provide the substrate for the local biogenesis of retinoic acid. Retinoid deficiency results first in a loss of goblet secretory cells, which can be detected histologically. Corneal xerosis and Bitôt's spots (foamy deposits) are strong signs of prolonged vitamin A deficiency (see 'Hypervitaminosis A and Vitamin A Toxicity'). Vitamin A must be administered immediately to prevent the progression of corneal xerosis to corneal ulceration, which causes life-long blindness.

Cell Morphology and Differentiation

Soon after the discovery of vitamin A, a light microscopic investigation of the tissues of vitamin A-deficient rats revealed marked abnormalities in many epithelial tissues. It is now recognized that essentially all organ systems require retinoids. Some epithelial tissues (skin, respiratory tract, the immune system, the reproductive organs, etc.) are especially sensitive to a lack, as well as an excess, of vitamin A. The systemic effects of vitamin A deficiency include dryness of the skin (follicular hyperkarotosis), loss of mucus-secreting goblet cells in the trachea and respiratory tract, and a generalized flattening of epithelia (squamous metaplasia, sometimes with keratinization) throughout the body. The hematopoietic system is also affected, as are reproductive organs. In the testes, spermatogenesis is inhibited by vitamin A deficiency. Although a lack or an excess of retinoids is recognized to affect many organ systems, the developing embryo and the functions of the immune system have been studied most intensively. Essentially all of the functions of vitamin A other than those involving the retina are mediated by its active metabolite, RA, in conjunction with nuclear retinoid receptors.

Nuclear Retinoid Receptors

The nuclear retinoid receptor proteins are synthesized in the cytoplasm but reside in the nucleus where they form dimers capable of binding to specific sequences of DNA in target genes (retinoid response elements, RAREs). The family consists of six retinoid receptors (RARα, β and □, and RXR□□ and □) that belong to the superfamily of steroid hormone receptors. The RAR and RXR function as ligand-activated transcription factors to either activate or repress the transcription of hormone-responsive genes. Two isomers of retinoic acid, all-*trans*-retinoic acid and 9-*cis*-retinoic acid, function as the major ligands for the RAR and RXR subfamilies, respectively. The binding of all-*trans*-retinoic acid to the RAR induces a conformational change in the receptor dimer pair, bound to its response elements (RARE) in the regulatory region of the DNA of target genes. This conformational change, in turn, promotes the interaction of the retinoid receptor dimer with other transcriptional regulators. Ligand binding may promote the dissociation of corepressor molecules and the binding of coactivator molecules, leading to gene activation when the basal transcriptional complex is recruited. This multiprotein complex then functions enzymatically to transcribe the DNA template into messenger RNA. Additionally, some receptor functions appear

to be ligand independent. Similarly to all-*trans*-retinoic acid, 9-*cis*-retinoic acid has been shown to bind to nuclear receptors of the RXR family. However, the physiological role of 9-*cis*-retinoic acid *in vivo* is currently unclear and, moreover, other ligands besides 9-*cis*-RA (such as polyunsaturated fatty acids) may also activate the RXR. Besides forming dimers with the RAR, the RXR bind in a similar manner with the nuclear receptors for vitamin D, thyroid hormone, and several other lipophilic hormones and xenobiotic agents.

Embryonic Development

Vitamin A is essential in appropriate amounts for normal embryogenesis. Retinoids are required from the early, postgastrulation stage of embryonic development. The requirement for retinoids has been deduced from molecular developmental studies in mice, and other species. These studies have consistently shown a highly regulated pattern of expression of the genes for nuclear retinoid receptors, retinoid-binding proteins, and enzymes of retinoid production and catabolism. It is likely, based on the expression of retinoid biosynthetic enzymes, that maternally derived retinol is metabolized by the embryo to produce retinoic acid at specific times in specific cells, and that retinoic acid is also catabolized in a highly regulated, tissue-specific manner. Retinoic acid has been proposed to be an essential morphogen whose concentration, or concentration gradient, is a key determinant of the expression of one or more families of genes, particularly the *Hox* gene family. This gene family is crucial in determining the formation of the anterior-posterior body axis. Some of these key genes contain a RARE.

In animals, both vitamin A deficiency and an excess of dietary vitamin A or retinoid analogs, at specific critical periods of development, can result in severe developmental defects, and may be lethal to the embryo. The differentiation of cells in the neural crest and the development of the head and sensory organs, nervous system, heart, limbs, and skeletomuscular system are often affected. Birth defects of a similar nature have occurred in women exposed to excessive dietary vitamin A, or to pharmacologic retinoids for treatment of skin diseases, in early pregnancy.

Immunity

Impaired immunity was one of the earliest effects described for vitamin A deficiency. Numerous abnormalities have been described. A dysregulation of T cell functions has been implicated in many abnormal immune responses, as vitamin A-deficient animals often have reduced T cell counts and an altered pattern of differentiation markers on T cell subsets. The response of T cells to antigens and mitogens tends to be low. Similarly, the functional capacity of cytotoxic cells, such as cytotoxic T cells and natural killer cells, and macrophages is often low. Numerous alterations have been documented in the production of cytokines that regulate T cell immunity and antibody production by B cells. Because the immune response elicited by pathogens, vaccines, or other experimental treatments can differ significantly depending on the type of stimuli, it is not surprising that the effect of vitamin A status has also varied depending on the type of natural infection or experimental challenge. Consistently, however, the administration of vitamin A, or therapeutic doses of retinoic acid, has restored a more normal pattern of T cell-dependent immune responses, often quite rapidly, to previously vitamin A-deficient hosts.

In children at risk of vitamin A deficiency, vitamin A supplementation, given prophylactically or as therapy during illness, has significantly reduced the severity of measles and measles-related mortality (see 'Hypervitaminosis A and Vitamin A Toxicity').

Recommended Dietary Allowances and Tolerable Upper Intake Levels for Vitamin A

Recommended dietary allowances (RDA) for the US and Canada were recently revised by the Institute of Medicine (IOM). Owing to the serious, potentially irreversible, effects caused by an excess of vitamin A, guidelines were also established for a tolerable upper intake level (UL), defined as the highest intake of a nutrient that is likely to pose no risk of adverse health effects in nearly all healthy individuals. The 2001 RDA and UL for vitamin A for various life stages are listed in **Table 2**.

Infancy and Childhood

Recommendations for an adequate intake (AI) of vitamin A for children 0–6 and 7–12 months of age are based on the vitamin A content of human breast milk and on usual milk intakes for these age groups. The vitamin A content of breast milk from well-nourished women was estimated to be $1.7\,\mu mol\,l^{-1}$; therefore, it provides approximately $400\,\mu g$ RAE day^{-1} to 0- to 6-month-old infants. This value was used by the IOM as the AI for infants in this age group. The AI is $500\,\mu g$ RAE day^{-1} for 7- to

Table 2 Recommended dietary allowances (RDA) for vitamin A in micrograms (μg), retinol activity equivalents (RAE) and international units (IU), and tolerable upper intake levels (UL, μg retinol day^{-1}) for children and adults

Age (years)	*Children*	*Men*	*Women*	*Pregnancy*	*Lactation*
RDA (μg RAE day^{-1})					
1–3	300 μg or 1000 IU				
4–8	400 μg or 1333 IU				
9–13	600 μg or 2000 IU				
14–18		900 μg or 3000 IU	700 μg or 2330 IU	750 μg or 2500 IU	1200 μg or 4000 IU
19+		900 μg or 3000 IU	700 μg or 2330 IU	770 μg or 2565 IU	1300 μg or 4335 IU
UL (μg retinol day^{-1})					
1–3	600 μg				
4–8	900 μg				
9–13	1700 μg				
14–18		3000 μg	2800 μg	2800 μg	2800 μg
19+		3000 μg	3000 μg	3000 μg	3000 μg

12-month-old infants, who are assumed to also consume complementary foods.

Physiological studies have shown that the placental transfer of vitamin A is limited in most mammals. Thus, it is normal for the liver and plasma vitamin A levels of newborns to be much lower than those in adults. In humans, premature infants often have lower plasma retinol levels than full-term infants. The period of breastfeeding is important for the accrual of vitamin A reserves as shown by animal studies in which liver vitamin A stores have increased rapidly in the suckling young of well-nourished healthy mothers. The importance of the neonatal period for establishing vitamin A reserves in young children is well recognized, and programs to promote maternal nutrition and breast feeding have become an integral component of public health programs to improve the vitamin A status of women and children worldwide.

For infants born prematurely, vitamin A (provided in enteral feeds or intramuscularly) is now recognized as an important component of medical care, and as a significant factor in reducing the risk of bronchopulmonary dysplasia and chronic lung disease.

Childhood and Adolescence

Little information specific to this age group is available, and recommendations are based on adult values, scaled down based on body weight.

Adulthood

Adults require a maintenance level of vitamin A. The RDA is based on maintaining an adequate level of vitamin A in liver while meeting normal tissue demands. In animals fed a normal vitamin A adequate diet, retinyl esters tend to accumulate as the animal ages, such that it becomes very difficult to induce vitamin A deficiency in adult animals, even by feeding them a diet free of vitamin A. These data imply that tissue reserves readily make up for lapses in the day-to-day intake of vitamin A. As is evident from **Table 1**, some foods contain an amount of vitamin A well in excess of 100% of the daily value (%DV).

Pregnancy and Lactation

The requirement for vitamin A is increased during pregnancy and lactation, but only to the extent needed for growth of maternal and fetal tissues. Nearly all of the vitamin A in breast milk is in the form of retinyl esters. Milk vitamin A concentration is influenced by recent maternal vitamin A intake. Physiological studies have shown that the lactating mammary glands derive retinol from holo-RBP and from the metabolism of chylomicra. As chylomicron vitamin A increases, milk vitamin A also increases. The RDA for lactation (**Table 2**) is calculated to provide sufficient vitamin A for the mother's needs and for the secretion of vitamin A in breast milk.

Upper Levels

There are three major adverse effects of hypervitaminosis A:

- birth defects;
- liver abnormalities; and
- reduced bone mineral density, which may result in osteoporosis.

The critical adverse effects used to establish the upper level (UL) were risk of teratogenicity for women of reproductive age, and liver abnormalities for all other life stage groups. For vitamin A, the UL applies specifically to chronic intakes of preformed vitamin A (not carotenoids) obtained from foods, fortified foods, and supplements. The UL is not meant to apply to individuals taking vitamin A under medical

supervision. For several life stage groups, the UL are less than three times the RDA (**Table 2**). Based on epidemiological studies of vitamin A intakes and birth outcomes in pregnant women, and on the well-documented teratogenic effects of excessive vitamin A in experimental animals, the UL for women of reproductive age is 2800–3000 $\mu g\,day^{-1}$. The UL based on risk of damage to the liver, although calculated in a different way to that for teratogenicity, has the same value of 2800–3000 $\mu g\,day^{-1}$. Like the RDA, the UL for younger age groups was scaled down based on body weight.

Risk of reduced bone mineral density was also considered in setting the UL, but dose-response data were insufficient to estimate a UL based on this effect. Nonetheless, there is increasing concern that bone health may be adversely affected by intakes of vitamin A not very much higher than the RDA. Epidemiological studies in Swedish men and women, and similar studies in the US, have provided evidence that a chronic intake of preformed vitamin A on the order of 2–3 times the recommended levels (near the UL) may increase the loss of bone mineral density and incidence of hip fracture. Although more research is needed, an upper intake of 2800–3000 μg preformed vitamin A per day also appears to be a prudent guideline for maintaining bone health.

Users of supplements that contain retinol or a retinyl ester should evaluate their average combined intake from diet (especially liver, milk, dairy products), fortified foods (including breakfast cereals), and all dietary supplements to ensure that it does not exceed the UL. Children's vitamin supplements should be checked to make sure that the amount of vitamin A is suitable for the child's age.

Hypervitaminosis A and Vitamin A Toxicity

Hypervitaminosis A is a rare but serious, sometimes fatal, condition. Hypervitaminosis A refers to high storage levels of vitamin A in the body that can lead to toxic symptoms. Toxic symptoms can arise after consuming very large amounts of preformed vitamin A over a short period of time, or they may develop slowly (chronic toxicity), depending on the duration and dose of vitamin A (retinol) consumed. Case reports of vitamin A toxicity include cases of excessive intakes of foods high in retinol such as liver (see **Table 1**), but most cases of vitamin A toxicity result from an excess intake of vitamin A in supplements. Symptoms resembling hypervitaminosis A have been reported in a few patients taking prescription retinoids for therapy. The clinical hallmarks of vitamin A toxicity include nausea and vomiting, headache, dizziness, blurred vision, muscular uncoordination, abnormal liver functions, and pain in weight-bearing bones and joints.

Besides eliminating intake of vitamin A, or retinoids, there is little that can be done, and no antidote, to treat hypervitaminosis A. Tissue retinoid levels fall gradually, but due to the body's tendency to conserve vitamin A, the loss is very slow. Thus, care should be exercised to avoid overconsumption or supplementation with preformed vitamin A (see 'Upper Levels', **Table 2**).

Excessive Consumption of β-Carotene

Beta-carotene and other carotenoids in foods, even when consumed at high levels, are believed to be nontoxic, and therefore no UL was established for β-carotene. Nonetheless, a 'safe range' of intracellular β-carotene has yet to be determined. Individuals who have consumed large amounts of carotenoid-rich foods, juices, or extracts containing a large amount of β-carotene over a prolonged period of time may show signs of carotene accumulation in fatty tissues, to the point where yellowing of the skin (carotenodermia) is apparent. This condition is not known to be harmful and the color subsides over time after carotene intake is reduced to normal levels. Nonetheless, epidemiological evidence suggests that the use of high-dose β-carotene as a dietary supplement should not be regarded as safe because the current knowledge of the metabolism of high doses is inadequate, and some epidemiological studies have indicated that high doses of β-carotene, at least in smokers, may be detrimental (see 00045).

See also: **Bioavailability**. **Carotenoids**: Chemistry, Sources and Physiology; Epidemiology of Health Effects. **Nutrient–Gene Interactions**: Molecular Aspects. **Vitamin A**: Biochemistry and Physiological Role.

Further Reading

Altucci L and Gronemeyer H (2001) Nuclear receptors in cell life and death. *Trends in Endocrinology and Metabolism* **12**: 460–468.

Collins MD and Mao GE (1999) Teratology of retinoids. *Annual Reviews of Pharmacology and Toxicology* **39**: 399–430.

De Luca LM, Kosa K, and Andreola F (1997) The role of vitamin A in differentiation and skin carcinogenesis. *Journal of Nutritional Biochemistry* **8**: 426–437.

Harrison EH and Hussain MM (2001) Mechanisms involved in the intestinal digestion and absorption of dietary vitamin A. *Journal of Nutrition* **131**: 1405–1408.

Li E and Tso P (2003) Vitamin A uptake from foods. *Current Opinions in Lipidology* **14**: 241–247.

Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*, ch. 8. Washington: National Academy Press.

Noy N (2000) Retinoid-binding proteins: mediators of retinoid action. *Biochemical Journal* **348**: 481–495.
Institute of Medicine (2001) *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc*, ch. 4. Washington: National Academy Press.
Ross AC (1996) Vitamin A deficiency and retinoid repletion regulate the antibody response to bacterial antigens and the maintenance of natural killer cells. *Clinical Immunology and Immunopathology* **80**(supplement): S36–S72.
Ross AC (1999) Vitamin A and retinoids. In: Shils ME, Olson JA, Shike M *et al.* (eds.) *Modern Nutrition in Health and Disease*, 9th edn, pp. 305–327. Baltimore: William & Wilkins.
Ross AC (2003) Retinoid production and catabolism: Role of diet in regulating retinol esterification and retinoic acid oxidation. *Journal of Nutrition* **133**(suppl): 291S–296S.
Ross AC, Zolfaghari R, and Weisz J (2001) Vitamin A: recent advances in the biotransformation, transport, and metabolism of retinoids. *Current Opinions in Gastroenterology* **17**: 184–192.
Olson JA (1984) Serum level of vitamin A and carotenoids as reflectors of nutritional status. *Journal of the National Cancer Institute* **73**: 1439–1444.
Saari JC (2000) Biochemistry of visual pigment regeneration: the Friedenwald lecture. *Investigative Ophthalmology and Visual Science* **41**: 337–348.
Sommer A and West KP Jr (1996) *Vitamin A Deficiency: Health, Survival, and Vision* New York: Oxford University Press.
Soprano DR and Blaner WS (1994) Plasma retinol-binding protein. In: Sporn MB, Roberts AB, and Goodman DS (eds.) *The Retinoids: Biology, Chemistry and Medicine*, 2nd edn., pp. 257–281. New York: Raven Press.
Stephensen CB (2001) Vitamin A, infection, and immune function. *Annual Reviews of Nutrition* **21**: 167–192.
Stoltzfus RJ and Humphrey JH (2002) Vitamin A and the nursing mother-infant dyad: evidence for intervention. *Advances in Experimental Medicine and Biology* **503**: 39–47.
Tanumihardjo SA (2002) Influencing the conversion of carotenoids to retinol: bioavailability to bioconversion to bioefficacy. *International Journal of Vitamin and Nutrition Research* **72**: 40–45.
Underwood BA and Smitasiri S (1999) Micronutrient malnutrition: Policies and programs for control and their implications. *Annual Reviews of Nutrition* **19**: 303–324.
van het Hof KH, West CE, Weststrate JA *et al.* (2000) Dietary factors that affect the bioavailability of carotenoids. *Journal of Nutrition* **130**: 503–506.
Von Reinersdorff D, Green MH, and Green JB (1998) Development of a compartmental model describing the dynamics of vitamin A metabolism in men. *Advances in Experimental Medicine and Biology* **445**: 207–223.

Biochemistry and Physiological Role

J L Napoli, University of California, Berkeley, CA, USA

In 1913, E.V. McCollum isolated a yellow, fat-soluble substance from egg yolks that was critical for animal growth. He called it fat-soluble A, indicating the first isolated of several dietary microconstituents emerging as obligatory for vertebrate life and health. Later, fat-soluble A was renamed vitamin A, derived from the terminology 'vital amine,' coined by Casmir Funk to describe these obligatory micronutrients.

Currently, the term vitamin A refers to the specific organic compound all-*trans*-retinol (atROH). atROH, however, does not have biological activity in its own right. Rather, it serves as a circulating substrate for metabolism into the compounds that fulfill the biological functions attributed to vitamin A. These metabolites include, but may not be limited to, 11-*cis*-retinal (11cROH) and all-*trans*-retinoic acid (atRA). The term 'retinoids' describes all compounds that support vitamin A activity, both naturally occurring and synthetic, including atROH. **Figure 1** illustrates the structures of key carotenoids and retinoids.

Daily Recommended Dietary Allowance of Vitamin A

The Food and Nutrition Board of the Institute of Medicine revised the Recommended Dietary Allowance (RDA) of vitamin A in 2001 as 900 retinol activity equivalents (RAE) for men and 700 RAE for women. The RAE was introduced to avoid the ambiguity of international units (IU), which arises because 1 IU of vitamin A (0.3 μg) and 1 IU of the vitamin A precursor (provitamin A) all-*trans*-β-carotene (0.6 μg) do not have the same biological activity. Rather, 6 IU of β-carotene and 12 IU of mixed carotenoids provide the biological activity of 1 IU of vitamin A. The RAE refers to the amounts necessary for the same degree of biological activity: 1 μg atROH = 12 μg β-carotene = 24 μg mixed carotenoids = 1 RAE.

Liver, dairy products, and saltwater fish, including herring, sardines, and tuna, serve as dietary sources of vitamin A and its esters. Cod and halibut liver oil provide especially rich sources of vitamin A, as does the liver of the polar bear, which benefits from occupying the top of the marine food chain. Carrots, yellow squash, corn, and dark-green leafy vegetables serve as dietary sources of provitamin A carotenoids. Because less than 10% of the 600 naturally occurring carotenoids generate vitamin A, vegetable color does not necessarily indicate vitamin A value. In the United States and Europe, retinol and its esters serve as the chief sources of dietary vitamin A, but elsewhere carotenoids serve as the primary source. According to the World Health

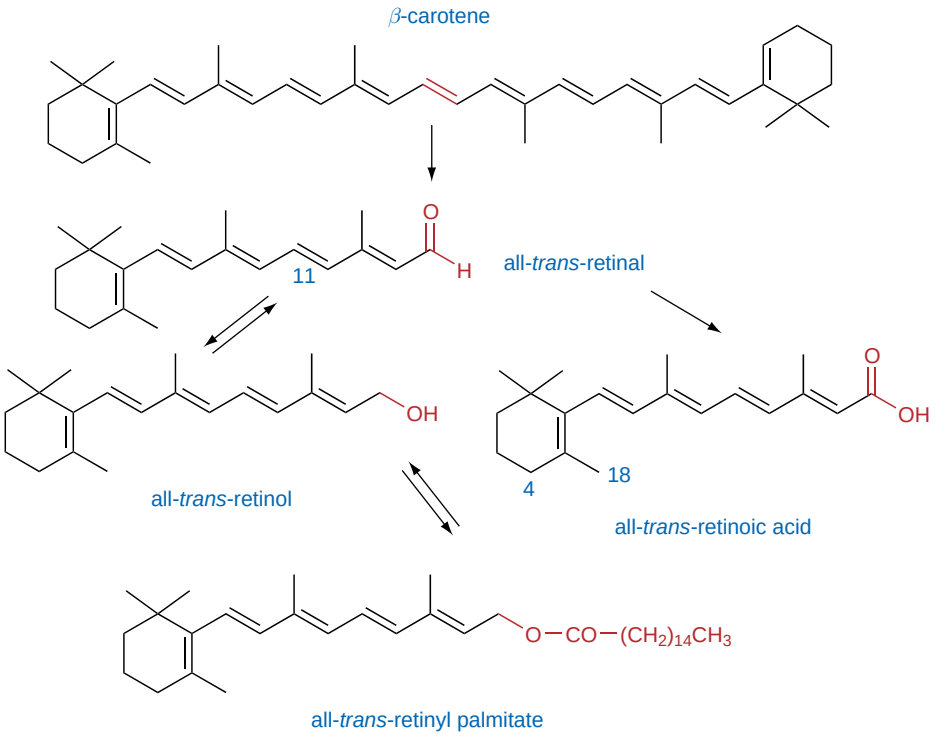


Figure 1 Structures of β-carotene and common endogenous retinoids. The numbers indicate the position of *trans* to *cis* isomerization of atRCHO (C11) and the positions of hydroxylation of atRA (C4 and C18).

Organization (WHO), the major inadequacies of dietary micronutrients involve vitamin A, iron, and iodine. Although vitamin A intake seems adequate in most populations in the United States, Canada, and Europe, this appears evolutionarily aberrant. The recurrent vitamin A-deficiency problem in Third World populations indicates that human diets in nonindustrialized countries are limited in vitamin A.

Vitamin A Status

atROH represents the quantitatively major circulating plasma retinoid in serum, but serum atROH does not reflect vitamin A status unless low (<0.35 μmol/l or 10 μg/dl) or unequivocally adequate (>1.1 μmol/l in children and >1.4 μmol/l in adults). Humans show a range of normal serum atROH values, with unknown factors contributing to the individual's normal value, and fever, infection, and/or inadequate intake of other nutrients (e.g., zinc and protein) can depress serum retinol. Liver reserves provide the most reliable measure of human vitamin A status. Noninvasive assessments of liver vitamin A reserves (the relative dose–response and modified relative dose–response tests) measure the amount of dose in serum relative to the original plasma atROH after a small oral dose of marker vitamin A. The larger the proportion of dose that appears in serum, the lower the liver reserves.

Binding Proteins

Several high-affinity (low K_d), soluble binding proteins seem crucial to vitamin A homeostasis and/or function. Two are widely distributed throughout many tissues and cell types, whereas others have limited expression loci (**Table 1**). These binding proteins occur in all vertebrates, have highly conserved amino acid sequences among orthologs, and show high specificity for distinct retinoids. Where measured, their concentrations exceed the concentrations of their ligands. Indeed, CRBP(II) accounts for ~1% of the soluble intestinal enterocyte proteins. These qualities and experimental data indicate that retinoids exist *in vivo* bound to specific proteins. For example, purification of CRBP(I) from tissues produces predominantly holoprotein, despite the capacity of membranes to sequester more atROH than occurs physiologically and the time-consuming isolation techniques originally used to isolate these proteins (including tissue homogenization, centrifugation, and several types of column chromatography). The locus of atROH at equilibrium would depend on both affinity for potential acceptors and acceptor capacity. Nature

Table 1 Examples of retinoid binding proteins

Retinoid binding protein	*Ligand(s)*	K_d *(nM)*	*Adult distribution*
CRBP (cellular retinol binding protein, type I)	atROH	~0.1	Nearly ubiquitous (low in intestine)
	atRCHO	10–50	
CRBP(II)	atROH	10–50	Intestine
	atRCHO	10–50	
CRABP (cellular retinoic acid binding protein, type I)	atRA	0.4	Widespread
CRABP(II)	atRA	2	Limited (e.g., skin, uterus, ovary)
CRALBP (cellular retinal binding protein)	11cROH	—	Eye, especially retinal pigment epithelium
	11cRCHO		
SRBP (serum retinol binding protein)	atROH	—	Serum

(*in vivo*) and the scientist (*in vitro*) provide plenty of opportunities for retinol to equilibrate between CRBP(I) and potential acceptors (membranes, lipid droplets, etc.). Evidently, the large capacity of membranes and other potential acceptors does not overcome the comparatively limited capacity of CRBP(I) to sequester retinol, consistent with tight binding.

CRBP (types I and II) and CRABP (types I and II) have molecular weights of ~15 kDa and belong to the intracellular lipid binding protein (iLBP) gene family, which includes the various fatty acid binding proteins. The family members have similar three-dimensional structures, despite low primary amino acid conservation among nonorthologous members. These proteins form globular but flattened structures of 10 antiparallel strands of β-sheets, 5 orthogonal to and above the other 5, referred to as a β-clam (**Figure 2**). The polar head group of atROH (i.e., the functional group that undergoes esterification or dehydrogenation) lies buried deep within CRBP, protected from the milieu of oxidants, nucleophiles, and enzymes.

CRBP(I) null mice are phenotypically normal until retinol depletion, but they eliminate retinol and its esters sixfold faster than wild-type mice, presumably through enhanced catabolism via enzymes that normally have limited access. In contrast, CRBP(II) null mice pups suffer 100% mortality by 24 h after birth when born to dams fed a vitamin A-marginal diet. Retinoid binding proteins apparently confer selective advantage to vertebrates by promoted sequestering, transport, and storage of vitamin A and limiting its catabolism. CRBP(III) has been detected in mouse heart and skeletal muscle, which express little or no CRBP(I) or CRBP(II), but not in other retinoid target tissues, such as liver, kidney, and brain. CRBP(III) seems to bind about equally well with atROH, 9cROH, and 13cROH (K_d ~80–110 nM). Humans express yet another CRBP, originally referred to as CRBP(III), but distinct from mouse

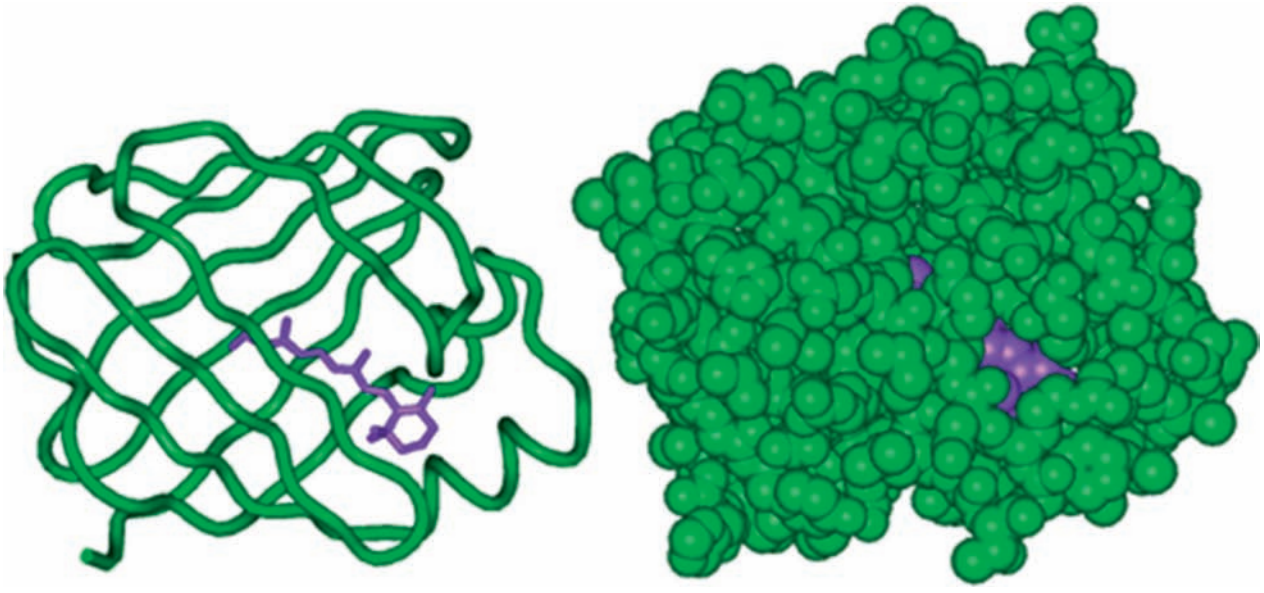


Figure 2 Ribbon (left) and space-filling (right) models of CRBP(I). (Courtesy of Marcia Newcomer (1995) Retinoid-binding proteins: Structural determinants important for function. *FASEB Journal* **9**: 229–239, Louisiana State University.)

CRBP(III) and therefore really CRBP(IV). CRBP(IV) mRNA is much more abundant in human liver and intestine than CRBP(I) mRNA, but the mouse does not encode a complete CRBP(IV) gene. CRBP(IV) binds atROH with a K_d of ~60 nM but does not bind *cis*-isomers. The precise functions of CRBP(III) and CRBP(IV) have not been clarified: Presumably, they moderate retinol metabolism, similar to CRBP(I) and CRBP(II).

CRABP(I) and -(II) do not have well-defined physiological functions. Mice doubly null in CRABP(I) and -(II) have an approximately fourfold higher rate of death from unknown causes by 6 weeks after birth than wild-type mice, but the survivors appear essentially normal, with one exception. The doubly null mouse as well as the CRABP(II)-only null mouse respectively show 83 and 45% incidence of a small outgrowth anomaly on the postaxial side of digit five, predominantly in the forelimbs. The double mutants do not exhibit enhanced sensitivity to atRA, suggesting that CRABP do not serve primarily to protect against atRA toxicity or teratogenicity.

CRALBP (~36 kDa) belongs to a gene family that includes the α-tocopherol transfer protein (TTP). *In vitro*, CRALBP sequesters 11cROH in the retinal pigment epithelium (RPE) of the rods, driving forward the *trans* to *cis* isomerization, and also facilitates dehydrogenation of 11cROH into 11cRCHO. Mutations in human CRALBP cause night blindness and photoreceptor degeneration.

SRBP (~20 kDa) belongs to the lipocalin family, which includes apolipoprotein D, β-lactoglobulin, odorant binding protein, and androgen-dependent secretory protein. SRBP has a globular structure formed by eight antiparallel β-sheets in two orthogonal sheets that mold a β-barrel. The β-ionone ring of atROH lies deep within SRBP, whereas the hydroxyl group lies closest to the opening. atROH in serum remains bound with SRBP, despite high concentrations of a potential alternative high-capacity carrier, albumin. This illustrates the affinity of SRBP for atROH and the importance of sequestering retinoids within specific proteins. Liver is the major site of SRBP synthesis: Accordingly, liver expresses SRBP mRNA most abundantly. Extrahepatic tissues, however, also express SRBP mRNA, including adipose and kidney, but the functions of SRBP produced extrahepatically remain unknown. Knocking out the SRBP gene produces a phenotypically normal mouse, except for impaired vision after weaning. Vision can be restored after months of feeding a vitamin A-adequate diet. Thus, the eye, which consumes (but does not store) the vast majority of vitamin A, normally relies on SRBP for retinol delivery. Retinol delivered by albumin and lipoproteins apparently supports the nonvisual functions of vitamin A, at least in the SRBP null mouse kept under laboratory conditions.

Vitamin A and the Visual Cycle

Figure 3 depicts a model of the functions of multiple proteins and forms of vitamin A that constitute the visual cycle. SRBP delivers atROH to the RPE, possibly through a plasma membrane SRBP receptor. No SRBP receptor has been isolated, however, and molecular characterization remains elusive. As in other tissues, CRBP(I) sequesters atROH and allows its esterification by the 25-kDa endoplasmic reticulum (ER) enzyme lecithin:retinol acyltransferase (LRAT). In fact, the amount of CRBP(I) may represent the controlling event in the rate of atROH uptake, and uptake may be coupled to retinyl ester (RE) formation. RPE65, an ~65-kDa RPE protein, binds the highly hydrophobic atRE with a K_d of ~20 pM. This accelerates mobilization of atRE and delivery to the next step, isomerization by an isomerohydrolase (IMH). The RPE65 null mouse cannot produce 11-*cis*-retinoids, attesting to its importance in mobilizing atRE in the quantity and rate necessary to support the visual cycle. The IMH both hydrolyzes atRE and isomerizes the double bond at C11 to produce 11cROH. Unfortunately, little is known about the IMH: It has not been isolated or cloned. The cytosolic protein CRALBP sequesters 11cROH and also 11cRCHO produced by the ER enzyme 11-*cis*-retinol dehydrogenase (11cRDH), and it seems to enhance isomerization. The 11cRDH of the RPE belongs to the SDR (short-chain dehydrogenase/reductase) gene family: RDH4 and RDH5 encode the murine and human 11cRDH genes, respectively. 11cRDH also has been referred to as 9cRDH (9-*cis*-RDH) or cRDH, but the eye expresses 11cRDH mRNA far more intensely than other tissues, and more efficient potential 9cRDHs have been cloned and characterized, namely the SDR isozymes CRAD1 and -3 (*cis*-retinol/androgen dehydrogenase). Mutations in RDH5 cause the rare autosomal recessive disorder fundus albipunctatus, which is a form of night blindness characterized by delayed regeneration of photopigments in rods and cones. Enzymes in addition to RDH4/5 probably contribute to dehydrogenation of 11cROH in the RPE, which explains why the RDH5 mutation does not cause blindness.

Interphotoreceptor retinoid binding protein (IRBP) occupies the space between the RPE and the rod outer segment (ROS), the interphotoreceptor matrix, and binds both 11cROH and 11cRCHO.

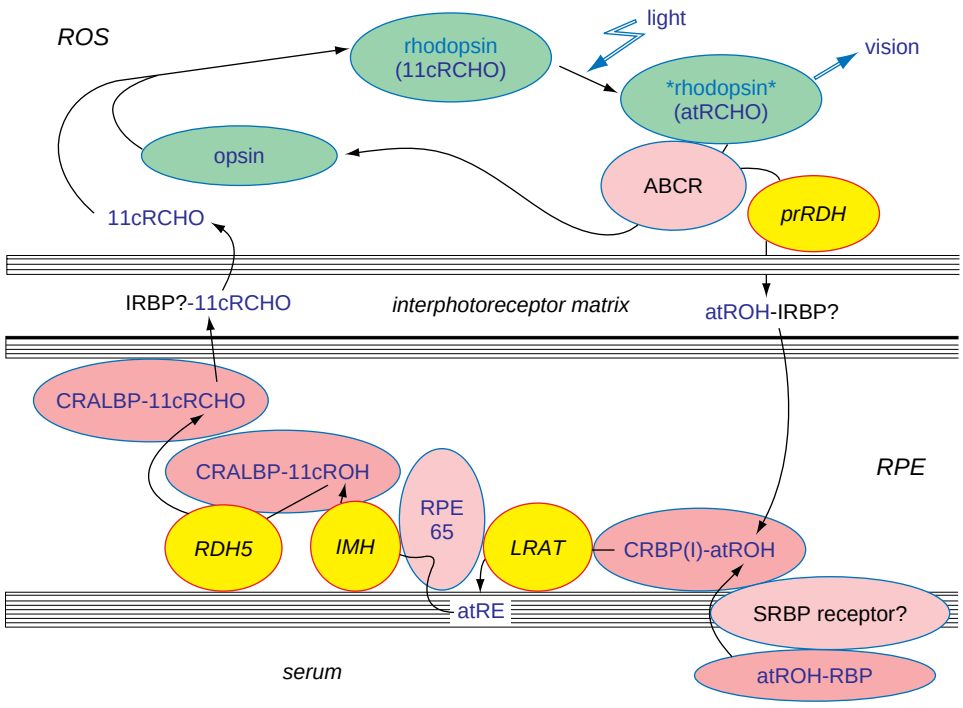


Figure 3 Model of the mammalian visual cycle. ABCR, ATP-binding cassette; atRCHO, all-*trans*-retinal; atROH, all-*trans*-retinol; atRE, all-*trans*-retinyl esters; 11cRCHO, 11-*cis*-retinal; 11cROH, 11-*cis*-retinol; CRALBP, cellular retinal binding protein; CRBP(I), cellular retinol binding protein, type I; IRBP, interphotoreceptor retinoid binding protein; IMH, isomerohydrolase; LRAT, lecithin:retinol acyltransferase; ROS, rod outer segment; RPE, retinal pigment epithelium; RPE65, RPE protein 65; SRBP, serum retinol binding protein.

Retinoids do not require IRBP for transfer between the two membranes, however, because IRBP null mice show no gross abnormalities of the visual cycle, even though they have severe retinal abnormalities. Although retinoids have very limited solubility in the aqueous phase, membranes of the RPE and ROS transfer retinoids between them in the absence of IRBP.

In the ROS, 11cRCHO forms a Schiff's base adduct with a lysine residue in the protein opsin to create rhodopsin. Light isomerizes the 11cRCHO of rhodopsin into atRCHO. The isomerization straightens the retinoid side chain, changing the conformation of rhodopsin. This conformation change initiates a nerve impulse through G proteins and releases atRCHO, regenerating opsin. An ATP transporter ABCR facilitates leaching of atRCHO from rhodopsin, and an ER enzyme, prRDH (photoreceptor retinol dehydrogenase), also an SDR, reduces atRCHO into atROH. atROH cycles back to the RPE and binds with CRBP(I).

Vitamin A Homeostasis and Activation into atRA

Intestinal absorptive cells absorb dietary carotenoids and retinol during the bile–acid-mediated process of lipid absorption. Within the enterocyte, central cleavage by a soluble 63-kDa carotene 15,15′-monooxygenase catalyzes the principal route of carotenoid metabolism (**Figure 4**). Carotene 15,15′-monooxygenase belongs to the same gene family as RPE65 (the mouse proteins have only 37% amino acid identity, however), suggesting a family of proteins/enzymes dedicated to transport/metabolism of highly hydrophobic substances. Intestine expresses the carotene 15,15′-monooxygenase mRNA, but kidney and liver show much more intense expression, and the testis curiously shows most intense expression, consistent with the ability of tissues other than the intestine to cleave carotenoids. Carotene 15,15′-monooxygenase also metabolizes carotenoids without provitamin A activity, such as lycopene, although with lower efficiency.

CRBP(II) sequesters atRCHO generated from carotenoids and allows its reduction into atROH, catalyzed by an ER retinal reductase (uncharacterized). In contrast to CRBP(I), CRBP(II) does not allow oxidation/dehydrogenation of its ligands. LRAT accesses the CRBP(II)–atROH complex and produces atRE for incorporation into chylomicrons. During conversion into remnants by lipoprotein lipase in adipose, chylomicrons retain most of their RE, as they do cholesterol esters.

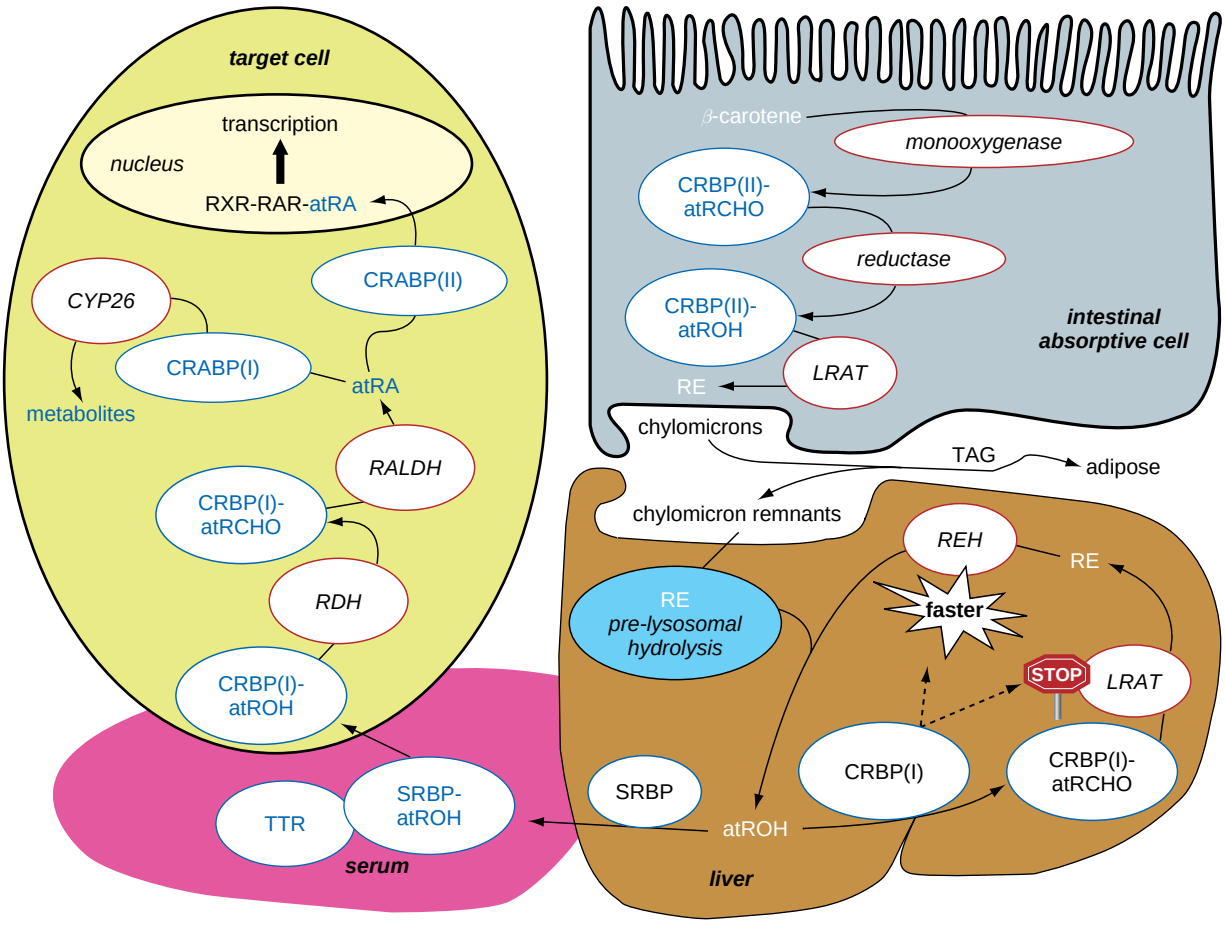


Figure 4 Model of atRA biogeneration in mammals. REH, retinyl ester hydrolase (e.g., ES4 and ES10); TTR, transthyretin; RAR-RXR, the heterodimer of retinoic acid receptors with retinoid X receptors; atRCHO, all-*trans*-retinal; atROH, all-*trans*-retinol; CRBP(I), cellular retinol binding protein, type I; LRAT, lecithin:retinol acyltransferase; SRBP, serum retinol binding protein. CRBP(I), CRABP(I), and CRABP(II) have been placed in the same cell for simplicity. This does not necessarily occur *in vivo*.

Hepatocytes sequester RE and cholesteryl esters by receptor-mediated endocytosis of chylomicron remnants. Substantial RE hydrolysis apparently occurs before engulfing of the remnants by lysosomes. CRBP(I) sequesters the atROH released and allows esterification by LRAT but protects from esterification via other acyltransferases, just like CRBP(II) functions in the intestine. Ultimately, liver stellate cells accumulate most of the RE. CRBP(I) seems necessary for retinoid transfer from hepatocytes to stellate cells because the CRBP(I) null mouse does not accumulate RE in stellate cells. The mechanism of transfer, however, has not been established.

Liver senses local and extrahepatic need for atRA biosynthesis by an unknown signal (possibly atRA) and need for atROH in the visual cycle, and it responds by mobilizing RE to maintain serum retinol levels. ES-10 and ES-4, two neutral ER-localized, bile salt-independent carboxyesterases, provide at least 94% of the RE hydrolysis activity in liver. Kidney also expresses ES-4, and kidney, testis, lung, and skin express ES-10. Either SRBP or CRBP(I) sequesters the atROH released. SRBP delivers atROH to serum, whereas CRBP(I), unlike CRBP(II), allows dehydrogenation into atRCHO to support atRA biosynthesis. (**Figure 4** shows atRA biogeneration only in target cells, but it also occurs in liver; likewise, RE storage also occurs in target cells in animals exposed to higher dietary atROH, suggesting that the liver does not initially sequester all dietary retinoids.) What keeps the CRBP(I)-bound atROH from undergoing futile cycling back to atRE? Apo-CRBP(I) stimulates endogenous microsomal RE hydrolysis and inhibits LRAT (**Figure 5**). Note that apo-CRBP(I) exerts potent effects at concentrations of ~2.5 μM—well within the range of the CRBP(I) expressed in liver. Thus, the ratio apo-CRBP/holo-CRBP signals atROH status and directs atROH flux into or out of atRE.

SRBP–atROH circulates as a complex with transthyretin (TTR) that protects it from degradation. The mechanism of atROH delivery from SRBP into

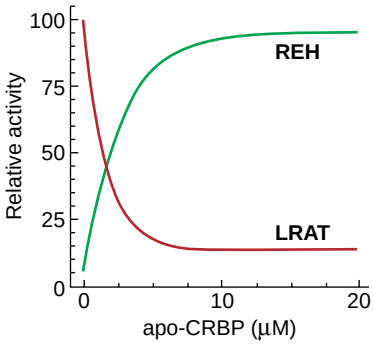


Figure 5 The effect of apo-CRBP(I) on the rates of retinol esterification (lecithin:retinol acyltransferase (LRAT)) and retinyl ester hydrolysis (REH).

cells has not been established. Some data suggest a specific SRBP membrane receptor, whereas other data indicate that CRBP(I) pulls atROH transfer from SRBP through the membrane. A third hypothesis is that an SRBP receptor is mainly in the eye, the quantitatively major site of atROH consumption.

Extrahepatic cell atROH supports atRA biosynthesis but also undergoes esterification if delivered in sufficient quantities. CRBP(I) allows dehydrogenation of atROH by an ~35-kDa ER retinol dehydrogenase (RDH) but protects atROH from dehydrogenation via other dehydrogenases. RDH (*Rdh1*) belongs to the SDR gene family. The SDR gene family consists of ~50 mammalian members that catalyze intermediary metabolism, and the metabolism of steroids and prostaglandins, in addition to retinoids.

The mechanism of atROH transfer from CRBP(I) or -(II) to RDH and LRAT has not been elucidated. Specific cross-linking of holo-CRBP(I) with both RDH1 and LRAT, however, indicates close proximity of CRBP(I) to the two enzymes. Notably, apo-CRBP(I) at concentrations of 1 μM prevents cytosolic dehydrogenation of atROH, even by soluble enzymes that access the CRBP(I)–atROH complex, while not impacting ER SDR until reaching much higher concentrations. This suggests the importance of membrane RDH, rather than soluble dehydrogenases, to atRA biosynthesis. Obviously, atRA biosynthesis *in vivo* occurs in the absence of CRBP(I), as indicated by lack of morphological pathology in the CRBP(I) null mouse. This was predicted by *in vitro* experiments that showed that neither RDH nor LRAT require presentation of atROH by CRBP(I). Rather, CRBP(I) operates as a molecular chaperone that restricts metabolism to enzymes that can access its atROH.

In the early 1950s, cytosolic alcohol dehydrogenases (ADHs) were suggested to metabolize atROH. This was an attempt to explain atRCHO generation and the controversial and poorly understood putative occurrence (at the time) of atRA in tissues, when other families of dehydrogenases remained anonymous. ADHs do recognize atROH *in vitro*, albeit only when presented free of CRBP(I) and with comparatively low efficiencies. However, enzymes have poor substrate discrimination *in vitro* and do not metabolize many of the same substrates *in vivo* because of intracellular constraints. Mice null in both ADH class I (*Adh1*) and ADH class IV (*Adh4*) show no vitamin A-deficiency phenotype, nor do mice null in ADH class III (*Adh3*). *Adh1* null mice show a decreased rate of metabolism of 50–100 mg/kg atROH, but this demonstrates only that extraordinary high exposure can defeat physiological controls imposed by retinoid binding proteins. Vitamin A excess has not been a problem throughout evolution (mice do not usually eat polar bear or marine fish liver): No pressure forced evolution of protective mechanisms against such exposure. Natural selection exerted the opposite pressure (i.e., evolution of retinoid binding proteins to conserve vitamin A).

Retinal dehydrogenases (RALDHs) catalyze the irreversible conversion of atRCHO into atRA and can do so in the presence of CRBP(I) using atRCHO generated *in situ* from CRBP(I)–atROH and RDH. These ~54-kDa soluble enzymes belong to the ADLH gene family. RALDH1 (*Aldh1a1*), -2 (*Aldh1a2*), and -3 (*Aldh1a3*) contribute most to atRA generation, whereas RALDH4 has much more efficient activity with 9-*cis*-retinal. The RALDH1 null mouse remains fertile and healthy but may have decreased ability to produce atRA in the liver. The RALDH2 null mouse dies *in utero* by midgestation, demonstrating its unique contribution to atRA synthesis during embryogenesis. The situation may differ in the adult. RALDH1–3 show overlapping expression patterns in the adult, with RALDH1 expressed most intensely. Interestingly, atRA regulates mRNA levels of RALDH1 differently in different tissues. For example, vitamin A sufficiency increases kidney and liver RALDH1 mRNA, whereas vitamin A insufficiency increases testis RALDH1 mRNA. This may represent a mechanism to divert atROH from liver to testis for atRA production during vitamin A scarcity.

Other Active Retinoids

Discrete loci synthesize 3,4-didehydro-atRA, such as skin. 3,4-Didehydro-atRA binds RAR with high affinity, similar to atRA. The purpose of creating a signaling molecule that functions similar to atRA

in specialized loci, which also biosynthesize atRA, has not been clarified.

Although 9-*cis*-RA was reported as a hormone *in vivo*, at best, its concentrations are much lower than atRA. Its putative function as a physiological ligand that controls RXR is controversial.

atRA Catabolism

atRA induces its own metabolism via cytochrome P450 (CYP) into a variety of initial catabolites, including 5,6-epoxy-atRA, 18-hydroxy-atRA, and 4-hydroxy-atRA. Metabolism of atRA limits its activity; conversely, inhibitors of atRA metabolism enhance atRA potency. CYP26A1 may catalyze the major degree of atRA catabolism, as evidenced by null mice dying in mid- to late gestation with serious morphological defects. Two other P450's, CYP26B1 and CYP26C1, also catabolize atRA, but null mice have not been reported. Several other CYPs reportedly catabolize atRA, but these candidates (CYP1A1/2, CYP2A6, CYP2C8/9, CYP2E1, and CYPP3A4/5) are not induced by atRA—in contrast to the well-established ability of atRA to induce its own metabolism as well as CYP26A1 transcription—and most have inefficient kinetics *in vitro* with atRA.

Presenting atRA to microsomes bound with CRABP(I) enhances kinetic efficiency of catabolism sevenfold. There seems to be little doubt that CRABP(I) sequesters atRA: Delivering the sequestered atRA for efficient catabolism seems to be a logical mechanism to discharge the ligand without releasing it back into the cell. Unfortunately, this insight does not reveal the primary purpose for CRABP(I) impounding atRA in the first place.

Regulation of Retinoid Homeostasis

atRA regulates retinoid homeostasis. It induces LRAT transcription ~100-fold and induces CRBP(I) transcription. These may be housekeeping functions that maintain retinol esterification and handling during vitamin A sufficiency. On the other hand, induction of CYP26A1 by atRA can be viewed both as a general mechanism of clearance and as a site-specific mechanism to maintain an environment during specific stages of development.

Several xenobiotics, including ethanol and polychlorinated biphenyls, reduce RE stores, possibly through enhancing atRA catabolism by inducing CYP. In addition, clofibrate, a ligand of PPARα (peroxisome proliferator activated receptor), causes a remarkably rapid depletion of liver retinoids and decreases TTR mRNA fourfold. In contrast, PPARγ, expressed in liver stellate cells among others, induces transcription of carotene 15,15′-monooxygenase. PPARγ also induces expression of CRBP(I), CRBP(II), LRAT, and RARβ. These actions of PPARα and -γ with retinoids reflect their effects on lipid metabolism: PPARα enhances fatty acid catabolism, whereas PPARγ enhances adipogenesis and fatty acid storage.

Mechanism of atRA Action

atRA induces transcription by binding with three ligand-activated transcription factors, RARs (retinoic acid receptors) α, β, and γ, each encoded by a distinct member of the nuclear hormone receptor gene superfamily. RARs function *in vivo* bound with three additional distinct members of the nuclear receptor gene family: RXR (retinoid X receptor) α, β, and γ. Forty-eight RXR–RAR combinations can occur because each receptor gene can express two to four protein isoforms, stemming from differential promoter use and alternative splicing. RXR in complex with RA acts 'silently' (i.e., it can function sans ligand). The RXR–RAR heterodimer recognizes several types of RARE (atRA response elements). The two most common consist of two direct repeats of (A/G)G(G/T)TCA separated by two (DR2) or five (DR5) nucleotides. This multiplicity of heterodimers and response elements suggests a mechanism for the pleiotropic functions of vitamin A.

atRA binding with RAR induces the receptor to convert from an open form to a more structured form that wraps the ligand. This conformation change leads to remodeling of coregulators bound to the RXR–RAR complex. Remodeling releases a complex array of co-corepressors and recruits a complex array of coactivators, accompanied by chromatin remodeling, histone acetylation, and recruitment of basal transcription factors, including TATA binding protein and RNA polymerase II. More than 500 genes have been reported as transcriptionally regulated by atRA either directly by this process or indirectly through events set into motion by this process.

RXR's influence extends beyond serving as an obligatory partner of RAR. RXR forms heterodimers with ~11 nuclear receptors, including the vitamin D receptor PPAR and the thyroid hormone receptors. 'The' RXR ligand, if not 9cRA, would have significant signaling impact.

CRABP(II), but not CRABP(I), delivers atRA to RAR and thereby induces transcription. The limited expression loci of CRABP(II), however, suggest that this action would be specialized rather than general.

Physiological Functions of atRA

All vertebrates require retinal for the visual cycle and atRA to support the systemic functions of vitamin A, including controlling the differentiation programs of epithelia cells and cells in nerve and bone and also the immune and reproductive systems. atRA frequently, but not always, induces terminal differentiation. atRA also regulates expression of genes in differentiated cells and genes crucial to spermatogenesis, hematopoiesis, estrus, placental development, embryogenesis, and apoptosis. atRA may also serve as a tumor suppressor.

The many RAR knockouts illustrate the physiological functions of vitamin A. Disruption of the RARα gene, the RAR with the most widespread, if not ubiquitous, expression in the embryo and adult, does not cause embryonic lethality but reduces the homozygous null population by 60% within 12–24 h after birth and by 90% within 2 months. The RARα null mice that survive 4 or 5 months have severe germinal epithelium degeneration and are sterile. RARβ gene null mice are fertile and viable and show no immediate signs of morphological abnormalities. Nevertheless, complementary data show that RARβ may mediate the antiproliferative function of atRA and as such may serve as a tumor suppressor. Moreover, RARβ null mice have virtually no hippocampal long-term potentiation or long-term depression, the forms of synaptic plasticity that provide a mechanism of short-term spatial learning and memory. This phenomenon can be reproduced by vitamin A depletion. RARγ null mice have an 86% incidence of skeletal abnormalities by embryonic day 18.5 but are born in Mendelian frequency. Postnatally, the homozygous null pups show retarded rates of growth, limiting them to 40–80% of the weight of wild-type pups. By 1–3 weeks old, 50% of RARγ null mice die; by 3 months, 80% die. Mature males also have squamous metaplasia of the seminal vesicles and prostate glands, similar to vitamin A deficiency, but other epithelia appear normal. These data reveal that even though RAR tend to express in site specific patterns in normal circumstances, they also exhibit a great degree of functional redundancy, enabling a mouse null in one receptor to (partially) compensate for the other two.

Numerous studies have correlated vitamin A insufficiency in laboratory animals with increased incidence of spontaneous and carcinogen-induced cancer. Chemopreventive trials in humans show some promise for retinoids in actinic keratoses, oral premalignant lesions, laryngeal leukoplakia, and cervical dysplasia. The US Food and Drug Administration has approved retinoids for acute promyelocytic leukemia and for non-life-threatening diseases, such as cystic acne and psoriasis. Retinoids also provide the active ingredients in agents to treat sun/age-damaged skin.

WHO recognizes vitamin A deficiency as a mortality factor for childhood measles. Two large doses (60,000 RAE each) of a water-soluble vitamin A formulation given to children on each of 2 days decrease the risk of death from measles 81% in areas of prevalent vitamin A deficiency.

See also: **Fish**. **Fruits and Vegetables**. **Nutrient–Gene Interactions**: Molecular Aspects.

Further Reading

Ahuja HS, Szanto A, Nagy L, and Davies PJ (2003) The retinoid X receptor and its ligands: Versatile regulators of metabolic function, cell differentiation and cell death. *Journal of Biological Regulatory and Homeostatic Agents* **17**: 29–45.

Aranda A and Pascual A (2001) Nuclear hormone receptors and gene expression. *Physiological Reviews* **81**: 1269–1304.

Blomhoff R, Green MH, Green JB, Berg T, and Norum KR (1991) Vitamin A metabolism: New perspectives on absorption, transport, and storage. *Physiological Reviews* **71**: 951–990.

Harrison EH (1998) Lipases and carboxyesterases: Possible roles in the hepatic metabolism of retinol. *Annual Reviews of Nutrition* **18**: 259–276.

Maden M (2001) Role of retinoic acid in embryonic and post-embryonic development. *Proceedings of the Nutrition Society* **59**: 65–73.

Mark M, Ghyselinck NB, Wendling O *et al.* (1999) A genetic dissection of the retinoid signaling pathway in the mouse. *Proceedings of the Nutrition Society* **58**: 609–613.

Napoli JL (2000) Retinoic acid: Its biosynthesis and metabolism. *Progress in Nucleic Acids Research* **63**: 139–188.

Newcomer ME (1995) Retinoid-binding proteins: Structural determinants important for function. *FASEB Journal* **9**: 229–239.

Olson JA (1994) Needs and sources of carotenoids and vitamin A. *Nutrition Reviews* **52**: S67–S73.

Saari JC (2000) Biochemistry of visual pigment regeneration: The Friedenwald lecture. *Investigative Ophthalmology and Visual Science* **41**: 337–348.

Stephensen CB (2001) Vitamin A, infection, and immune function. *Annual Reviews of Nutrition* **21**: 167–192.

Sun SY and Lotan R (2002) Retinoids and their receptors in cancer development and chemoprevention. *Critical Reviews in Oncology and Hematology* **41**: 41–55.

Wolf G (1984) Multiple functions of vitamin A. *Physiological Reviews* **64**: 873–937.

Deficiency and Interventions

K P West Jr, Johns Hopkins University, Baltimore, MD, USA

Vitamin A (VA) deficiency is the leading cause of pediatric blindness, increases risk of severe infection, and is an underlying cause of child mortality in many developing countries. Night blindness, the mildest ocular manifestation of VA deficiency, has been recognized since antiquity, with the condition depicted in bas-relief on the Egyptian pyramid in Sakura, dating to the Middle Kingdom, and Hippocrates in the fourth century BC recognizing and treating the condition with animal liver. Corneal destruction and consequent blindness, as well as milder conjunctival lesions of xerophthalmia, were linked to dietary insufficiency in the eighteenth and nineteenth centuries, with cod liver oil emerging as recommended treatment for the various conditions of night blindness, Bitot's spots, and corneal necrosis (keratomalacia) more than a century ago. Discovery in the early twentieth century of 'fat-soluble A,' an ether-soluble factor in butter and egg yolk critical for sustaining growth, health, and vision in animals, accelerated recognition and treatment of xerophthalmia in children as well as decades of subsequent research that led to the synthesis of vitamin A and its analogues, an understanding of the vitamin's roles in the visual cycle, and discovery of the vitamin's involvement in maintaining epithelial, immune, hematopoietic, and osteoid function and multiple facets of human health.

Vitamin A Deficiency Disorders

Vitamin A is essential for maintaining normal retinal function and differentiation of rapidly dividing, bipotential cells. These regulatory roles give rise to specific manifestations of hypovitaminosis A, such as poor photoreceptor function leading to night blindness, metaplasia, and keratinization of mucosal epithelial surfaces leading to clinical abnormalities of conjunctival and corneal xerosis as well as epidermoid metaplasia and other epithelial defects throughout the respiratory, genitourinary, and gastrointestinal tracts and glandular ducts. Deficiency can also impair development or functioning of multiple arms of the immune system that can weaken host defenses against infection. Collectively, all pathophysiological consequences attributed in varying degrees to VA depletion are termed 'vitamin A deficiency disorders (VADD) (**Figure 1**).

Biochemical Depletion

Tissue depletion of vitamin A, although not a disorder per se, precedes the functional consequences of deficiency. In uncomplicated hypovitaminosis A, plasma retinol tends to be homoeostatically controlled until body (primarily liver) stores are low, after which plasma concentration declines. Plasma retinol may also decline during states of chronic inflammation and clinically significant infection, in parallel with raised circulating concentrations of acute phase proteins, likely reflecting increased tissue delivery, reduced hepatic mobilization via retinol binding protein, and increased urinary loss of vitamin A. Plasma retinol gradually normalizes during recovery from infection if there are adequate hepatic stores of the vitamin. If not, infection can leave the host more tissue depleted and at risk. Despite nondietary influences, plasma or serum retinol measurement remains the most common biochemical index of vitamin A status. Vitamin A deficiency is generally diagnosed at a serum retinol concentration below a cutoff of 0.70 μmol/l (20 μg/dl), below which 20 to >50% of concentrations occur in a VA-deficient population compared to <3% of well-nourished societies. A serum retinol concentration of <0.35 μmol/l is indicative of severe deficiency. Decrements in serum retinol concentration below these cutoffs are associated with marked increases in risk of xerophthalmia and infection. Other indices of tissue retinol depletion include the relative dose–response, a before–after test dose difference in serum retinol that indirectly reflects hepatic retinol adequacy, breast milk retinol concentration for assessing both maternal status

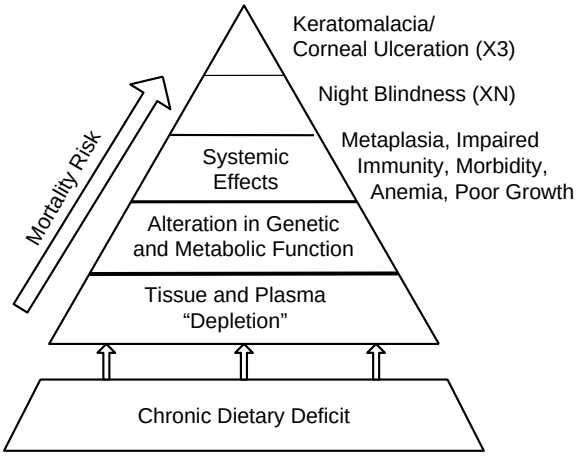


Figure 1 Concept of vitamin A deficiency disorders (VADD), due primarily to underlying chronic dietary deficit in preformed vitamin A and provitamin A carotenoids. From West KP Jr (2002) Extent of vitamin A deficiency among preschool children and women of reproductive age. *Journal of Nutrition* **132**: 2857S–2866S.

and intake adequacy of breast-fed infants, stable isotopic dilution to assess the total body vitamin A pool, impression cytology that detects early or mild metaplasia on the bulbar conjunctiva, and clinical stages of xerophthalmia.

Xerophthalmia

Conjunctival and corneal epithelium deprived of vitamin A undergoes keratinizing metaplasia. Columnar epithelial cells become squamous and mucus-producing goblet cells disappear, providing the histopathologic mechanisms for deficiency-induced xerotic (drying) changes to the ocular surfaces. VA deficiency is also required for rod vision in dim light. VA deficiency-induced night blindness often occurs with histopathologic changes on the ocular surface. Thus, night blindness and clinical eye signs are both listed under one xerophthalmia classification scheme (**Table 1**).

Night blindness Vitamin A, as retinaldehyde, is an essential cofactor in the generation of rhodopsin. This is a photosensitive pigment in rod photoreceptors of the retina that responds to light (it is 'bleached') by releasing vitamin A and initiating neural impulses to the brain that permit vision under conditions of low illumination. The utilization and recycling of vitamin A in this process is known as the 'visual cycle.' Hypovitaminosis A restricts rhodopsin production, which in turn raises the scotopic (low light) visual threshold. Gradually, a perceptive threshold is reached that leads to recognition of night blindness (XN), the earliest symptom of xerophthalmia. It is marked by an inability to move about in the dark. Children between 1 and 5 years of age and pregnant women appear to be at greatest risk of XN. Where endemic, there is often a local term for XN that translates into 'evening' or 'twilight' blindness or 'chicken eyes' (lacking rod cells, chickens cannot see at night), making the condition readily detectable by history. Typically, gestational night blindness resolves spontaneously with child birth and expulsion of the placenta, likely relieving maternal metabolic demands for vitamin A.

Table 1 WHO and IVACG classification and minimum prevalence criteria for xerophthalmia and vitamin A deficiency as a public health problem

Definition (code)	*Minimum prevalence (%)*	*Highest risk period*
Children 1–5 years of age		
Night blindness (XN)	1.0	2–6 years
Conjunctival xerosis (XIA)	—	—
Bitot's spots (X1B)	0.5	2–6 years
Cornea xerosis (X2)/corneal ulceration (X3A)/keratomalacia (X3B)	0.01	1–3 years
Xerophthalmic corneal scar (XS)	0.05	>1 year
Deficient serum retinol (<0.70 μmol/l)	15.0	<6 months; 1–5 years
Pregnant/lactating women		
Night blindness (XN) during most recent pregnancy	5.0	3rd trimester
Low serum retinol (<1.05 μmol/l)	20.0	

Adapted from Sommer A and Davidson FR (2002) Assessment and control of vitamin A deficiency: The Annecy Accords. *Journal of Nutrition*. **132**: 28455–28505.

Conjunctival xerosis and Bitot's spots Early xerosis of the conjunctiva can be detected subclinically by filter paper impression cytology, showing distorted, enlarged, and noncontiguous sheaths of epithelial cells and the disappearance of goblet cells. In advanced vitamin A deficiency, xerosis appears clinically as a dry, unwetable surface of the bulbar conjunctiva (X1A). The affected areas are usually overlaid with superficial white, cheesy, or foamy patches of triangular or oval shape that consist of desquamated keratin and bacteria (often the xerosis bacillus). These are known as Bitot's spots (X1B). They are nearly always bilateral, found temporal (and, in more advanced cases, also nasal) to the corneal limbus, and more reliably diagnosed than X1A. Bitot's spots are not blinding but are reflective of chronic moderate to severe systemic depletion of vitamin A.

Corneal xerophthalmia Corneal xerophthalmia is manifested in increasingly severe stages. The earliest corneal lesions appear as superficial punctate defects, evident with a slit lamp, that with advanced deficiency become more numerous and concentrated. The cornea is considered xerotic (X2) when punctate keratopathy covers large areas of the surface, rendering a hazy, nonwetable, lusterless, and irregular appearance on handlight examination. Stromal edema may be present. In more severe cases, thick, elevated xerotic plaques may form. Usually, both eyes are affected. Corneal ulcers (X3A) can be sharply demarcated, round or oval defects that are usually shallow but may also perforate the cornea. Healed ulcers form a leukoma (scar) or adherent leukoma if the iris has plugged the perforated ulcer. Most ulcers occur peripheral to the visual axis and thus may not threaten central vision if treated promptly. Keratomalacia (X3B) refers to a full-thickness softening and necrosis of the corneal stroma that can cause protruding, opaque, yellow to gray lesions to form (**Figure 2**). These tend to collapse or slough off, leaving a

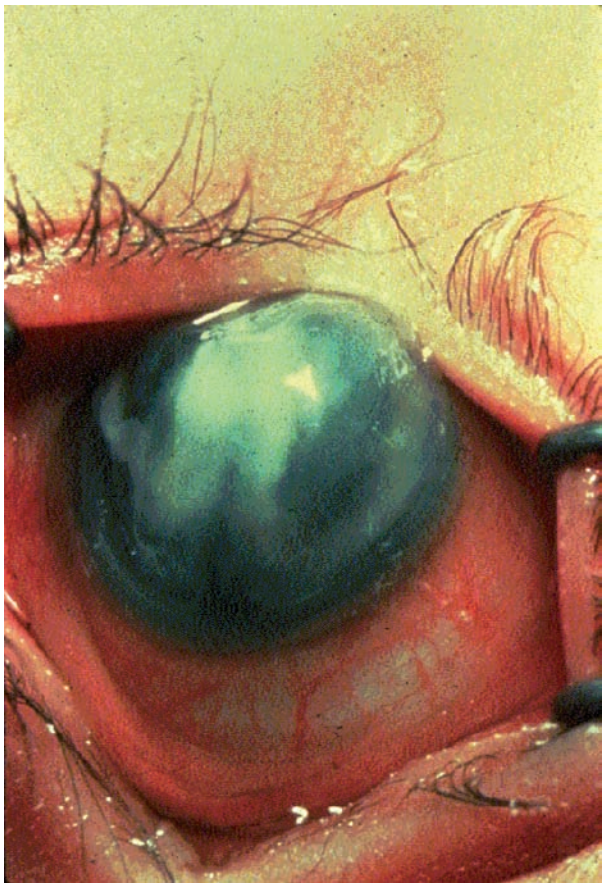


Figure 2 Keratomalacia. From Sommer A (1995) Vitamin A deficiency and its consequences. A Field Guild to detection and control. 3rd ed. Geneva, WHO.

descemetocele following VA treatment. Keratomalacia usually impairs vision in the involved eye, although the degree of visual loss depends on the location, thickness, and extent of corneal necrosis and resultant scar. Due to the generally malnourished and ill state of children with corneal xerophthalmia, the mortality rate of hospitalized cases is 4–25%.

Other VADD

Infection A bidirectional relationship exists between hypovitaminosis A and infection, each exacerbating the other, representing a classic 'vicious cycle.' Thus, infection may be considered both a cause of VA deficiency and, in terms of severity and sequellae, a 'disorder' as well. Cross-sectionally, xerophthalmia or severe hyporetinolemia has been consistently associated with higher frequencies of diarrhea, fever, and other infections, although directionality is difficult to parse from such evidence.

VA deficiency raises the risk of infection presumably due to compromised 'barrier' epithelial function and impaired innate, cell-mediated, and humoral immune mechanisms. VA-deficient Southeast Asian preschoolers (i.e., with mild xerophthalmia) were twice as likely to develop acute respiratory infection and (in Indonesia) three times more likely to develop diarrhea over subsequent 3- to 6-month periods. Deficient children are also more likely to die. This was so among Indonesian preschool children, whose risk of mortality increased with increased severity of mild eye signs (**Figure 3**). In Nepal, siblings of patients were more likely to develop the eye lesions but were also at a twofold higher risk of dying than children living in unaffected households, reflecting a clustering of child mortality risk within VA-deficient households.

Data from children and animals support the plausibility of these findings. VA-deficient children show increased bacterial adherence to respiratory epithelium, low lymphocyte counts and T helper to cytotoxic/suppressor cell ratios, and a weaker delayed-type hypersensitivity response compared to nonxerophthalmic children. In animals, VA deficiency produces keratinizing metaplasia of epithelial linings that may affect 'barrier' defenses. It also compromises acquired immunity, indicated by lymphoid atrophy, reduced numbers of circulating lymphocytes, impaired blast transformation responses to antigen, T cell-dependent antibody responses, and natural killer cell activity, and a greatly increased risk of infection and death.

Anemia and Poor Growth Children with xerophthalmia and night blind mothers tend to be anemic relative to peers without eye disease. VA-supplemented trials often show improvement in indicators of iron status, including reductions in anemia. Mechanisms involved in this interaction are not clear but may involve enhanced iron absorption, storage,

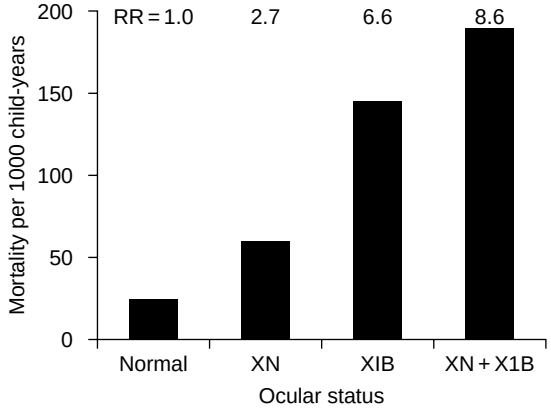


Figure 3 Risk of mortality among ~3500 Indonesian preschool children by ocular status at the outset of each 3-month interval. RR, relative risk of mortality. Adapted from Sommer A *et al.* (1983) Increased mortality in children with mild vitamin A deficiency. *Lancet* **2**: 585–588.

and transport as well as direct effects on hematopoiesis in the presence of adequate iron stores.

VA deficiency decelerates growth in animals and has been associated with both stunting and wasting malnutrition in children, possibly reflecting roles for the vitamin in osteogenesis and protein metabolism. Trials, however, have shown inconsistent effects of VA supplementation on child growth, possibly due to variations in the extent of infection, seasonality in dietary protein and energy adequacy, exclusion criteria, and levels of VA status among study children. It appears that VA supplementation can influence ponderal and linear growth, as well as body composition, in children for whom VA deficiency is a 'growth limiting' nutritional deficit.

Epidemiology

The epidemiology of VA deficiency provides a basis for quantifying its extent and health consequences; describing its demographic patterns, geocultural context, and dietary causes; and with which to target interventions for treatment and prophylaxis. It is mostly understood in relation to hyporetinolemia and xerophthalmia (**Table 1**), especially the noncorneal stages involving night blindness and Bitot's spots due to conjunctival xerosis. The latter eye signs are common, specific, and likely to exhibit risks relevant to more widespread hypovitaminosis A. Importantly, although both conditions represent 'mild' non-blinding xerophthalmia, they typically occur in moderate to severe systemic VA deficiency.

Magnitude

Current estimates suggest that VA deficiency (serum retinol $<0.70\,\mu$mol/l) afflicts 25% or 127 million preschool-aged children in the developing world, of whom 4–5 million have xerophthalmia. The number of potentially blinding corneal cases occurring annually is less well-known but, based on historic data, efficacy of VA interventions, and successes in controlling precipitating factors (such as measles), it is likely to be ~250,000 cases globally per year. The geographic distribution of VA deficiency, based on joint distributions of preschool xerophthalmia and hyporetinolemia, is presented in **Figure 4**. However,

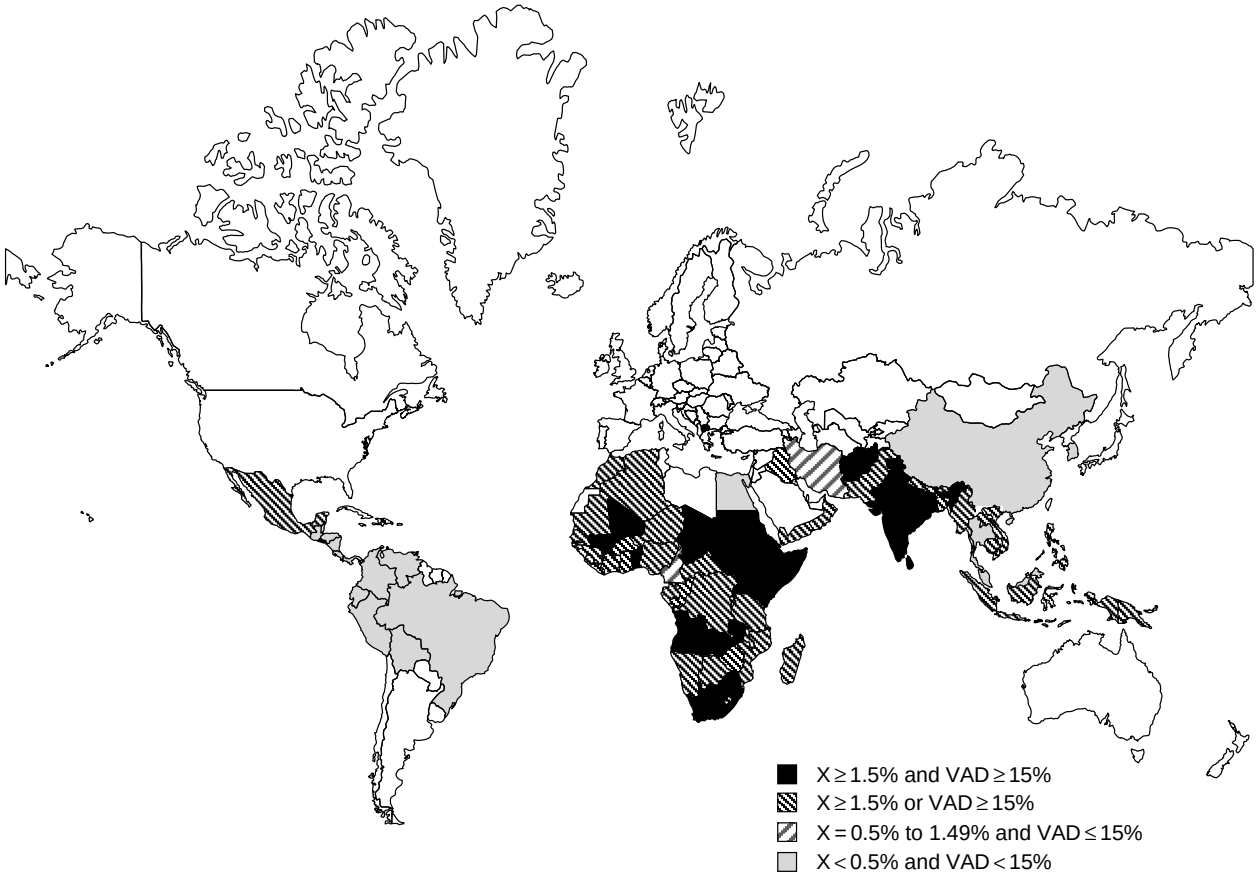


Figure 4 Global geographic distribution of xerophthalmia (X, all clinical stages) and vitamin A deficiency (VAD, serum retinol concentrations $<0.70\,\mu$mol/l) in preschool-aged children. From West KP Jr (2002) Extent of vitamin A deficiency among children and women of reproductive age. *Journal of Nutrition*.

increasing evidence shows that VA deficiency extends beyond the preschool years. For example, in Southeast Asia, >80 million children 5–15 years of age are thought to be hyporetinolemic, 9 million of whom have mild xerophthalmia. The burden of preadolescent school-aged VA deficiency in Africa and Latin America remains unestimated due to lack of population data. Adult VA deficiency is most notable in women living in undernourished conditions during and after pregnancy. Globally, approximately 20 million pregnant women have low vitamin A status (serum retinol concentration <1.05 μmol/l), of whom 7 million are thought to be deficient (serum retinol <0.70 μmol/l) and 6 million night blind, typically in the latter half of pregnancy. Regions of the world at greatest risk of VA deficiency are Southeast Asia and sub-Saharan Africa, where ~45 and 30% of all affected children and pregnant women reside, respectively.

High-Risk Groups

Preschool children and women during and following pregnancy are at highest risk for developing vitamin A deficiency and suffering its consequences. Infants in most poor populations are born with low VA status, which may predispose them to risk of blinding keratomalacia at <6 months of age. Without adequate VA from breast milk and complementary foods, VA stores can remain low to deficient throughout the first year of life and beyond. Risk of potentially blinding corneal xerophthalmia (**Figure 2**) peaks in the second through fourth years of life, typically following epidemics of acute infection such as severe measles, which can exacerbate a chronically wasted and vitamin A-deficient state. Milder stages of xerophthalmia, night blindness and Bitot's spots typically affect 1–5% of young children, with prevalence increasing from ~1 year of age through, in some cultures, the school-aged years. Boys tend to have higher rates of mild xerophthalmia than girls, possibly reflecting differences in dietary practices. In traditional societies, the increase follows weaning of children from VA-containing breast milk to a household diet chronically lacking preformed vitamin A or its carotenoid precursors.

VA deficiency persists through adolescence and into adulthood, as indicated by frequent reports of low serum retinol level (<0.70 μmol/l) in ~25% and night blindness in 5–20% of pregnant or lactating women in endemically deficient areas. Deficiency presumably results from heightened nutritional demand from late gestation through lactation superimposed on a chronic, poor dietary intake and preexisting low VA status of women. Although rarely blinding, maternal night blindness can identify high-risk women in populations in which it is associated with hyporetinolemia, anemia, wasting malnutrition, increased infant mortality, and markedly increased risks of maternal morbidity and mortality (**Figure 5**). In recent years, widespread maternal VA deficiency has been observed to coexist with HIV infection and AIDS throughout sub-Saharan Africa. However, modest, absent, or inconsistent effects of maternal VA supplementation on infant and maternal health, survival, or rates of transmission in trials suggest that VA deficiency may be more a consequence than a determinant of HIV-induced disease severity.

Geographic Clustering

Patterns of childhood hyporetinolemia or abnormal impression cytology plus xerophthalmia define the geographic risk of VA deficiency as a periequatorial nutritional problem throughout the world (**Figure 4**). Within regions, VA deficiency tends to cluster in parallel with indices of underdevelopment, low availability of food sources of vitamin A, and

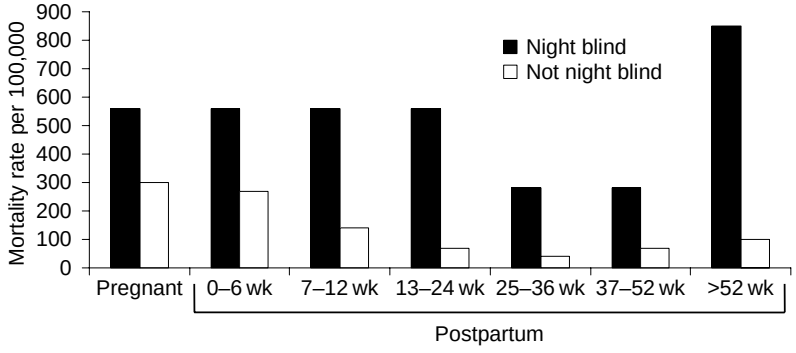


Figure 5 Mortality rates of rural Nepalese women (per 100,000 pregnancies) during and for up to 2 years following pregnancy according to whether mothers experienced night blindness ($n = 361$) or not ($n = 3052$) during pregnancy. From Christian P *et al.* (2000) Night blindness during pregnancy and subsequent mortality among women in Nepal: Effects of vitamin A and B-carotene supplementation. *American Journal of Epidemiology*.

Table 2 Household characteristics of xerophthalmia cases, controls, and the remaining Aceh study population

Household characteristic	*Cases (%) (N = 466)*	*Village-matched controls (%) (N = 466)*	*Aceh study households (%) (N = 15,915)*
Unprotected water source	47.5	43.8	41.1[a]
No private latrine	86.7	83.6	71.3[a]
Bamboo house walls	47.1	33.5	31.6[a]
Household head farms	57.3	55.5	53.4
Mother has <6 years of education	94.3	86.6	80.3[a]
History of child death	12.1	9.7	7.5[a]

[a]Significant linear trend in proportions ($p < 0.001$).
Adapted from Mele L *et al.* (1991) Nutritional and household risk factors for xerophthalmia in Aceh, Indonesia: A case-control study. *American Journal of Clinical Nutrition* **53**: 1460–1465.

disease patterns. In a large study in Aceh, Indonesia, villages in which cases of xerophthalmia were detected tended to be poorer than xerophthalmia-free communities and, within a village, cases arose more frequently from households of low socioeconomic standing than did children with normal eyes (controls) (**Table 2**). Also, in multiple surveys in Africa and Asia, preschool children incur an approximately twofold higher risk of having or developing xerophthalmia in villages in which at least one other child has been diagnosed compared to villages in which xerophthalmia has not been previously seen (**Table 3**). More striking is a 7- to 13-fold higher risk of xerophthalmia in siblings of cases compared to children in homes with no previous history of xerophthalmia. Maternal night blindness and childhood xerophthalmia tend to coexist at both household and community levels, reflecting the chronicity of dietary inadequacy in undernourished communities. Spatial clustering of deficiency risk seems to arise mostly from shared dietary practices in homes and villages rather than other exposures that lead to common infections. The exception to this observation is likely to be in households and communities afflicted by HIV/AIDS.

Periodicity

The occurrence of xerophthalmia can follow predictable, although not parallel, seasonal patterns in different areas of the world. Typically, a seasonal peak in VA deficiency emerges from a convergence of causal risk factors. In Southeast Asia, for example, a distinct peak in the incidence of mild xerophthalmia occurs during the late dry and early monsoon seasons (April–July). This peak follows a postharvest growth spurt in the cool dry season. It also coincides with a general scarcity of pro-vitamin A-rich vegetables and fruits and a seasonal increase in the incidence of diarrhea, respiratory infection, and measles. In this area of the world, the seasonal peak is often curbed abruptly midway through the 'mango season,' reflecting a likely impact of widespread consumption of this β-carotene-rich fruit. Periodicity, where it exists, can help identify causes and target prevention to specific times of the year.

Causes

Breast-feeding and diet Dietary risk in children refers to inadequate breast-feeding combined with low intakes of VA-rich foods from the household diet. A low dietary fat intake (e.g., $\leq$5% of calories) may restrict absorption of pro-vitamin A carotenoids from vegetables and fruits and thus also predispose children to deficiency.

Breast milk is the most important initial dietary source of vitamin A for an infant. Commonly, breast milk from marginally nourished mothers in developing countries contains ~500 μg of retinol activity equivalents (RAE) per liter, delivering 325 μg RAE per day to infants consuming a typical intake of ~650 ml per day. In the absence of a recommended dietary allowance for vitamin A, an 'adequate

Table 3 Age-adjusted village and household odds ratios for risk of xerophthalmia among preschool children[a]

	Malawi		*Zambia*		*Indonesia*		*Nepal*	
	n	*OR*[b]	*n*	*OR*	*n*	*OR*	*n*	*OR*
Village	50	1.2 (1.0–1.5)[c]	110	1.7 (0.9–3.2)	460	1.8 (1.4–2.2)	40	2.3 (1.6–3.4)
Household	2899	7.3 (3.2–16.7)	2449	7.9 (3.5–17.8)	16,337	10.5 (7.0–15.7)	2909	13.2 (6.0–29.0)

[a]Numbers of children <6 years of age in each country: Malawi, $n = 5441$; Zambia, $n = 4316$; Indonesia, $n = 28{,}586$; and Nepal, $n = 4764$.
[b]Pairwise odds ratio based on alternating logistic regression.
[c]95% confidence intervals in parentheses.
Adapted from Katz J *et al.* (1993) Clustering of xerophthalmia within households and villages. *International Journal of Epidemiology* **22**: 709–715.

intake' of 400–500 μg RAE has been set as a guide for infants in their first year, making infant VA intakes from breast milk marginal and, beyond infancy, marginally above an estimated deficient threshold thereafter (i.e., 210 μg RAE). Decrements in intake or concentration below this minimum increasingly place breast-fed children at risk, with greatest risk for those fully weaned and lacking a nutritious home diet. Studies in Asia and Africa show that breast-fed infants and toddlers are 65–90% less likely to have or develop xerophthalmia than non-breast-fed peers of the same age. Xerophthalmic children have been shown to begin weaning earlier (by ~1 month) and to be weaned ~6 months earlier than nonxerophthalmic children. Even when breast-fed, the more frequent the daily feeds, the greater the reduction in risk of xerophthalmia, reflecting the potential benefit achieved by promoting breast-feeding. However, the nutritional margin is still evident in the protective effects of vitamin A in reducing mortality in late infancy and early childhood, even where breast-feeding is ubiquitous for the first 2 or 3 years of life.

Complementary feeding affects childhood risk of VA deficiency. Indonesian preschoolers were at a two- to sixfold higher risk of xerophthalmia if food sources of vitamin A, such as dark green leaves, mango or papaya, egg, meat or fish with liver, and milk and other dairy products, were not routinely given during the first year of complementary feeding, instilling a pattern that may place children at risk throughout their first several years of life. Across undernourished regions of the world, studies reveal less frequent dietary intakes of foods rich in preformed vitamin A and carotenoid precursors by preschool-aged children with a history of xerophthalmia than children who remain free from the eye disease (**Figure 6**). Similar dietary differences are evident among women with versus those without a history of maternal night blindness.

Figure 6 Foods that protect against hypovitaminosis A (xerophthalmia), based on numerous studies. Dark line, strong evidence; dashed line, suggestive evidence. From Sommer A and West KP Jr (1996) Vitamin A Deficiency: Health, Survival and Vision. New York: Oxford University Press.

Infection As noted previously, a vicious cycle exists between VA deficiency and infection; thus, infection can be viewed as a cause of deficiency. Prospective studies show that severe infections, such as measles, chicken pox, diarrhea, and acute respiratory illness, decrease serum as well as apparent hepatic levels of retinol and increase the risk of xerophthalmia. In some settings, measles has been observed to increase the risk of children developing corneal xerophthalmia by >13-fold. In Indonesia, young children with diarrhea and acute respiratory infections were also twice as likely to develop mild xerophthalmia (XN or X1B) than apparently disease-free children. Similar patterns have been observed in undernourished populations of women, whereby maternal infection early in gestation raises the risk of becoming night blind later in pregnancy. Explanations for a role of infection as a cause of VA deficiency include decreased absorption of vitamin A, increased metabolic requirements, impaired retinol transport, greatly increased renal excretion during the acute phase response, and slow normalization of these mechanisms coupled with a chronically decreased dietary VA intake during extended recovery or repeated illness.

Impact of Interventions

VA deficiency can be prevented through direct supplementation, fortification of commonly eaten food items, or other food-based interventions that include home gardening, nutrition education, and agronomic approaches. Most evaluations have assessed the impact of direct supplementation and, occasionally, fortification on vitamin A status, xerophthalmia, survival, and other health outcomes. Data on the efficacy of dietary regimens are limited to change in vitamin A status.

Vitamin A Status

The impact of vitamin A prophylaxis on status varies by indicator, dosage and mode of delivery of the supplement, level of initial deficiency, and other risk factors. A single, high-potency supplement (210 μmol, 60 mg retinol activity equivalents or 200,000 IU) has been shown to elevate serum retinol in deficiency-prone populations for periods of 1–6 months, with most data suggesting protection from hyporetinolemia for only a few months. Continuous intake of 50–100% of the recommended allowance of vitamin A through fortified foods gradually improves and sustains adequate serum and breast milk retinol concentrations or, when assessed by indirect means, hepatic retinol

adequacy. Regular consumption of pro-vitamin A food sources (dark green leaves and yellow vegetables and fruits) has variable, although generally positive, effects on vitamin A status. Dietary carotenoid intake appears most efficacious in raising serum retinol from deficient concentrations to minimally adequate levels in children and women but often fails to optimize vitamin A status. Variations in food matrix, methods of storage and preparation, amounts of preformed vitamin A and fat in the diet, gut integrity and function, protein energy and VA status of the host, and other factors may affect dietary efficacy. Among these, food matrix factors may be most important in determining bioavailability of pro-vitamin A carotenoid for uptake, conversion, and absorption. The belief has long been held that β-carotene, the most ubiquitous pro-vitamin A carotenoid in the diet, can be converted from dietary sources to vitamin A in the body at a ratio of 6:1, but it is now recognized that conversion is far less efficient, conservatively set at 12:1 and 24:1 for other pro-vitamin A carotenoids (e.g., α-carotene and β-cryptoxanthin) for a mixed vegetarian diet. The change effectively reduces by half previous estimates of vitamin A activity in the developing world's nonanimal food supply and begins to address persistent questions of how VA deficiency can coexist amid at least a seasonal abundance of vegetables and fruits.

Xerophthalmia

Practically any intervention that delivers adequate amounts of VA will control VA deficiency. High-potency vitamin A delivered to preschool children every 4–6 months is ~90% efficacious in preventing both corneal and noncorneal xerophthalmia. Prophylactic failure (~10%) may reflect inadequacy of dosage for some children who are severely VA deficient or become ill. Xerophthalmia, on the other hand, virtually disappears in child populations consuming adequate amounts of vitamin A-fortified foods. There is less experience with regard to preventing night blindness in women, other than in large trials that suggest supplementation at recommended dietary levels may be insufficient to prevent all xerophthalmia, depending on background severity. Supervised dietary treatment has been reported to cure or improve noncorneal xerophthalmia; however, population trials to assess the impact of dietary change in preventing xerophthalmia have yet to be carried out.

Mortality

There has been extensive investigation into the effects of VA on mortality during the past two decades in preschool-aged children and, more recently, in young infants and women of reproductive age. The impact of vitamin A supplementation on preschool child mortality has been firmly established through eight controlled community trials performed in the 1980s and early 1990s involving ~160,000 children on three subcontinents (**Table 4**). In six trials, children 6 months to 6 years of age were supplemented every 4–6 months with an oral dose of vitamin A containing 60 mg retinol equivalents (RE) (or 200,000 IU). Half this dosage was provided to children <12 months of age. One study, in India, provided a small weekly dose to children and the other, in Indonesia, supplied half of a recommended allowance of vitamin A to children in treatment villages through a routinely marketed fortified mono-sodium glutamate product (a meal flavor enhancer). Rates of mortality in supplemented groups were compared to rates among children in concurrent control groups. Six of the eight trials showed reductions of 19–54% in preschool child mortality beyond either 6 or 12 months of age. Meta-analyses of data from these trials have estimated the reduction in mortality to range from 23 to 34%, with the latter value likely applicable to Southeast Asia. The estimates are remarkably consistent given differences in study designs and analytic approaches. Cumulative mortality curves from trials with positive results show a characteristic departure in mortality experience from control groups soon after initiation of vitamin A supplementation (**Figure 7**). Notably, the largest mortality impacts occurred in the two trials that mimicked a normal dietary intake of vitamin A compared to its periodic delivery as a bolus. Investigations of illnesses and events prior to death suggest that vitamin A reduced mortality associated with measles, diarrhea, and acute wasting malnutrition. A lack of apparent impact on mortality attributable to acute respiratory infection has been a perplexing, although consistent, finding across these and other trials. In contrast, administration of vitamin A, in the community prior to measles epidemics or on hospital admission as treatment for severe measles infection, has been shown to lower case fatality by ~50%.

High-potency VA (50,000–100,000 IU) may also substantially reduce early infant mortality, but this appears to depend almost entirely on the timing or age of supplement receipt. That is, in several trials involving infants of various ages through 6 months, during 4- to 6-monthly community dosing, or at the times of DPT vaccinations (typically at 6, 10, and 14 weeks of age), little or no effect in reducing mortality has been shown. However, two trials to date, in Indonesia and India, have reported 64 and 22% reductions, respectively, in mortality for infants <6 months of age when they were dosed at or soon

Table 4 Vitamin A mortality prevention trials

Location/target group	*Vitamin A dosage*[a]	*N*	*% change*[b]
Infants <1 month			
Bandung Indonesia	15 mg RAE at birth	2067	↓64*
Madurai, India	7.2 mg RAE at birth	11,619[c]	↓22*
Sarlahi, Nepal	15 mg RAE ~1–3 weeks of age	1621	↑7
Infants 1–5 months			
Sarlahi, Nepal	30 mg RAE	4617	↑4
Jumla, Nepal	15 mg RAE	1058	↑1
Infants <6 months			
Sarlahi, Nepal	7 mg RAE/week to mothers	15,987	↑4
Children 6–72 months			
Aceh, Indonesia	60 mg RAE/6 months	29,236	↓34*
West Java, Indonesia	0.81 mg RAE/day	11,220	↓46*
Tamil Nadu, India	2.5 mg RAE/week	15,419	↓54*
Hyderabad, India	60 mg RAE/6 months	15,775	↓6
Sarlahi, Nepal	60 mg RAE/4 months	28,640	↓30*
Jumla, Nepal	60 mg RAE/5 months	7197	↓29*
Khartoum, Sudan	60 mg RE/6 months	29,615	↑6
Northern Ghana	60 mg RE/4 months	21,906	↓19*
Pregnant/lactating women			
Sarlahi, Nepal	7 mg RAE/week	22,189[d]	↓44

[a]RAE, Retinol activity equivalents; trials providing 60 mg RE gave a half dose to infants <12 months.
[b]Percentage change in mortality rate among vitamin A recipients compared to controls of similar age or lifestage for all trials.
[c]Child-years of observation.
[d]Number of pregnancies.
*Indicates statistically significant differences ($p < 0.05$).
From Sommer and West (1996) Vitamin A Deficiency: Health, Survival and Vision. New York: Oxford University Press and West (2003) Vitamin A Deficiency Disorders in Children and Women Food and Nutrition Bulletin. **24**: S78–S90.

after birth with 45,000–50,000 IU of VA (**Table 4**). In the trial reported from India, the effect was restricted to low-birth-weight infants (<2500 g), suggesting an impact on growth-restricted or preterm newborns. Explanations include plausible maturational effects on an immature immune system, gut, and airway that may enhance resistance to infection months later. Additional confirmatory trials are under way in Southeast Asia and sub-Saharan Africa.

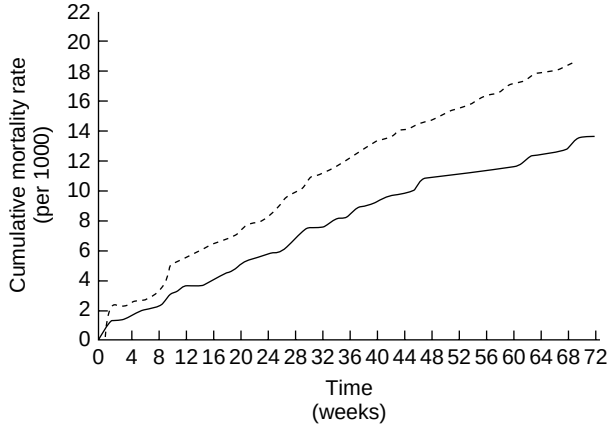


Figure 7 Cumulative mortality of children randomized to 4-monthly placebo control (dashed line) versus 200,000 IU of vitamin A (solid line) during a large community trial in Sarlahi District, rural Nepal. From Sommer A and West KP Jr (1996) Vitamin A Deficiency: Health, Survival and Vision. New York: Oxford University Press.

Finally, improving intake of VA, either preformed or as pro-vitamin A β-carotene in amounts approximating a recommended dietary intake, may reduce the risk of pregnancy-related death where maternal mortality is high and deficiency evident by widespread night blindness during pregnancy. In rural Nepal, supplementing women with VA lowered mortality related to pregnancy from 704 to 385 deaths per 100,000 pregnancies (44%), likely due to less severe infection, eclampsia, anemia, and possibly other obstetric causes. Malnourished women (e.g., those with night blindness) may likely benefit most from supplemental VA intake. Additional trials addressing this question are under way in Bangladesh and Ghana.

Morbidity

Direct effects of VA on morbidity have been difficult to establish, possibly due to variation in disease sensitivity to VA and inherent problems in measuring the incidence, duration, and severity of morbidity in community studies. Vitamin A interventions exert modest, if any, impact on the prevalence of common childhood morbidities typically obtained by history. In contrast,

VA appears to reduce the severity of potentially fatal infections, such as measles, persistent diarrhea, dysentery, and falciparum malaria, especially in the presence of wasting malnutrition. The protective effect becomes stronger with episode severity. Thus, febrile illnesses appear to be more responsive to vitamin A than nonfebrile events. Illnesses for which care is sought show a response to VA. In a large trial in Ghana, VA supplementation decreased childhood clinic visits for illness (RR = 0.88), hospitalization rates for severe disease (RR = 0.62), and severity of illness among children admitted for diarrhea compared to placebo recipients. In Brazil, prior VA receipt had no effect on children's diarrheal episodes of 1 or 2 days' duration (RR = 0.97) but was increasingly protective against episodes lasting $\geq$3 days with four or more stools per day (RR = 0.91) and episodes of $\geq$3 days with five or more stools per day (RR = 0.80). Vitamin A treatment of measles has led to fewer and less severe complications and enhanced immunologic and clinical recovery. However, multiple treatment trials report little effect of vitamin A on recovery from childhood pneumonia. This remains a paradox given decades of experimental animal evidence linking VA deficiency to extensive metaplasia and keratinization and, presumably, greater susceptibility to pathogen invasion and infection of the respiratory tract.

Management

Treatment

Children with xerophthalmia and measles should be treated immediately with oral, high-potency vitamin A (200,000 IU) according to WHO and IVACG guidelines (**Table 5**) and provided other supportive nutritional and medical therapy as indicated. Corneal lesions should be topically treated with a suitable antibiotic (e.g., tetracycline or chloramphenicol) to prevent bacterial infection. Corneal xerophthalmia typically improves with VA treatment within 1 week, with complete resolution within 4 weeks, depending on the size, thickness, and location of the lesion and nutritional and health status of the patient. Night blindness is typically cured within 24 h of VA treatment. Most Bitot's spots begin to respond within 2–5 days and disappear within 2 weeks, although some may persist, particularly in older children. High-potency vitamin A is indicated for women of reproductive age with corneal disease. For milder lesions, smaller daily (10,000 IU) or weekly (25,000 IU) doses are recommended for at least 3 months. Children presenting with severe wasting malnutrition should be given a single large oral dose (200,000 IU). It is also judicious to give cases of severe diarrhea, dysentery, respiratory infection, and exanthematous infections the same single, large oral dose of vitamin A.

Prevention

Hypovitaminosis A is prevented by increasing intakes of preformed vitamin A or pro-vitamin A carotenoids to levels that maintain adequate status. This can be done through direct supplementation of targeted risk groups, food fortification, or a number of dietary approaches that protect breast-feeding and improve the quality of the home diet.

Administration of high-potency, oral vitamin A (200,000 IU), adjusted for age (**Table 5**), on a 4- to 6-monthly basis is a common preventive approach in many developing countries. Half a dose is dispensed to infants 6–11 months and a quarter dose is given to younger infants to minimize risk of toxicity in high-risk areas. The intervention is based on the principle that a large dose of vitamin A is stored primarily in the liver, from where it is mobilized as needed. Beyond treatment, supplements can be provided during routine health care (e.g., for growth monitoring, immunization, other extension services) or more extensively and systematically in targeted populations on a regular basis (e.g., semiannual or every

Table 5 Vitamin A treatment and prevention schedules

Age	*Treatment at diagnosis*[a]	*Prevention*	
		Dosage (IU)	*Frequency*
<6 months	50,000 IU	50,000	With each of 3 doses of DPT/polio vaccine
6–11 months	100,000 IU	100,000	Every 4–6 months
12–59 months	200,000 IU	200,000	Every 4–6 months
Women	By severity of eye signs[b]	200,000	2 doses 24 h apart $\leq$6 weeks after delivery

[a]Treat all cases of xerophthalmia and measles on days 1 and 2; give an additional dose for xerophthalmia on day 14. For severe malnutrition give one dose on day 1.
[b]For women of reproductive age, give 200,000 IU only for corneal xerophthalmia on days 1, 2 and 14; for night blindness or Bitot's spots, give 10,000 IU per day or 25,000 IU per week for $\geq$3 months.
Based on Ross D (2002) Recommendations for vitamin A Supplementation. *Journal of Nutrition* **131**: 2902S–2906S.

4 months). Scaling up semiannual delivery of high-potency VA through national campaigns, such as National Immunization Days, and other mechanisms that have routinely achieved 80% or greater coverage has probably been decisive in reducing preventable, VA deficiency-induced child deaths annually from an estimated 1.7 million in 1991 to approximately 0.7 million in 2002. With accelerated immunization nearing completion, momentum established for periodic VA delivery has, in many countries, been transferred to campaigns such as National Child Health Days that, in early reports, are achieving comparable rates of coverage. Nearly four decades of distributing billions of high-potency VA supplements for population prophylaxis attests to the acceptance, effectiveness, and safety of this approach. However, it is essential to adequately inform the health and lay communities of potential benefits and risks, the latter including a 2–4% rate of (mild and self-limiting) bulging fontanel in young infants and ~5% rate of nausea, vomiting, irritability, or diarrhea following high-potency VA receipt.

Providing mothers two sequential doses of 200,000 IU each as soon after birth as possible, but always within 6 weeks to avoid excessive risk of periconceptional exposure, is a safe and effective way to improve vitamin A status of mothers and their breast-fed infants. Otherwise, supplements of up to 10,000 IU per day or 25,000 IU per week offer safe prophylactic regimens to women against VA deficiency during reproductive years.

Increasingly, developing countries are fortifying staple food items with a quarter to a full day's recommended allowance of VA to prevent deficiency in high-risk populations. Potential food vehicles should be technically fortifiable at required concentrations and consumed within a range that may be both effective in target groups and safe in the entire population. Fortification has been effectively carried out on a scaled-up basis with a limited number of products in a limited number of countries, including nonfat milk powder and vegetable oils in food assistance programs and sugar in Central America. Other food carriers have been successfully demonstrated, including monosodium glutamate (flavor enhancer) in Indonesia and the Philippines, nonrefrigerated margarine and wheat flour in the Philippines, and a powdered beverage in Tanzania. Additional foods are being fortified with VA each year. This trend will likely continue with increased use of processed foods and as the food industry becomes engaged in solving VA and other micronutrient deficiency problems.

Dietary diversification is widely held to be the most culturally appropriate and potentially sustainable approach to preventing VA deficiency. Although pilot trials show efficacy of a variety of dietary approaches for improving VA intake and status, data on the effectiveness and cost of population food-based interventions are generally lacking. Dietary intakes can be improved through home and school gardening initiatives, nutrition education, and social marketing of locally available food sources of vitamin A. However, effective dietary change requires a thorough understanding of local cultural, food system, and behavioral factors that increase the risk of VA deficiency.

See also: **Anemia**: Iron-Deficiency Anemia. **Breast Feeding**. **Children**: Nutritional Requirements; Nutritional Problems. **Fruits and Vegetables**. **Malnutrition**: Primary, Causes Epidemiology and Prevention. **Pregnancy**: Safe Diet for Pregnancy.

Further Reading

Beaton GH, Martorell R, Aronson KJ *et al.* (1993) *Effectiveness of Vitamin A Supplementation in the Control of Young Child Morbidity and Mortality in Developing Countries*, ACC/SCN State of the Art Series Nutrition Policy Discussion Paper No. 13. Geneva: World Health Organization.

Christian P, West KP Jr, Khatry SK *et al.* (1998) Night blindness of pregnancy in rural Nepal—Nutritional and health risks. *International Journal of Epidemiology* **27**: 231–237.

Christian P, West KP Jr, Khatry SK *et al.* (2000) Night blindness during pregnancy and subsequent mortality among women in Nepal: Effects of vitamin A and β-carotene supplementation. *American Journal of Epidemiology* **152**: 542–547.

De Benoist B, Martines J, and Goodman T (eds.) (2001) Special issue on vitamin A supplementation and the control of vitamin A deficiency. *Food and Nutrition Bulletin Supplement* **22**(3): 213–340.

Gillespie S and Mason J (1994) *Controlling Vitamin A Deficiency. A Report Based on the ACC/SCN Consultative Group Meeting on Strategies for the Control of Vitamin A Deficiency*. Geneva: United Nations Administrative Committee on Coordination, Subcommittee on Nutrition.

Institute of Medicine (2001) Vitamin A. In *Dietary Reference Intakes*, pp. 82–161. Washington, DC: National Academy of Medicine.

Sommer A (1995) *Vitamin A Deficiency and Its Consequences: A Field Guide to Detection and Control*, 3rd edn. Geneva: World Health Organization.

Sommer A and West KP Jr (1996) *Vitamin A Deficiency: Health, Survival and Vision*. New York: Oxford University Press.

West KP Jr (2003) Public health impact of preventing vitamin A deficiency in the first six months of life. In: Delange FM and West KP Jr (eds.) *Micronutrient Deficiencies in the First Months of Life*, Nestle Nutrition Workshop Series Pediatric Program, vol. 52, pp. 103–127. Vevey/Basel: Nestec/Karger.

West KP Jr and Sommer A (1987) *Delivery of Oral Doses of Vitamin A to Prevent Vitamin A Deficiency and Nutritional Blindness. A State-of-the-Art Review*, Nutrition Policy Discussion Paper No. 2. Rome: United Nations Administrative Committee on Coordination, Subcommittee on Nutrition.

WHO/UNICEF/IVACG (1997) *Vitamin A Supplements: A Guide to Their Use in the Treatment and Prevention of Vitamin A Deficiency and Xerophthalmia*, 2nd edn. Geneva: World Health Organization.

Vitamin B_1 *see* **Thiamin**: Physiology; Beriberi

Vitamin B_2 *see* **Riboflavin**

VITAMIN B_6

D A Bender, University College London, London, UK

Vitamin B_6 has a central role in amino acid metabolism as the coenzyme for a variety of reactions, including transamination and decarboxylation. It is also the coenzyme of glycogen phosphorylase and acts to modulate the activity of steroid and other hormones (including retinoids and vitamin D) that act by regulation of gene expression.

Severe deficiency disease has only been reported in a single outbreak in infants fed overheated formula. However, a significant proportion of people in developed countries have marginal vitamin B_6 status, and this may be associated with enhanced responsiveness to steroid hormone action and may be a factor in the development of hormone-dependent cancer of the breast, uterus, and prostate. A number of drugs have antivitamin activity, and prolonged use may lead to secondary development of pellagra, as a result of impaired tryptophan metabolism.

Estrogens do not cause vitamin B_6 deficiency. However, there is evidence that high doses of vitamin B_6 may overcome some of the side effects of estrogenic steroids used in contraceptives and as menopausal hormone replacement therapy. At very high levels of intake, supplements may cause sensory nerve damage.

Absorption and Metabolism

The main form of vitamin B_6 in foods is pyridoxal phosphate, bound as a Schiff base to lysine in dietary proteins. There is also a small amount of pyridoxamine phosphate. In plant foods a significant amount of the vitamin is present as pyridoxine, and a number of plants contain pyridoxine glycosides, which have limited availability. Heating foods can lead to the formation of (phospho)pyridoxyllysine, which has antivitamin activity.

Pyridoxal phosphate bound to proteins is released on digestion of the protein. The phosphorylated vitamers are dephosphorylated by membrane-bound alkaline phosphatase in the intestinal mucosa; all three vitamers are absorbed by carrier-mediated diffusion, followed by oxidation and phosphorylation, so there is accumulation of pyridoxal phosphate, which does not cross cell membranes, by metabolic trapping.

Both pyridoxal and the phosphate circulate in the bloodstream; the phosphate is dephosphorylated by extracellular alkaline phosphatase, and tissues take up pyridoxal by carrier-mediated diffusion, followed by metabolic trapping as phosphate esters. Pyridoxine and pyridoxamine phosphates are oxidized to pyridoxal phosphate (**Figure 1**).

Tissue concentrations of pyridoxal phosphate are controlled by the balance between phosphorylation and dephosphorylation. The activity of the phosphatases is greater than that of the kinase in most tissues so that pyridoxal phosphate that is not bound to enzymes will be dephosphorylated. Free pyridoxal either leaves the cell or is oxidized to 4-pyridoxic acid by aldehyde dehydrogenase, which is present in all tissues, and also by hepatic and renal aldehyde oxidase. 4-Pyridoxic acid is the main excretory product.

Approximately 80% of total body vitamin B_6 is in muscle, associated with glycogen phosphorylase. This does not seem to function as a true reserve of the vitamin and is not released from muscle in times of deficiency.

Metabolic Functions of Vitamin B_6

The metabolically active vitamer is pyridoxal phosphate, which is involved in many reactions of amino acid metabolism, where the carbonyl group is the reactive moiety, in glycogen phosphorylase, where it is the phosphate group that is important in catalysis, and in the release of hormone receptors from tight

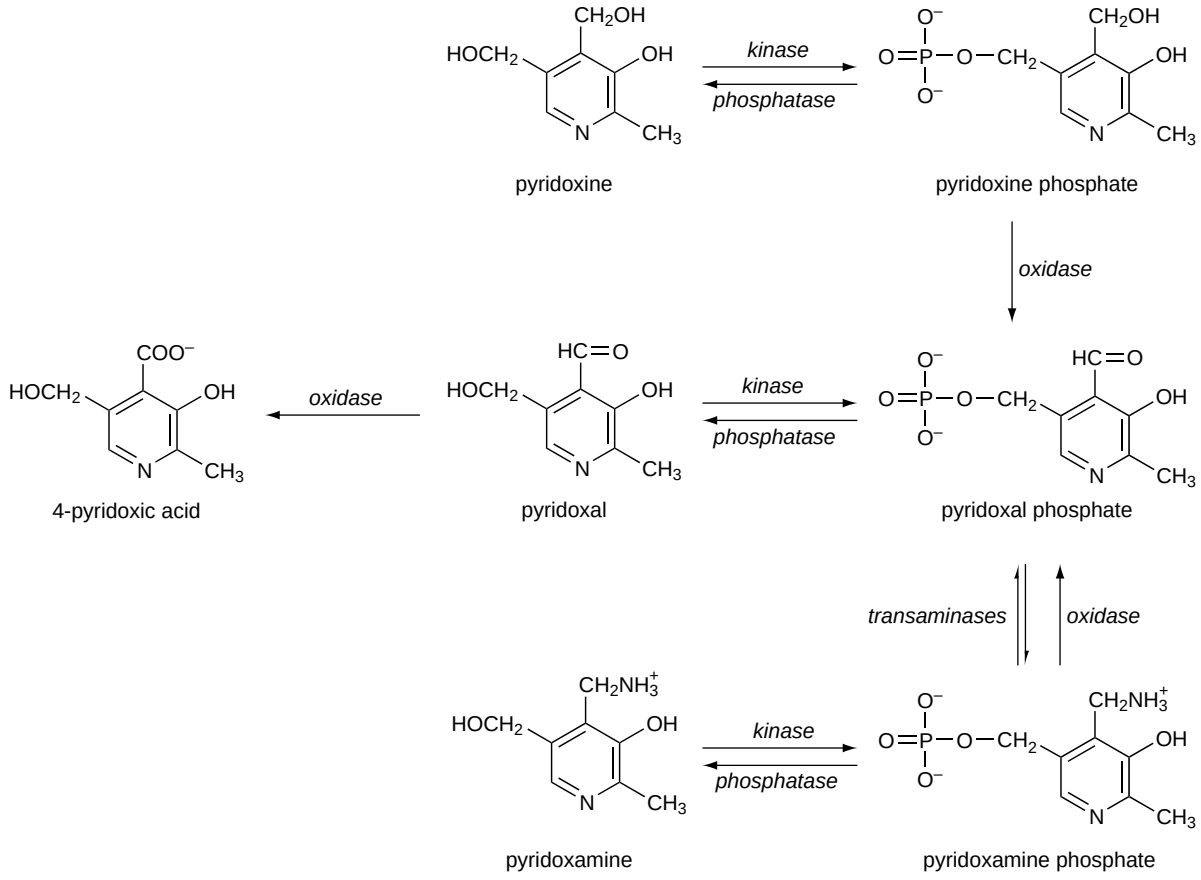


Figure 1 Metabolism of vitamin B_6.

nuclear binding, where again it is the carbonyl group that is important.

The Role of Pyridoxal Phosphate in Amino Acid Metabolism

The various reactions of pyridoxal phosphate in amino acid metabolism (**Figure 2**) all depend on the same chemical principle—the ability to stabilize amino acid carbanions, and hence to weaken bonds about the α-carbon of the substrate. This is achieved by reaction of the α-amino group with the carbonyl group of the coenzyme to form a Schiff base (aldimine).

Pyridoxal phosphate is bound to enzymes, in the absence of the substrate, by the formation of an internal Schiff base to the ϵ-amino group of a lysine residue at the active site. Thus the first reaction between the substrate and the coenzyme is transfer of the aldimine linkage from this ϵ-amino group to the α-amino group of the substrate.

The ring nitrogen of pyridoxal phosphate exerts a strong electron-withdrawing effect on the aldimine, and this leads to weakening of all three bonds about the α-carbon of the substrate. In nonenzymic model systems, all the possible pyridoxal-catalyzed reactions are observed: α-decarboxylation, aminotransfer, racemization, and side chain elimination and replacement reactions. By contrast, enzymes show specificity for the reaction pathway followed; which bond is cleaved will depend on the orientation of the Schiff base relative to reactive groups of the catalytic site.

α-Decarboxylation If the electron-withdrawing effect of the ring nitrogen is primarily centered on the α-carbon–carboxyl bond, the result is decarboxylation of the amino acid with the release of carbon dioxide. The resultant carbanion is then protonated, and the primary amine corresponding to the amino acid is displaced by the lysine residue at the active site, with reformation of the internal Schiff base.

A number of the products of the decarboxylation of amino acids are important as neurotransmitters and hormones—5-hydroxytryptamine, the catecholamines dopamine, noradrenaline, and adrenaline, and histamine and γ-aminobutyrate (GABA)—and as the diamines and polyamines involved in the regulation of DNA metabolism. The decarboxylation of phosphatidylserine to phosphatidylethanolamine is important in phospholipid metabolism.

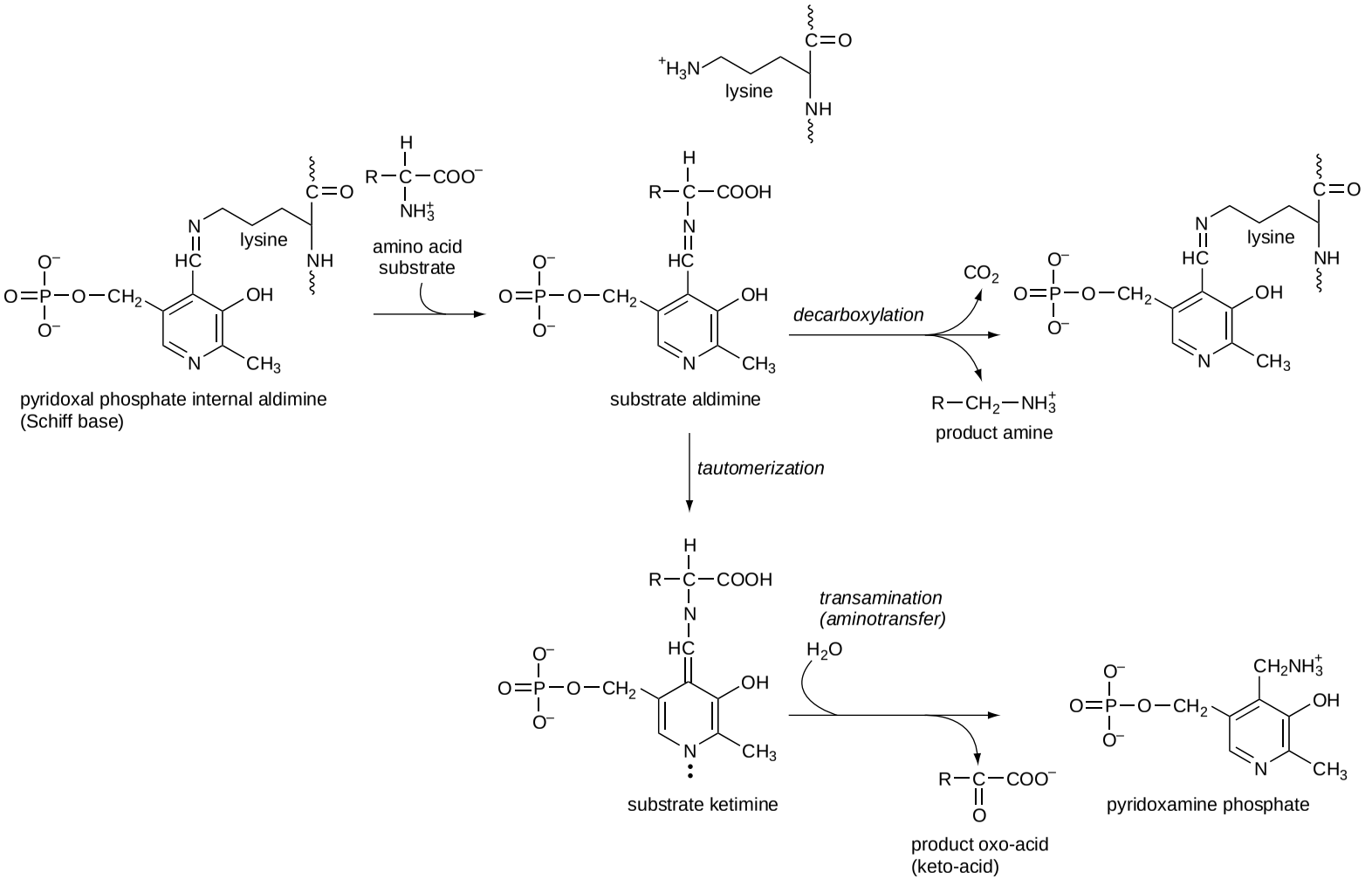


Figure 2 Roles of vitamin B_6 in amino acid metabolism.

Racemization of amino acids Deprotonation of the α-carbon of the amino acid leads to tautomerization of the Schiff base to yield a quinonoid ketimine. The simplest reaction that the ketimine can undergo is reprotonation at the now symmetrical α-carbon. Displacement of the substrate by the reactive lysine residue results in the racemic mixture of D- and L-amino acid.

Amino acid racemases have long been known to be important in bacterial metabolism since several D-amino acids are required for the synthesis of cell wall mucopolysaccharides. D-Serine is found in relatively large amounts in mammalian brain, where it acts as an agonist of the *N*-methyl-D-aspartate (NMDA) type of glutamate receptor. Serine racemase has been purified from rat brain and cloned from human brain.

Transamination Hydrolysis of the α-carbon–amino bond of the ketimine formed by deprotonation of the α-carbon of the amino acid results in the release of the 2-oxo-acid corresponding to the amino acid substrate and leaves pyridoxamine phosphate at the catalytic site of the enzyme. This is the half-reaction of transamination. The process is completed by reaction of pyridoxamine phosphate with a second oxo-acid substrate, forming an intermediate ketimine, followed by the reverse of the reaction sequence shown in **Figure 3**, releasing the amino acid corresponding to this second substrate after displacement from the aldimine by the reactive lysine residue to reform the internal Schiff base.

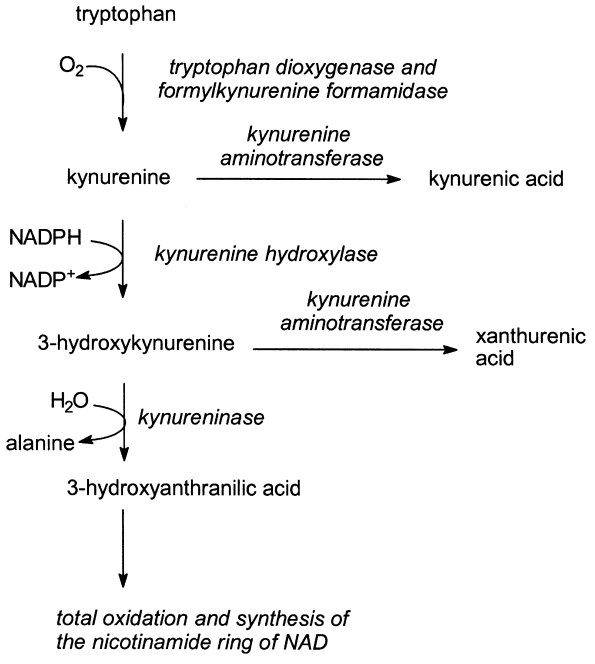


Figure 3 Tryptophan metabolism, the basis of the tryptophan load test for vitamin B_6 status.

Transamination is of central importance in amino acid metabolism, providing pathways for catabolism of most amino acids as well as the synthesis of those amino acids for which there is a source of the oxo-acid other than from the amino acid itself—the nonessential amino acids.

The Role of Pyridoxal Phosphate in Steroid Hormone Action

Pyridoxal phosphate has a role in controlling the action of hormones that act by binding to a nuclear receptor protein and modulating gene expression. Such hormones include androgens, estrogens, progesterone, glucocorticoids, calcitriol (the active metabolite of vitamin D), retinoic acid and other retinoids, and thyroid hormone. Pyridoxal phosphate reacts with a lysine residue in the receptor protein and displaces the hormone–receptor complex from DNA binding, so terminating the hormone action.

In experimental animals, vitamin B_6 deficiency results in increased and prolonged nuclear uptake and retention of steroid hormones in target tissues, and there is enhanced sensitivity to hormone action. In a variety of cells in culture that have been transfected with a glucocorticoid, estrogen or progesterone response element linked to a reporter gene, acute vitamin B_6 depletion (by incubation with 4-deoxypyridoxine) leads to a 2-fold increase in expression of the reporter gene in response to hormone action. Conversely, incubation of these cells with high concentrations of pyridoxal, leading to a high intracellular concentration of pyridoxal phosphate, results in a halving of the expression of the reporter gene in response to hormone stimulation.

Assessment of Vitamin B_6 Nutritional Status

The fasting plasma concentration of either total vitamin B_6 or, more specifically, pyridoxal phosphate is widely used as an index of vitamin B_6 nutritional status, as is the urinary excretion of 4-pyridoxic acid. The generally accepted criteria of adequacy are shown in **Table 1**.

Various pyridoxal phosphate dependent enzymes compete with each other for the available pool of coenzyme. Thus the extent to which an enzyme is saturated with its coenzyme provides a means of assessing the adequacy of the body pool of coenzyme. This can be determined by measuring the activity of the enzyme before and after the

Table 1 Indices of vitamin B_6 nutritional status

Index	*Adequate status*
Plasma total vitamin B_6	>40 nmol (10 μg)/l
Plasma pyridoxal phosphate	>30 nmol (7.5 μg)/l
Erythrocyte alanine aminotransferase activation coefficient	<1.25
Erythrocyte aspartate aminotransferase activation coefficient	<1.80
Erythrocyte aspartate aminotransferase	>0.13 units (8.4 μkat)/l
Urine 4-pyridoxic acid	>3.0 μmol/24 h >1.3 mmol/mol creatinine
Urine total vitamin B_6	>0.5 μmol/24 h >0.2 mmol/mol creatinine
Urine xanthurenic acid after 2 g tryptophan load	<65 μmol/24 h increase
Urine cystathionine after 3 g methionine load	<350 μmol/24 h increase

Data from Bitsch R (1993) Vitamin B_6. *International Journal of Vitamin and Nutrition Research* **63**: 278–282; Leklem JE (1990) Vitamin B-6: A status report. *Journal of Nutrition* **120**(supplement 11): 1503–1507; McChrisley B, Thye FW, McNair HM and Driskell JA (1988) Plasma B_6 vitamer and 4-pyridoxic acid concentrations of men fed controlled diets. *Journal of Chromatography* **428**: 35–42.

activation of any apoenzyme present in the sample by incubation with pyridoxal phosphate added *in vitro*. Erythrocyte aspartate and alanine transaminases are both commonly used; the results are usually expressed as an activation coefficient—the ratio of activity with added coenzyme to that without.

It seems to be normal for a proportion of pyridoxal phosphate-dependent enzymes to be present as inactive apoenzyme, without coenzyme. This may be a mechanism for metabolic regulation. It is possible that increasing the intake of vitamin B_6, so as to ensure complete saturation of pyridoxal phosphate-dependent enzymes, may not be desirable.

Tryptophan Load Test

The oxidative pathway of tryptophan metabolism is shown in **Figure 3**. Kynureninase is a pyridoxal phosphate-dependent enzyme, and in deficiency its activity is lower than that of tryptophan dioxygenase, so that there is an accumulation of hydroxykynurenine and kynurenine, resulting in greater metabolic flux through kynurenine transaminase and increased formation of kynurenic and xanthurenic acids. Kynureninase is exquisitely sensitive to vitamin B_6 deficiency because it undergoes a slow inactivation as a result of catalysing the half-reaction of transamination instead of its normal reaction. The resultant enzyme with pyridoxamine phosphate at the catalytic site is catalytically inactive and can only be reactivated if there is an adequate concentration of pyridoxal phosphate to displace the pyridoxamine phosphate.

The ability to metabolise a test dose of tryptophan has been widely adopted as a convenient and sensitive index of vitamin B_6 nutritional status. However, induction of tryptophan dioxygenase by glucocorticoid hormones will result in a greater rate of formation of kynurenine and hydroxykynurenine than the capacity of kynureninase, and will thus lead to increased formation of kynurenic and xanthurenic acids—an effect similar to that seen in vitamin B_6 deficiency. Such results may be erroneously interpreted as indicating vitamin B_6 deficiency in a variety of subjects whose problem is increased glucocorticoid secretion as a result of stress or illness, not vitamin B_6 deficiency.

Inhibition of kynureninase (e.g., by estrogen metabolites) also results in accumulation of kynurenine and hydroxykynurenine, and hence increased formation of kynurenic and xanthurenic acids, again giving results which falsely suggest vitamin B_6 deficiency. This has been widely, but incorrectly, interpreted as estrogen-induced vitamin B_6 deficiency: it is in fact simple competitive inhibition of the enzyme that is the basis of the tryptophan load test by estrogen metabolites.

While the tryptophan load test is a useful index of status in controlled depletion/repletion studies to determine vitamin B_6 requirements, it is not an appropriate index of status in population studies.

Methionine Loading Test

The metabolism of methionine, shown in **Figure 4**, includes two pyridoxal phosphate-dependent steps, catalysed by cystathionine synthetase and cystathionase. In vitamin B_6 deficiency there is an increase in the plasma concentration of homocysteine, and increased urinary excretion of cystathionine and homocysteine, both after a loading dose of methionine and under basal conditions. The ability to metabolize a test dose of methionine therefore provides an index of vitamin B_6 nutritional status.

Some 10–25% of the population have a genetic predisposition to hyperhomocysteinemia, which is a risk factor for atherosclerosis and coronary heart disease, as a result of polymorphisms in the gene for methylenetetrahydrofolate reductase. There is no evidence that supplements of vitamin B_6 reduce fasting plasma homocysteine in these subjects, and like the tryptophan load test, the methionine load test may be an appropriate index of status in

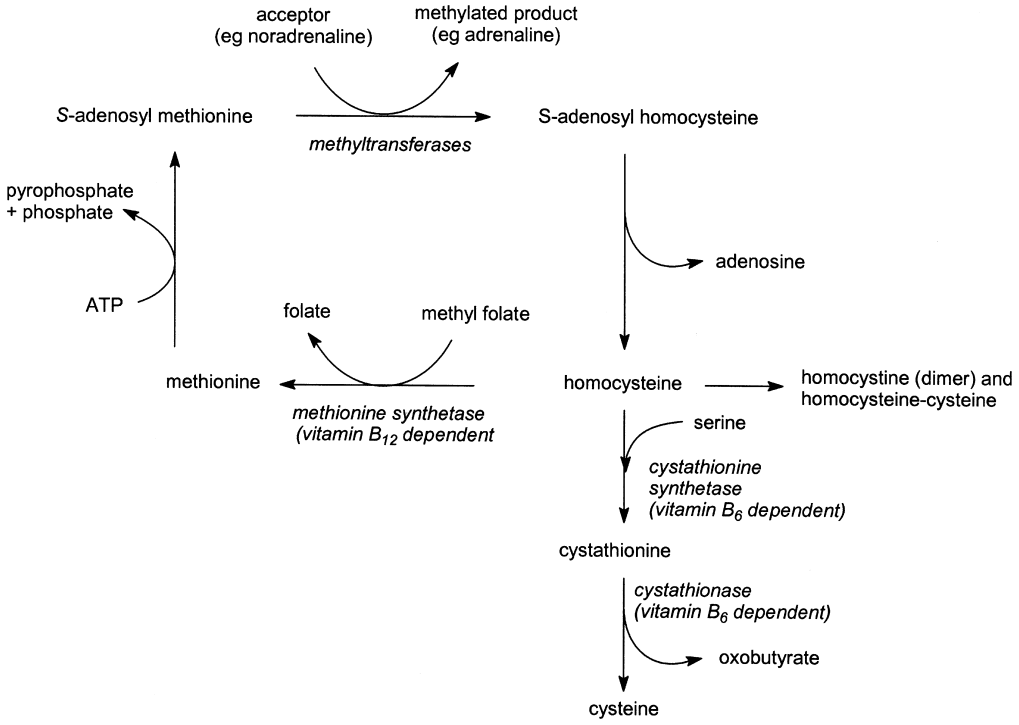


Figure 4 Methionine metabolism, the basis of the methionine load test for vitamin B_6 status.

controlled depletion/repletion studies to determine vitamin B_6 requirements, but not in population studies.

Requirements and Reference Intakes

The total body pool of vitamin B_6 is of the order of 15 μmol (3.7 mg) per kilogram body weight. Isotope tracer studies suggest there is turnover of about 0.13% per day, and hence a minimum requirement for replacement of 0.02 μmol (5 μg) per kilogram body weight—some 350 μg per day for a 70 kg adult. However, depletion/repletion studies suggest that requirements are higher than this.

Most studies of vitamin B_6 requirements have followed the development of abnormalities of tryptophan (and sometimes also methionine) metabolism during depletion and normalization during repletion with graded intakes of the vitamin.

Although some 80% of the total body pool of vitamin B_6 is associated with muscle glycogen phosphorylase, this pool turns over relatively slowly. The major metabolic role of the remaining 20% of total body vitamin B_6, which turns over considerably more rapidly, is in amino acid metabolism. Therefore, *a priori*, it seems likely that protein intake will affect vitamin B_6 requirements. People maintained on (experimental) vitamin B_6-deficient diets develop abnormalities of tryptophan and methionine metabolism faster, and their blood vitamin B_6 falls more rapidly, when their protein intake is high. Similarly, during repletion of deficient subjects, tryptophan and methionine metabolism and blood vitamin B_6 are normalized faster at low than at high levels of protein intake.

These studies suggest a mean requirement of 13 μg of vitamin B_6 per gram of dietary protein; reference intakes are based on 15–16 μg per gram of protein. At average intakes of about 100 g of protein per day, this gives an RDA of 1.4–1.6 mg of vitamin B_6. More recent depletion/repletion studies, using more sensitive indices of status, in which subjects were repleted with either a constant intake of vitamin B_6 and varying amounts of protein or a constant amount of protein and varying amounts of vitamin B_6, have shown average requirements of 15–16 μg/g of dietary protein, suggesting a reference intake of 18–20 μg/g protein.

In 1998 the reference intake in the United States and Canada was reduced from the previous RDA of 2 mg/day for men and 1.6 mg/day for women to 1.3 mg/day for both, compared with the UK RNI of 1.2 mg for women and 1.4 mg for men. The report cites six studies that demonstrated that this level of intake would maintain a plasma concentration of pyridoxal phosphate at least 20 nmol/l, although, as shown in **Table 1**, the more generally accepted criterion of adequacy is 30 nmol/l.

Possible Benefits of Higher Levels of Intake

The identification of hyperhomocysteinaemia as an independent risk factor in atherosclerosis and coronary heart disease has led to suggestions that higher intakes of vitamin B_6 may be beneficial. As shown in **Figure 4**, homocysteine may undergo either of two metabolic fates: remethylation to methionine (a reaction that is dependent on vitamin B_{12} and folate) or vitamin B_6-dependent trans-sulfuration to yield cysteine.

A number of studies have shown that while folate supplements lower fasting homocysteine in moderately hyperhomocysteinemic subjects, 10 mg/day vitamin B_6 has no effect, although they do reduce the peak plasma concentration of homocysteine following a test dose of methionine.

Vitamin B_6 Requirements of Infants

Estimation of the RDA for vitamin B_6 of infants presents a problem, and there is a clear need for further research to achieve a realistic estimate of infants' requirements. Human milk, which must be assumed to be adequate for infant nutrition, provides only some 40–100 μg per liter, or 3–8 μg of vitamin B_6 per gram of protein—very much lower than the apparent requirement for adults. There is no reason why infants should have a lower requirement than adults, and indeed since they must increase their total body pool of the vitamin as they grow, they might be expected to have a proportionally higher requirement than adults.

A first approximation of the vitamin B_6 needs of infants came from studies of those who convulsed as a result of gross deficiency caused by overheated infant milk formula in the 1950s. At intakes of 60 μg per day there was an incidence of convulsions of 0.3%. Provision of 260 μg per day prevented or cured convulsions, but 300 μg per day was required to normalize tryptophan metabolism. This is almost certainly a considerable overestimate of requirements since pyridoxyllysine, formed by heating the vitamin with proteins, has antivitamin activity, and would therefore result in a higher apparent requirement.

Based on the body content of 15 μmol (3.7 mg) of vitamin B_6 per kilogram body weight, and the rate of weight gain, the minimum requirement for infants over the first 6 months of life would appear to be 100 μg (417 nmol) per day to establish tissue reserves.

Pharmacological Uses and Toxicity of Vitamin B_6 Supplements

Supplements of vitamin B_6 ranging from 25 to 500 mg/day, and sometimes higher, have been recommended for treatment of a variety of conditions in which there is an underlying physiological or biochemical mechanism to justify the use of supplements, although in most cases there is little evidence of efficacy. Such conditions include postnatal depression, depression and other side effects associated with oral contraceptives, hyperemesis of pregnancy, and the premenstrual syndrome.

Supplements have also been used empirically, with little or no rational basis, and little or no evidence of efficacy, in the treatment of a variety of conditions, including acute alcohol intoxication, atopic dermatitis, autism, carpal tunnel syndrome, dental caries, diabetic neuropathy, Down's syndrome, Huntington's chorea, schizophrenia, and steroid-dependent asthma.

Doses of 50–200 mg per day have an antiemetic effect, and the vitamin is widely used, alone or in conjunction with other antiemetics, to minimize the nausea associated with radiotherapy and to treat pregnancy sickness. There is no evidence that vitamin B_6 has any beneficial effect in pregnancy sickness, nor that women who suffer from morning sickness have lower vitamin B_6 nutritional status than other pregnant women.

Vitamin B_6 and the Side Effects of Oral Contraceptives

Although oestrogens do not cause vitamin B_6 deficiency, the administration of vitamin B_6 supplements has beneficial effects on some of the side effects of both administered and endogenous oestrogens. The supplements act in two main areas: in normalizing glucose tolerance and as an antidepressant.

Impairment of glucose tolerance is common in pregnancy and may indeed be severe enough to be classified as gestational diabetes mellitus, which generally resolves at parturition, although in some subjects it may persist. High-estrogen oral contraceptives may also cause impaired glucose tolerance. This seems to be the result of increased tissue and blood concentrations of xanthurenic acid, because of the inhibition of kynureninase by estrogen metabolites. Xanthurenic acid forms a complex with insulin which has little or no hormonal activity. Vitamin B_6 supplements may have a beneficial effect by activating apokynureninase or kynureninase that has been inactivated by undergoing transamination.

One of the relatively common side effects of oestrogenic oral contraceptives is depression, affecting about 6% of women in some studies. This frequently responds well to the administration

of relatively large amounts of vitamin B_6 (generally in excess of 40 mg per day). Postnatal depression also responds to similar supplements in some studies.

Again, this does not seem to be due to correction of vitamin B_6 deficiency, but rather to a direct effect of pyridoxal phosphate on the metabolism of tryptophan. High concentrations of pyridoxal phosphate attenuate the response to glucocorticoid hormones; tryptophan dioxygenase is a glucocorticoid-induced enzyme, and thus its synthesis and activity will be reduced by high intakes of vitamin B_6. This reduces the oxidative metabolism of tryptophan and increases the amount available for synthesis of 5-hydroxytryptamine in the brain.

Vitamin B_6 in the Premenstrual Syndrome

The studies showing a beneficial action of vitamin B_6 in overcoming depression associated with oral contraceptives have led to the use of the vitamin in depression and other pathology associated with endogenous estrogens in the premenstrual syndrome. There is no evidence of poorer vitamin B_6 nutritional status in women who suffer from the premenstrual syndrome.

There are few well-controlled studies of the effects of vitamin B_6 in premenstrual syndrome. In general, those that have been properly controlled report little benefit from doses between 50 and 200 mg per day compared with placebo, although some studies do claim a beneficial effect. Interestingly, meta-analysis of controlled crossover trials shows that whichever treatment is used second, active vitamin or placebo, is (marginally) more effective. There is no obvious explanation for this observation.

Despite the lack of evidence of efficacy, vitamin B_6 is widely prescribed (and self-prescribed) for the treatment of premenstrual syndrome.

Vitamin B_6 for Prevention of the Complications of Diabetes Mellitus

A number of studies have suggested that vitamin B_6 (and specifically pyridoxamine) may be effective in preventing the adverse effects of poor glycemic control that lead to the development of the complications of diabetes mellitus, many of which are mediated by nonenzymic glycation of proteins. Pyridoxamine is a potent inhibitor of the rearrangement of the immediate product of protein glycation to the 'advanced glycation end-product.'

Toxicity of Vitamin B_6

Animal studies have demonstrated the development of signs of peripheral neuropathy, with ataxia, muscle weakness, and loss of balance, in dogs given 200 mg pyridoxine per kilogram body weight for 40–75 days, and the development of a swaying gait and ataxia within 9 days at a dose of 300 mg per kilogram body weight. At a dose of 50 mg per kilogram body weight, there are no clinical signs of toxicity, but histologically there is a loss of myelin in dorsal nerve roots. At higher doses there is more widespread neuronal damage, with loss of myelin and degeneration of sensory fibers in peripheral nerves, the dorsal columns of the spinal cord, and the descending spinal tract of the trigeminal nerve. The clinical signs of vitamin B_6 toxicity in animals regress after withdrawal of these massive doses, but sensory nerve conduction velocity, which decreases during the development of the neuropathy, does not recover fully. The mechanism of the neurotoxic action of vitamin B_6 is unknown.

The development of sensory neuropathy has been reported in patients taking 2–7 g of pyridoxine per day. Although there was residual damage in some patients, withdrawal of these extremely high doses resulted in a considerable recovery of sensory nerve function.

There is little evidence that intakes of up 200 mg vitamin B_6 per day for prolonged periods are associated with any adverse effects. The US Food and Nutrition Board set a tolerable upper level for adults of 100 mg/day; the EU Scientific Committee on Food set 25 mg/day.

Vitamin B_6 Deficiency

Gross clinical deficiency of vitamin B_6 is more or less unknown. The vitamin is widely distributed in foods, and intestinal flora synthesize relatively large amounts, at least some of which is believed to be absorbed.

In vitamin B_6-deficient experimental animals there are more or less specific skin lesions (e.g., acrodynia in the rat) and fissures or ulceration at the corners of the mouth and over the tongue, as well as a number of endocrine abnormalities, defects in the metabolism of tryptophan, methionine, and other amino acids, hypochromic microcytic anemia (the first step of heme biosynthesis is a pyridoxal phosphate-dependent reaction), changes in leucocyte count and activity, a tendency to epileptiform convulsions, and peripheral nervous system damage resulting in ataxia and sensory neuropathy.

Much of our knowledge of human vitamin B_6 deficiency is derived from an outbreak in the early 1950s, which resulted from an infant milk preparation which had undergone severe heating in manufacture. The probable result of this was the

formation of pyridoxyllysine by reaction between pyridoxal phosphate and the ϵ-amino groups of lysine in proteins. In addition to a number of metabolic abnormalities, many of the affected infants convulsed. They responded to the administration of vitamin B_6 supplements.

Investigation of the neurochemical basis of the convulsions in vitamin B_6 deficiency helped to elucidate the role of GABA as a neurotransmitter; GABA is synthesized by the decarboxylation of glutamate. More recent studies have suggested that the accumulation of hydroxykynurenine in the brain may be the critical factor precipitating convulsions in deficiency; GABA is depleted in the brains of deficient adult and neonate animals, while hydroxykynurenine accumulation is considerably more marked in neonates than adults—only neonates convulse in vitamin B_6 deficiency. GABA depletion may be a necessary but not sufficient condition for convulsions in vitamin B_6 deficiency.

Groups at Risk of Deficiency

A number of studies have shown that between 10 and 20% of the apparently healthy population have low plasma concentrations of pyridoxal phosphate or abnormal erythrocyte transaminase activation coefficient, suggesting vitamin B_6 inadequacy or deficiency. In most studies, only one of these indices of vitamin B_6 nutritional status has been assessed. Where both have been assessed, while each shows some 10% of the population apparently inadequately provided with vitamin B_6, few of the subjects show inadequacy by both criteria.

There is a decrease in the plasma concentration of vitamin B_6 with increasing age, and some studies have shown a high prevalence of abnormal transaminase activation coefficient in elderly subjects, suggesting that the elderly may be at risk of vitamin B_6 deficiency. It is not known whether this reflects an inadequate intake, a greater requirement, or changes in the tissue distribution and metabolism of the vitamin with increasing age.

Drug-Induced Vitamin B_6 Deficiency

A number of drugs that react with carbonyl compounds are capable of causing vitamin B_6 deficiency on prolonged use. These include the antituberculosis drug isoniazid (iso-nicotinic acid hydrazide), penicillamine, and the anti-Parkinsonian drugs, benserazide and carbidopa. In general, the main effect is impairment of tryptophan metabolism by inhibition of kynureninase, and hence the development of the niacin-deficiency disease, pellagra. The condition therefore responds to the administration of either vitamin B_6 or niacin.

See also: **Amino Acids**: Metabolism. **Diabetes Mellitus**: Dietary Management. **Homocysteine**. **Infants**: Nutritional Requirements. **Niacin**. **Pellagra**.

Further Reading

Bender DA (1987) Oestrogens and vitamin B_6—Actions and interactions. *World Review of Nutrition and Dietetics* **51**: 140–188.

Bender DA (1989) Vitamin B_6 requirements and recommendations. *European Journal of Clinical Nutrition* **43**: 289–309.

Bender DA (1999) Non-nutritional uses of vitamin B_6. *British Journal of Nutrition* **81**: 7–20.

Bender DA (2003) In *Nutritional Biochemistry of the Vitamins*, 2nd edn, pp. 232–269. New York: Cambridge University Press.

Coburn SP (1994) A critical review of minimal vitamin B_6 requirements for growth in various species with a proposed method of calculation. *Vitamins and Hormones* **48**: 259–300.

Coburn SP (1996) Modelling vitamin B_6 metabolism. *Advances in Food and Nutrition Research* **40**: 107–132.

Fasella PM (1967) Pyridoxal phosphate. *Annual Review of Biochemistry* **36**: 185–210.

Ink SL and Henderson LM (1984) Vitamin B_6 metabolism. *Annual Review of Nutrition* **4**: 455–470.

Kruger WD (2000) Vitamins and homocysteine metabolism. *Vitamins and Hormones* **60**: 333–352.

Oka T (2001) Modulation of gene expression by vitamin B_6. *Nutrition Research Reviews* **14**: 257–265.

Wiss O and Weber F (1964) Biochemical pathology of vitamin B_6 deficiency. *Vitamins and Hormones* **22**: 495–501.

Vitamin B_{12} *see* **Cobalamins**

Vitamin C *see* **Ascorbic Acid**: Physiology, Dietary Sources and Requirements; Deficiency States

VITAMIN D

Contents

Physiology, Dietary Sources and Requirements

M F Holick, Boston University Medical Center, Boston, MA, USA

Introduction

Vitamin D is a fat-soluble vitamin that is recognized for its importance for bone health. Vitamin D is neither a vitamin nor a nutrient because exposure to sunlight can provide the body's requirement for vitamin D. We take vitamin D for granted because it is casual exposure to sunlight that provides most humans with their vitamin D requirement. This recognition and the fortification of milk and other foods including some margarines, cereals, and orange juice with vitamin D has eradicated vitamin D deficiency rickets as a significant health problem for children in the US and countries that practice this fortification process. However, it is now recognized that both children and adults are at risk for developing vitamin D deficiency. The recommended adequate dietary intake for vitamin D is $200\,\text{IU}\,\text{day}^{-1}$ for children and adults <50 years old, $400\,\text{IU}\,\text{day}^{-1}$ for adults 51–70 years, and $600\,\text{IU}\,\text{day}^{-1}$ for those >70 years. However, without adequate sun exposure, $1000\,\text{IU}\,\text{day}^{-1}$ is needed. Once vitamin D is formed in the skin or ingested in the diet, it enters the bloodstream and travels to the liver and kidney where it is hydroxylated on carbons 25 and 1 to form 25-hydroxyvitamin D (25(OH)D) and 1,25-dihydroxyvitamin D (1,25(OH)D), respectively. 25-Hydroxyvitamin D is the major circulating form of the vitamin that is measured to determine the vitamin D status of patients. 1,25-Dihydroxyvitamin D is the biologically active form of vitamin D that is responsible for maintaining calcium homeostasis and bone health. It is now recognized that vitamin D deficiency increases the risk of many chronic diseases including cancers of the breast, prostate, and colon, type 1 diabetes mellitus, multiple sclerosis, rheumatoid arthritis, and heart disease.

Origin and Structure of Vitamin D

As the industrial revolution began to take hold in Northern Europe in the 17th century, it was quickly associated with a new disease that caused severe growth retardation and bony deformities in young children (**Figure 1**). This disease was commonly known as rickets or 'English disease' and plagued the children of the industrialized cities in Europe and North America for more than 250 years. Although Sniadecki in 1822 and Palm in 1890 both recognized that it was lack of exposure to sunlight that was the likely cause of rickets in children, Huldschinsky, in 1919, was the first to prove that exposure of the skin to ultraviolet radiation could cure rickets. Within 2 years, Hess and Unger reported that exposure of several rachitic children to sunlight was adequate for curing this bone-deforming disease.

Steenbock and Black and Hess independently recognized that exposure of animals and their food to ultraviolet radiation imparted antirachitic activity. This led to the recommendation for the ultraviolet irradiation of foods as a means of fortifying them with vitamin D. This resulted in the addition of provitamin D to milk followed by ultraviolet irradiation. As soon as it was possible to commercially synthesize vitamin D in large quantities, it was added directly to milk and other foods.

The first vitamin D was isolated from the irradiation of the yeast sterol ergosterol (**Figure 2**). This vitamin D was thought to be identical to that produced in the skin of animals and humans. However, studies revealed that when vitamin D produced from yeast was fed to chickens, they were unable to utilize it and developed rickets. When chickens were fed natural vitamin D from fish liver oil, rickets was prevented. This led to the conclusion that vitamin D originating from yeast was different from that in fish liver oil and animal and human skin. In 1937, this mystery was solved when the structure of provitamin D from pig skin was determined. A structural analysis revealed that provitamin D derived from ergosterol differed from that derived from pig skin. The provitamin D (ergosterol; provitamin D_2) that came from yeast had a double bond between carbons 22 and 23 and a methyl group on carbon 24.

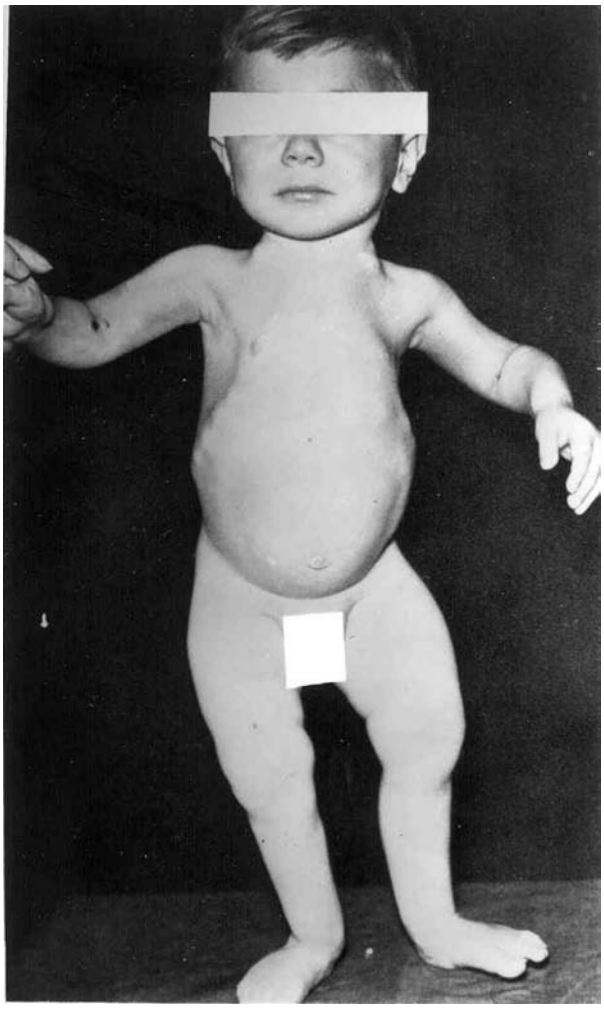


Figure 1 This is a typical presentation of a child with rickets. The child is suffering from severe muscle weakness, has bony deformities including bowed legs, and knob-like projects in the middle of his ribcage called the rachitic rosary. (Reproduced with permission from Fraser D and Scriver CR (1979) Disorders associated with hereditary or acquired abnormalities of vitamin D function: hereditary disorders associated with vitamin D resistance or defective phosphate metabolism. In: De Groot LJ *et al.* (eds.) *Endocrinology*, pp. 797–808. New York: Grune and Stratton.)

The provitamin D in animal skin had a side-chain that was identical to cholesterol, i.e., it did not contain either a double bond or methyl group on carbons 22–23 and 24, respectively, and was identified as 7-dehydrocholesterol (provitamin D_3) (**Figure 2**). The vitamin Ds generated from ergosterol and 7-dehydrocholesterol were called ergocalciferol (vitamin D_2) and cholecalciferol (vitamin D_3), respectively.

Production of Vitamin D in the Skin

During exposure to sunlight, the ultraviolet B photons with energies between 290 and 315 nm are absorbed by provitamin D_3 (7-dehydrocholesterol) in the skin. This absorption results in a photolysis of the B-ring of provitamin D_3 resulting in the formation of previtamin D_3 (**Figure 3**). However, since previtamin D_3 is thermodynamically unstable, it quickly undergoes an isomerization (rearrangement) of its triple bond system to form vitamin D_3. This isomerization process is enhanced in skin cells because the previtamin D_3 is synthesized in the cell membrane, which restricts its movement thereby accelerating the transformation of previtamin D_3 to vitamin D_3. Once vitamin D_3 is formed in the skin cell membrane, it is no longer restricted in its movement and freely translocates into the extracellular space to find its way into the dermal capillary bloodstream where it is bound to a specific vitamin D-binding protein (**Figure 3**).

An increase in skin pigmentation and zenith angle of the sun (change in latitude, season, and time of day) and the topical application of a sunscreen can markedly diminish or even prevent the production of vitamin D_3 in the skin. Over the age of ~65 years, there is a three- to fourfold decline in the synthetic capacity of the skin to produce vitamin D_3. Excessive exposure to sunlight cannot cause vitamin D_3 intoxication because once previtamin D_3 and vitamin D_3 are made in the skin, excessive quantities are rapidly destroyed by sunlight (**Figure 3**).

Absorption, Metabolism, and Excretion of Vitamin D

Vitamin D (vitamin D without a subscript represents either vitamin D_2 or D_3) is fat soluble and, therefore, once ingested vitamin D_2 and vitamin D_3 are incorporated into the chylomicron fraction and absorbed in the small intestine into the lymphatic system. Both dietary vitamin D_2 and vitamin D_3, and cutaneous vitamin D_3 enter the circulation and are bound to a specific α_1-globulin known as the vitamin D-binding protein. It is believed that this protein acts as a buffering system whereby it helps maintain circulating concentrations of 25(OH)D so that the free unbound form of 25(OH)D can enter into the renal tubular cells to be metabolized.

Neither vitamin D_2 nor vitamin D_3 possess any intrinsic biologic activity on calcium metabolism. They both require a hydroxylation on carbon 25 to form 25(OH)D (**Figure 4**). 25(OH)D is the major circulating form of vitamin D and at physiologic concentrations it has little biologic activity on calcium metabolism. It must undergo a hydroxylation

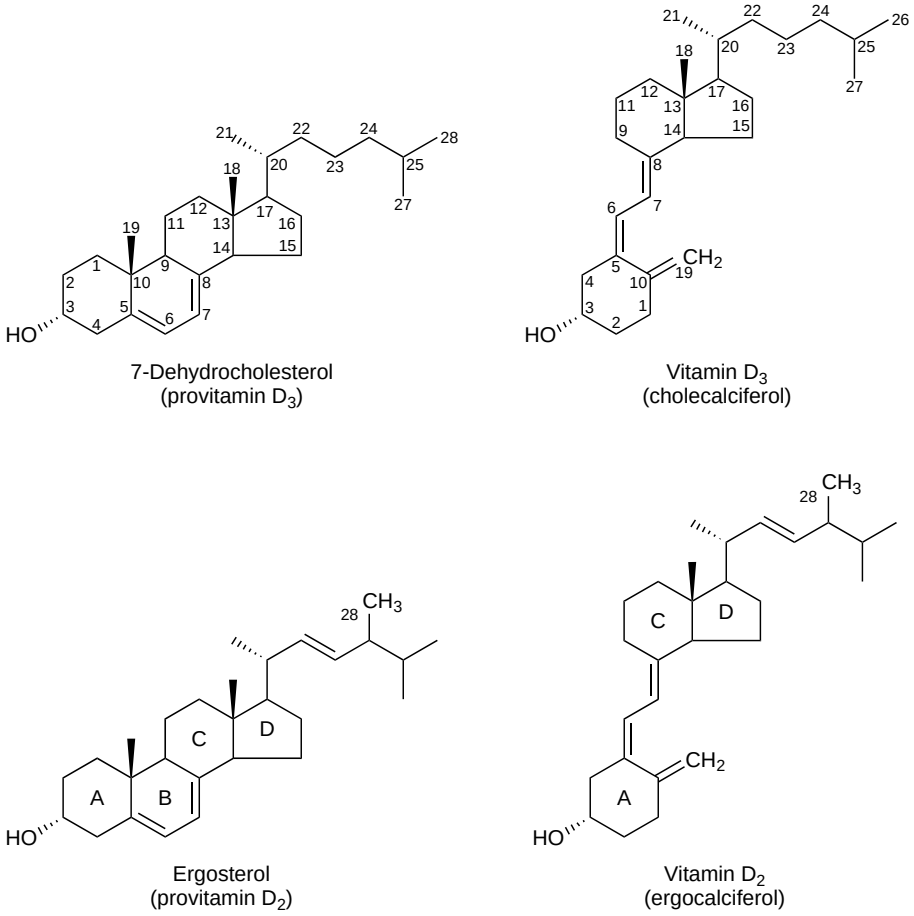


Figure 2 Structures for 7-dehydrocholesterol (provitamin D_3), ergosterol (provitamin D_2), vitamin D_3 (cholecalciferol), and vitamin D_2 (ergocalciferol). The carbons are numbered and the ring systems are labeled.

on carbon 1 in the kidney to form $1,25(OH)_2D$, the biologically active form of vitamin D (**Figure 4**). The metabolism of 25(OH)D to $1,25(OH)_2D$ is tightly regulated by parathyroid hormone (PTH) and serum phosphorus levels (**Figure 5**). PTH and low serum phosphorus levels increase the production of $1,25(OH)_2D$.

25(OH)D and $1,25(OH)_2D$ act as substrate for a 24-hydroxylase (an enzyme that attaches an hydroxyl on carbon-24), which is found in the kidney and other target tissues for $1,25(OH)_2D$. Once $1,25(OH)_2D$ is hydroxylated on carbon 24, this is the first step in its degradation to a water-soluble acid, calcitroic acid (**Figure 4**). Whereas, vitamin D is excreted in the bile, calcitroic acid is excreted by the kidney.

There continues to be speculation and controversy as to whether the 24-hydroxylation of 25(OH)D and $1,25(OH)_2D$ to 24, 25-dihydroxyvitamin D and 1,24,25-trihydroxyvitamin D, respectively, has important physiologic functions other than simply initiating the degradation of both metabolites.

Biologic Functions of Vitamin D on Calcium Metabolism

$1,25(OH)_2D$ interacts with a specific nuclear receptor that is commonly known as the vitamin D receptor (VDR) and is one of the many members of the super family of steroid hormone receptors that includes retinoic acid, thyroid hormone, glucocorticoids, and sex steroids. Once $1,25(OH)_2D$ interacts with the VDR, the complex forms a heterodimer with retinoic acid X receptor (RXR) (**Figure 6**). This new complex sits on specific segments of vitamin D responsive genes known as vitamin D responsive elements (VDREs) to either increase or decrease transcriptional activity of the vitamin D-sensitive genes such as osteocalcin, calcium binding protein (calbindin), PTH, and osteonectin (**Figure 6**).

In the intestine, $1,25(OH)_2D$ enhances the absorption of dietary calcium and phosphorus across the microvilli of the small intestinal absorptive cells (**Figure 5**). $1,25(OH)_2D$ also interacts with

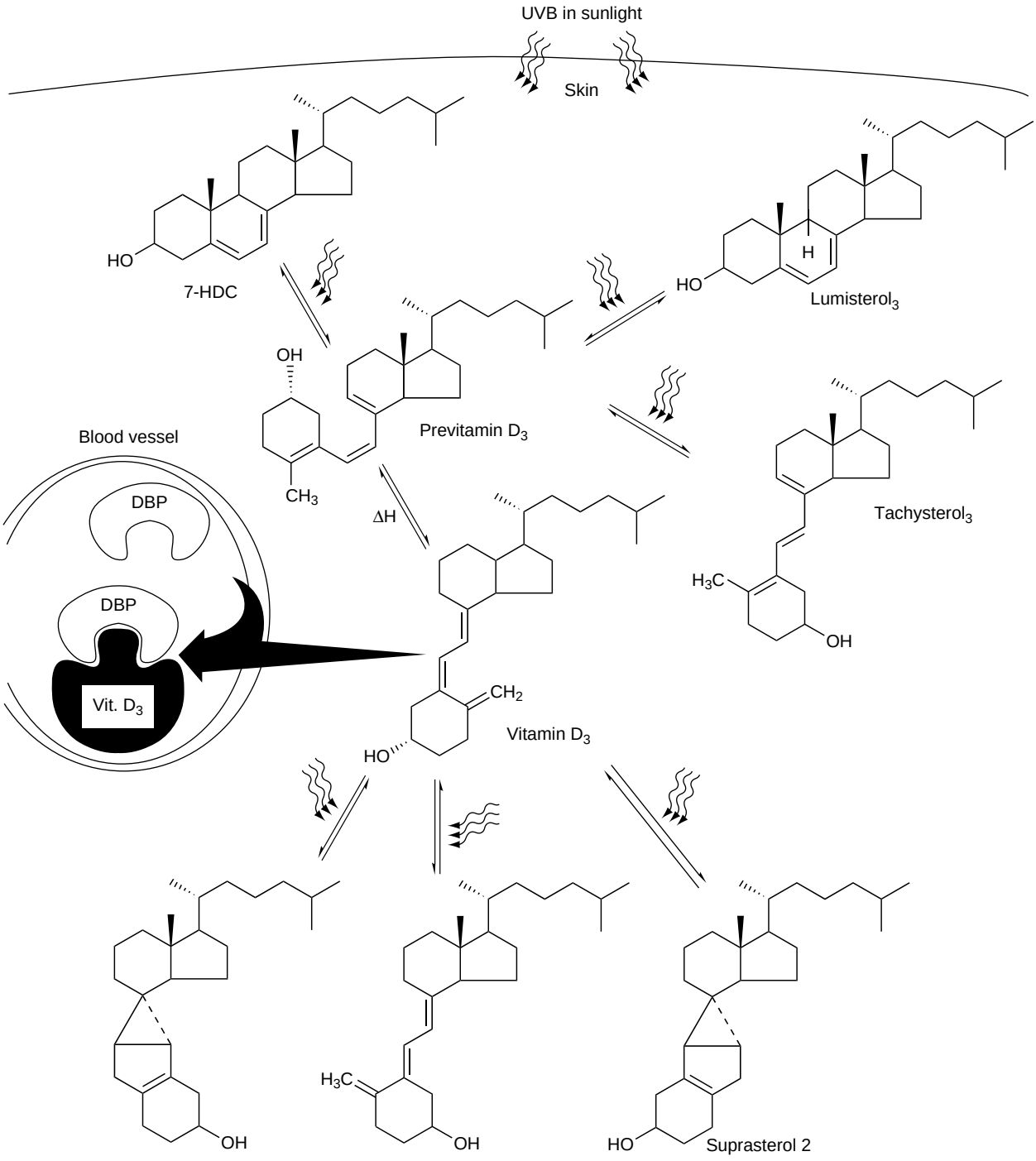


Figure 3 A schematic representation of the photochemical and thermal events that result in the synthesis of vitamin D_3 in the skin, and the photodegradation of previtamin D_3 and vitamin D_3 to biologically inert photoproducts. 7-Dehydrocholesterol (7-DHC) in the skin is converted to previtamin D_3 by the action of solar ultraviolet B radiation. Once formed, previtamin D_3 is transformed into vitamin D_3 by a heat-dependent (ΔH) process. Vitamin D_3 exits the skin into the dermal capillary blood system and is bound to a specific vitamin D-binding protein (DBP). When previtamin D_3 and vitamin D_3 are exposed to solar ultraviolet B radiation, they are converted to a variety of photoproducts that have little or no activity on calcium metabolism. (Reproduced with permission from Holick MF (1995) Vitamin D: Photobiology, Metabolism, and Clinical Applications. In: DeGroot LJ *et al.* (eds.) *Endocrinology*, 3rd edn, pp. 990–1013. Philadelphia: W.B. Saunders.)

monocytic stem cells in the bone marrow to initiate their transformation into mature osteoclasts (**Figure 5**). Thus, $1,25(OH)_2D_3$ regulates serum calcium levels by enhancing the efficiency of intestinal calcium absorption and stimulating resorption of calcium from the bone. It remains

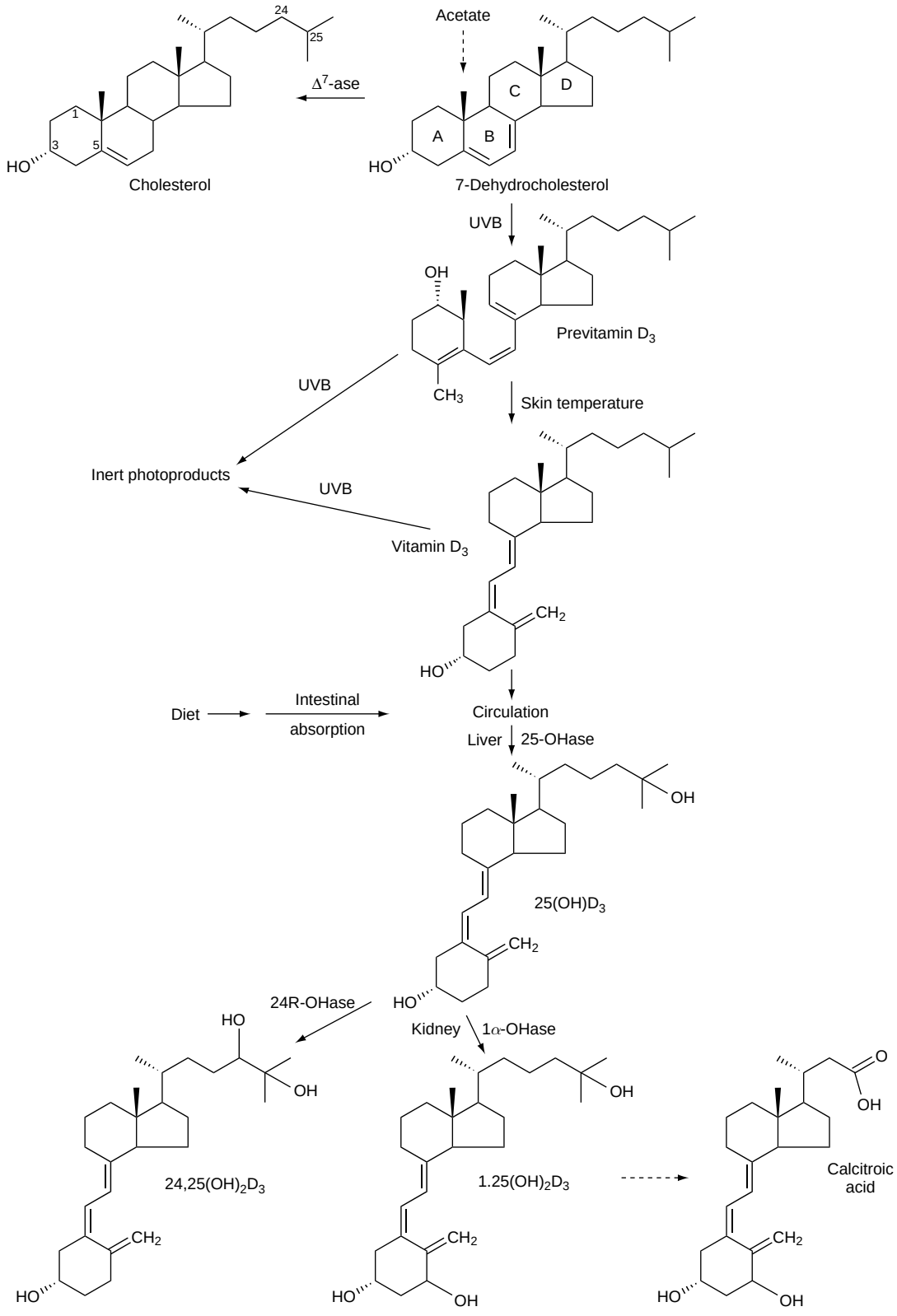


Figure 4 A schematic representation of the origin of vitamin D_3 and its metabolism in the liver by the hepatic vitamin D-25-hydroxylase. Once formed, the 25-hydroxyvitamin D_3 (25(OH)D_3) is metabolized by either a 25(OH)D-1α-hydroxylase or a 25(OH)D-24-hydroxylase. 1,25-Dihydroxyvitamin D_3 (1,25(OH)$_2D_3$) can either go to its target tissues to carry out its biologic function(s), or it can be metabolized in its side-chain and degraded to calcitroic acid. (Reproduced with permission from Holick MF (1995) Vitamin D: Photobiology, Metabolism, and Clinical Applications. In: DeGroot LJ *et al.* (eds.) *Endocrinology*, 3rd edn, pp. 990–1013. Philadelphia: W.B. Saunders.)

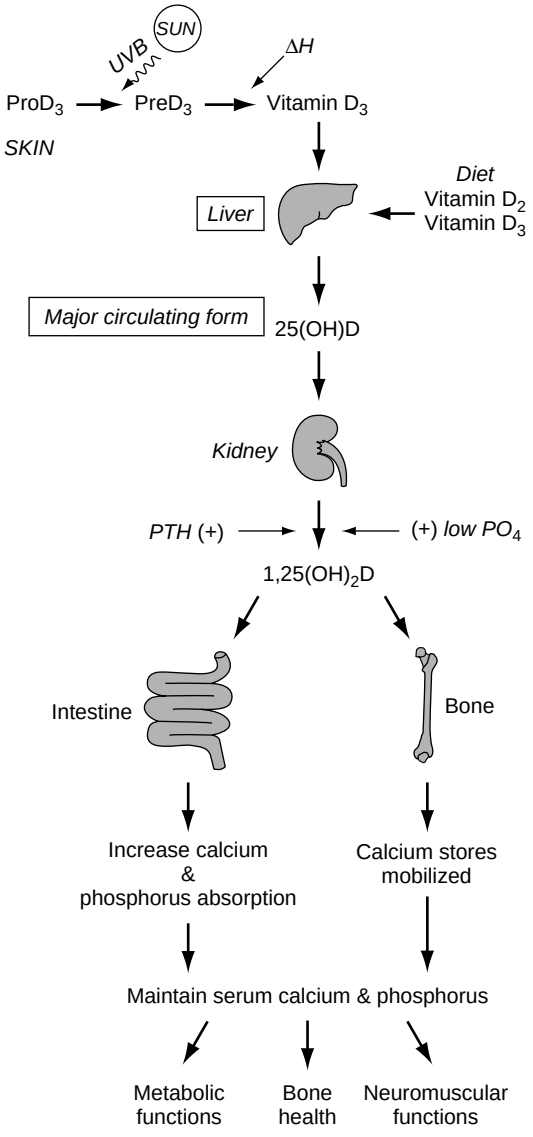


Figure 5 A schematic representation of various factors that regulate the metabolism of vitamin D to 1,25-dihydroxyvitamin D_3. (Copyright Michael F Holick (2004) Vitamin D: Importance in the prevention of cancers, type 1 diabetes, heart disease, and osteoporosis. *American Journal of Clinical Nutrition* **79**: 362–371, used with permission.)

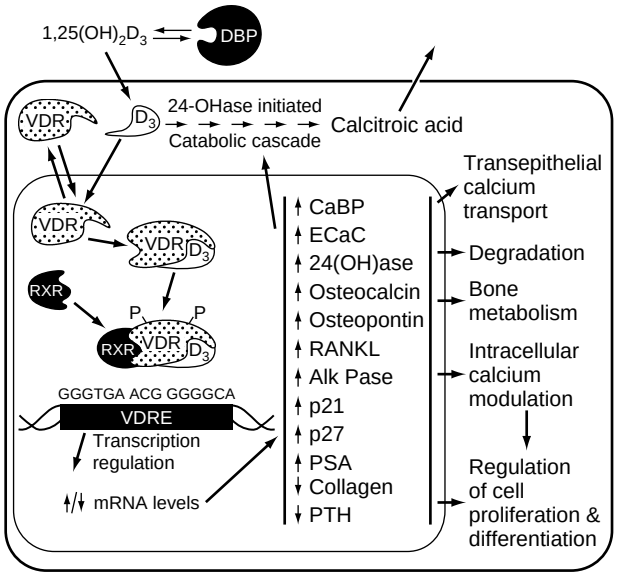


Figure 6 A schematic representation of the mechanism of action of $1,25(OH)_2D$ in various target cells resulting in a variety of biological responses. The free form of $1,25(OH)_2D_3$ enters the target cell and interacts with its nuclear vitamin D receptor (VDR), which is phosphorylated (Pi). The $1,25(OH)_2D$–VDR complex combines with the retinoic acid X receptor (RXR) to form a heterodimer, which, in turn, interacts with the vitamin D responsive element (VDRE), causing an enhancement or inhibition of transcription of vitamin D-responsive genes including calcium-binding protein (CaBP), ECaC, 24-OHase, RANKL, alkaline phosphatase (alk Pase), prostate-specific antigen (PSA), and PTH.

controversial as to whether $1,25(OH)_2D$ has any direct action on the renal handling of either calcium or phosphorus.

There are a variety of other tissues including the brain, gonads, pancreas, stomach, activated T and B lymphocytes, monocytes, and skin that have nuclear VDR. Although the exact physiologic function of $1,25(OH)_2D$'s interaction with these VDRs is not well understood, it is known that *in vivo* and *in vitro* $1,25(OH)_2D_3$ can inhibit proliferation and induce terminal differentiation of various normal and tumor cells including normal human keratinocytes. This is the reason why activated vitamin D compounds are now routinely used for the treatment of the hyperproliferative skin disorder psoriasis.

Evaluation for and Consequences of Vitamin D Deficiency

Vitamin D deficiency in young children causes rickets. As a child becomes vitamin D deficient, this results in a decrease in the efficiency of intestinal calcium absorption. There is a decline in blood-ionized calcium, which causes the parathyroid glands to produce and secrete more parathyroid hormone (PTH). PTH tries to conserve calcium by enhancing tubular reabsorption of calcium in the kidney. However, in the face of developing hypocalcemia, which could disturb neuromuscular function and a wide variety of metabolic and cellular processes, the body calls upon $1,25(OH)_2D$ and PTH to mobilize stem cells to become functional osteoclasts, which, in turn, mobilize calcium from the skeleton. In addition, PTH causes a loss of phosphorus into the urine causing hypophosphatemia. Thus, in early vitamin D deficiency the serum calcium is normal; it is the low serum phosphorus that causes the extracellular $CaXPO_4$ to be too low for normal mineralization of

bone matrix. This causes a disruption in the orderly sequence of events in the differentiation of hypertrophied chondrocytes in the epiphyseal plates resulting in their disorganization causing a widening of the epiphyseal plates (end of long bones), demineralization of the skeleton, and bony deformities (**Figure 1**).

Once the epiphyseal plates are closed later in adolescence, vitamin D deficiency can no longer cause bone deformities. Instead, there is an inability to mineralize newly deposited bone matrix leading to wide osteoid seams within the trabecular and cortical bone causing the bone disease commonly known as osteomalacia. In addition, the secondary hyperparathyroidism that results from vitamin D deficiency results in the mobilization of precious calcium stores from the bone thereby exacerbating bone loss and causing osteoporosis. This can increase a person's risk for fracture.

The hallmark for determining the vitamin D status is the measurement of the circulating concentration of 25(OH)D. The 25(OH)D is low or undetectable in vitamin D deficiency and markedly elevated in vitamin D intoxication. Measurement of $1,25(OH)_2D$ is of little value for determining the vitamin D nutritional status because its synthesis is tightly regulated. Indeed, as a person becomes vitamin D deficient, there is an increase in the secretion of PTH which, in turn, increases the production of $1,25(OH)_2D$. Thus, early in vitamin D deficiency one can see a normal fasting serum calcium, low-normal to low phosphorus, low 25(OH)D, and elevated PTH, $1,25(OH)_2D$ and alkaline phosphatase. In chronic vitamin D deficiency, all the above are seen with the exception that $1,25(OH)_2D$ is low-normal or low.

Nonskeletal Consequences of Vitamin D Deficiency

As early as 1941, it was appreciated that if you lived at higher latitudes in the US you were at higher risk of dying of cancer. A multitude of epidemiologic studies clearly show that if you live at higher latitudes and are more prone to vitamin D deficiency then you are at higher risk of dying of colon, prostate, breast, ovarian, and a variety of other cancers. It is also known that living at higher latitudes increases risk of having high blood pressure and heart disease as well as autoimmune diseases including multiple sclerosis and type I diabetes.

Essentially every cell and organ in the body requires vitamin D, i.e., they all have a VDR. It is also known that most tissues in the body can activate vitamin D. Thus, maintaining adequate levels of 25(OH)D in the circulation of at least $20\,ng\,ml^{-1}$ and preferably $30\,ng\,ml^{-1}$ is necessary for various organs including colon, breast, and prostate to convert it to $1,25(OH)_2D$, which in turn can help regulate various genes responsible for cell growth and differentiation. This could be the explanation for how vitamin D sufficiency is protective against most common cancers. The immune cells also recognize $1,25(OH)_2D_3$. This may explain why children who at 1 year of age had received 2000 IU of vitamin D a day decreased their risk of developing type I diabetes by 80%. Increasing intake of vitamin D and sun exposure has now been associated with decreased risk of developing multiple sclerosis, rheumatoid arthritis, and even Crohn's disease.

The relationship of vitamin D to cardiovascular disease is finally being understood. $1,25(OH)_2D$ inhibits the production of the blood pressure hormone renin. It also alters cardiomyocyte growth and modulates the inflammatory response of atherosclerosis (**Figure 7**).

Recommended Dietary Intake of Vitamin D

Vitamin D is very rare in foods naturally, with the exception of fatty fish and some fish liver oils. Although milk in the US is fortified with 400 IU of vitamin D/quart, several surveys during the past decade have demonstrated that approximately 80% of milk in the US contained less than 300 IU/quart. Fifty pecent of the milk samples contained less than 200 IU/quart and 15% had no detectable vitamin D. Some orange juice and other juice products are fortified with calcium and 100 IU of vitamin D_3/8 oz. Multivitamin preparations that contain vitamin D are a good source of vitamin D as are pharmaceutical preparations.

In 1997, the Institute of Medicine and the National Academy of Sciences reviewed the recommended dietary intake for several nutrients and vitamins including vitamin D. The recommended dietary allowance (RDA) was defined as the daily intake level that is sufficient to meet nutrient requirements for nearly all (97–98%) individuals in life-stage and gender group. The RDA was meant to apply to individuals and not groups. When sufficient scientific evidence was not available to calculate an estimated average requirement (EAR), i.e., a nutrient value that was estimated to meet the requirement defined by a specified indicator of adequacy in 50% of individuals in a life-stage and gender group, the Committee recommended using an adequate intake (AI). The AI is based on the observation of experimentally determined approximations of average nutrient intake by a defined population or subgroup that appears to sustain a defined nutritional state

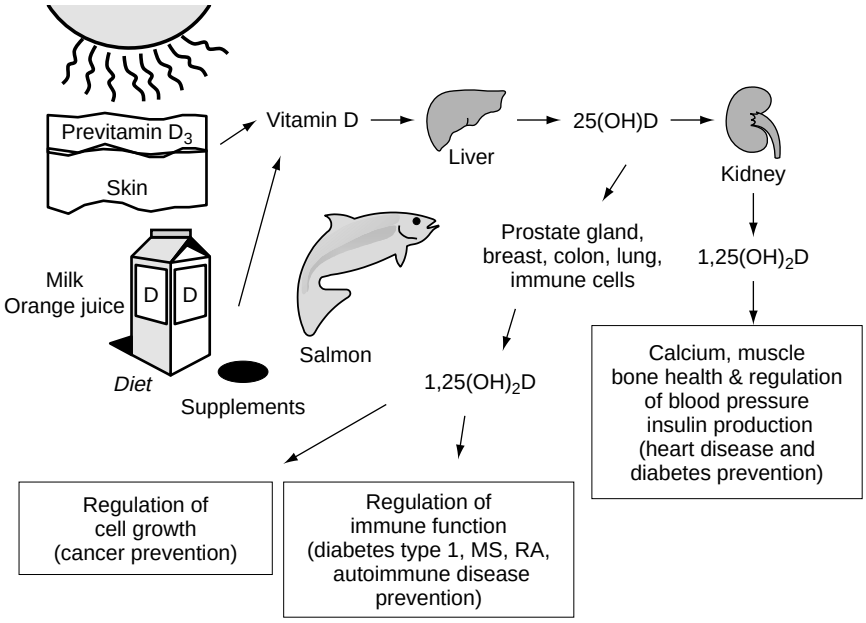


Figure 7 Photoproduction and sources of vitamin D. Vitamin D is metabolized in the liver to 25-hydroxyvitamin D [25(OH)D], which is responsible for maintaining calcium homeostasis. 25(OH)D is also converted to $1,25(OH)_2$ D in a variety of other cells and tissues for the purpose of regulating cell growth, immune function, as well as a variety of other physiologic processes that are important for the prevention of many chronic diseases. MS, multiple sclerosis; RA, rheumatoid arthritis. (Copyright Michael F Holick (2004) Vitamin D: Importance in the prevention of cancers, type 1 diabetes, heart disease, and osteoporosis. *American Journal of Clinical Nutrition* **79**: 362–371, used with permission.)

such as normal circulation nutrient values or growth. Because sunlight played such an important role in providing humans with their vitamin D requirement and, therefore, was a variable that was difficult to quantify in most studies that were reviewed by the Committee, it was concluded that an AI rather than an RDA should be used for vitamin D (**Table 1**).

Adequate Intake for Ages 0–6 Months

It is well documented that human and cows' milk has very little vitamin D naturally. Human milk contains on average between 10 and 50 IU l^{-1} (0.25–1.25 μg). This is dependent on the mother's exposure to sunlight and her vitamin D intake. Several studies have suggested that infant intakes of vitamin D of between 8.5 and 15 μg day^{-1} would provide the maximum effect on their linear growth. A study in infants from Northern China (40–47 °N) found that vitamin D supplements of 2.5, 5, or 10 μg day^{-1} resulted in 36, 29, and 2% of the infants being vitamin D deficient with 25(OH)D levels of less than 25 nmol l^{-1} (10 ng ml^{-1}). None of the infants, however, had manifestations of rickets. Chinese infants from two southern cities (22 °N and 30 °N) maintained normal vitamin D status on as little as 2.5 μg day^{-1} of vitamin D.

Table 1 Adequate Intake (AI) and Tolerable Upper Limit (UL) for Vitamin D

Age	*AI μg (IU)/d*	*UL μg (IU)/d*
0–6 m	5 (200 IU)	25 (1000 IU)
6 m–12 yr	5 (200 IU)	25 (1000 IU)
1 yr–18 yr	5 (200 IU)	50 (2000 IU)
19 yr–50 yr	5 (200 IU)	50 (2000 IU)
51 yr–70 yr	10 (400 IU)	50 (2000 IU)
71+ yr	15 (600 IU)	50 (2000 IU)

There was a seasonal variation of vitamin D status of infants when they were fed human milk only and did not receive vitamin D supplements; their 25(OH)D levels decreased in the winter due to less exposure to sunlight. However, this decrease did not occur in infants receiving a vitamin D supplement of 10 μg/day beginning at 3 weeks of age.

Therefore, based on the available literature, it was concluded that a minimum intake of 2.5 μg day^{-1} of vitamin D was adequate to prevent rickets. However, at this intake and in the absence of sunlight, infants are at risk for developing vitamin D insufficiency; therefore, it was recommended that an AI of 5 μg day^{-1} (200 IU) was prudent. 10 μg day^{-1} (400 IU), the current amount in 1 l of standard infant formula or one quart of commercial cows' milk, was not considered to be excessive.

Adequate Intake for Ages 6–12 Months

Infants between 6 and 12 months of age who were fed human milk and exposed to an average of 35 min day^{-1} of sunshine had similar 25(OH)D concentrations at 1 year of age whether the infants received 400 IU of vitamin D or no vitamin D supplementation. However, in Norway, in the winter, older infants who received an average of 5 μg day^{-1} of vitamin D had normal 25(OH)D levels that were intermediate between those of infants studied at the end of the summer and formula-fed infants.

Therefore, in the absence of any sunlight exposure, an AI of 5 μg day^{-1} was recommended. However, an intake of 10 μg day^{-1} was not considered to be excessive.

Adequate Intake for Ages 1–18 Years

There are no studies in the scientific literature that systematically evaluated the influence of different amounts of vitamin D on either serum 25(OH)D or bone mineral content in this age group. Sunlight exposure is very important for this age group to obtain its required vitamin D. In South Africa, children aged 1–8 years of mixed race showed no evidence of vitamin D deficiency. A longitudinal study in Norway, where sun exposure was presumed to vary widely over a year, an intake of vitamin D of about 2.5 μg day^{-1} from fortified margarine in children aged 8–18 years was adequate to prevent vitamin D deficiency.

During puberty, there is a need to increase the efficiency of dietary calcium absorption in order to satisfy the rapid growth of the skeleton. As a result, there is an increase in the metabolism of 25(OH)D to $1,25(OH)_2D$. Because the blood levels of $1,25(OH)_2D$ are approximately 1000 times less than 25(OH)D, this increase in metabolism does not appear to increase the requirement of vitamin D for either boys or girls between the ages of 8 and 18 years. An average daily intake of 2.5 μg day^{-1} prevented any evidence of vitamin D deficiency in Scandinavian children in this age group. However, intakes less than 2.5 μg day^{-1} in Turkish children aged 12–17 years resulted in a decrease in 25(OH)D levels consistent with vitamin D deficiency.

Therefore, based on the available literature, it appears that children between 1 and 18 years obtain most of their vitamin D from exposure to sunlight and do not normally need to ingest vitamin D. However, for children who live in far northern and southern latitudes, vitamin D supplementation may be necessary. An AI of 5 μg day^{-1} (200 IU) was recommended to maintain vitamin D sufficiency in this age group regardless of exposure to sunlight.

Adequate Intake for Ages 19–50 Years

There is only sparse literature regarding the role that sunlight and diet play in maintaining an adequate vitamin D status for men and women in this age group. This age group depends on sunlight for most of its vitamin D requirement. Regardless of exposure to sunlight, it was estimated that an AI of 5 μg day^{-1} is sufficient for preventing vitamin D deficiency in this age group.

Adequate Intake for Ages 51–70 Years

The Committee recommended a doubling of the dietary intake of vitamin D for this age group. This was based on several studies that demonstrated the importance of increasing dietary intakes of vitamin D to maximize bone health. An evaluation of 333 ambulatory Caucasian women (mean age 58 ± 6 years) found that serum PTH concentrations were elevated in the winter (between March and May) in women consuming less than 5.5 μg (220 IU) day^{-1} of vitamin D. There was no seasonal variation in serum PTH concentrations when vitamin D intakes were greater than 5.5 μg (220 IU) day^{-1}. When bone loss was evaluated between seasons in women (62 ± 0.5 years) who had a usual vitamin D intake of 2.5 μg day^{-1}, a dietary supplement of 10 μg day^{-1} decreased spinal and hip-bone density loss.

Thus, since this age group does not obtain as much of its vitamin D from exposure to sunlight, it is at more risk for developing vitamin D deficiency. Therefore, in the absence of exposure to sunlight, there appears to be an increased requirement for vitamin D in this age group and an AI of 10 μg (400 IU) day^{-1} was recommended. This is twice the previous RDA for this age group.

Adequate Intake for Ages Greater Than 70 Years

There was strong evidence-based literature that demonstrated a decrease in the circulating concentration of 25(OH)D, and an increase in the PTH level correlated with an increased risk of skeletal fractures in both the hip and spine in this age group. Studies in both men and women supplemented with 10–25 μg day^{-1} of vitamin D demonstrated reduced bone resorption, increased bone mineral content, and a decrease in vertebral and nonvertebral fractures. Therefore, because this age group is even less likely to receive an adequate amount of exposure to sunlight than adults aged 50–70 years and because they have a reduced capacity to

produce vitamin D in their skin, it was recommended that men and women in this age group, regardless of exposure to sunlight, have an AI of 15 μg (600 IU) day^{-1}, which is three times the previous RDA for vitamin D for this age group.

Adequate Intake for Pregnancy and Lactation

Although there is an increase in the metabolism of 25(OH)D to 1,25$(OH)_2$D during the last trimester of pregnancy and during lactation there is nothing in the evidence-based literature to suggest that there is an increased vitamin D requirement for pregnant and lactating women. Therefore, it was recommended that the AI of vitamin D for pregnancy and lactation follow that recommended for their age group, i.e., 5 μg (200 IU)/day^{-1}. However, the 400 IU of vitamin D found in prenatal supplements was not considered to be excessive.

Healthy Vitamin D Intakes

Since the publication of these recommendations there have been a multitude of studies that suggest that the AIs for vitamin D are inadequate if there is no exposure to sunlight. In the absence of sunlight, children above 1 year and all adults need 1000 IU of vitamin D to maintain a healthy level of 25(OH)D (above 20 ng ml^{-1}) in their circulation.

Tolerable Upper Intake Levels and Vitamin D Intoxication

An excessive intake of vitamin D can lead to vitamin D intoxication. This is characterized by a marked increase in serum 25(OH)D that is usually greater than 375 nmol l^{-1} (150 ng ml^{-1}), and is associated with hypercalciuria and hypercalcemia. This can lead to soft tissue calcification and increased risk of kidney stones. The safe upper limit for vitamin D, as recommended by the Committee, is found in **Table 1**.

Vitamin D intoxication usually occurs when a person ingests more than 5000 IU of vitamin D daily for several months. A person does not need to be concerned about becoming vitamin D intoxicated if they take a multivitamin that contains 400 IU of vitamin D, drink a quart of milk that contains 400 IU of vitamin D, and are exposed to sunlight.

See also: **Calcium**. **Lactation**: Dietary Requirements. **Pregnancy**: Nutrient Requirements. **Vitamin D**: Rickets and Osteomalacia.

Further Reading

Aksnes L and Aarskog D (1982) Plasma concentrations of vitamin D metabolites in puberty: effect of sexual maturation and implications for growth. *Journal of Clinical Endocrinology and Metabolism* **55**: 94–101.

Apperly FL (1941) The relation of solar radiation to cancer mortality in North America. *Cancer Research* **1**: 191–195.

Bouillon R, Okamura WH, and Norman AW (1995) Structure-function relationships in the vitamin D endocrine system. *Endocrine Reviews* **16**: 200–257.

Chapuy MC, Arlot M, Duboeuf F *et al.* (1992) Vitamin D_3 and calcium to prevent hip fractures in elderly women. *New England Journal of Medicine* **327**: 1637–1642.

Darwish H and DeLuca HF (1993) Vitamin D-regulated gene expression. *Critical Reviews in Eukaryotic Gene Expression* **3**: 89–116.

DeLuca H (1988) The vitamin D story: A collaborative effort of basic science and clinical medicine. *Federal Proceedings of the American Society of Experimental Biology* **2**: 224–236.

Fieser LD and Fieser M (1959) Vitamin D. In: Fieser LD and Fieser M (eds) *Steroids*, pp. 90–168. New York: Reinhold.

Grant WB (2002) An estimate of premature cancer mortality in the U.S. due to inadequate doses of solar ultraviolet-B radiation. *Cancer* **70**: 2861–2869.

Heaney RP, Dowell MS, Hale CA, and Bendich A (2003) Calcium absorption varies within the reference range for serum 25-hydroxyvitamin D. *Journal of American College of Nutrition* **22**(2): 142–146.

Holick MF (1994) Vitamin D: new horizons for the 21st century. *American Journal of Clinical Nutrition* **60**: 619–630.

Holick MF (2003) Vitamin D: photobiology, metabolism, mechanism of action, and clinical application. In: Favus MJ (ed.) *Primer on the Metabolic Bone Diseases and Disorders of Mineral Metabolism*, 5th edn, pp. 129–137. Washington DC: American Society for Bone and Mineral Research.

Holick MF (2004) Vitamin D: Importance in the prevention of cancers, type 1 diabetes, heart disease, and osteoporosis. *American Journal of Clinical Nutrition* **79**: 362–371.

Holick MF and Jenkins M (2004) *The UV Advantage* New York: iBooks.

Hypponen E, Laara E, Jarvelin M-R, and Virtanen SM (2001) Intake of vitamin D and risk of type 1 diabetes: a birth-cohort study. *Lancet* **358**: 1500–1503.

Krall E, Sahyoun N, Tannenbaum S, Dallal G, and Dawson-Hughes B (1989) Effect of vitamin D intake on seasonal variations in parathyroid hormone secretion in postmenopausal women. *New England Journal of Medicine* **321**: 1777–1783.

Luscombe CJ, Fryer AA, French ME, Liu S, Saxby MF, and Jones PW (2001) Exposure to ultraviolet radiation: association with susceptibility and age at presentation with prostate cancer. *Lancet* **192**: 145–149.

Malabanan A, Veronikis IE, and Holick MF (1998) Redefining vitamin D insufficiency. *Lancet* **351**: 805–806.

Markestad T and Elzouki AY (1991) Vitamin D deficiency rickets in northern Europe and Libya. In: Glorieux FH (ed.) *Rickets Nestle Nutrition Workshop Series*, vol. 21, pp. 203–213. New York: Raven Press.

Specker BL, Valanis B, Hertzberg V, Edwards N, and Tsang (1985) Sunshine exposure and serum 25-hydroxyvitamin D concentrations in exclusively breast-fed infants. *Journal of Pediatrics* **107**: 372–376.

Rickets and Osteomalacia

J J B Anderson, University of North Carolina, Chapel Hill, NC, USA

Introduction

Rickets and osteomalacia, diseases of impaired mineralization of bone tissue that occur among infants/young children or during adulthood, result from nutritional deficiencies of vitamin D (cholecalciferol) and/or calcium. Rickets has more severe deformities of bones because of the continued growth of the skeleton, but osteomalacia, especially less severe disease, commonly coexists with osteoporosis in many older adults. The latter comorbidities have been increasing, at least in part, as a result of both low dietary intakes of vitamin D and calcium and insufficient skin production of vitamin D because of limited skin exposure to ultraviolet (UV) light. This problem is likely to be more widespread than currently acknowledged because of behavioral changes in technologically advanced societies.

The last few years have witnessed a resurgence of rickets and osteomalacia in the US and possibly in other technologically advanced nations. The prevalence of rickets in the US has occurred primarily among young children of color, i.e., African-Americans and Hispanics, after cessation of breastfeeding and the failure to provide adequate amounts of vitamin D-fortified milk. The prevalence of osteomalacia has been suspected to be increasing because of low serum 25-hydroxycholecalciferol (25HCC) measurements in several studies of adults. In addition, new recommendations for healthy 25-hydroxyvitamin D ($25(OH)D_3$) concentrations have emerged from interpretation of the results of studies assessing the sufficiency of vitamin D to meet requirements across the life cycle.

The terminology of the vitamin D metabolites has not changed, and relatively little new resulted from basic research on vitamin D or cholecalciferol skin biosynthesis and subsequent biotransformations in the liver and kidney (**Figure 1**). Understandings of the role of the hormonal form of vitamin D, 1,25-dihydroxyvitamin D ($1,25(OH)_2D_3$), in intestinal absorbing cells have been expanded. In addition, new information suggests that the consumption of dietary calcium at adequate levels may reduce the critical need for vitamin D for the maintenance of serum calcium concentration. New information is also emerging on the role of vitamin D in patients with chronic renal failure and in the prevention of colon cancer.

This article reviews vitamin D status and highlights new research findings on vitamin D.

Dietary Vitamin D Intakes and Low Vitamin D Status in the US

Vitamin D intakes have not been assessed in national surveys and only rarely in research investigations involving smaller sample sizes. The few studies that have estimated vitamin D intakes typically find them to be below recommended amounts, especially among the elderly and, more recently, among adults. In the US, at least, most experts think that both intakes are too low and exposures of skin to sunlight are inadequate.

Role of the Diet in Providing Vitamin D

The few sources of vitamin D consumed in the diets of North Americans are fortified milks, fortified ready-to-eat breakfast cereals, and fish. For infants and young children who develop rickets, it has been established that they consume little milk and fish, but some cereals. No supplements containing calcium and vitamin D are ingested. For adults and the elderly, similar low consumption patterns of vitamin D-rich foods exist. Therefore, evidence strongly supports low intake of vitamin D as a major determinant of rickets and osteomalacia.

Role of Skin Biosynthesis of Vitamin D

The other major determinant is poor skin exposure to sunlight, mainly to UV-B that is responsible for the conversion of 7-dehydrocholesterol to $25(OH)D_3$ in the dermis layer of the skin. In the US, inadequate exposure has become a major contributor over the last few decades because of concerns about skin cancer and because of increased indoor activities, including television and computers. (This poor dietary consumption and poor skin production of vitamin D seems to be paralleling the increase in overweight.) Because it is even more difficult to assess skin exposure for vitamin D synthesis, it has been extremely difficult to estimate with accuracy the additional need for dietary vitamin D. Seasonal variations yield wide swings or oscillations in skin production, depending on the position of the sun. For example, in the northern hemisphere, the highest skin production rates occur in the late spring, summer, and early autumn months (May to October), whereas in the southern hemisphere, November to April are the months of the highest

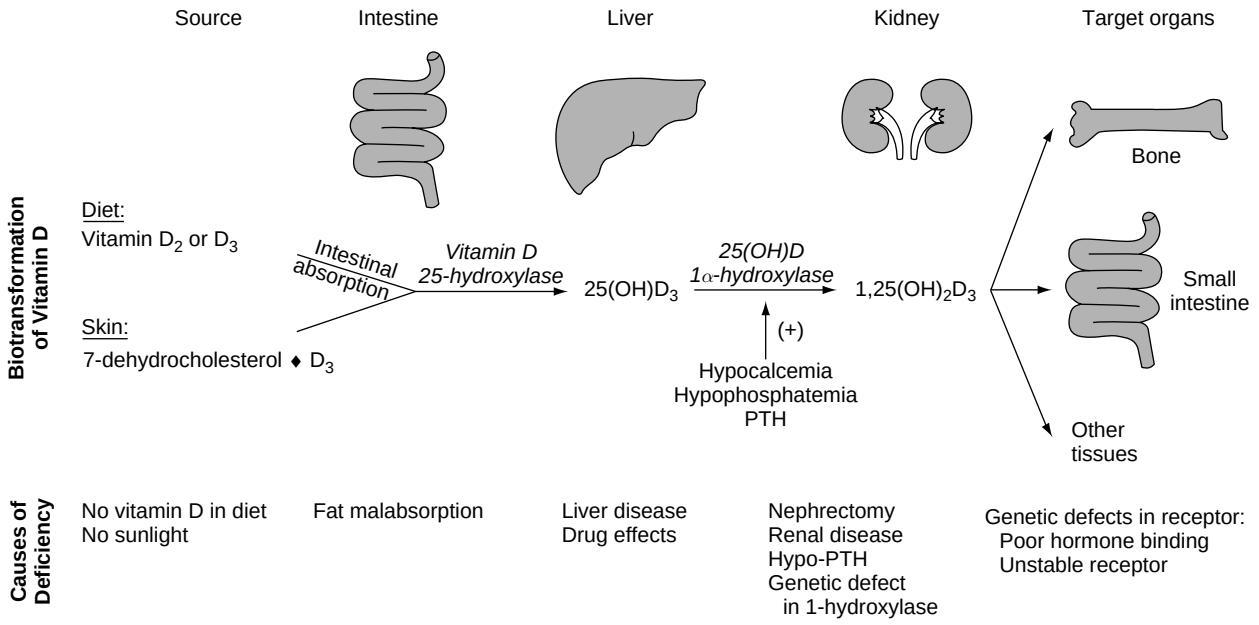


Figure 1 Causes of vitamin D deficiency. PTH, parathyroid hormone.

vitamin D production. Living near the equator extends these periods of optimal production. It is the winter months when low or even zero skin production occurs that are most problematic for the development of rickets or osteomalcia and, in the elderly, osteoporosis.

Vitamin D Recommendations

The current recommendations for consumption of vitamin D in the US are given in **Table 1**. The recommendations are defined as adequate intakes (AIs) rather than as recommended dietary allowances (RDAs) because insufficient data have been accumulated to determine estimated average requirements (EARs) across the life cycle. The AIs listed in **Table 1** are totally independent of skin production of vitamin D. Because of the uncertainty of knowing the amount of endogenous vitamin D, it is very difficult to determine, with accuracy, the amounts needed from the diet (see below).

Table 1 Recommended adequate intakes (AIs) of vitamin D in the US

Life stage group	*Age, years*	*AI, mcg/day*
Infants and Children	0–8	5
Males, Females	9–50	5
	51–70	10
	>70	15
Pregnancy	–	5
Lactation	–	5

From: Institute of Medicine, Food and Nutrition Board (1997) *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride*. Washington, DC: National Academy Press.

Primary Causes and Abnormalities of Rickets and Osteomalacia

The primary cause of rickets and osteomalacia is vitamin D deficiency and the clinical characteristics of these diseases depend on age at onset. The biochemical patterns of too low a serum concentration of $25(OH)D_3$, however, remain quite similar (**Table 2**) even though the structural effects on the skeleton differ. One common microscopic feature of

Table 2 Characteristic clinical features and blood serum measurements in rickets and osteomalacia

Age	*Clinical features*	*Blood serum measurements*
Children	Skeletal deformations (rickets) Impaired growth Undermineralized bone	Hypocalcemia Hypophosphatemia Secondary hyperparathyroidism Low 25-hydroxyvitamin D Elevated alkaline phosphatase
Adults	Undermineralized bone (osteomalacia) Fractures	Hypocalcemia Hypophosphatemia Low 25-hydroxyvitamin D Elevated alkaline phosphatase Elevated osteocalcin[a]

[a]This finding has not been consistently reported.

the skeleton is that both rickets and osteomalacia have unmineralized bone matrix (osteoid), also known as widened osteoid seams.

In rare cases, e.g., in central Nigeria, rickets may occur despite elevated serum $25(OH)D_3$ concentrations. The rickets in these cases results from severely inadequate dietary calcium. It is likely that most cases of rickets result from a combination of insufficient skin biosynthesis and inadequate dietary intake of vitamin D.

Other biochemical changes include depressed serum concentrations of calcium and inorganic phosphate, largely because of insufficient intestinal absorption directly relating to too little of these ions in the usual diet. Serum alkaline phosphatase, especially bone-specific alkaline phosphatase, is elevated because of osteoblastic cell overproduction when these cells attempt to form new bone tissue.

Radiographic changes, the primary diagnostic evidence for many years, show widened growth plates of the long bones and reduced bone density (translucence) in rickets. In osteomalacia, unmineralization is evident and pseudofractures may be visible in nonweight-bearing bones.

Secondary Causes and Abnormalities of Rickets and Osteomalacia

The secondary causes of rickets and osteomalacia that result from vitamin D deficiency are illustrated in **Figure 1**. For example, in liver disease, serum concentrations of $25(OH)D_3$ are invariably too low, and in renal disease too little of the hormonal form of vitamin D, $1,25(OH)_2D_3$, is produced. Other causes relate to reduced cell receptor responsiveness to the hormone because of genetic mutations and, hence, inappropriate adaptations that normally contribute to conservation of calcium and/or phosphate ions.

Public Health Consequences of Vitamin D Deficiency

Despite the earlier belief that these classical defiency diseases had been eliminated, the surprising increase in incidence of rickets and osteomalacia places the burden on society to be ever more vigilant in assessing for vitamin D deficits. Two obvious explanations exist for the rise in vitamin D deficiency in the US: reduced consumption of vitamin D-fortified milks and reduced skin exposures to sunlight because of greater indoor activities. The deficit in vitamin D has possibly even greater consequences as the average age at death is extending in our populations. Low vitamin D intakes in later life, often coexisting with low dietary calcium, may increase the risk of osteoporotic fractures.

Supplements of vitamin D (400 IU or greater) are recommended for the elderly by health professionals in order to ensure adequate intakes, but even higher amounts are considered safe. The tolerable upper level of safety (UL) established in the US by the Institute of Medicine has been set at $50\,mg\,day^{-1}$ (equivalent to $2000\,IU\,day^{-1}$). Supplemental vitamin D becomes increasingly important as skin biosynthesis capability declines with age. Oral capsules of vitamin D containing up to 100 000 IU have been found to be safe and effective in reducing fractures. An alternative approach is injection of a depot of vitamin D ($\sim$200 000 IU) in the late autumn for slow release over the winter months when sunlight, especially UV-B, is limited or unavailable.

Rising Prevalence of Rickets

The increase in rickets in the US is occurring primarily in African-American and Hispanic children who have gone off breast-feeding and are not getting sufficient calcium and vitamin D in their diets. This problem has been more common in the southern US despite greater availability of sunlight. In large part, rickets is an educational issue that requires input from both medical and public health professionals.

Rising Prevalence of Osteomalacia

The increase in low serum $25(OH)D_3$ found in so-called ordinary adults in a hospital survey of surgical patients in Massachusetts opened the eyes of health authorities who did not expect to find such low blood concentrations, which indicate future osteomalacia and also osteoporosis. This evidence suggests that many adults in the US are not consuming adequate amounts of vitamin D and calcium in their usual patterns of food selection and that supplementation of these two nutrients is probably inadequate among adults. Low intakes among the elderly result from the same type of eating pattern, but typically with even lower caloric consumption (see below).

Life Cycle Changes in Vitamin D Production and Metabolism

Production of vitamin D by the skin typically declines during late adulthood or the early elderly period because of changes in the skin *per se*, i.e., reduction in thickness, reduction in circulating 7-dehydrocholesterol, and decline in the rate of conversion of 7-dehydrocholesterol to cholecalciferol. When these physiological decrements are combined

with less direct exposure to sun by the elderly, even in Florida, they have little opportunity to make the vitamin in their skin. Sun-screens and broad-brimmed hats and other protective clothing complete this scenario, which is aimed at preventing skin cancer.

Special Populations at Risk of Low Vitamin D Intakes and Low Status

Several at-risk subpopulations for poor vitamin D status have been identified. Except for migrant populations, these groups have already been mentioned previously.

The Elderly

Elderly individuals prefer to stay indoors and many are actually 'shut-ins' who have little opportunity for direct sun exposure (UV-B does not penetrate glass windows in rooms or solaria). The shut-ins are most likely to be deficient in vitamin D and calcium and at increased risk for fractures of the hip, especially with increasing age. They need supplementation on a daily basis with a calcium salt (~1000 mg of elemental calcium) plus vitamin D (400 IU or more) to counter not only hypovitaminosis D but also secondary hyperparathyroidism and possibly osteoporosis. Two large prospective trials have demonstrated efficacy of such therapy in reducing hip and other nonverteabral fractures.

Vegetarians

Vegans are at risk of low serum $25(OH)D_3$ and the pathologic changes mentioned above for the elderly unless they consume a supplement of vitamin D and calcium because plant foods are typically low in each nutrient. Fortification of plant foods, such as of soy milk, may overcome this concern.

Long-Term Breast-Fed Infants and Young Children of Dark Skin

Although breast-feeding is strongly recommended and lauded, the switching of infants from breast-milk to other beverages does not always include cows' milk or other calcium-rich drinks. In the southern US, this switching has led to a modest epidemic of rickets, which should not occur with our established knowledge about causation. Dark pigmentation reduces the efficiency of the skin to produce vitamin D and because many children with such skin coloration do not tolerate milk (lactose) well, they consume too little vitamin D and calcium. Supplementation and/or alternate food sources should easily correct these nutrient deficits.

Migrant Populations

Migration of dark-skinned people, especially Muslims, from the Middle East and other Asian nations to the UK and other Northern European nations has led to decreased skin production of vitamin D, especially in the winter months, and to reported cases of rickets and osteomalacia. Cultural practices, i.e., limited food selections and clothing that covers the bodies of women and children, contribute to the etiology of these diseases. Supplementation with calcium and vitamin D should prove effective in promoting bone health of affected individuals.

Excessive Consumption of Vitamin D and Toxic Effects

Because the hormonal form of vitamin D, 1,25DHCC, acts on the DNA in the genome, high intakes or excessive drug dosages of the vitamin lead to synthetic overexpression of proteins that contribute to toxic effects. In recent years, the amounts of vitamin D needed to cause toxic effects has been found to be considerably higher than previously thought; but caution is needed via monitoring of serum 25HCC to ensure that blood concentrations do not get too high. Concern is expressed here that excessive treatment of those with rickets or osteomalacia with high doses of vitamin D may result in toxicity.

Summary

The continuing discovery of low serum $25(OH)D_3$ concentrations among adults suggests that under-recognition of osteomalacia among adults and the elderly exists in the US and possibly in other technologically advanced nations. The hidden nature of the disease has resulted from poor diagnostic criteria; new criteria with lower cut-off points are under review. The surprising resurgence of rickets in the southern US has resulted from too little guidance by health professionals, poor nutrition knowledge of mothers, or other aspects, such as poverty. Correction of both rickets and osteomalacia can be simply achieved with vitamin D and calcium supplements and increased numbers of servings of calcium-rich foods. Public health agencies need to become more active in this regard.

See also: **Bone**. **Calcium**. **Older People**: Nutrition-Related Problems. **Vegetarian Diets**. **Vitamin D**: Physiology, Dietary Sources and Requirements.

Further Reading

Anderson JJB (1999) Plant-based diets and bone health: nutritional implications. *American Journal of Clinical Nutrition* **70**(supplement): 539S–542S.

Chapuy MC, Arlot ME, Delmas PD, and Meunier PJ (1994) Effect of calcium and cholecalciferol treatment for three years on hip fractures in elderly women. *British Medical Journal* **308**: 1081–1082.

Chapuy MC, Arlot ME, Duboeuf F *et al.* (1992) Vitamin D3 and calcium to prevent hip fractures in elderly women. *New England Journal of Medicine* **327**: 1637–1642.

Chapuy MC, Pamphile R, Paris E *et al.* (2002) Combined calcium and vitamin D_3 supplementation in elderly women: confirmation of reversal of secondary hyperparathyroidism and hip fracture risk: The Decalyos II Stusy. *Osteoporosis International* **13**: 257–264.

Cheng S, Tylavsky F, Kroger H *et al.* (2003) Association of low 25-hydroxyvitamin D concentrations with elevated parathyroid hormone concentrations and low cortical bone density in early pubertal and prepubertal Finnish girls. *American Journal of Clinical Nutrition* **78**: 485–492.

Dawson-Hughes B, Harris SS, and Dallal GE (1997) Plasma calcidiol, season, and serum parathyroid hormone concentrations in healthy elderly men and women. *American Journal of Clinical Nutrition* **65**: 65–71.

Delucia MC, Mitnick ME, and Carpenter TO (2003) Nutritional rickets with normal circulating 25-hydroxyvitamin D: a call for reexamining the role of dietary calcium intake in North American infants. *Journal of Clinical Endocrinology and Metabolism* **88**: 3539–3545.

Gartner LM and Greer FR (2003) Prevention of rickets and vitamin D deficiency; new guidelines for vitamin D intake. *Pediatrics* **111**: 908–910.

Gloth FM III, Gundberg CM, Hollis BW *et al.* (1995) Vitamin D deficiency in homebound elderly persons. *Journal of the American Medical Association* **274**: 1683–1686.

Heaney RP (1999) Lessons for nutritional science from vitamin D. *American Journal of Clinical Nutrition* **69**: 825–826.

Heaney RP, Davies KM, Chen TC *et al.* (2003) Human serum 25-hydroxycholecalciferol response to extended oral dosing with cholecalciferol. *American Journal of Clinical Nutrition* **77**: 204–210.

Henry A and Bowyer L (2003) Fracture of the neck of the femur and osteomalacia in pregnancy. *BJOG: An International Journal of Obstetrics and Gynaecology* **110**: 329–330.

Holick MF (1995) Vitamin D and bone health. *Journal of Nutrition* **126**: 1159S–1164S.

Holick MF (2001) Sunlight "D"ilemma: risk of skin cancer or bone disease and muscle weakness. *Lancet* **357**: 4–6.

Institute of Medicine, Food and Nutrition Board (1997) *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride*. Washington, DC: National Academy Press.

Jacques PF, Felson DT, Tucker KL *et al.* (1997) Plasma 25-hydroxyvitamin D and its determinants in an elderly population. *American Journal of Clinical Nutrition* **66**: 929–936.

Kreiter SR, Schwartz RP, Kirkman HN Jr *et al.* (2000) Nutritional rickets in African American breast-fed infants. *Journal of Pediatrics* **137**: 153–157.

Lamberg-Allardt C, Karkkainen M, Seppanen R, and Bistrom H (1993) Low serum 25-hydroxyvitamin D concentrations and secondary hyperparathyroidism in middle-aged white strict vegetarians. *American Journal of Clinical Nutrition* **58**: 684–689.

Norman AW (2001) Vitamin D. In: Bowman BA and Russell RM (eds.) *Present Knowledge in Nutrition*, 8th edn, pp. 146–155. Washington DC: ILSI Press.

Oginni LM, Sharp CA, Worsfold M *et al.* (1999) Healing of rickets after calcium supplementation. *Lancet* **353**: 296–297.

Okonofua F, Gill DS, Alabi ZO *et al.* (1991) Rickets in Nigerian children: a consequence of calcium malnutrition. *Metabolism* **40**: 209–213.

Panunzio MF, Pisano A, Telesfor P, and Tomaiuolo P (2003) Diet can increase 25-hydroxyvitamin-D3 plasma levels in the elderly: a dietary intervention trial. *Nutrition Research* **23**: 1177–1181.

Thomas MK, Lloyd-Jones DM, Thadani RM *et al.* (1998) Hypovitaminosis D in medical inpatients. *New England Journal of Medicine* **338**: 777–783.

Trivedi DP, Doll R, and Khaw KT (2003) Effect of four monthly oral vitamin D_3 (cholecalciferol) supplementation on fractures and mortality in men and women living in the community: randomized double blind controlled trial. *British Medical Journal* **326**: 469–474.

Utiger RD (1998) The need for more vitamin D. *New England Journal of Medicine* **338**: 828–829.

Vieth R, Chan PCR, and MacFarlane GD (2001) Efficacy and safety of vitamin D3 input exceeding the lowest observed adverse effect concentration. *American Journal of Clinical Nutrition* **73**: 288–294.

Webb AR, Kline L, and Holick MF (1988) Influence of season and latitude on the cutaneous synthesis of vitamin D_3: exposure to winter sunlight in Boston and Edmonton will not promote vitamin D_3 synthesis in human skin. *Journal of Clinical Endocrinology and Metabolism* **67**: 373–378.

VITAMIN E

Contents
Metabolism and Requirements
Physiology and Health Effects

Metabolism and Requirements

M G Traber, Oregon State University, Corvallis, OR, USA

Introduction

Vitamin E is the most potent fat-soluble antioxidant in human plasma. Although vitamin E was first discovered in 1922, its metabolic function remains an enigma. There are eight different molecular forms with vitamin E antioxidant activity, yet the body preferentially retains α-tocopherol. This preference for α-tocopherol has led the Food and Nutrition Board in its 2000 Dietary Reference Intakes (DRIs) for vitamin E to recommend that only α-tocopherol, not the other forms, meets human requirements for vitamin E. Moreover, only α-tocopherol is recognized by the hepatic α-tocopherol transfer protein (α-TTP). This protein regulates plasma α-tocopherol concentrations and genetic abnormalities in the protein (or its absence) leads to vitamin E deficiency in humans.

General Description and Scientific Name

Dietary components with vitamin E antioxidant activity include α-, β-, γ-, and δ-tocopherols and α-, β-, γ-, and δ-tocotrienols. These compounds all have a chromanol ring with a phytyl tail (tocopherols) or an unsaturated tail (tocotrienols) (**Figure 1**) and vary in the number of methyl groups on the chromanol ring: α-tocopherol or α-tocotrienol has three methyl groups, β- or γ- have two, and δ- tocopherol and δ-tocotrienol have one.

The naturally occurring form of α-tocopherol is called *RRR*-α-tocopherol; on labels it is called *d*-α-tocopherol and it is more formally known as 2,5,7,8-tetramethyl-2*R*-(4′*R*,8′*R*,12 trimethyltridecyl)-6-chromanol. At positions 2, 4′, and 8′of α-tocopherol are chiral carbon centers that are in the *R*-conformation in naturally occurring α-tocopherol, but theoretically can take on either the *R*- or the *S*-conformation. Position 2 is the most important for biologic activity. Therefore, the DRIs for vitamin E are given in milligrams of 2*R*-α-tocopherol (**Table 1** see below for discussion).

The chemical synthesis of α-tocopherol results in an equal mixture of eight different stereoisomers (*RRR, RSR, RRS, RSS, SRR, SSR, SRS, SSS*) or, more formally, 2,5,7,8-tetramethyl-2*RS*-(4′*RS*,8′*RS*,12 trimethyltridecyl)-6-chromanol. To indicate that synthetic α-tocopherol is a racemic mixture, it is called *all-rac*-α-tocopherol, or on labels, *dl*-α-tocopherol. The first letter of the three-letter combination is the 2 position; therefore, only half of the synthetic α-tocopherol is in the 'active' 2*R*-α-tocopherol conformation. **Table 2** lists the factors used to convert international units (IU) to milligrams. For example, if a vitamin E supplement is labeled 400 IU and it is *dl*-α-tocopheryl acetate, then 400 times 0.45 equals 180 mg 2*R*-α-tocopherol, but if it is labeled *d*-α-tocopheryl acetate, then 400 times 0.67 equals 268 mg 2*R*-α-tocopherol.

Vitamin E Supplements

Most vitamin E supplements and food fortificants contain *all rac*-α-tocopherol, but can contain mixtures of tocopherols or tocotrienols. Supplements often are sold as esters, which protect α-tocopherol from oxidation. These can be acetates, succinates, or nicotinates of α-tocopherol. Either the natural stereoisomer (*RRR*-α-tocopherol) or the synthetic (*all rac*-α-tocopherol) can be sold as an ester, e.g., *d*- or *dl*-α-tocopheryl acetate, respectively.

Dietary Vitamin E

Vitamin E can be readily obtained from food. Generally, the richest sources are vegetable oils. Wheat germ oil, safflower oil, and sunflower oil contain predominantly α-tocopherol, while soy and corn oils contain predominantly γ-tocopherol. All of these oils are polyunsaturated. Good sources of monounsaturated oils, such as olive or canola oils, also contain predominantly α-tocopherol. Whole grains and nuts are also good sources of vitamin E. Fruits and vegetables, although rich in water-soluble antioxidants, are not good sources of vitamin E.

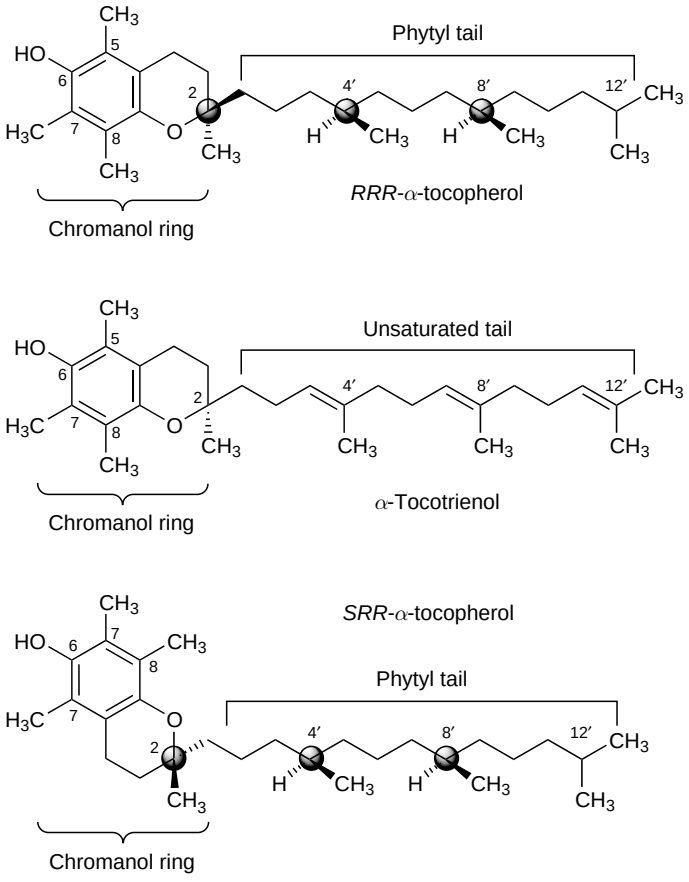


Figure 1 Structures of *RRR*-α-tocopherol, α-tocotrienol, and *SRR*-α-tocopherol.

α-Tocopherol equivalents

It is often assumed for the purpose of calculating vitamin E intakes from food in α-tocopherol equivalents (α-TEs) that γ-tocopherol can substitute for α-tocopherol with an efficiency of 10%. However, functionally γ-tocopherol is not equivalent to α-tocopherol and some caution should be used in applying α-TEs to estimates of α-tocopherol intakes when corn or soybean oils (hydrogenated vegetable oils) represent the major oils present in foods. These oils have high γ-tocopherol contents and if food tables reporting α-TEs are used to estimate dietary α-tocopherol, α-tocopherol intakes are overestimated. α-TEs are no longer recommended in the Food and Nutrition Board of the Institute of Medicine, National Academy of Sciences 2000 DRIs; only milligrams of α-tocopherol

Table 1 Estimated average requirements (EARs), recommended dietary allowances (RDAs), and average intakes (AIs) (mg day^{-1}) for α-tocopherol in adults and children

Lifestage	*EAR*	*RDA*	*AI*
0–6 months			4
7–12 months			6
1–3 years	5	6	
4–8 years	6	7	
9–13 years	9	11	
14–18 years	12	15	
Adult (male or female)	12	15	
Pregnant	12	15	
Lactation	16	19	

Adapted from Food and Nutrition Board and Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. Washington, DC: National Academy Press.

Table 2 Factors to convert international units (IU) of vitamin E to milligrams of 2*R*-α-tocopherol

	mg/IU[a]
***all rac*-α-Tocopherol and esters**	
dl-α-Tocopheryl acetate	0.45
dl-α-Tocopheryl succinate	0.45
dl-α-Tocopherol	0.45
***RRR*-α-Tocopherol and esters**	
d-α-Tocopheryl acetate	0.67
d-α-Tocopheryl succinate	0.67
d-α-Tocopherol	0.67

[a]Multiply the IU in foods or supplements times the indicated factor to obtain the milligrams of active vitamin E.
Adapted from Food and Nutrition Board and Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. Washington, DC: National Academy Press.

and 2*R*-α-tocopherol (for synthetic vitamin E) should be included in estimates of vitamin E intakes.

Vitamin E Actions and Metabolism

Antioxidant Activity

Vitamin E is the most potent, lipid-soluble antioxidant in human plasma and tissues. Thus, vitamin E protects polyunsaturated fatty acids within membrane phospholipids and plasma lipoproteins. When a peroxyl radical forms in a membrane, it is 1000 times more likely to attack a vitamin E molecule than a polyunsaturated fatty acid (**Figure 2**). The hydroxyl group on the chromanol ring of vitamin E reacts with the peroxyl radical to form the corresponding lipid hydroperoxide and tocopheroxyl radical. Thus, vitamin E acts as a chain-breaking antioxidant, preventing further auto-oxidation of lipids.

The tocopheroxyl radical has a number of possible fates. It can react with another radical to form nonreactive products. Alternatively, it can be further oxidized to the tocopheryl quinone, a two-electron oxidation product. Another possibility is 'vitamin E recycling,' where the tocopheroxyl radical is restored to its unoxidized form by other antioxidants such as vitamin C, ubiquinol, or thiols, such as glutathione. This process will deplete these other antioxidants. For this reason, it is important to maintain a good intake of other dietary antioxidants.

Biologic Activity

Biologic activity is a term that has been used historically to indicate a disconnection between vitamin E antioxidant activities and *in vivo* activities. Observations in rodent experiments carried out in the 1930s formed the basis for determining the 'biologic activity' of vitamin E. Although the various vitamin E forms had somewhat similar structures and antioxidant activities, they differed in their abilities to prevent or reverse specific vitamin E deficiency symptoms (e.g., fetal resorption, muscular dystrophy, and encephalomalacia). α-Tocopherol with three methyl groups and a free hydroxyl group on the chromanol ring with the phytyl tail meeting the ring in the *R*-orientation (**Figure 1**) had the highest biological activity. This specific structural requirement for biological, but not chemical, activity is now known to be dependent upon the hepatic α-tocopherol transfer protein (α-TTP), as discussed below. α-TTP maintains plasma and, indirectly, tissue α-tocopherol concentrations.

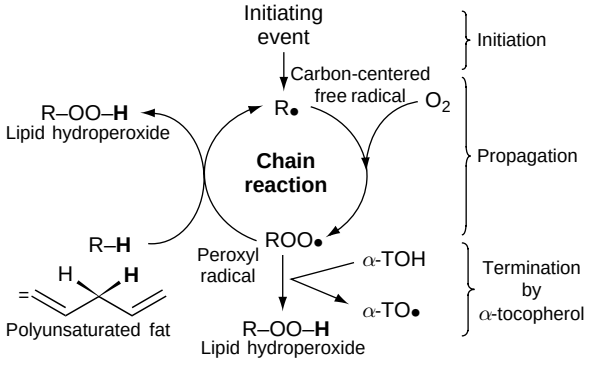


Figure 2 Vitamin E: chain-breaking antioxidant activity. Adapted from Burton GW and Traber MG (1990) Vitamin E: antioxidant activity, biokinetics, and bioavailability. *Annu Rev Nutr* **10**: 357–382.

Molecular Function

In addition to antioxidant activity, there are specific α-tocopherol-dependent functions that normalize cellular functions in a variety of cells. α-Tocopherol plays a critical role through its ability to inhibit the activity of protein kinase C, a central player in many signal transduction pathways. Specifically, it modulates pathways of platelet aggregation, endothelial cell nitric oxide production, monocyte/macrophage superoxide production, and smooth muscle cell proliferation. Regulation of adhesion molecule expression and inflammatory cell cytokine production by α-tocopherol has also been reported. However, most of the information in this area has been obtained from *in vitro* studies. More studies in humans are needed to relate α-tocopherol intakes and tissue concentrations to optimal tissue responses.

Vitamin E metabolism

α- and γ-tocopherols, as well as α-and γ-tocotrienols, are metabolized to α- and γ-CEHCs (2,5,7,8-tetramethyl- and 2,7,8-trimethyl-2-(2′ carboxyethyl)-6-hydroxychromans), respectively. About 1% of a dose of α-tocopherol or tocotrienol, or 5% of a dose of γ-tocopherol or tocotrienol is excreted in the urine as CEHCs. The importance of vitamin E metabolism in the regulation of vitamin E status is unknown.

Recommended Intake Levels

In 2000, the Food and Nutrition Board of the Institute of Medicine, National Academy of Sciences published the DRIs for vitamin C, vitamin E, selenium, and the carotenoids. Their recommendations for vitamin E appear in **Table 1**.

The requirements for vitamin E intakes are based primarily on long-term (5–7 years) depletion and repletion studies in humans. Serum α-tocopherol concentrations and corresponding hydrogen peroxide-induced erythrocyte hemolysis were determined

at various intervals. Serum concentrations necessary to prevent *in vitro* erythrocyte hemolysis in response known levels of vitamin E intake in subjects who had undergone experimentally induced vitamin E deficiency were used to determine estimated average requirements (EARs) for vitamin E. The recommended dietary allowances (RDAs) are levels that represent the daily α-tocopherol intakes required to ensure adequate nutrition in 95–97.5% of the population and are an overestimation of the level needed for most people in any given group.

Vitamin E Units

According to the US Pharmacopoeia (USP), 1 IU of vitamin E equals 1 mg *all rac* α-tocopheryl acetate, 0.67 mg *RRR*-α-tocopherol, or 0.74 mg *RRR*-α-tocopheryl acetate. These conversions were estimated on the relative 'biologic activities' of the various forms when tested in the rat assay for vitamin E deficiency, the fetal resorption assay. These USP IUs are currently used in labeling vitamin E supplements and food fortificants. It should be noted that the current RDA does not use vitamin E USP units but rather the recommendation for adults is set at 15 mg of *RRR*-α-tocopherol or 2*R*-α-tocopherols. Most foods contain *RRR*-α-tocopherol naturally, but foods that have been fortified with vitamin E contain the synthetic form, e.g., fortified breakfast cereals. If the amount of vitamin E on the label is given in international units, then this must be multiplied by the factors given in **Table 2** to obtain the amount of 2*R*-β-tocopherol.

Overdosage

In 2000 the Food and Nutrition Board of the Institute of Medicine, National Academy of Sciences recommended 1000 mg as an upper limit (UL) of all forms of α-tocopherol in supplements taken by adults 19 years and older, including pregnant and lactating women. ULs were set for children and adolescents by adjusting the adult limit on the basis of relative body weight. **Table 3** gives the α-tocopherol UL by age group. No UL was set for infants due to lack of adequate data. In 2000 the Food and Nutrition Board did recommend that food be the only source of vitamin E for infants. However, a UL of 21 mg day^{-1} was suggested for premature infants with birth weights of 1.5 kg, based on the adult UL.

Table 3 Upper limits (UL) for α-tocopherol intakes

Age (years)	*UL (mg day^{-1})*
1–3	200
4–8	300
9–13	600
14–18	800
>19	1000

Adapted from Food and Nutrition Board and Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. Washington, DC: National Academy Press.

Table 4 Upper limits (UL) reported in IU for α-tocopherol-containing supplements

	Number of IU that equal the UL
***all rac*-α-Tocopherol and esters**	
dl-α-Tocopheryl acetate	1100
dl-α-Tocopheryl succinate	1100
dl-α-Tocopherol	1100
***RRR*-α-Tocopherol and esters**	
d-α-Tocopheryl acetate	1500
d-α-Tocopheryl succinate	1500
d-α-Tocopherol	1500

Adapted from Food and Nutrition Board and Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. Washington, DC: National Academy Press.

The vitamin E UL was set for supplements because it is almost impossible to consume enough α-tocopherol-containing foods to achieve a daily 1000 mg intake for prolonged periods of time. The UL was defined for all forms of α-tocopherol, not just the 2*R* forms, because all of the forms in *all rac*-α-tocopherol are absorbed and delivered to the liver. The appropriate conversion factors are different from those shown in **Table 2**, and necessary to estimate the UL for supplements containing either *RRR*- or *all rac*-α-tocopherol supplements. The ULs given in IU are shown in **Table 4**. The UL for *RRR*-α-tocopherol is apparently higher because each capsule contains less α-tocopherol than those containing *all rac*-α-tocopherol.

Precautions and Adverse Reactions

High vitamin E intakes are associated with an increased tendency to bleed. It is not known if this is a result of decreased platelet aggregation caused by an inhibition of protein kinase C by α-tocopherol, some other platelet-related mechanism, or decreased clotting due to a vitamin K and E interaction causing abnormal blood clotting.

Individuals who are deficient in vitamin K or who are on anticoagulant therapy are at increased risk of uncontrolled bleeding. Patients on anticoagulant therapy should be monitored when taking vitamin E supplements to ensure adequate vitamin K intakes.

Adverse Effects of Drugs on Vitamin E Status

Drugs intended to promote weight loss by impairing fat absorption, such as Orlistat or sucrose polyester,

can also impair vitamin E and other fat-soluble vitamin absorption. Therefore, multivitamin supplementation is recommended with these drugs. Vitamin supplements should be taken with meals at times other than when these drugs are taken to allow adequate absorption of the fat-soluble vitamins.

Vitamin E Bioavailability

Absorption and Plasma Transport

Intestinal absorption of vitamin E is dependent upon normal processes of fat absorption. Specifically, both biliary and pancreatic secretions are necessary for solubilization of vitamin E in mixed micelles containing bile acids, fatty acids, and monoglycerides (**Figure 3**). α-Tocopheryl acetates (or other esters) from vitamin E supplements are hydrolyzed by pancreatic esterases to α-tocopherol prior to absorption. Following micellar uptake by enterocytes, vitamin E is incorporated into chylomicrons and secreted into the lymph. Once in the circulation, chylomicron triglycerides are hydrolyzed by lipoprotein lipase. During chylomicron catabolism in the circulation, vitamin E is nonspecifically transferred both to tissues and to other circulating lipoproteins.

It is not until the vitamin E-containing chylomicrons reach the liver that discrimination between the various dietary vitamin E forms occurs. The hepatic α-TTP preferentially facilitates secretion of α-tocopherol, specifically *2R*-α-tocopherols, and not other tocopherols or tocotrienols from the liver into the plasma in very low-density lipoproteins (VLDLs). In the circulation, VLDLs are catabolized to low-density lipoproteins (LDL are also known as the 'bad cholesterol' because high LDL levels are associated with increased risk of heart disease). During this lipolytic process, all of the circulating lipoproteins become enriched with α-tocopherol.

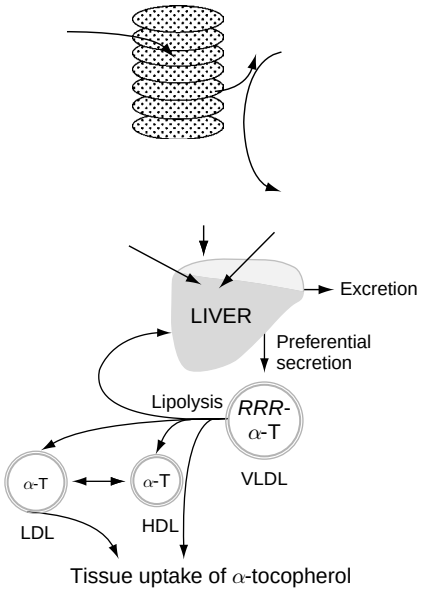


Figure 3 Intestinal vitamin E absorption and plasma lipoprotein transport. (Adapted from Traber MG (1998) Vitamin E. In: Shils ME, Olson JA, Shike M, and Ross AC (eds.) *Modern Nutrition in Health and Disease,* pp. 347–362. Baltimore: Williams & Wilkins.)

There is no evidence that vitamin E is transported in the plasma by a specific carrier protein, but rather it is nonspecifically transported in lipoproteins. An advantage of vitamin E transport in lipoproteins is that easily oxidizable lipids are protected by the simultaneous transport of this lipid-soluble antioxidant. Similarly, delivery of vitamin E to tissues is dependent upon lipid and lipoprotein metabolism. Thus, as peroxidizable lipids are taken up by tissue, the tissues simultaneously acquire a lipid-soluble antioxidant.

Plasma Concentrations, Kinetics, and Tissue Delivery

Plasma α-tocopherol concentrations in normal humans range from 11 to 37 $\mu mol\,l^{-1}$. When plasma lipids are taken into account the lower limits of normal are 1.6 µmol α-tocopherol/mmol lipid or 2.5 µmol α-tocopherol/mmol cholesterol. α-Tocopherol is transported in plasma lipoproteins, so if lipid concentrations are extraordinarily high or low, then correction for lipid levels are helpful to determine adequacy of vitamin E status. Additionally, α-tocopherol concentrations in erythrocytes, adipose tissue, or even peripheral nerves have been used to assess vitamin E status.

The apparent half-life of *RRR*-α-tocopherol in plasma of normal subjects is approximately 48 h, while that of *SRR*-α-tocopherol or γ-tocopherol is only 15 h.

Vitamin E is delivered to tissues by three mechanisms: transfer from triglyceride-rich lipoproteins during lipolysis; as a result of tissue lipoprotein uptake by various receptors that mediate lipoprotein uptake; and as a result of vitamin E exchange between lipoproteins or tissues. The regulation of tissue vitamin E is not well understood, but α-tocopherol is the predominant form in tissues as a result of its dominance in plasma.

Human Vitamin E Deficiency

Vitamin E deficiency was first described in children with fat malabsorption syndromes, principally abetalipoproteinemia, cystic fibrosis, and cholestatic liver disease. Subsequently, humans with severe vitamin E deficiency with no known defect in lipid or

lipoprotein metabolism were described to have a defect in the α-TTP gene.

Erythrocyte fragility, hemolysis, and anemia were described as vitamin E deficiency symptoms in various animals fed diets devoid of vitamin E. Additionally, studies in experimental animals have shown that a deficiency of both selenium (a required component of glutathione peroxidases) and vitamin E causes a more rapid and severe onset of debilitating deficiency symptoms. Hypothetically, a deficiency of both vitamins E and C should also cause more severe antioxidant deficiency symptoms, but most animals make their own vitamin C, so this interaction has not been unequivocally demonstrated in humans or animals.

In contrast to experimental vitamin E deficiency in rodents, in humans the major vitamin E deficiency symptom is a peripheral neuropathy characterized by the degeneration of the large caliber axons in the sensory neurons.

Vitamin E deficiency occurs only rarely in humans and almost never as a result of inadequate vitamin E intakes, therefore, interactions with other nutrients have not been well studied. There have been reports of vitamin E deficiency symptoms in persons with protein-calorie malnutrition. Vitamin E deficiency does occur as a result of genetic abnormalities in α-TTP and as a result of various fat malabsorption syndromes. Vitamin E supplementation halts the progression of the neurologic abnormalities caused by inadequate nerve tissue α-tocopherol and, in some cases, has reversed them.

Patients with these disorders require daily pharmacologic vitamin E doses for life to overcome the mechanisms leading to deficiency. Generally, patients with 'ataxia with vitamin E deficiency' are advised to consume 1000 mg *RRR*-α-tocopherol per day in divided doses, patients with abetalipoproteinemia 100 mg per kg body weight, and cystic fibrosis sufferers 400 mg day^{-1}. However, patients with fat malabsorption due to impaired biliary secretion generally do not absorb orally administered vitamin E. These patients are treated with special forms of vitamin E, such as α-tocopheryl polyethylene glycol succinate, that spontaneously form micelles, obviating the need for bile acids.

Chronic Disease Prevention

The frequency of human vitamin E deficiency is very rare. In individuals at risk, it is clear that vitamin E supplements should be recommended to prevent deficiency symptoms. What about vitamin E supplement use in normal individuals? Dietary changes such as decreasing fat intakes, substituting fat-free foods for fat-containing ones, and increased reliance on meals away from the home have resulted in decreased consumption of α-tocopherol-containing foods. Therefore, intakes of the vitamin E RDA of 15 mg α-tocopherol, may be difficult. Special attention to consuming nuts, seeds, and whole grains will improve α-tocopherol intakes; alternatively, multivitamin pills can be consumed.

Importantly, vitamin E's potential role in preventing or ameliorating chronic diseases associated with oxidative stress leads us to ask whether vitamin E supplements might be beneficial. For many vitamins, when 'excess' amounts are consumed, they are excreted and provide no added benefits. Antioxidant nutrients may, however, be different. Heart disease and stroke, cancer, chronic inflammation, impaired immune function, Alzheimer's disease – a case can be made for the role of oxygen free radicals in the etiology of all of these disorders, and even in aging itself. Do antioxidant nutrients counteract the effects of free radicals and thereby ameliorate these disorders? And, if so, do large antioxidant supplements have beneficial effects beyond 'required' amounts? The 2000 Food and Nutrition Board and Institute of Medicine DRI Report on Vitamin C, Vitamin E, Selenium, and Carotenoids stated that there was insufficient proof to warrant advocating supplementation with antioxidants. But, they also stated that the hypothesis that antioxidant supplements might have beneficial effects was promising. This remains a very controversial area in vitamin E research.

See also: **Antioxidants**: Diet and Antioxidant Defense; Observational Studies; Intervention Studies. **Ascorbic Acid**: Physiology, Dietary Sources and Requirements; Deficiency States. **Fats and Oils**. **Nuts and Seeds**. **Vitamin E**: Physiology and Health Effects.

Further Reading

Food and Nutrition Board and Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*. Washington, DC: National Academy Press.

Keaney JF Jr, Simon DI, and Freedman JE (1999) Vitamin E and vascular homeostasis: implications for atherosclerosis. *FASEB Journal* **13**: 965–975.

Ouahchi K, Arita M, Kayden H, Hentati F, Ben Hamida M, Sokol R, Arai H, Inoue K, Mandel JL, and Koenig M (1995) Ataxia with isolated vitamin E deficiency is caused by mutations in the alpha-tocopherol transfer protein. *Nature Genetics* **9**: 141–145.

Pryor WA (2000) Vitamin E and heart disease: basic science to clinical intervention trials. *Free Radical Biology and Medicine* **28**: 141–164.

Traber MG Vitamin E. In: Shils ME, Olson JA, Shike M, and Ross AC (eds.) *Modern Nutrition in Health and Disease*, vol. 10. Baltimore: Williams & Wilkins (in press).

Physiology and Health Effects

P A Morrissey and M Kiely, University College Cork, Cork, Ireland

In 1922, Evans and Bishop discovered a fat-soluble dietary constituent that was essential for the prevention of fetal death and sterility in rats accidentally fed a diet containing rancid lard. This was originally called 'factor X' and 'antisterility factor' but was later named vitamin E. Subsequently, the multiple nature of the vitamin began to appear when two compounds with vitamin E activity were isolated and characterized from wheat germ oil. These compounds were designated α- and β-tocopherol, derived from the Greek 'tokos' for childbirth, 'phorein' meaning to bring forth, and 'ol' for the alcohol portion of the molecule. Later, two additional tocopherols, γ- and δ-tocopherol, as well as four tocotrienols were isolated from edible plant oils. After the initial discovery, more than 40 years passed before it was proved that vitamin E deficiency could cause disease in humans and was associated with antioxidant functions in cellular systems. It took another 25 years before the non-antioxidant properties of the vitamin were highlighted.

This article reviews the chemistry of the tocopherols; their dietary sources, absorption, transport, and storage; and their metabolic function. In addition, the potential role of dietary or supplemental tocopherol intake in the prevention of chronic disease and possible mechanisms for observed protective effects are discussed. Finally, a summary of the assessment of tocopherol status in humans, intake requirements, and an overview of the safety of high intakes is provided.

Chemistry

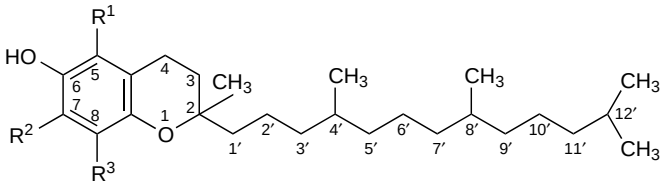
The chemistry of vitamin E is rather complex because there are eight structurally related forms—four tocopherols (α, β, γ, and δ) and four tocotrienols (α, β, γ, and δ)—that are synthesized from homogentisic acid and isopentenyl diphosphate in the plastid envelope of plants. The structures of α-, β-, γ-, and δ-tocopherols are shown in **Figure 1**. α-Tocopherol is methylated at C5, C7, and C8 on the chromanol ring, whereas the other homologs (β, γ, and δ) have different degrees of methylation (**Figure 1**). Tocopherols have a saturated phytyl side chain attached at C2 and have three chiral centers that are in the R configuration at positions C2, C4^1, and C8^1 in the naturally occurring forms, which are given the prefix 2R, 4^{1}R, and 8^{1}R (designated RRR). The members of the tocotrienols are unsaturated at C3^1, C7^1, and C11^1 in the isoprenoid side chain and possess one chiral center at C2 in addition to two sites of geometric isomerism at C3^1 and C7^1. Vitamin E biological activity is expressed as mg RRR-α-tocopherol equivalents (α-TE) whenever possible. The activity of RRR-α-tocopherol is 1. The activities of RRR-β-, RRR-γ-, and RRR-δ-tocopherol are 0.5, 0.1, and 0.03, respectively.

Dietary Sources

The composition and content of the different tocopherol components in plant tissue vary considerably, ranging from extremely low levels found in potato tubers to high levels found in oil seeds. α-Tocopherol is the predominant form in photosynthetic tissues and is mainly localized in plastids. The particular enrichment in the chloroplast membranes is probably related to the ability of tocopherols to quench or to scavenge reactive oxygen species and lipid peroxy radicals by physical or chemical means. In this way, the photosynthetic apparatus can be protected from oxygen toxicity and lipid peroxidation. In nonphotosynthetic tissues, γ-tocopherol frequently predominates and can be involved in the prevention of autoxidation of polyunsaturated fatty acids.

Most of the tocopherol content of wheat germ, sunflower, safflower, and canola and olive oils is in the form of α-tocopherol, and these oils contain approximately 1700, 500, 350, 200, and 120 mg α-TE kg^{-1}, respectively. Vegetable oils (e.g., corn, cottonseed, palm, soybean, and sesame) and nuts (e.g., Brazil nuts, pecans, and peanuts) are rich sources of γ-tocopherol. Corn and soybean oils contain 5–10 times as much γ-tocopherol as α-tocopherol-rich sources of γ-tocopherol, and each contains approximately 200 mg α-TE kg^{-1}. Because of the widespread use of these plant products, γ-tocopherol is considered to represent $\sim$70% of the vitamin E consumed in the typical US diet. The level of vitamin E in nuts ranges from 7 mg α-TE kg^{-1} in coconuts to 450 mg α-TE kg^{-1} in almonds. Cereals are moderate sources of vitamin E, providing between 6 (barley) and 23 mg α-TE kg^{-1} (rye). Fresh fruit and vegetables generally contain approximately 1–10 mg α-TE kg^{-1}. The concentration of vitamin E (α-tocopherol is the predominant form) in animal products is usually low, but these may be significant dietary sources because of their high consumption.

Mean dietary intakes of 6.3–13.0 mg α-TE per day have been reported in various European and US population studies. Data from the Third National Health and Nutrition Examination Survey



Compound	R^1	R^2	R^3
α-Tocopherol	CH_3	CH_3	CH_3
β-Tocopherol	CH_3	H	CH_3
γ-Tocopherol	H	CH_3	CH_3
δ-Tocopherol	H	H	CH_3

Figure 1 The four major forms of vitamin E (α-, β-, γ-, and δ-tocopherols) differ by the number and positions of methyl groups on the chromonol ring. In α-tocopherol, the most biologically active form, the chromonol ring is fully methylated. In β- and γ-tocopherols, the ring contains two methyl groups, whereas δ-tocopherol is methylated in one position. The corresponding tocotrienols have the same structural arrangement except for the presence of double bonds on the isoprenoid side chain of $C3^1$, $C7^1$, and $C11^1$.

(NHANES III) (1988–1994) in the United States indicate a median total intake (including supplements) of α-TE of 12.9 mg day^{-1} and a median intake from food only of 11.7 mg day^{-1} in men aged 31–50 years. In women in this age range, the median total intake (including supplements) of α-TE was 9.1 mg day^{-1} and the median intake from food only was 8.0 mg day^{-1}. In the United States, fats and oils used in spreads, etc. contribute 20.2% of the total vitamin E intake; vegetables, 15.1%; meat, poultry, and fish, 12.6%; desserts, 9.9%; breakfast cereals, 9.3%; fruit, 5.3%; bread and grain products, 5.3%; dairy products, 4.5%; and mixed main dishes, 4.0%.

The North/South Ireland Food Consumption Survey, published in 2001, reported that the median daily intake of vitamin E from all sources was 6.3 mg in men and 6.0 mg in women aged 18–64 years. The largest contributors of vitamin E to the diet were vegetables and vegetable dishes (18.9%) and potatoes and potato products (12.4%), most likely as a result of the oils used in composite dishes. Nutritional supplements contributed 5.5% of the vitamin E intake in men and 11.9% in women overall. In the subgroup that regularly consumed nutritional supplements (23% of total), vitamin E was the nutrient most frequently obtained in supplemental form in men (78%) and women (73%). In these people, supplements made a larger contribution to total vitamin E intakes than did food.

Absorption Metabolism and Excretion

Because of its hydrophobicity, vitamin E requires special transport mechanisms in the aqueous environment of plasma, body fluids, and cells. In humans, vitamin E is taken up in the proximal part of the intestine depending on the amount of food lipids, bile, and pancreatic esterases that are present. It is emulsified together with the fat-soluble components of food. Lipolysis and emulsification of the formed lipid droplets then lead to the spontaneous formation of mixed micelles, which are absorbed at the brush border membrane of the mucosa by passive diffusion. Both α- and γ-tocopherol and dietary fat are taken up without preference by the intestine and secreted in chylomicron particles together with triacylglycerol and cholesterol (**Figure 2**). The nearly identical incorporation of α- and γ-tocopherol in chylomicrons after supplementation with equal amounts of the two tocopherols indicates that their absorption is not selective (**Figure 2**). The chylomicrons are stored as secretory granula and eventually excreted by exocytosis to the lymphatic compartment, from which they reach the bloodstream via the *ductus thoracicus*. The exchange between the apolipoproteins of the chylomicrons (types AI, AII, and B_{48}) and high-density lipoprotein (HDL) (types C and E) triggers the intravascular degradation of the chylomicrons to remnants by the endothelial lipoprotein lipase (LPL) and is a prerequisite for the hepatic uptake of tocopherols (**Figure 2**). During LPL-mediated catabolism of chylomicron particles, some of the chylomicron-bound vitamin E appears to be transported and transferred to peripheral tissues, such as muscle, adipose, and brain (**Figure 2**). The formation of remnants favors the rapid uptake of the tocopherols via the hepatic receptors for apo-E and apo-B.

The chylomicron remnants are subsequently taken up by the liver, where α-tocopherol is preferentially

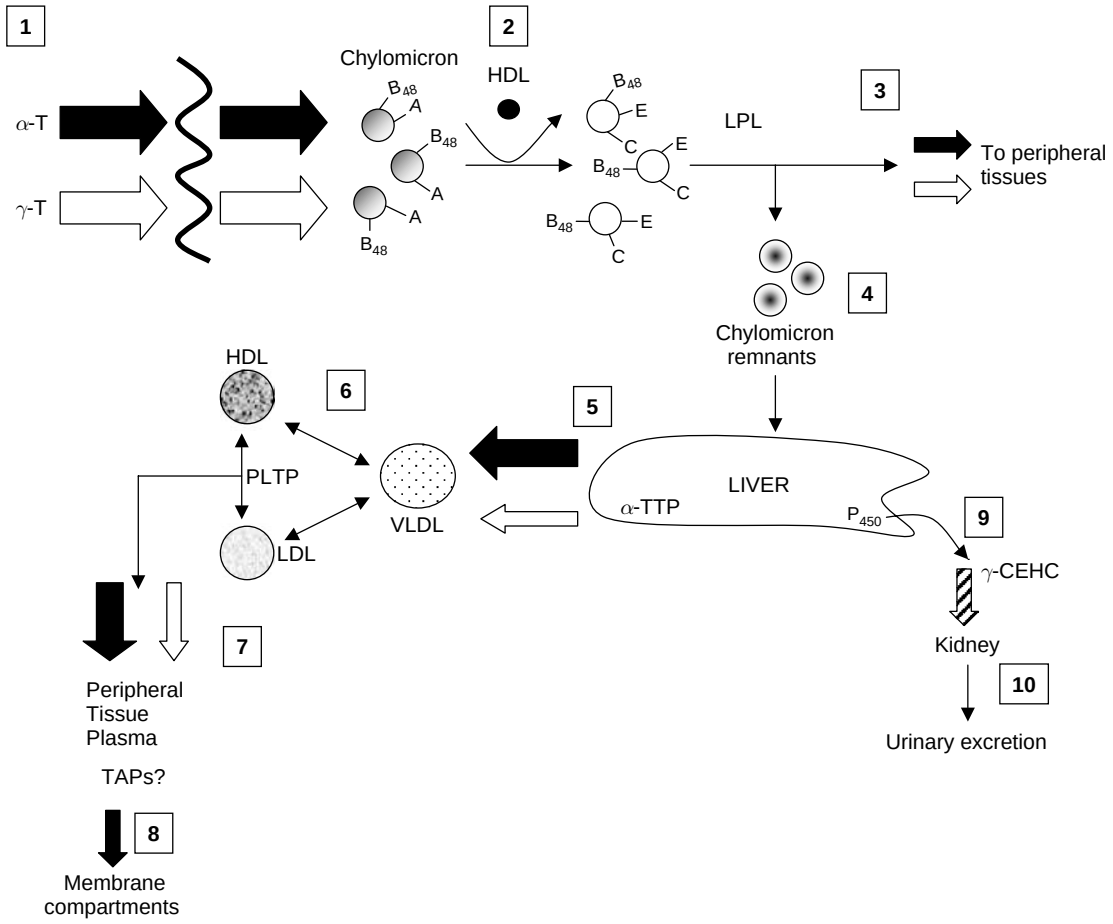
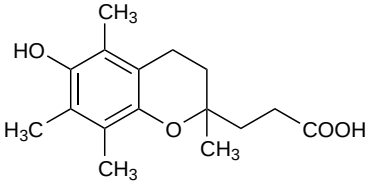


Figure 2 Absorption, transport, and metabolism of α-tocopherol (α-T) and γ-tocopherol (γ-T) in peripheral tissues. 1: Both α-T and γ-T are absorbed without preference by the intestine along with lipid and reassembled into chylomicrons. 2: Exchange between apolipoproteins of the chylomicrons (types AI, AII, and B_{48}) and high-density lipoprotein (HDL) (types C and E) occurs. 3: Chylomicrons are degraded to remnants by lipoprotein lipase (LPL) and some α-T and γ-T are transported to peripheral tissues. 4: The resulting chylomicron remnants are then taken up by the liver. 5: In the liver, most of the remaining α-T, but only a small fraction of γ-T, is reincorporated in nascent very low-density lipoproteins (VLDLs) by α-tocopherol transfer protein (α-TTP). 6: Plasma phospholipid transfer protein (PLTP) facilitates the exchange of tocopherol between HDL and LDL for delivery to tissues. 7: Plasma tocopherols are delivered to tissues by LDL and HDL. 8: Tocopherol-associated proteins (TAPs) probably facilitate intracellular tocopherol transfer between membrane compartments. 9: Substantial amounts of γ-T are degraded by a cytochrome P450-mediated reaction to 2,7,8-trimethyl-2-(β-carboxyethyl-6-hydroxychroman (γ-CEHC). 10: γ-CEHC is excreted into urine. Adapted from Azzi A and Stocker A (2000) Vitamin E: Non-antioxidant roles. *Progress in Lipid Research* **39**: 231–255; and from Jiang Q, Christen S, Shigenaga MK and Ames BN (2001) γ-Tocopherol, the major form of vitamin E in the US diet, deserves more attention. *American Journal of Clinical Nutrition* **74**: 714–722.

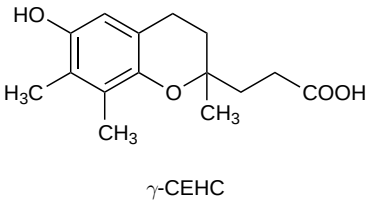
incorporated into nascent very low-density lipoprotein (VLDL) by a specific 32-kDa α-tocopherol transfer protein (α-TTP), which enables further distribution of α-tocopherol to peripheral cells (**Figure 2**). α-TTP is mainly expressed in the liver, in some parts of the brain, in the retina, in low amounts in fibroblasts, and in the placenta. α-TTP possesses stereospecificity as well as regiospecificity toward the most abundant isomer of vitamin E, (RRR)-α-tocopherol. The sorting process does not tolerate alteration at C2. As a consequence of the selective transfer mechanism, major parts of the natural homologs and nonnatural isomers of α-tocopherol are excluded from the plasma and secreted with the bile. Relative affinities of tocopherols for α-TTP are as follows: α-tocopherol, 100; β-tocopherol, 38; γ-tocopherol, 9; and δ-tocopherol, 2. A 75-kDa plasma phospholipid transfer protein (PLTP), which is known to catalyze the exchange of phospholipids and other amphipatic compounds between lipid structures, has been shown to facilitate the exchange of α-tocopherol from VLDL to HDL and LDL for further delivery to tissues (**Figure 2**).

A family of cellular tocopherol-associated proteins (TAPs) with the ability to bind and redistribute α-tocopherol has been identified. TAPs bind to α-tocopherol but not to other isomers of tocopherol. Present in all cells, TAPs may be specifically involved in intracellular α-tocopherol movement, for example, between membrane compartments and plasma membranes, or in optimizing the α-tocopherol content of membranes.

γ-Tocopherol appears to be mainly degraded to its hydrophilic 3′-carboxychromanol metabolite, 2,7,8-trimethyl-2-(β-carboxyethyl)-6-hydroxychroman (γ-CEHC) (**Figure 3**), and excreted in the urine. The mechanism of γ-tocopherol metabolism involves terminal cytochrome P450 (CYP)-mediated ω-hydroxylation of the tocopherol phytyl side chain, oxidation to the corresponding terminal carboxylic acid, and sequential removal of two- or three-carbon moieties by β-oxidation, ultimately yielding the hydrophilic 3′-carboxychromanol metabolite of the parent tocopherol that is excreted in the urine. Functional analysis of several recombinant human liver P450 enzymes revealed that tocopherol ω-hydroxylase activity was associated only with the cytochrome P450 isoform 4F2 (CYP4F2). Kinetic analysis of the tocopherol ω-hydroxylase activity in recombinant human CYP4F2 microsomal systems revealed similar K_m values (37 and 21 μM) but notably different V_{max} values (1.99 vs 0.16 nmol/nmol of P450/min) for γ- and α-tocopherol, respectively. The data suggest a role for the CYP-mediated ω-hydroxylase pathway in the preferential physiological retention of α-tocopherol and elimination of γ-tocopherol. In nonsupplemented individuals, a substantial proportion of the estimated daily intake of γ-tocopherol is excreted in human urine as its γ-CEHC metabolite, but a much smaller proportion of α-tocopherol is excreted as 2,5,7,8-tetramethyl-2-(β-carboxyethyl)-6-hydroxychroman (α-CEHC) (**Figure 3**). α-CEHC is excreted in large amounts only when the daily intake of α-tocopherol exceeds 150 mg or plasma concentrations of α-tocopherol are above a threshold of 30–40 μmol l^{-1}. Even then, urinary excretion of α-CEHC is lower than that of γ-CEHC.



α-CEHC



γ-CEHC

Figure 3 Chemical structures of 2,5,7,8-tetramethyl-2-(β-carboxyethyl)-6-hydroxychroman(α-CEHC) and 2,7,8-trimethyl-2-(β-carboxyethyl)-6-hydroxychroman (γ-CEHC).

It is likely that it is the capacity of α-TTP rather than the plasma α-tocopherol concentration that determines α-tocopherol degradation. Overall, hepatic catabolism of γ-tocopherol appears to be responsible for the relatively low preservation of γ-tocopherol in plasma and tissues, whereas α-TTP-mediated α-tocopherol transfer plays a key role in the preferential enrichment of α-tocopherol in most tissues. Supplementation with α-tocopherol depletes plasma and tissue γ-tocopherol levels. This is likely due to the preferential affinity of α-TTP for α-tocopherol. However, the depletion of γ-tocopherol may also occur because an increase in α-tocopherol may further reduce the incorporation of γ-tocopherol into VLDL, which leaves more γ-tocopherol to be degraded by CYP. On the other hand, γ-tocopherol supplementation may spare α-tocopherol from being degraded.

Plasma (RRR)-α-tocopherol incorporation is a saturable process. Plasma concentrations of α-tocopherol reach a threshold of 30–40 μmol l^{-1} despite supplementation with high levels (400 mg or greater) of (RRR)-α-tocopherol. Dose–response studies showed that the limitation in plasma α-tocopherol concentration appears to be a result of rapid replacement of circulating with newly absorbed α-tocopherol. Kinetic analysis has shown that the entire plasma pool of α-tocopherol is replaced daily. The highest concentrations of α-tocopherol in the body are in adipose tissues and adrenal glands. Adipose tissues are also a major store of the vitamin, followed by liver and skeletal muscle. The rate of uptake and turnover of α-tocopherol by different tissues varies greatly. Uptake is most rapid into lungs, liver, spleen, kidney, and red cells (in rats, $t_{1/2} < 15$ days) and slowest in brain, adipose tissues, and spinal cord ($t_{1/2} < 30$ days). Likewise, depletion of α-tocopherol from plasma and liver during times of dietary deficiency is rapid, whereas adipose tissue, brain, spinal cord, and neural tissues are much more difficult to deplete.

The major route for the elimination of tocopherol from the body is via the feces. Fecal tocopherol arises from incomplete absorption, secretion from mucosal cells, and biliary excretion. Excess

α-tocopherol as well as forms of vitamin E not preferentially used, such as synthetic racemic isomer mixtures, or γ-tocopherol are eliminated during the process of nascent VLDL secretion in the liver and are probably excreted into bile. In addition to the urinary excretion of γ-tocopherol as γ-CEHC, biliary excretion is an alternative route for elimination of excess γ-tocopherol. This is confirmed by the fact that the ratio of γ- to α-tocopherol in bile is sevenfold higher than in plasma.

Tocopherols as Antioxidants

Under normal physiological conditions, cellular systems are incessantly challenged by stressors arising from both internal and external sources. The most important potential stressors are reduced derivatives of oxygen, which are classified as reactive oxygen species (ROS), and include the superoxide anion ($O_2^{\cdot-}$), hydroxyl radical ($^{\cdot}OH$), and oxygen-centered radicals of organic compounds (peroxyl ($ROO^{\cdot}$) and alkoxyl ($RO^{\cdot}$)) together with other nonradical reactive compounds, such as hydrogen peroxide (H_2O_2). In addition, reactive nitrogen species such as nitric oxide ($NO^{\cdot}$), nitrogen dioxide ($NO_2^{\cdot}$), peroxynitrite ($ONOO^-$), and hypochlorous acid are involved.

Cellular systems have evolved a powerful and complex antioxidant defence system to limit inappropriate exposure to these stressors. α-Tocopherol is quantitatively the most important chain-breaking antioxidant in plasma and biological membranes. The antioxidant activities of chain-breaking antioxidants are determined primarily by how rapidly they scavenge peroxyl radicals, thereby preventing the propagation of free radical reactions. When the chromanol phenolic group of α-tocopherol (TOH) encounters a $ROO^{\cdot}$ it forms hydroperoxide (ROOH), and in the process a tocopheroxyl radical ($TO^{\cdot}$) is formed:

$$TOH + ROO^{\cdot} \xrightarrow{k_1} ROOH + TO^{\cdot}$$

The rate constant (k_1) for hydrogen abstraction from α-tocopherol is $2.35 \times 10^6\,M^{-1}\,s^{-1}$, which is higher than that for the other tocopherols and related phenols. Because the rate constant (k_2) for the chain propagation reaction between $ROO^{\cdot}$ and an unsaturated fatty acid (RH) ($ROO^{\cdot} + RH \rightarrow ROOH$) is much lower than k_1, at approximately $10^2\,M^{-1}\,s^{-1}$ α-tocopherol outcompetes the propagation reaction and scavenges the $ROO^{\cdot}$ $\sim 10^4$ times faster than RH reacts with $ROO^{\cdot}$. Thus, the kinetic properties of antioxidants, in particular α-tocopherol, require that only relatively small concentrations are required for them to be effective. The concentration of α-tocopherol in biological membranes is approximately 1 mol per 1000–2000 mol phospholipids (i.e., $\sim 1{:}10^3$). Ascorbic acid can reduce the tocopheroxyl radical ($TO^{\cdot}$) to its native state, and it has been concluded that part of the reason why low concentrations of α-tocopherol are such efficient antioxidants in biological systems is because of this capacity to be regenerated by intracellular reductants such as ascorbic acid.

The heteroxyclic chromanol ring of α-tocopherol has an optimised structure for resonance stabilization of the unpaired electron of the α-tocopheroxyl radical, and the electron-donating substituents (e.g., the three methyl groups) increase this effect. Because γ-tocopherol lacks one of the electron-donating methyl groups on the chromanol ring, it is somewhat less potent in donating electrons than α-tocopherol and is thus a slightly less powerful antioxidant. However, the unsubstituted C5 position on γ-tocopherol allows it to trap lipophilic electrophiles such as peroxynitrite, thereby protecting macromolecules from oxidation.

Vitamin E Deficiency

Vitamin E deficiency is seen rarely in humans. However, there may be a risk of vitamin E deficiency in premature infants because the placenta does not transfer α-tocopherol to the fetus in adequate amounts. When it occurs in older children and adults, it is usually a result of lipoprotein deficiencies or a lipid malabsorption syndrome. These include patients with abetalipoproteinemia or homozygous hypobetalipoproteinemia, those with cholestatic disease, and patients receiving total parenteral nutrition. There is also an extremely rare disorder in which primary vitamin E deficiency occurs in the absence of lipid malabsorption. This disorder is a rare autosomal recessive neurodegenerative disease caused by mutations in the gene for α-TTP. This disorder is known as ataxia with vitamin E deficiency (AVED). Patients with AVED have extraordinary low plasma vitamin E concentrations ($<5\,\mu g\,ml^{-1}$) and have an onset between 4 and 18 years, with progressive development of peripheral neuropathy, spinocerebellar ataxia, dysarthria, the absence of deep tendon reflexes, and vibratory and proprioceptive sensory loss. Patients with an α-TTP defect have enhanced urinary excretion of α-CEHC despite having much lower plasma α-tocopherol concentrations than healthy subjects. Therapeutic and prophylactic vitamin E supplementation (up to $2000\,mg\,day^{-1}$) prevents the onset of the disease before irreversible neurological damage develops.

Tocopherols and Low-Density Lipoprotein Modification

The hypothesis that oxidative stress plays an important role in the pathogenesis of atherosclerosis is generally accepted. Substantial *in vitro* evidence indicates that oxidized LDL is the component central to the initiation and/or progression of atherogenesis at the molecular and cellular level. The typical LDL particle is not only rich in cholesterol but also contains approximately 1300 molecules of RH, which are very sensitive to oxidation. Vitamin E, mainly α-tocopherol, is quantitatively the most important lipophilic antioxidant present in LDL particles. On average, each LDL particle is protected by ~6 mol α-tocopherol (range, 3–15 mol), 1 mol of γ-tocopherol, and small amounts of carotenoids.

All major cells of the artery wall, such as monocyte macrophages, endothelial cells, and smooth muscle cells, can modify LDL oxidatively *in vitro*. Monocytes have been shown to induce peroxidation of lipids such as those in LDL by the generation of reactive species, including superoxide anion, hydrogen peroxide, and hydroxyl radicals. Other oxidants have been implicated, including 15-lipoxygenase, myeloperoxidase-generated hypochlorous acid, and reactive nitrogen species such as peroxynitrite. *In vivo*, oxidized LDL particles are recognized by macrophage scavenger receptors and taken up by macrophages, forming lipid-laden foam cells in the fatty streak lesions. The free radical oxidation of LDL results in numerous structural changes that all depend on a common event—the peroxidation of polyunsaturated fatty acids in the LDL particle.

In vitro studies have indicated that increasing the vitamin E content of LDL particles increases their resistance to oxidation and decreases their uptake by macrophages. Vitamin E supplementation has also been reported to suppress macrophage uptake of oxidized LDL in human arterial lesions and decrease urinary F_2-isoprostane (a 'footprint' of free radical-mediated oxidation of arachidonic acid) concentrations. Reactive nitrogen species are also implicated in aortic oxidation of LDL and therefore potentially in atherosclerosis. Because of the nonsubstituted 5-position, γ-tocopherol reacts with peroxynitrite and other electrophilic mutagens generated during inflammation and forms a stable carbon-centered adduct, 5-nitro-γ-tocopherol. This mechanism of protecting LDL may be significant when γ-tocopherol constitutes a major portion of vitamin E in the diet. It is worth noting that the ability of γ-tocopherol to attenuate oxidative damage produced by these reactive species may prevent or delay the progression of other diseases as well as cardiovascular disease (CVD), in which inflammation plays a role, such as cancer, rheumatoid arthritis, inflammatory bowel disease, and neurodegenative disorders. In addition, γ-CEHC has natriuretic activity and functions in the kidney to control sodium excretion, and it regulates the body's extracellular fluid volume, an important determinant in hypertension and congestive heart failure.

Tocopherols and Other Metabolic Functions

Vitamin E, in addition to having a protective role in the oxidative modification of LDL, may affect or limit the progression of atherosclerosis and a number of other conditions in ways that are unrelated to its antioxidant activity. Some of these effects appear to stem from the ability of α-tocopherol, at physiological concentrations of vitamin E, to activate protein phosphatase 2A, which inhibits the activity of protein kinase C (PKC), a biological indicator of inflammation, by dephosphorylating the protein. PKC is an important element in the signal transduction cascade mediated by growth factors, such as platelet-derived growth factors, which are necessary for the progression and completion of the cell proliferation cycle.

The cellular effects of α-tocopherol-mediated inhibition of PKC depend on the cell type in question, but the cumulative effect is highly protective against the progression of atherosclerosis. PKC inhibition results in reduced smooth muscle cell proliferation, inhibition of platelet aggregation, and thus delayed intra-arterial thrombus formation. Endothelial cell function is preserved by the downregulation of adhesion molecule (ICAM-1 and VCAM-1) expression (possibly by downregulation of nuclear factor-κB) and hence prevention of monocyte and neutrophil adhesion, which is an important early event in the initiation of fatty streak formation and atherogenesis. In addition, PKC inhibition in monocytes reduces the production of reactive oxygen species by impairment of NADPH-oxidase assembly, which may help to reduce LDL oxidation.

The release of proinflammatory cytokines in monocytes, such as interleukin-1β and tumour necrosis factor-α, is impeded by α-tocopherol-mediated inhibition of the 5-lipoxygenase pathway, and production of eicosanoids, such as prostaglandin E_2 and thromboxane A_2, is impeded by γ-tocopherol-mediated inhibition of the cycloxygenase pathway. Lower circulating levels of inflammatory mediators, which are aggregatory and

vasoconstrictive, as well as inhibition of monocyte chemoattractant protein-1 (MCP-1) production, reduces the attraction of monocytes to inflammatory sites at the arterial wall and prevents the formation of foam cells. Furthermore, α-tocopherol increases production of prostacyclin, which has anti-aggregatory and vasodilatory properties, thereby reducing the risk of a coronary event. There is evidence that in a formed atherosclerotic plaque, vitamin E may have a stabilizing effect and prevent its rupture and subsequent clot formation. This may be an important contributor to the prevention of heart disease because plaque types that are most subject to rupture present the greatest threat.

Nitric oxide (NO) produced by NO synthase in the endothelium is important in the maintenance of vascular tone; it suppresses the expression of proinflammatory cytokines, adhesion molecules, and MCP-1. It also inhibits platelet adhesion, maintains the integrity of the arterial wall, and acts as an antioxidant. Vitamin E can reduce the inhibition of NO synthase by reactive oxygen species, thus maintaining NO production, either through its antioxidant activity or perhaps by suppressing PKC activity in smooth muscle.

Tocopherols and Cardiovascular Disease—Epidemiological Evidence

The effects of dietary vitamin E have been examined in several studies, many of which have reported a clear association between the reduction in the relative risk of CVD and high intake or supplement of vitamin E, although some have shown no such association. The Vitamin Substudy of the WHO/MONICA Project showed that in European populations whose classical risk factors for CVD were very similar, the 7-fold differences in CVD mortality could be explained at least to approximately 60% by differences in the plasma levels of vitamin E and up to 90% by the combination of vitamins E, A, and C. The Edinburgh Case Control Study and Basel Prospective Study consistently revealed an increased risk of ischemic heart disease and stroke for low plasma levels of vitamin E. However, other European population studies have not found an association between blood levels of vitamin E and end points of CVD. In the EURAMIC study, the adipose levels of vitamin E did not correlate with the relative risk of myocardial infarction.

A number of prospective studies have examined the association between vitamin E intake and risk of CHD. The Nurses' Health Study, conducted on 87 245 women, showed a 34% reduction in CHD in women who had consumed vitamin E supplements containing more than 67 mg α-TE daily for more than 2 years. However, there was no significant effect of vitamin E obtained from food sources. The Established Populations for Epidemiolodic Studies of the Elderly (EPESE) trials showed that the use of vitamin E supplements significantly decreased risks for all-cause-mortality and mortality from heart disease. Another prospective study, performed in Canada, reported a consistent inverse association between CVD and vitamin E supplement usage. The Health Professionals Study, conducted on 39 910 men aged 40–75 years, also showed that dietary intakes of vitamin E were not significantly correlated with reduced risk of CHD or death. A protective effect was seen in those who took 67–160 mg supplemental α-TE daily for more than 2 years. In contrast, the Iowa Women's Health Study reported that dietary vitamin E (mainly γ-tocopherol) was inversely associated with the risk of death from CVD. This association was particularly striking in the subgroup of women who did not consume vitamin supplements. There was little evidence that the intake of vitamin E from supplements (mainly α-tocopherol) was associated with a decreased risk of death from CVD. The reasons for the differences between dietary and supplemental vitamin E are not clear. However, some epidemiological studies point to the potential importance of γ-tocopherol in preventing heart disease. High dietary intake of nuts, an excellent source of γ-tocopherol, lowered serum cholesterol, improved plasma lipid profiles, and was inversely associated with the risk of death from heart disease.

The ability of α-tocopherol supplementation to prevent cardiovascular events in different populations was tested in four larger prospective clinical trials: The α-Tocopherol, β-Carotene Cancer Prevention (ATBC) study, the Cambridge Heart Antioxidant Study (CHAOS), the Gruppo Italiano per lo studio della Sopravvivenza nell'Infarto Miocardito (GISSI) trial, and the Heart Outcome Prevention Evaluation (HOPE) study. In addition, at least two smaller prospective clinical trials have been completed: the Secondary Prevention with Antioxidants of Cardiovascular Disease in Endstage Renal Disease (SPACE) study and the Antioxidant Supplementation in Atherosclerosis Prevention Study (ASAP).

In the ATBC study, the subjects who were supplemented with 50 mg *all rac*-α-tocopheryl acetate day^{-1} for 5–8 years had only a moderately lower incidence (4%) of angina pectoris than did the control subjects, and among male smokers,

cardiovascular mortality did not differ significantly between those who received supplementation and those who did not. However, subjects who received supplementation had a significantly higher incidence of haemorrhagic stroke than did the control subjects. Note that the ATBC study was not designed to investigate cardiovascular disease development. The results of the CHAOS trial, the first prospective trial with cardiovascular disease as an end point, were encouraging. The risks of nonfatal myocardial infarction declined 77% and total (fatal plus nonfatal) myocardial infarction declined 47% when patients with established coronary artery disease were treated with 268 or 536 mg α-TE daily for approximately 500 days. The GISSI study showed that feeding 211 mg α-TE day^{-1} for 3.5 years did not significantly reduce the rate of all-cause death, nonfatal myocardial infarction, or nonfatal stroke. However, in a later four-way reanalysis in which each individual variable was considered as an end point, there were significantly fewer (20%) cardiovascular deaths in the α-tocopherol group than in the control group. The HOPE study reported that vitamin E (400 IU (268 mg) day^{-1} RRR-α-tocopherol) treatment of CVD patients had no effect on reducing the primary end points, which included nonfatal myocardial infarction, stroke, and cardiovascular death. In the SPACE trial, haemodialysis patients with preexisting cardiovascular disease received 536 mg (RRR)-α-tocopherol or placebo day^{-1}. Patients who received vitamin E had a striking 54% reduction in cardiac events compared with control subjects.

In the ASAP study, men and women (all subjects had hypercholesterolemia at entry) were given vitamin E (91 mg twice daily), slow-release vitamin C (250 mg twice daily), a combination of both, or placebo for 3 years. The progression of atherosclerosis (the mean intima-media thickness of the common carotid artery measured) was significantly retarded only in the men who smoked and took both vitamins. It is important to note that, in general, women develop fewer cardiovascular events than do men. Thus, women may profit less from vitamin E treatment than men. In studies in which many women are enrolled, the low incidence of CVD may weaken the statistical power of the overall trial.

Tocopherols and Cancer—Epidemiological Evidence

Clinical and epidemiological data, together with evidence from experimental models, support a role for the involvement of free radicals throughout the cancer process. Attempts to prevent cancer using vitamin E are based on the rationale that oncogenesis results from free radicals attacking DNA. As an antioxidant, vitamin E may inhibit cancer formation by scavenging reactive oxygen or nitrogen species. Several studies of oral, pharyngeal, and cervical cancer found a relationship between vitamin E status and cancer risk. The evidence for stomach and pancreatic cancers has not been consistent, and no association with breast cancer has been found.

The Linxian, China, intervention trial provided evidence that nutritional supplementation may lower the risk of certain cancers. A modest but significant reduction in cancer mortality was observed in a general population trial in those receiving daily (for 5.25 years) a combination of β-carotene (15 mg), vitamin E (30 mg), and selenium (50 μg). The subjects who received this mixture had a 13% lower incidence of cancer and a 10% lower mortality from stomach and oesophageal cancer than did the subjects who did not receive the mixture. In the ATBC study, male smokers who took vitamin E supplements had a 34% lower incidence of prostate cancer and 41% lower mortality from prostate cancer than did those who did not take the supplements. In the United States, in a nested case–control study conducted to examine the association of α-tocopherol, γ-tocopherol, and selenium with the incidence of prostate cancer, a striking fivefold reduction in risk was observed for the men in the highest quintile of γ-tocopherol compared with those in the lowest. Overall, evidence for the protection from cancer by vitamin E is not compelling.

Tocopherols and Other Diseases—Epidemiological Evidence

Vitamin E appears to act as an immunosuppressant due to its ability to suppress both humoral and cellular immune responses. Tocopherol supplementation significantly enhances lymphocyte proliferation, interleukin-2 production, and delayed-type hypersensitivity skin response and decreases prostaglandin E_2 production by inhibiting cyclooxygenase activity. There appears to be compelling evidence that intervention with dietary antioxidants, such as vitamin E, may help maintain the well-preserved immune function of 'very healthy' elderly, restore the age-related decrease in immune function, and reduce the risk of several age-associated chronic diseases. Epidemiological evidence suggests an

association between the incidence of cataract and vitamin E status. In a prospective study, the sum of serum α- and γ-tocopherol, but neither tocopherol alone, was inversely associated with the incidence of age-related nuclear cataracts.

Among the most common neurologic diseases are neurodegenerative disorders such as Alzheimer's and Parkinson's disease, which may be caused by oxidative stress and mitochondrial dysfunction leading to progressive neural death. An increasing number of studies show that antioxidants (vitamin E and polyphenols) can block neuronal death *in vitro*. In a 2-year, double-blind, placebo-controlled, randomised trial of patients with moderately severe impairment from Alzheimer's disease, treatment with 1340 mg day^{-1} α-TE significantly slowed the progression of the disease. Clinical treatment of Alzheimer's patients with large doses of vitamin E (670 mg α-TE twice daily) is one of the key therapeutic guidelines of the American Academy of Neurology. In a multicentre, double-blind trial, vitamin E (1340 mg α-TE day^{-1}) was not beneficial in slowing functional decline or ameliorating the clinical features of Parkinson's disease. Administration of vitamin E significantly relieved symptoms in patients suffering from several types of acute or chronic inflammatory conditions, such as acute arthritis, rheumatoid arthritis, and osteoarthritis.

Vitamin E Status and Requirements

Interest in the role of vitamin E in disease prevention has encouraged the search for reliable indices of vitamin E status. Most studies in human subjects make use of static biomarkers of status, usually α-tocopherol concentrations in plasma, serum, erythrocytes, lymphocytes, platelets, lipoproteins, adipose tissues, buccal mucosal cells, and LDL, and the α-tocopherol:γ-tocopherol ratio in serum or plasma. Other markers of vitamin E status include susceptibility of erythrocyte or plasma LDL to oxidation, breath hydrocarbon exhalation, and the concentration of α-tocopherol quinone in cerebrospinal fluid. There is no consensus as to the threshold concentration of plasma or serum α-tocopherol at which a person can be defined as having inadequate tocopherol status, but values of $<$11.6, 11.6–16.2, and $>$16.2 μmol^{-1} are normally regarded as indicating a deficient, low, and acceptable vitamin E status, respectively. It is recommended that plasma or serum α-tocopherol concentrations be lipid-corrected (i.e., expressed relative to either the sum of cholesterol and triacylglycerol or cholesterol alone). For convenience, α-tocopherol:cholesterol is the simplest to obtain and probably the most useful, with values below 2.2 μmol α-tocopherol/mmol cholesterol indicating a risk or deficiency and an optimal value $>$5.2. It has been estimated that an average daily dietary intake of 15–30 mg α-tocopherol would be required to maintain this plasma level, an amount that could be obtained from dietary sources if a concerted effort were made to eat foods rich in vitamin E.

The US Institute of Medicine Food and Nutrition Board set an estimated average requirement (EAR) of 12 mg α-tocopherol for adults $>$19 years on the criterion of vitamin E intakes that were sufficient to prevent hydrogen peroxide-induced hemolysis in men. The same value was set for men and women on the basis that although body weight is smaller on average in women than men, fat mass as a percentage of body weight is higher on average in women. Because information is not available on the standard deviation of the requirement for vitamin E, the recommended dietary allowance (RDA) was established for men and women as the EAR (12 mg) plus twice the coefficient of variation (assumed to be 10%), rounded up, giving a value of 15 mg day^{-1}. In Europe, the Scientific Committee for Food did not set a population reference intake (PRI) for vitamin E on the basis that there is no evidence for deficiency from low intakes, and the frequency of distribution of intakes is skewed to the right, making it difficult to set a PRI that is not inappropriately high, especially for those with a low consumption of polyunsaturated fatty acid (PUFA), whose requirements are lower than those with a high consumption of PUFA.

It has been suggested that the optimum concentration of α-tocopherol in plasma for protection against cardiovascular disease and cancer is $>$30 μmol l^{-1}, given normal plasma lipid levels and in conjunction with a plasma vitamin C concentration $>$50 μmol l^{-1} and a β-carotene level $>$0.4 μmol l^{-1}. This has not been proven in large-scale human intervention trials, but even in the absence of conclusive evidence for a prophylactic effect of vitamin E on chronic disease prevention, some experts believe that a recommendation of a daily intake of 87–100 mg α-tocopherol is justifiable based on current evidence. Realistically, these levels can be achieved only by using nutritional supplements. The tolerable upper intake level for vitamin E is 1000 mg day^{-1}, based on studies showing hemorrhagic toxicity in rats, in the absence of human dose–response data.

See also: **Antioxidants**: Diet and Antioxidant Defense. **Fats and Oils**. **Fatty Acids**: Omega-3 Polyunsaturated; Omega-6 Polyunsaturated. **Lipoproteins**. **Nuts and Seeds**.

Further Reading

Azzi A and Stocker A (2002) Vitamin E: Non-antioxidant roles. *Progress in Lipid Research* **39**: 231–255.

Brigeluis-Flohe R, Kelly FJ, Salonen JT *et al.* (2002) The European perspective on vitamin E: Current knowledge and future research. *American Journal of Clinical Nutrition* **76**: 703–716.

Esposito E, Rotilio D, Di Matteo V *et al.* (2002) A review of specific dietary antioxidants and the effects on biochemical mechanisms related to neurodegenerative processes. *Neurobiology of Ageing* **23**: 719–735.

Frei B (1994) In *Natural Antioxidants in Human Health and Disease*. London: Academic Press.

Halliwell B (1996) Antioxidants in human health and disease. *Annual Review of Nutrition* **16**: 33–50.

Institute of Medicine (2000) *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium and Carotenoids*. Washington, DC: National Academy Press.

Jiang Q, Christen S, Shigenaga MK, and Ames BN (2001) γ-Tocopherol, the major form of vitamin E in the US diet, deserves move attention. *American Journal of Clinical Nutrition* **74**: 712–722.

Machlin LJ (1984) Vitamin E. In: Machlin LJ (ed.) *Handbook of Vitamins: Nutritional Biochemical and Clinical Aspects*, pp. 99–145. New York: Marcel Dekker.

Morrissey PA and Kiely M (2002) Vitamin E, nutritional significance. In: Roginski H, Fuguay JW, and Fox PF (eds.) *Encyclopedia of Dairy Science*, pp. 2670–2677. London: Elsevier.

Neuzil J, Weber C, and Kontush A (2001) The role of vitamin E in atherogenesis: Linking the chemical, biological and clinical aspects to the disease. *Atherosclerosis* **157**: 257–283.

Packer L and Fuchs J (eds.) (1993) *Vitamin E in Health and Disease*. New York: Marcel Dekker.

Pryor WA (2000) Vitamin E and heart disease: Basic science to clinical intervention trials. *Free Radical Biology and Medicine* **28**: 141–164.

Rimbach G, Minihane AM, Majewicz J *et al.* (2002) Regulation of cell signalling by vitamin E. *Proceedings of the Nutrition Society* **61**: 415–425.

Thomas SR and Stocker R (2000) Molecular action of vitamin E in lipoprotein oxidation: Implications for atherosclerosis. *Free Radical Biology and Medicine* **28**: 1795–1805.

Traber MG and Sies H (1996) Vitamin E in humans: Demand and delivery. *Annual Review of Nutrition* **16**: 321–347.

VITAMIN K

C J Bates, MRC Human Nutrition Research, Cambridge, UK

The discovery of vitamin K as an essential nutrient arose in the late 1920s from Henrik Dam's studies of sterol metabolism. He observed that chicks fed a fat-free diet developed subcutaneous hemorrhages and anemia. A lipid extract of liver or of certain plant tissues was curative, and by 1935 he claimed discovery of a new vitamin in these extracts that he named 'vitamin K' from the German *Koagulation*. By the late 1930s, two chemically similar forms of the vitamin from different sources were recognized, namely phylloquinone or K_1 and menaquinone or K_2, which had been isolated from alfalfa and from putrefied fish meal, respectively (**Figure 1**). Phylloquinone, with its saturated phytyl side chain, is now understood to be the sole representative of vitamin K that occurs in plant tissues, especially in green leafy ones, where it acts as a component of the electron transport chain. The menaquinones, or MK-n, by contrast, comprise a broad family of representatives that have a variable length, unsaturated side chain, and are composed of one or more (sequential) isoprene units in place of the saturated phytyl side chain. These menaquinones can be produced by certain types of bacteria, both in the large bowel of animals and at other locations where they may contribute to human food sources of menaquinones. Germ-free rats become vitamin K deficient more readily than their conventional counterparts, and they can develop very low hepatic MK-4 levels. The specific menaquinone with the same side chain length as phylloquinone is called menatetranone, or MK-4, and this is produced commercially for human medication, especially in Japan. There is evidence that phylloquinone can be converted to MK-4 in animals and humans. Most bacterially synthesized menaquinones have longer side chains, typically 7–9 isoprene units and up to 13, which are indicated by 'n' in the MK-n shorthand notation. A synthetic homolog of phylloquinone, $K_{1(25)}$, is not found in nature and can therefore be used as an internal standard in the chromatographic separation and quantitation of vitamin K. Menadione, a water-soluble form of the vitamin that has a single methyl group in place of the side chain, has vitamin K activity (it can be converted to menatetranone *in vivo*) and is used in animal feeds, but it is not used in humans because of its toxicity at high doses.

Food Sources, Absorption, Distribution, and Turnover

Food sources of phylloquinone for man (**Table 1**) include green leafy vegetables as the major

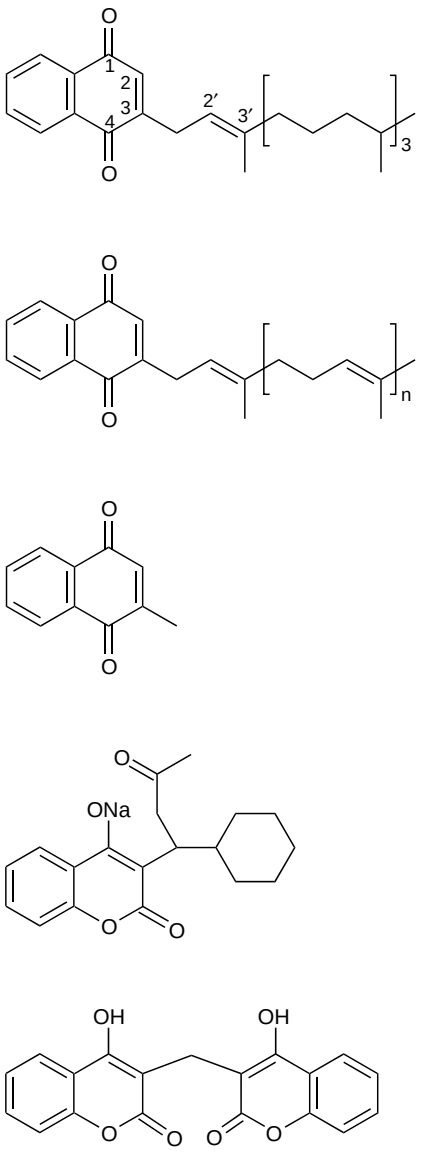


Figure 1 Chemical structures of phylloquinone, menaquinones, menadione, warfarin, and dicumarol.

Table 1 Mean estimate food contents of phylloquinone and selected menaquinones

Food item	*μg/100 g wet weight or μg/100 ml*[a]				
	Phylloquinone (vitamin K_1)	*MK-4*	*MK-7*	*MK-8*	*MK-9*
Kale	817	—	—	—	—
Spinach	387	—	—	—	—
Broccoli	156	—	—	—	—
Peas	36	—	—	—	—
Apples	3.0	—	—	—	—
Chicken	—	8.9	—	—	—
Pork	0.3	2.1	—	0.5	1.1
Luncheon meat	3.9	7.7	—	—	—
Mackerel	2.2	0.4	—	—	—
Plaice	—	0.2	0.1	1.6	
Milk	0.5	0.8	—	—	—
Hard cheese	10.4	4.7	1.3	16.9	51.1
Soft cheese	5	4	1	10	40
Natto[b]	34.7	—	998	84.1	
Olive oil	53.7	—	—	—	—
Margarine	93.2	—	—	—	—
Butter	14.9	15.0	—	—	—
Corn oil	2.9	—	—	—	—
Bread	1.1	—	—	—	—

[a]A dashed line means not detectable. Values obtained for MK-5 and MK-6 are omitted from this summary. The data demonstrate clearly (i) the huge difference in vitamin K contents between different foods and (ii) the preponderence of phylloquinone in some foods and of menaquinones (of several different chain lengths) in others.
[b]A Japanese food made from fermented soya bean curd.
Data from Schurgers LJ and Vermeer C (2000) Determination of phylloquinone and menaqunones in food. Haemostasis **30**: 298–307, **Table 2**.

quantitative source; however, its availability for absorption from these foods is thought to be relatively poor. Certain plant-derived oils, notably soya and canola oils, are also rich in the vitamin, which is probably much more readily available from such sources than it is from leaves. Menaquinones are typically obtained from foods, such as cheeses or Japanese 'natto' (fermented bean curd), in which bacterial fermentation has occurred. Smaller amounts of both phylloquinones and menaquinones are obtained from liver and other animal-derived foods.

Phylloquinone is highly lipophilic; however, at low concentrations it is transported by a saturable, energy-dependent transport system across the gut wall, mainly in the upper small intestine. Phylloquinone in foods consisting of plant tissues is much less readily bioavailable for absorption than the pure vitamin since it is tightly bound to the thylakoid membranes of the chloroplasts, and the absorption of vitamin K from plant foods is considerably improved by including additional fat in the meal. Its absorption also depends on the stimulation of bile salt and pancreatic lipase secretions. The long-chain menaquinones, which are even more lipophilic, are only passively absorbed and are much less bioavailable for absorption than phylloquinone. However, if given by injection (e.g., intracardially), they can be even more functionally active than phylloquinone. The relative bioavailability and bioactivity of the different forms and food sources of vitamin K need more research. Preliminary studies with

deuterium-labeled broccoli suggest that the bioavailability of endogenous vitamin K can be studied in humans by intrinsic stable isotope-labeling procedures.

Once absorbed, vitamin K is transported to the liver in the chylomicrons, where it becomes distributed among the triglyceride-rich chylomicron remnants (ca. 50%) and the low-density lipoprotein and high-density lipoprotein fractions of plasma (ca. 25% each). Plasma vitamin K concentrations, which are typically in the low nanomolar range in humans, are much lower than for the other fat-soluble vitamins (A, D, and E), and they are strongly correlated with the triglyceride content of the plasma. Indeed, some authorities prefer to express plasma vitamin K as a ratio to triglycerides instead of as a simple concentration. Differences between the apoE lipoprotein genetic variants affect plasma vitamin K, according to their different triglyceride clearance profiles. There is evidence for a major diurnal cycle of plasma vitamin K, with peak concentrations of both vitamin K and its associated triglycerides occurring in late evening and with lowest values in the morning. A kinetic study using radioactive vitamin K indicated that the turnover time of the exchangeable pool of the vitamin is quite short, approximately 1.5 days, and the first and second exponential decay curves had half-lives of 0.5–1 and 25–78 h, respectively. The exchangeable body pool size was only approximately 1 μg/kg body weight. The liver is an important repository of the vitamin for both plant-derived phylloquinone and the bacterially derived menaquinones. Depletion studies have indicated that the hepatic phylloquinone stores seem much more labile than the menaquinone stores, and that a functional deficiency accompanies the loss of the phylloquinone, which the remaining nondepleted menaquinones cannot prevent. Despite this, if menaquinones are given exogenously, they can be curative. Different tissues have different relative avidities for phylloquinone and menaquinones, and it has been suggested that they may have a different spectrum of functions from each other. Thus, in humans, phylloquinone is concentrated in liver, heart, and pancreas. The longer chain menaquinones, MK-6 to -11, are found mainly in liver with traces in heart and pancreas, but MK-4 is found especially in brain and kidney, where it exceeds phylloquinone concentrations. The tissue distribution in humans is similar to that in the rat.

The turnover of phylloquinone results in *ca.* 40–50% of the exchangeable body pool being transferred via the bile into the feces and 20% being excreted into the urine, the latter including the excretion of oxidized products that become conjugated as glucuronides.

Physiological Functions of Vitamin K: Interaction with Antagonists

Blood Coagulation Proteins

The principal physiological function that led to the discovery of vitamin K, and its confirmation as an essential vitamin for higher vertebrates, was its unique role in the blood clotting cascade. This cascade comprises a complex series of linked proenzyme-to-enzyme conversions, which leads eventually to a fibrin clot (**Figure 2**). Central to this process is the activation by calcium of gamma-carboxylated glutamyl (Gla) residues in some of the members of the cascade series: factors VII, IX, and X and factor II (prothrombin). In addition, there is an inhibitory level of control by proteins C, S, and possibly Z. All seven of these Gla proteins have Gla clusters that interact specifically with calcium so as to alter their polypeptide conformations and to permit their interaction with other members of the coagulation cascade (by exposing a phospholipid-binding domain) and hence leading either to activation or to inhibition of individual components. The Gla moieties of these and indeed all the vitamin K-dependent Gla proteins are formed by a post-translational carboxylation reaction catalyzed by the single enzyme, 'carboxylase,' at the

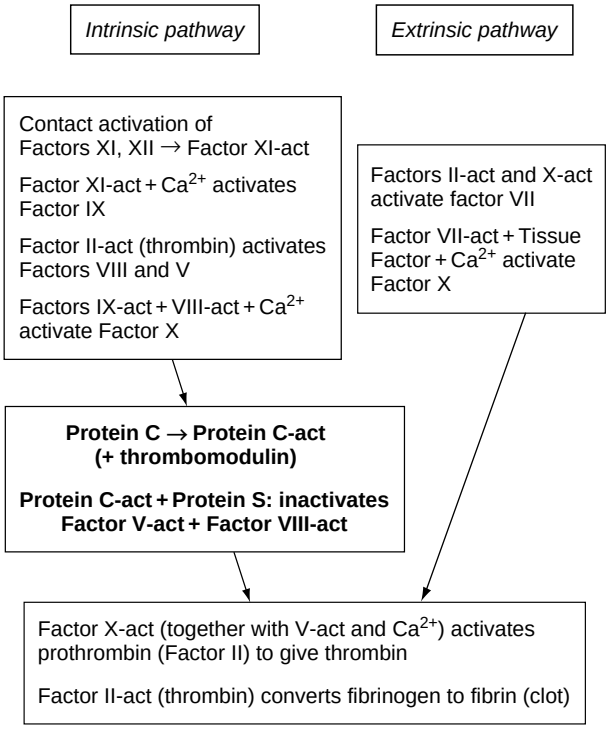


Figure 2 Vitamin K-dependent clotting factors. Factors II (prothrombin), VII, IX, and X and proteins C and S are all Gla proteins. The functions of proteins C and S, shown in bold, are inhibitory to the clotting cascade, whereas the other factors all form part of the cascade mechanism.

endoplasmic reticulum sites of Gla protein synthesis. In the case of the blood coagulation proteins, the sole site of synthesis is the liver. Each carboxylated protein has a C-terminal 'propeptide' sequence that binds the carboxylase enzyme, and directs a coordinated series of carboxylations of the recipient glutamyl residues, before the propeptide is removed and the fully carboxylated protein is then secreted into the extracellular space for transport into the plasma.

Vitamin K acts as the essential recycling cofactor (or cosubstrate) for all protein carboxylation, Gla-forming reactions (**Figure 3**). In its dihydro or quinol form, the vitamin reacts with molecular oxygen, thereby creating a highly reactive, high-energy carbanion at the Glu site for insertion of carbon dioxide, creating a new Gla residue. This vitamin K quinol oxidation step provides the essential energy for the endothermic carboxylation step. The other product of the reaction is the epoxide of vitamin K, comprising a three-membered carbon–oxygen ring. Since the oxidized vitamin needs to be recycled back to the quinol form before the next protein carboxylation cycle, a two-stage reduction process ensues, forming first vitamin K quinone and then the original quinol (**Figure 3**). Both of these reduction steps can be catalyzed by the enzyme vitamin K epoxide reductase, which is linked to a dithiol–disulfide reducing couple and which is highly sensitive to inhibition by the coumarin class of drugs, of which warfarin (**Figure 1**) is the best known and most commonly used member. The reduction of the intermediate vitamin K quinone to its quinol form can also be catalyzed by another, NAD(P)H-dependent, quinone reductase that is warfarin resistant, and for this reason the inhibition of carboxylation by warfarin can be reversed or antagonized by large doses of vitamin K provided exogenously in its normal quinone form. A severe deficiency of vitamin K, or treatment with coumarin drugs (for the control of excessive blood clotting tendency in humans), results in prolonged clotting times that can be detected by the standardized 'one stage prothrombin time' test, in which citrated or oxalated (i.e., calcium-complexed) blood is treated with tissue factor plus additional calcium so as to initiate the clotting process. However, a much more sensitive test for mild vitamin K deficiency is the PIVKA test (Proteins Induced by Vitamin K Absence or Antagonism), which is an immunological enzyme-linked immunosorbent assay (ELISA) test that specifically recognizes undercarboxylated blood clotting proteins and particularly des-gamma-carboxy prothrombin.

Proteins C and S, and possibly also Z, function differently from the other Gla-containing blood clotting factors that are an integral part of the fibrin-forming cascade. Protein C has a regulatory role, inactivating factors V and VIII, and in conjunction with protein S it also acts as a cofactor to enhance the rate of fibrinolysis of blood clots in locations where they are unwanted and potentially harmful. The exact function of protein Z remains unresolved, although interactions with thrombin and factor X have been reported. Clearly, there is a delicate balance of

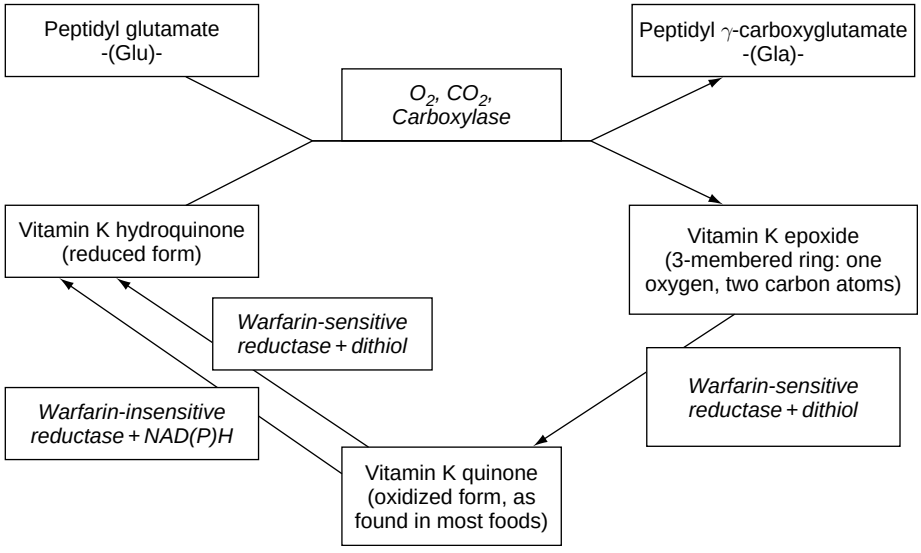


Figure 3 Vitamin K oxidation–reduction cycle during Gla formation. Oxidation of vitamin K hydroquinone (reduced vitamin) to vitamin K epoxide by molecular oxygen provides the energy needed to drive the carboxylation of peptidyl-Glu to peptidyl-Gla (i.e., gamma-carboxy glutamate). The vitamin K epoxide is then recycled by reduction with dithiols in two stages. The first stage requires a reductase enzyme that is coumarin drug (e.g., warfarin) inhibitable. The second stage can be catalyzed by either of two reductases, one of which is NAD(P)H dependent and is not warfarin inhibited.

pro- and anti-clot formation and removal activities among the vitamin K-dependent Gla proteins of the cascade, although the net effect of a deficiency of the vitamin or of its antagonism by drugs appears to be a reduction of the clotting tendency.

Bone Gla proteins

Protein S together with two other Gla proteins, osteocalcin (OC; or bone Gla protein) and matrix Gla protein (MGP), play a variety of only partly understood roles in bone and other mineralized tissues. Of these proteins, only OC is produced solely and specifically by mineralized tissue, whereas the other two (or at least their mRNA templates) are more widespread and occur also in soft tissues.

OC is synthesized specifically by osteoblasts and odontoblasts, and it accounts for ca. 15–20% of the noncollagen protein of the bone matrix. Approximately 20% is secreted into blood plasma, where it has no obvious function, but it has frequently been measured as an index of bone-forming (osteblastic) activity, and is present in increased amounts in plasma of people with certain bone diseases and of young infants. It is a small protein, MW 5700, with just three Gla residues. Unlike the blood coagulation Gla proteins, which in most people not severely vitamin K deficient and not vitamin K antagonist treated are almost completely carboxylated, circulating OC is at least 5–10% undercarboxylated in many population groups, as measured by assays that depend on the affinity of the undercarboxylated form for hydroxyapatite or a specific ELISA assay for the undercarboxylated form. Since vitamin K supplements can reduce its degree of undercarboxylation in many people, it has been proposed as a new and highly sensitive functional test of vitamin K status in man.

Despite the growing level of interest in its practical use as a status index, our understanding of the essential function of OC remains incomplete. Its affinity for calcium is less strong than that of the larger Gla proteins, but it binds avidly to hydroxyapatite and is chemotactic for osteoclasts and their progenitors. Moreover, it can enhance the differentiation of osteoclast progenitor cells in culture, which has been interpreted as implying a possible role in bone resorption. Transgenic mice that lack the gene for OC have increased bone mass, despite an increased number of osteoclasts. In humans, however, underhydroxylation of OC especially in postmenopausal women has been linked to low vitamin K intakes, reduced bone mineral density, and increased risk of fracture. Intervention with high-dose MK-4, mainly in Japan, has been reported to improve bone mineral density and decrease fracture risk. Although a single study in the United Kingdom suggested that a combination of vitamin K_1 and vitamin D supplements may benefit bone mineral density in postmenopausal women, considerably more research is needed in this area. The separate roles of OC and other vitamin K-dependent proteins also need to be clarified.

The second vitamin K-dependent Gla protein in bone, MGP, has a MW of 9600 and five Gla residues and is highly insoluble. Unlike OC, it is also found in cartilage, and, significantly, its mRNA occurs in several soft tissues including artery walls. Its synthesis is modulated by 1,25-dihydroxy vitamin D and by retinoic acid. Mice lacking the gene for MGP quickly developed calcified arteries and died of aortic rupture before 2 months of age. For this reason, MGP is believed to antagonize the pathological calcification of soft tissues and thus to protect them. The absence of MGP also led to inappropriate calcification of growth plate cartilage, reduced growth, osteopenia, and fracture in the MGP gene knockout mice. In humans, defects in the MGP gene are associated with Keutel's syndrome and chondroplasia punctata, in which cartilage calcification is abnormal. Similar abnormalities have been observed in infants whose mothers were treated with warfarin during the first trimester of pregnancy. In one study, low vitamin K intake was associated with atherosclerotic calcification of the aorta in postmenopausal women. Also, circulating MGP levels were found to be raised in severe atherosclerosis and in type 1 diabetes in humans. A specific immunoassay for MGP has been developed that should assist further research on this potentially important regulatory protein.

The third bone-associated Gla protein, protein S, is also involved with blood clotting. It is synthesized by osteoblast-like and osteblastoma cells in culture, and it has been detected in bone matrix. It is also synthesized by hepatocytes, megakaryocytes, and endothelial cells. Children with an inborn deficiency of it developed osteopenia and bone lesions; however, its precise functional role is unknown.

All three bone Gla proteins (and probably most other Gla proteins) have 'leader' or 'pre'-peptides when first formed on the endoplasmic reticulum (ER) that are required for translocation across the ER and are removed during this process. OC, protein S, and most other Gla proteins also have a pro-peptide sequence that is removed during secretion and that directs the action of the carboxylase enzyme before secretion. MGP differs from the other Gla proteins in that its carboxylase recognition sequence is not removed; instead, only a short (five-residue) carboxy-terminal sequence is removed from it. All known mammalian Gla proteins contain the characteristic amino acid sequence

Gla-X-X-X-Gla-X-Cys, where X represents an undefined amino acid. If vitamin K is in short supply or antagonized, certain Gla residues escape gamma-glutamyl formation more than others. Thus, in a study of OC, the Glu residue at position 17 was typically only 67% carboxylated, that at position 21 was 88% carboxylated, and that at position 24 was 93% carboxylated.

Surprisingly, in a meta-analysis of studies on warfarin-treated adult patients, no evidence of any increase in bone disorders was found.

Gas6 and Other Vitamin K-Requiring Gla Proteins

A Gla protein that is associated with the central nervous system, rather than with liver or bone, was discovered in 1993. In tissue culture models it had the properties of a growth arrest-specific (GAS) cell-signalling gene product. It acts as a ligand for a number of receptor protein kinases; it potentiates the growth of vascular smooth muscle cells, Schwann cells, and the neurons that synthesize gonadotropin-releasing hormones; and it can prevent apoptotic cell death. Knockout mice in which three Gas6 receptors are mutated had major neurological and spermatogenic abnormalities. There is interest in potential roles for Gas6 in Alzheimer's disease and Parkinson's disease. Clearly, these properties and emerging roles have helped to confirm the growing suspicion that vitamin K-dependent Gla proteins possess key functions beyond blood clotting and even bone remodelling. Gas6 has a MW of 75,000 with 11 or 12 Gla residues, and its structure is partly homologous with protein S.

Even less well characterized are several other Gla proteins from a variety of tissues. Kidney contains 'nephrocalcin,' with just two or three Gla residues, which may be involved in renal calcium transport (another important function that may be impaired by vitamin K deficiency in man). Atherocalcin, or plaque Gla protein, may be related or even identical to MGP. Proline-rich Gla proteins PRGP-1 and PRGP-2 are found predominantly in the spinal cord and thyroid gland, respectively, but their functions are unknown. Gla proteins occur in most vertebrates and also in molluscs, so their evolutionary appearance in the animal kingdom is probably quite ancient in origin.

Other, Probably Non-Gla Functions of Vitamin K

Vitamin K is thought to be involved in sphingolipid metabolism in certain bacteria by modulating serine palmitoyl transferase, and warfarin treatment decreased brain levels of sulfatides and galactocerebroside sulfotransferase activity in animals, which was reversible by vitamin K (either K_1 or MK-4). Therefore, it is now thought that vitamin K may be involved in sphingolipid metabolism, and this in turn has implications for its action as a second messenger as well as being a structural component. There are several functions of MK-4 that are shared by the isolated geranyl-geraniol side chain, which involve the induction of apoptosis of osteoclasts and of certain cancer cells in culture. Depriving certain tumours of vitamin K, both *in vitro* and *in vivo*, seemed to inhibit their growth and metastasis. Patients receiving warfarin for cardiovascular disease seem to have a reduced incidence of tumors, and warfarin may also suppress delayed-type hypersensitivity reactions.

Recent studies have suggested that MK-4, in particular, has a transcriptional regulatory function, for example, in osteosarcoma cell cultures, in which it binds to and activates the SXR steroid and xenobiotic receptor. This in turn increases mRNA levels for osteoblast markers: bone alkaline phosphatase, osteoprotogerin, osteopontin, and MGP. MK-4 and its isolated geranyl-geraniol side chain was also able to suppress the synthesis of prostaglandin E_2, which is a potent bone resorption catalyst. These observations have led to speculation (i) that some of the menaquinones may possess some functions that are not shared by phylloquinone, and (ii) that there may be implications for cell proliferation and for cancer risk from variations in the supply of vitamin K and in its speciation.

Population Groups at Risk of Vitamin K Deficiency

Because of the minimal extent of transfer of vitamin K across the placenta, the fetus and newborn infant have much lower circulating vitamin K than adults (typically 30-fold lower). In addition, human milk has a lower concentration of the vitamin than that of most other mammalian species. Although low vitamin K levels have not been found to affect the developing fetus in a functionally deleterious way, it is clear that the newborn, and especially the solely breast-fed infant, is at higher risk of functional deficiency than older infants and adults. In a minority of cases, this can lead to life-threatening or long-term damage associated with intracranial bleeding. Hemorrhagic disease of the newborn (HDN) is classified as early (first 24 h of life), classic (days 1–7), or late (2–12 weeks). Of these, the third category is most likely to involve dangerous intracranial bleeding. Risk factors for HDN include intestinal fat malabsorption and hepatic disease. In Western countries, since the 1950s, it has been routine practice to give prophylactic phylloquinone in a 1 or 2 mg dose at birth, and this has been found to considerably reduce the risk of HDN. An intramuscular depot dose was found to be

highly effective; however, a study in the United Kingdom in the 1990s suggested a possible link with childhood cancer. Despite little subsequent support for this contraindication, the adverse publicity led to a shift in practice toward oral dosing. An oral micellar preparation containing glycholate and lecithin has been developed that has improved absorption characteristics. Another approach toward the avoidance of late HDN is vitamin K supplementation of breastfeeding mothers since breast milk vitamin K levels can be increased substantially by dosage to the mother. Modern commercial formula feeds typically contain 50–125 μg phylloquinone/l.

Antibiotic-treated patients may be at increased risk of developing vitamin K deficiency. Some antibiotics may reduce the production of usable menaquinones by gut bacteria; others, such as cephalosporin, may exert vitamin K epoxide reductase inhibitory effects. Vitamins A and E in large doses may increase the risk of vitamin K deficiency and/or its sequelae in susceptible people. Thus, in one study, patients receiving anticoagulant drugs exhibited a further reduction of prothrombin levels if they were given 400 IU α-tocopherol per day for 4 weeks. The microsomal vitamin K-dependent carboxylase enzyme was found to be inhibited by α-tocopheryl quinone and, to a lesser extent, by α-tocopherol. It is also inhibited by other oxygen free radical antagonists. Control of blood clotting with warfarin-type drugs thus requires control of intakes of vitamins A and E as well as vitamin K so as to achieve consistent results.

As noted previously, some older people, especially postmenopausal women, seem to be at increased risk of developing marginal vitamin K deficiency, which manifests itself, for instance, by an increased percentage of undercarboxylated osteocalcin (ucOC) in the circulation. The sequelae of such marginal deficiency, and in particular its implications for bone health, are currently the subject of considerable research effort (**Table 2**). Several epidemiological cross-sectional studies have noted an association between higher vitamin K intakes and higher bone mineral density or lower fracture risk. One study reported that a subgroup of postmenopausal women who were 'fast losers' of calcium responded to vitamin K supplements by reduced calcium and hydroxyproline excretion. Although vitamins D and K have distinct functions in calcium absorption, and its distribution, deposition, and excretion, there is evidence that synergistic interactions can occur between them, and that both can affect the same cell-signalling pathways. Osteocalcin and MGP synthesis is stimulated by 1,25-dihydroxy vitamin D in cell culture.

MK-4 in large doses has been used for prophylaxis and treatment of osteoporosis, especially in Japan. A study in The Netherlands reported reduced bone loss after 2 years of treatment of postmenopausal women with amounts of phylloquinone that are achievable from dietary sources. More long-term intervention trials are needed.

Table 2 Studies (1985–2001) linking vitamin K intake, status, or effects of supplementation with bone health in humans

Nature of evidence	*No. of studies*
Serum vitamin K positively correlated with BMD	4
Serum vitamin K lower in people with hip or vertebral fractures	3
Vitamin K intake directly correlated with BMD	2
ucOC directly correlated with risk of hip fracture	5
ucOC inversely correlated with velocity of ultrasound (a measure of bone quality)	1
ucOC inversely correlated with BMD	2
Supplementation with phylloquinone increased carboxylation of osteocalcin	7
Supplementation with phylloquinone or menaquinone reduced calcium loss	3
Supplementation with phylloquinone increased markers of bone formation and reduced markers of bone resorption	1
Supplementation with phylloquinone (+vitamin D) increased BMD	1
Supplementation with menaquinone (+vitamin D) increased BMD	2
Supplementation with menaquinone alone increased BMD and/or decreased bone loss	6
Supplementation with menaquinone reduced fracture risk	2

BMD, bone mineral density; ucOC, undercarboxylated osteocalcin. Data from Weber P (2001) Vitamin K and bone health. *Nutrition* **17**: 880–887, and S. Karger AG, Basel.

Status, Requirements, and Recommended Intakes

Vitamin K status can be measured either by its concentration in plasma or by its efficacy in ensuring optimal carboxylase function, as indicated by specific carboxylated plasma proteins. Accurate assay of the very low concentrations of vitamin K that are present in plasma was a considerable analytical challenge, which was eventually solved by high-performance liquid chromatography (HPLC) followed by high-sensitivity coulometric or fluorometric detection. A popular method uses organic solvent extraction, a cartridge cleanup step, an HPLC separation followed by postcolumn reduction of the vitamin K quinone to the reduced quinol form by metallic zinc or other reductant, and finally fluorometric quantitation of the fluorescent quinol. A useful internal standard, not found in nature, is the homolog of phylloquinone, vitamin $K_{1(25)}$. With

modern detectors, analysis is possible with only 0.25 ml plasma. A published 'normal' range in the United States is 0.25–2.7 nmol/l, corresponding to approximate average daily intakes of 100 μg/day in men and 80 μg/day in women. As noted earlier, the phylloquinone content of plasma has a short half-life and is strongly correlated with plasma triglycerides. It is therefore not ideal as a long-term index of status. Alternatives include functional indices such as plasma prothrombin time (increased only by severe vitamin K deficiency), PIVKA, (which is more sensitive to marginal deficiency), and ucOC (which is the most sensitive functional indicator). These functional indices are not totally specific for vitamin K deficiency, although ucOC (for which monoclonal antibodies now exist) does appear to possess reasonably good specificity. Unfortunately, the different commercial kit assays measure different epitopes of OC, which makes harmonization difficult. Urinary total Gla is sensitive to vitamin K status, but it varies with age and has not yet proved to be very useful as a status indicator. Functional indices that are based on impaired carboxylase activity affecting other Gla proteins may be developed in the future.

Most estimates of the amount of phylloquinone needed to correct clotting changes suggest that adult human requirements are between 0.5 and 1 μg/kg/day. There are no reference nutrient intakes defined for vitamin K in the United Kingdom, although a 'safe intake' for adults was set in 1991 at 1 μg/kg/day and for infants 10 μg/day. In the United States, the Food and Nutrition Board of the National Academy of Sciences has defined an Adequate Intake (AI) of phylloquinone of 90 μg/day for adult women and 120 μg/day for adult men, with proportionately smaller values for children. For infants aged 0–6 months, the AI is only 2 μg/day, and it is 2.5 μg/day at 7–12 months, thus creating a larger proportional difference between infants and older age groups than for most micronutrients.

Both phylloquinone and the menaquinones appear to be nontoxic, even in multimilligram amounts. However, menadione, the water-soluble form of vitamin K, was found to cause hemolytic anemia, hyperbilirubinemia, and kernicturus in infants when >5 mg was given. Therefore, it is not currently used for human prophylaxis or treatment.

Since vitamin K is thought to have a wide range of functions in the body in addition to blood clotting, and some of these may have long-term health implications, research on requirements and optimal intakes, with multiple end points, is needed. Metabolic and health-related differences between the menaquinones and phylloquinone also need to be defined.

See also: **Bone**. **Fruits and Vegetables**. **Infants**: Nutritional Requirements. **Pregnancy**: Safe Diet for Pregnancy. **Vitamin A**: Biochemistry and Physiological Role. **Vitamin E**: Physiology and Health Effects.

Further Reading

Binkley NC and Suttie JW (1995) Vitamin K nutrition and osteoporosis. *Journal of Nutrition* **125**: 1812–1821.

Bugel S (2003) Vitamin K and bone health. *Proceedings of the Nutrition Society* **62**: 839–843.

Ferland G (1998) The vitamin K-dependent proteins: An update. *Nutrition Review* **56**: 223–230.

Greer FR (1999) Vitamin K status of lactating mothers and their infants. *Acta Paediatrica Supplement* **430**: 95–103.

Nelsestuen GL, Shah AM, and Harvey SB (2000) Vitamin K-dependent proteins. *Vitamins and Hormones* **58**: 355–389.

Saxena SP, Israels ED, and Israels LG (2001) Novel vitamin K-dependent pathways regulating cell survival. *Apoptosis* **6**: 57–68.

Shearer MJ (1997) The roles of vitamins D and K in bone health and osteoporosis prevention. *Proceedings of the Nutrition Society* **56**: 915–937.

Shearer MJ (2000) Role of vitamin K and Gla proteins in the pathophysiology of osteoporosis and vascular calcification. *Current Opinion in Clinical Nutrition and Metabolic Care* **3**: 433–438.

Suttie JW (1992) Vitamin K and human nutrition. *Journal of the American Dietetic Association* **92**: 585–590.

Suttie JW (1995) The importance of menaquinones in human nutrition. *Annual Review of Nutrition* **15**: 399–417.

Tsaioun KI (1999) Vitamin K-dependent proteins in the developing and aging nervous system. *Nutrition Review* **57**: 231–240.

Vermeer C, Jie K-SG, and Knapen MHJ (1995) Role of vitamin K in bone metabolism. *Annual Review of Nutrition* **15**: 1–22.

Vermeer C and Schurgers LJ (2000) A comprehensive review of vitamin K and vitamin K antagonists. *Hematology/Oncology Clinics of North America* **14**: 339–353.

Weber P (2001) Vitamin K and bone health. *Nutrition* **17**: 880–887.

W

Water *see* **Thirst**

WEIGHT MANAGEMENT

Contents

Approaches

N Finer, Luton and Dunstable Hospital NHS Trust, Luton, UK

Weight loss and weight loss maintenance require a decrease in energy intake (diet), an increase in energy expenditure (exercise and physical activity), or both. Dietary management should encourage healthy eating, that is, an appropriately balanced intake of macro- and micronutrients. For most obese individuals this will entail not just a decrease in total energy intake, but specifically a decrease in fat intake, together with an increase in complex carbohydrates, fruit, and vegetables. Myriad diets have been popularized as a means to reducing energy intake, but few are recommended as meeting the overall nutritional needs of an obese individual, and many are so restrictive that they clearly could not be followed for more than a few weeks. Increasing exercise and physical activity has benefits beyond those that result from the relatively modest amounts of extra energy expended. These include a beneficial protection from excessive loss of lean body tissue during dieting, improved fitness and psychological health, and a greater likelihood of long-term weight maintenance. Diet and exercise are core components of behavioral treatments; such treatments, based on learning theories, also aim to help individuals become aware of the behaviors that have led to their weight gain, and to develop strategies to alter them. Weight loss can be achieved successfully with all strategies; behavioral therapies that include a strong focus on increasing exercise and activity seem to offer the best chances of long-term success.

The Concept of Desirable Weight

Body weight reflects the additive mass of the various tissues that make up the organism, and is a function of energy and nutrient balance over a prolonged period. Positive energy balance will result in weight gain (mainly from deposition of lipid in adipose tissue), while prolonged undernutrition will lead to weight loss. For most of human history, the dominant disorder of body weight has been thinness. Thinness, whether from malnutrition or disease, was associated with illness and was often a prelude to death; in societies where food supplies are scarce or seasonal, a high body weight may be seen as a desirable sign of health, and probably wealth. In contrast, in developed societies where levels of activity are low and food is plentiful, the growing prevalence of over-weight and obesity has been clearly linked to illness and premature mortality. The concept of a desirable weight at which health is optimal and the risk of disease minimal has not been easy to define, largely because of the effects of many other factors such as age, sex, social status, and smoking.

Dietary Management

Dietary management of obesity aims to reduce fat stores by changing eating habits to reduce energy intake below that required for weight maintenance. The term 'reducing diet' has been coined to describe such diets used to treat the obese. Since many obese individuals may eat a nutritionally inadequate (apart from energy) diet, it is important that advice on energy restriction is accompanied by the prescription of a 'healthy' diet that contains adequate protein, vitamins, calcium, trace elements, and a desirable ratio of complex carbohydrate to fat. Weight loss *per se* is of no medical benefit unless it is maintained, and this will require the obese individual to adhere to a permanent change in eating habits. Many think of a 'diet' as a temporary change in eating habits (often extreme or quirky), a view encouraged by many of the diet books that hold out the promise of easy and instant success. It is essential that the concept of a long-term change in dietary habits be accepted at the start of treatment.

The energy value of weight gained or lost is approximately 31 MJ kg^{-1} (7500 kcal kg^{-1}) since it is composed approximately of 3 parts fat to 1 part lean. Thus a daily energy deficit of 2.1 MJ (500 kcal) will produce a weight loss of about 2 kg per month. For the average man or woman this represents a 20–30% reduction in energy intake, although for the obese the percentage reduction will be smaller. Thus the severely obese, for example with a body mass index (BMI) of 35 or more, will need to follow an energy-restricted diet for months rather than weeks to reverse their obesity. As weight is lost, energy requirements fall, in part because of the reduced energetic mass of the person, and also because of adaptive changes in energy expenditure. For this reason, the rate of weight loss will eventually slow and reach a plateau for any fixed level of dietary energy restriction (**Figure 1**).

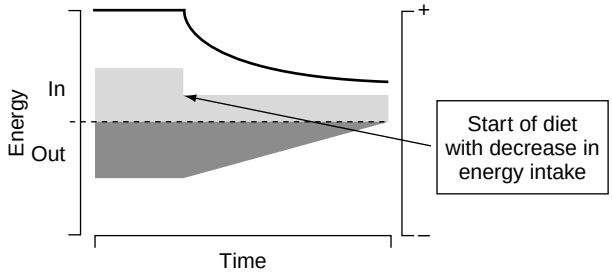


Figure 1 Fall in body weight (solid line) resulting from a fixed decrease in energy intake. Note that the rate of weight loss slows as the gap between energy expenditure (darker shading) and energy intake (lighter shading) narrows.

Myriad diets have been popularized and promoted directly to the public, reflecting every possible permutation of increasing or decreasing the major macronutrients. Fashion and commercialism have dictated many of them. **Table 1** shows the variety of diets that have been suggested, and used, for treating obesity. Many of these diets fail to focus on long-term dietary change, and the quirkiness of many makes it unlikely that they would be followed for long.

Current ideas on a reasonable reducing diet are that it should contain at least 100 g carbohydrate to prevent glycogen depletion and ketosis. High-carbohydrate diets are composed of complex carbohydrates and are thus of low energy density, which may aid management of hunger. Since high-carbohydrate diets are low in fat, they have the theoretical advantage of directly reducing the risk of cardiovascular disease. The energetic efficiency with which carbohydrate is converted and stored as fat is lower than that of dietary fat, providing a further advantage. Protein intake must be adequate to maintain lean body mass. Although there is an inevitable fall with weight loss, 0.8 g per kg per day + 1.75 g per 100 calorie deficit of protein (about 44 g daily for women and 56 g daily for men) should be consumed, and fat restricted to less than 30% of total energy. The diet should contain recommended daily intakes of vitamins, minerals, and electrolytes, if necessary by supplementation; 20–30 g daily of fiber should also be consumed.

Many diets prescribe an energy intake that is based on a generalized rather than an individualized assessment of energy needs. The common prescription of 4.2–5.0 MJ (1000–1200 kcal) daily may be problematic and inappropriate. Weight loss in men will be faster and greater compared to women of equal BMI, because of the relatively greater metabolic rate per kilogram of body weight of men. The very obese, whose daily energy requirements can be as high as 12.6 MJ (3000 kcal), may lose weight at an excessive rate and develop symptoms of ketosis, postural hypotension, or excessive hunger. Many obese patients fail to register or admit to the amount of food they consume, and claim that such a diet is more than their habitual intake.

One principle of energy prescription that has proved easy to administer and successful in outcome is to calculate energy requirements from standard formulas (**Table 2**), and prescribe a diet that provides a fixed energy deficit of 2.1 MJ (500 kcal). Compliance and weight loss were better with this approach than with a fixed 5 MJ (1200 kcal) diet.

A diametrically opposite approach is the use of very low-energy liquid diets. These were originally

Table 1 Types of diet used for treating obesity

Generic name for diet	*Typical dietetic modification*	*Popular example of diet*
Starvation diet	Less than 1.2 MJ (300 kcal) per day	Grapefruit and Black Coffee
Very low-energy (protein-sparing) diets[a]	About 2 MJ (500 kcal) per day with >50 g high-quality protein; usually liquid	Cambridge Diet Modifast
Low-energy diet[a]	5–7.5 MJ (1200–1800 kcal) per day often from menus, recipes	Weight Watchers TOPS
Fixed energy deficit diet[a]	Nutritionally balanced, individually tailored to produce fixed energy deficit (e.g., 2 MJ or 500 kcal per day) based on measured or predicted energy needs	Prescribed by dietitian
High-protein diet	Over 40% protein, thus low in carbohydrate and fat	Scarsdale Medical Diet
Low-protein diet		Beverly Hills Diet
High-fat diet	Restricted carbohydrate and protein	Drinking Man's Diet
Low-fat diet[a]	Restrict fat to <20% energy	Prescribed by dietitian Pritikin Diet
High-carbohydrate diet	Effectively low fat, may be high in fiber	F-Plan Diet
Low-carbohydrate diet	Limits carbohydrate to maximum <50 g daily	Yudkin Diet
Macronutrient choice[a]	Choice from lists of macronutrients to encourage intake of foods high in complex carbohydrates	No Counting Diet
Meal replacement	Liquid formula meals of about 1.7 MJ (400 kcal) to replace 1–2 meals daily	SlimFast
Fad diets	Varied; e.g., food combining diets that require macronutrients to be eaten separately and separated by time	Hay Diet

[a]Diets considered medically reasonable under defined circumstances.

Table 2 Formula for estimating resting metabolic rate (RMR) for men and women. The energy expenditure over 24 h can be estimated by multiplying by a factor related to activity levels (1.3, sedentary; 1.5, moderate activity; 1.8, physically very active)

Age (years)	*RMR*
Men	
18–30	0.063 × weight in kg + 2.896 MJ daily
31–60	0.048 × weight in kg + 3.653 MJ daily
>60	0.049 × weight in kg + 2.459 MJ daily
Women	
18–30	0.062 × weight in kg + 2.036 MJ daily
31–60	0.034 × weight in kg + 3.538 MJ daily
>60	0.038 × weight in kg + 2.755 MJ daily

developed in the 1960s to provide a nutritionally complete intake in terms of protein, vitamins, and micronutrients, but provide as little as 1.4 MJ (350 kcal) daily. The inclusion of sufficient high-quality protein was designed to prevent the excessive loss of lean body mass seen with starvation or other ketotic diets, hence the alternative term 'protein-sparing modified fast.' Appropriately selected, well-motivated patients are highly compliant with such diets, and their weight loss can be very rapid. Paradoxically, perhaps, patients seem to find it easier to mount levels of near-total restraint than more moderate restriction. It appears that withdrawing all solid or 'proper' food helps the patient to define himself or herself as 'not eating', in the same way that some quitting smokers find it easier to abstain completely from cigarettes rather than to cut down.

In the 1970s a commercial very low-energy diet formulation (the Last Chance Diet) was marketed, and was associated with a number of deaths from cardiac arrhythmia. This diet was deficient in essential amino acids and in minerals such as magnesium and potassium. It was withdrawn. In the 1980s newer, better formulated diets were commercially marketed. Concerns about their inappropriate use by already slim women, often with an eating disorder, forced governmental health agencies to issue guidelines on their use. In the US, a task force recommended that such diets contain at least 3.3 MJ (800 kcal), be supervised by experienced physicians, and be used only by those with a BMI more than 30, for less than 16 weeks. In the UK a report from the Committee on Medical Aspects of Food Policy suggested such diets should provide a minimum of 1.7 MJ (400 kcal) and 40 g protein daily for

women, and 2.1 MJ (500 kcal) and 50 g protein daily for men and tall women. They were recommended for use only by those with a BMI more than 25 and under medical supervision, for no longer than 4 weeks. The drawback of such diets is that unless they are combined with, or followed by, some other treatment (pharmacological or behavioral), weight regain, often soon and rapid, is almost universal.

More recently, low-energy liquid diets of around 3 MJ (750 kcal) daily have been popularized, often as part of an overall behavior modification program (see later), or in the form of sachets intended to be used as meal replacements. Both approaches have been shown to have potential for success in short-term studies lasting up to 1 year.

Exercise and Physical Activity

The term 'physical activity' refers to bodily movement produced by skeletal muscle that results in energy expenditure; it thus includes activities of daily living, as well as leisure activity from sport and exercise. The term 'exercise' refers to planned or structured bodily movements, usually undertaken in leisure time in order to improve fitness (e.g., aerobics), while 'sport' is physical activity usually in structured competitive situations (e.g., football). Physical activity at recommended levels (moderate intensity for 30 min for 5 days each week) is associated with many health benefits; these include lower all-cause mortality rates, fewer cardiovascular events such as myocardial infarction and stroke, and a lower incidence of metabolic disorders including non-insulin-dependent diabetes mellitus and osteoporosis. Levels of activity have been falling in Westernized societies largely because of a decrease in physical activity at work (from increasing mechanization) and increasingly sedentary leisure-time pursuits (such as television viewing). The Allied Dunbar National Fitness Survey of the UK showed that 70% of the population are insufficiently active, and a separate UK government survey showed that 1 in 3 adults could be classified as sedentary, i.e., taking less than half an hour of continuous moderate-intensity physical activity each week (**Figure 2**). Both cross-sectional data and prospective studies confirm an inverse relationship between physical activity and weight gain. The finding that in many countries such as the UK, average energy intake has fallen over the time that obesity has been increasing, emphasizes the importance of inactivity as a cause of obesity. These secular changes of inactivity are most marked in children who now spend much of their leisure time watching television or in other sedentary pursuits. Health authorities in many countries now

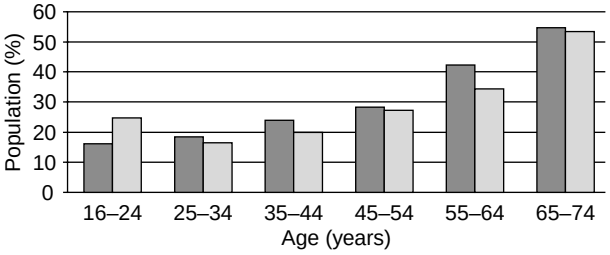


Figure 2 Percentage of adults in England by age and sex (1990–1991) with a sedentary life style; dark bars, men; light bars, women. Data from Fentem and Walker (1995) Setting targets for England: challenging, measurable and achievable. In: Killoran A (ed.) *Moving On: International Perspectives on Promoting Physical Activity*. London: Health Education Authority.

advocate an increase in physical activity as a means of preventing obesity and improving health and fitness. While there is agreement that such measures may be useful in preventing obesity, the role of exercise in treating obesity is less clear. Potential mechanisms linking exercise and activity with weight loss and weight loss maintenance are shown in **Figure 3**. Like dietary change, increasing time spent on exercise and activity can be seen as part of a generalized behavioral change, which can be self-reinforcing.

Exercise and activity raise energy expenditure over and above the resting metabolic rate. Under some circumstances, such as prolonged vigorous exercise in trained individuals, rates of energy expenditure remain elevated for some time after the cessation of exercise. Logically, therefore, exercise should be a useful way to treat obesity. However, the amounts of exercise-induced energy expenditure are small in comparison with potential changes in energy intake.

The energy cost of activity and exercise can be expressed as a multiple of resting metabolic rate, termed a MET; the term 'physical activity level' (PAL) represents the total daily energy expenditure divided by the resting energy expenditure; it typically averages 1.5. The energy costs of walking are about 2.0 MET – for a 70 kg individual this is about 0.5 MJ h^{-1} (120 kcal h^{-1}) – while gentle running costs about 8 MET or 2 MJ h^{-1} (480 kcal h^{-1}). A moderately fit individual would only be able to maintain a level of exercise of 7 MET for about 30 min, representing an additional energy expenditure of about 1.5 MJ (360 kcal) resulting, if energy intake were maintained, in a weight loss of about 0.3 kg per week.

Energy expenditure remains above baseline for some time after exercise has stopped; this is termed 'post-exercise energy expenditure.' The effect is small and only produced by very high levels of activity, capable of achievement only by elite

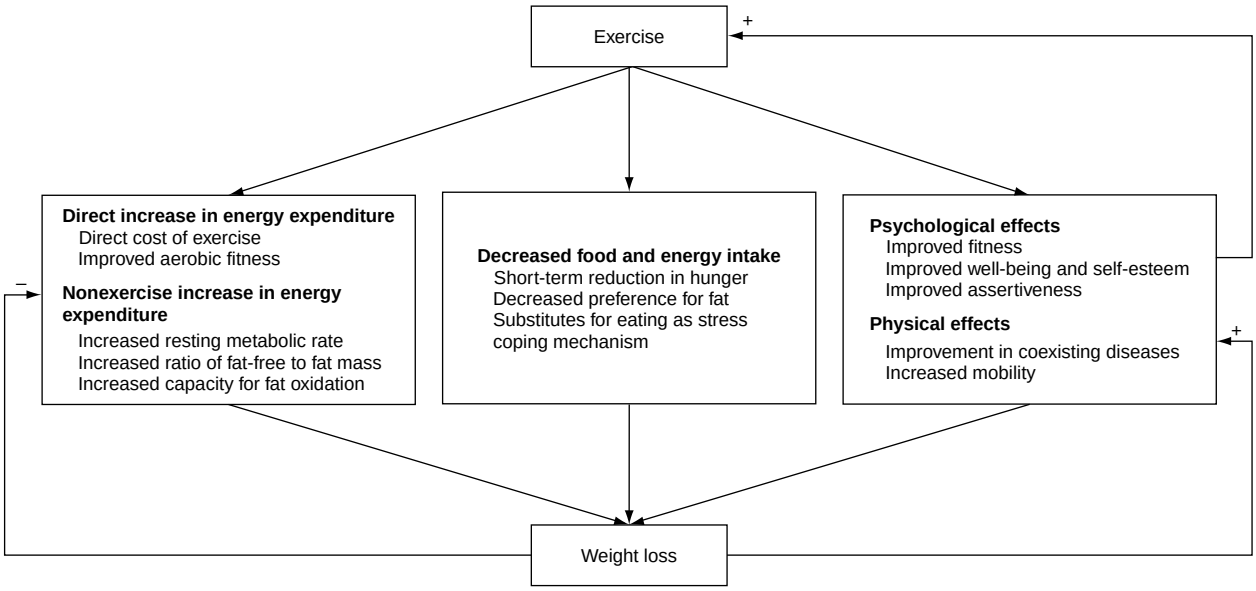


Figure 3 Mechanisms linking exercise with weight loss and weight loss maintenance.

athletes. The mechanism for this effect is unknown. Moderate intensity exercise programs, of the sort prescribed to the obese, are unlikely to raise energy expenditure by more than about 0.2 MJ (50 kcal) per exercise session.

Regular exercise does, however, elevate long-term energy expenditure by its effect on altering body composition. Resting metabolic rate is proportional to the fat-free mass. Exercise increases muscle development and bone mass, so directly raising metabolic rate. The purpose of weight loss is to reduce fat mass, with as little loss of fat-free mass (FFM) as possible. The loss of fat to meet the extra energy requirements of regular exercise will decrease the ratio of fat to FFM and thus indirectly favor an increase in resting metabolic rate for any given body weight. These effects are modest, and mainly only seen from the sort of high-intensity excercise achieved by athletes. Even endurance-level training over periods of up to 12 weeks increases nonexercising daily energy expenditure by less than 0.8 MJ (190 kcal).

The effects of exercise are thus quantitatively small. The relatively small potential for exercise to reduce body weight is borne out by the results of trials of exercise in obesity treatment, which suggest that exercise programs achieve weight losses of less than 0.1 kg per week, and that total weight loss averages about 3 kg. In one meta-analysis of five controlled trials of exercise without dietary restriction, mean weight loss in 95 men was 2.6 kg over 30 weeks, compared with a gain of 0.4 kg in the control group.

Programes that combine dietary and exercise interventions can be more successful, but it is often difficult to separate the effects of one from the other. In order to explore the effect of exercise on the composition of weight loss during dieting, Garrow analyzed data from 21 randomized, controlled studies. All trials that combined exercise and diet and included information about weight and FFM loss were included (**Figure 4**). A small reduction in the percentage of FFM lost is observed if exercise is included with the dietetic intervention. Thus, for example, in a woman losing 15 kg, exercise would reduce her FFM loss from 3.6 kg (24%) to 3.0 kg (20%). Similar but quantitatively greater benefits are seen in men: for a 15 kg weight loss, exercise reduced FFM loss from 3.6 kg (24%) to 2.5 kg (17%).

Activity and exercise are strong predictors for successful weight loss maintenance. A number of studies have shown that obese women who have lost weight and continue to undertake regular exercise are 3–4 times more likely to maintain their weight loss over a follow-up period of 2–3 years. The amount of exercise also correlates with the degree of success. In one study of about a hundred obese men and women who had lost about 27 kg, those with high levels of exercise were maintaining an average of 18 kg loss at 3 years, compared with 9 kg in the moderate exercise group and no weight loss in the nonexercisers. The importance of exercise and weight loss maintenance is demonstrated by a 2-year study of obese subjects treated by either diet, exercise, or a combination of the two. Weight loss in the diet group at 1 year was 6.8 kg, in the exercise group 2.9 kg, and 8.9 kg in the combination treatment group. However, after 2 years

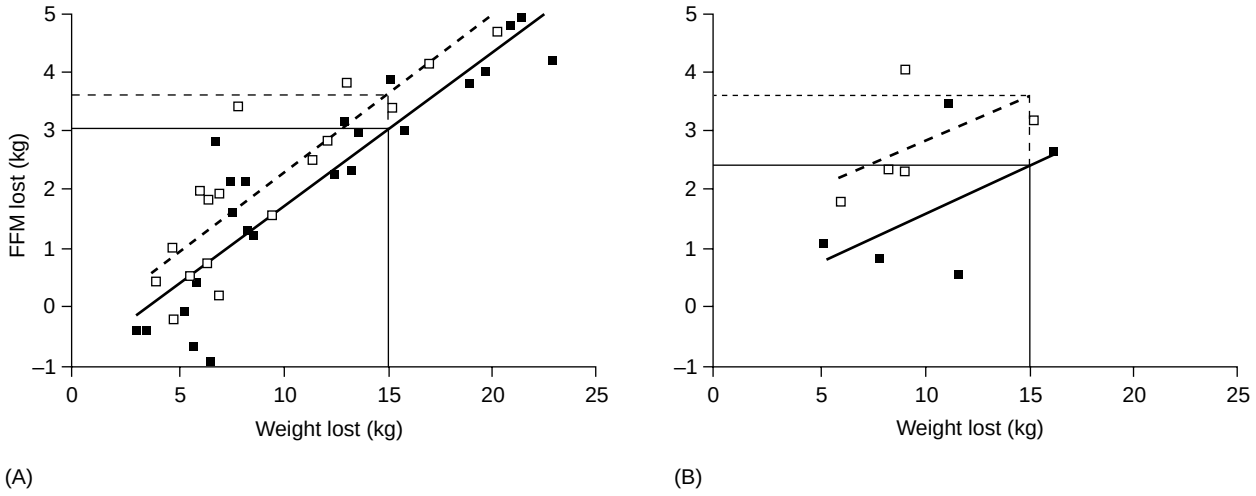


Figure 4 Relationship of total weight loss to fat-free mass loss in women (A) and men (B) undertaking a diet with exercise (solid squares, solid line) or without exercise (open squares, broken line). Data from 21 randomized controlled studies, collated by Garrow JS, Summerbell CD (1995) Meta-analysis: effect of exercise, with or without dieting, on the body composition of overweight subjects. *Eur J Clin Nutr* **49**: 1–10.

the groups that had included exercise were maintaining losses of 2.2–2.7 kg while those on diet alone had only managed to maintain a 0.9 kg loss. Similar findings have been seen in dieters from commercial slimming groups.

Behavior Modification

Behavioral modification is seen as the cornerstone of any treatment program that seeks to empower and enable obese individuals to make voluntary changes in life style. Any therapy relies to a greater or lesser extent on such a principle. For example, treating hypertension should be an apparently straight-forward clinical management issue, but patient non-compliance with medication is common. The skilled clinician will often include the principles of behavior therapy in consultations to help the patient understand and put into practice the new 'life style' of taking their drugs regularly. The approach in obesity is firmly based on theories of learning, and relies on the concept that behaviors associated with weight gain and weight maintenance are to a significant extent learned and subject to modification. Such a behavioral theory is not undermined by the knowledge that genetic and environmental factors are also important in determining the predisposition to obesity. A prerequisite for successful behavior change is that the individual must be 'ready' and motivated to change. It is common practice to assess this aspect of 'readiness' prior to enrolling patients in behavioral programs, and a number of standardized and validated questionnaires are available. Because behavioral programs are intensive of therapist time, patients are often treated in groups, often with manuals, which allow for individual study. These groups are usually 'closed;' that is, a small group of patients start the program simultaneously and go through it together. This contrasts with many commercial diet groups, in which patients are free to join or leave at any time. More recently computeraided interventions have been developed, but as yet results are not promising.

The components of a typical behavior modification programme are shown in **Table 3**. For each area, patients need to learn the underlying concepts, recognize the importance to their own situation, and practise strategies to change their behavior. The results of a large number of programs have been published, either as audit outcome or as comparative trials. Programs vary in duration from 12 weeks to 52 weeks (there has been a trend since the 1970s to lengthen treatment time). Drop-out rates are clearly biased by selection procedures, but are typically 10–20%. Weight loss during treatment is typically 10–15% of initial weight, at a rate of about 0.5 kg per week. In order to strengthen the impact of the intervention on weight loss, many programs have included a period of time on very low-energy or liquid-based diets. This approach of a complete withdrawal for a time from established (abnormal) eating habits can be usefully integrated into a model of behavior change, and is well and positively tolerated by obese patients. Although data suggest that the greater weight loss induced by very low-energy diets has little effect on the long-term results in terms of weight loss maintenance, these diets do represent a practical and pragmatic initial approach to treating patients in a group, especially when many individuals within such a group may resist

Table 3 The components of a typical behavior modification program

Domain	*Intervention strategy*	*Example*
Self-monitoring	Food intake diaries	Food diaries
	Exercise and activity	Activity logs
	Weight change	Regular weighing and recording on weight charts
Nutrition	Nutrition knowledge	Energy, macronutrients, understanding food labeling
	Healthy eating	Low fat, high complex carbohydrate, adequate fruit and vegetable intake
Exercise and activity	Increasing daily energy-using activities	Using stairs not escalators
	Decreasing sedentariness	Decrease television viewing
	Formal exercise	Group workouts at sports centers
Goal setting	Realistic rates of weight loss	Aim for 0.5–1.0 kg weekly
	Realistic target weight	10% weight loss as initial goal
	Weight maintenance	
Problem solving	Identifying conflicts with aims	Holidays, parties, restaurant meals
	Interpersonal conflicts	The unhelpful relative or friend
	Stimulus control and negative feelings	Hunger on returning home from work
Cognitive change	Modifying thoughts about and responses to food cues	Good and bad foods; food as a reward; coping with 'highly desirable' foods
	Self-esteem and assertiveness training	Recognizing and exerting choice
	Preventing relapse	Acceptance of occasional small weight gains

the idea that they are able to lose weight on conventional reduced-energy diets.

Research is now directed towards finding ways of improving the results of such programs in terms of long-term weight loss maintenance. An increased focus on weight-maintaining behavior rather than weight loss, a stronger emphasis on increasing activity and exercise, and better relapse strategies are being evaluated. Targeting the needs of specific subgroups, for example those with binge eating disorders or dysfunctional family circumstances, is another way in which behavioral therapy may be improved.

See also: **Eating Disorders**: Bulimia Nervosa. **Energy Expenditure**: Indirect Calorimetry. **Exercise**: Diet and Exercise. **Obesity**: Definition, Etiology and Assessment; Fat Distribution; Childhood Obesity; Prevention; Treatment. **Starvation and Fasting**.

Further Reading

Activity and Health Research, Allied Dunbar National Fitness Survey (1992) *A Report on Activity Patterns and Fitness Levels: Main Findings*. London: Sports Council and Health Education Authority.

Brownell K (1997) *The Learn Programme for Weight Control*, 7th edn. American Health Publishing: Dallas Texas.

Fentem P and Walker A (1995) Setting targets for England: challenging, measurable and achievable. In: Killoran A (ed.) *Moving On: International Perspectives on Promoting Physical Activity*. London: Health Education Authority.

Finer N (ed.) (1997) Obesity: a series of expert reviews. *British Medical Bulletin* **53**(2): 229–450.

Frost G, Masters K, King C *et al.* (1991) A new method of energy prescription to improve weight loss. *Journal of Human Nutrition and Dietetics* **4**: 369–373.

Thomas PR (ed.) (1995) *Weighing the Options. Criteria for Evaluating Weight-management Programs*. Washington: National Academy Press.

Tremblay A, Bouchard C, and Despres JP (eds.) (1995) Proceedings of a satellite symposium of the 7th ICO on Exercise and Obesity: Morphological, metabolic and clinical implications. *International Journal of Obesity* (supplement 4): S1–S129.

Scottish Intercollegiate Guideline Network (1996) *Obesity in Scotland*, Integrating prevention with weight management. A national clinical guideline recommended for use in Scotland. Edinburgh: SIGN.

Wing RR (1997) Behavioural approaches to the treatment of obesity. In: Bray GA, Bouchard C, and James WPT (eds.) *Handbook of Obesity*, pp. 855–873. New York: Marcel Dekker.

Weight Maintenance

H A Raynor and R R Wing, Brown University, Providence, RI, USA

Most of the developed world is in the midst of an obesity epidemic. In the United States, more than 60% of adults are overweight and obese. The negative impact of obesity on health outcomes and health care costs has heightened awareness of the importance of achieving successful weight loss maintenance. This article reviews information obtained from observational and experimental studies on successful weight loss maintenance. First, the definition

and prevalence of successful weight loss maintenance are presented, and why weight loss maintenance may be difficult is discussed. Next, factors identified in research examining weight loss maintenance, obtained from the National Weight Control Registry (NWCR) in the United States and from randomized controlled trials examining long-term weight loss and weight loss maintenance, are described. Finally, general recommendations for achieving successful weight loss maintenance are provided.

Definition of Successful Weight Loss Maintenance

There is no universally accepted definition of successful weight loss maintenance. We recommend the following definition: an intentional weight loss of $\geq$10% of initial body weight that is maintained for >1 year. Several points in this definition should be noted. First, the definition requires that the weight loss be intentional; this is important since several studies suggest that unintentional weight loss occurs frequently and may likely have different causes and consequences than intentional weight loss. Second, criteria for both magnitude and duration of the weight loss are set. The criterion of 10% weight loss is recommended since weight losses of this magnitude have been shown to have positive health consequence. The 1-year duration is selected in keeping with the US Institute of Medicine definition. However, examining both 1-year and 5-year maintenance of weight loss is suggested.

Data on Prevalence of Long-Term Maintenance of Weight Loss

Most information about long-term maintenance of weight loss comes from obesity treatment studies. In such studies, overweight individuals who receive a lifestyle intervention achieve a weight loss of approximately 7–10 kg at 6 months. Typically, the maximum weight loss occurs at 6 months, followed by weight maintenance for the next 6 months and then gradual weight regain. This pattern of weight change is illustrated in the Diabetes Prevention Program (DPP). DPP is a multicenter clinical trial of the effects of lifestyle intervention, metformin, and placebo on the development of diabetes in more than 3000 individuals with impaired glucose tolerance. **Figure 1** shows that participants in lifestyle intervention achieved an average weight loss of 7 kg (7% of initial body weight) at 6 months, maintained this weight loss through 12 months, and then gradually regained weight.

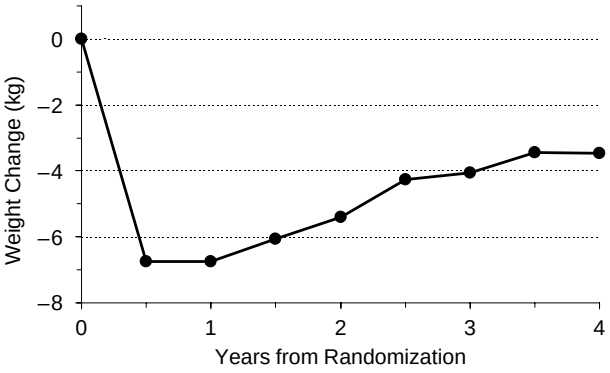


Figure 1 Average weight loss achieved in the lifestyle intervention of the Diabetes Prevention Program.

Few lifestyle treatment programs provide follow-up beyond 1 or 2 years. One study reported that at 5-year follow-up 13% of participants remained >5 kg below their baseline weight. Likewise, 22% of participants were >5 kg below baseline weight at 5 years in another lifestyle intervention.

These studies may underestimate the prevalence of successful long-term weight loss because they are based on a single episode of weight loss and likely involve a selected sample who find weight loss most problematic. For example, a random digit dialing telephone survey of 500 adults in the United States found that 228 of these adults reported being overweight (body mass index >27) at their heaviest weight. Sixty-nine of the 228 individuals were currently at least 10% below their highest body weight and had maintained at least a weight loss of $\geq$10% for at least 1 year (mean weight loss was 19.1 kg, maintained for 7 years). When successful weight losers were further restricted to those who reported intentional weight loss of >10% maintained for >1 year, 47 (20.6%) of the 228 overweight participants met this criterion. Thus, 20% of overweight individuals appear to meet the criteria specified for "success."

Why is Weight Loss Maintenance Difficult?

Long-term weight loss maintenance may be difficult due to a combination of physiological, environmental, and psychological factors. Proposed physiological factors contributing to weight regain include reduced resting metabolic rate and insulin and leptin resistance. However, investigations examining metabolic factors in individuals who have lost weight have not been able to consistently document changes in physiological characteristics that would explain the tendency for weight regain to occur. Environmental

factors may affect energy balance by promoting increased intake and/or reduced energy expenditure, causing weight regain to occur. The strong impact that environmental cues have on energy intake and expenditure have recently been acknowledged, as Americans are now described to be living in an "obesogenic environment." This environment provides greater exposure to a variety of highly palatable, energy-dense foods and expanding portion sizes that potentially increase intake. Additionally, the environment is filled with products of convenience and efficiency that promote decreased energy expenditure. The psychological self-control needed to override these environmental cues may be difficult for most people to sustain over long periods. Finally, during obesity treatment, weight loss can provide reinforcement for adherence to eating and activity prescriptions that promote weight loss. During weight loss maintenance, weight loss no longer occurs; therefore, there is less reinforcement of healthy eating and activity behaviors, causing motivation for sustaining these behaviors to decrease.

Research on Successful Weight Loss Maintenance

Although many individuals have difficulty sustaining weight loss, some are able to maintain a substantial amount of weight loss over a long period of time. To increase the prevalence of successful weight loss maintenance, two types of research investigating weight loss maintenance have been conducted: observational and experimental. In observational research, successful weight loss maintainers are identified and information about how they maintain their weight loss is collected. With experimental research, variables that are believed to affect weight status are manipulated and weight change over time is measured.

The National Weight Control Registry

The largest observational study of successful weight losers is the NWCR in the United States. The NWCR is a registry of individuals who have lost at least 13.6 kg and kept it off at least 1 year. On average, these participants have lost more than 27.3 kg and kept it off more than 6 years. Information that registry members have provided has aided in learning about the weight loss maintenance process.

Registry participants are recruited through newspaper and magazine articles and thus are a self-selected population. The registry members are primarily female (80%) and Caucasian (97%). Many have a strong genetic predisposition to obesity, with 73% having one or both parents with obesity and 46% having been overweight as a child.

Participants in the registry are asked to indicate how they lost their weight in this successful effort. Approximately half say they lost the weight entirely on their own, whereas the other half reported receiving some type of assistance from a physician, dietician, or commercial program. The combination of diet plus exercise was used by 89%, with the most common dietary strategy being restricting intake of certain types of foods.

Although there is marked heterogeneity in the approaches used for weight loss, there appear to be some common themes for weight loss maintenance. The first common element is consumption of a low-calorie, low-fat diet. Registry participants are consuming an average of 1381 kcal/day, with 24% of calories from fat, 19% from protein, and 56% from carbohydrates. Very few (less than 1%) report consuming a low-carbohydrate diet (less than 24% of calories from carbohydrates).

Registry members report eating an average of 4.87 meals or snacks per day. More than three-fourths of the sample report eating breakfast every day of the week, whereas less than 5% report never eating breakfast. Consuming breakfast may be an important behavioral characteristic of successful weight losers.

These long-term changes in diet in registry members are accompanied by long-term changes in physical activity. Women in the registry report 2545 kcal/week of physical activity and men report 3293 kcals/week. This is equivalent to 1 h per day of brisk activity. Approximately half of registry members engage in walking plus another form of physical activity, including cycling, weight lifting, aerobics, running, or stair climbing. Only 9% of registry members maintain weight loss without physical activity.

The final characteristic of registry members is that they weigh themselves regularly. More than 44% weigh themselves daily and 31% weigh themselves at least once a week. Frequent monitoring of weight may allow these individuals to quickly catch small weight gains and institute corrective actions. See **Table 1** for a summary of the strategies that registry

Table 1 Strategies used by successful weight loss maintainers in the National Weight Control Registry

Area	*Strategy*
Diet	Consuming a low-calorie, low-fat diet
	Consuming breakfast regularly
Physical activity	Engaging in 1 h of moderate–intense physical activity per day
Behavioral tools	Self-monitoring of weight

members have reported as being helpful for successful weight loss maintenance.

Experimental Studies Examining Weight Loss Maintenance

Our understanding of weight loss maintenance comes not only from the study of successful weight losers but also from randomized clinical trials evaluating specific treatment components. These trials are stronger scientifically because participants are randomly assigned to treatment conditions and all aspects of the intervention are kept constant except the factor under investigation. However, these studies are limited by their short duration (typically 1 or 2 years) and their relatively small sample size (usually 100–200 participants).

In experimental studies of weight loss maintenance, the primary focus is usually on overall weight loss (from baseline to the end of the study), usually defined as long-term weight loss, rather than on maintenance of weight loss from the end of the initial treatment (typically 6 months) to study end. Overall weight loss is selected as the variable of interest because it is most strongly associated with health impact. In addition, focusing on weight change from end of treatment to follow-up would make those who lost small amounts of weight but maintained that weight loss in full appear to be more successful than those who lost large amounts of weight and regained some.

Experimental research has focused on three main ways to increase long-term weight loss (**Figure 2**). The first is to increase the rate of initial weight loss so that a greater amount of weight loss occurs during the first 6 months of treatment. A second focus is to improve maintenance of weight loss achieved after the first 6 months of treatment. Finally, combining both of these approaches is considered the ideal approach.

Several strategies have been tested in experimental studies of weight loss maintenance, including focusing on energy balance, in which changes in diet and/or physical activity are used to create larger energy deficits that produce greater weight loss, or focusing on intensifying behavioral components of interventions so that skills necessary for sustaining weight loss can be maintained over a longer period. These strategies have been implemented during the initial weight loss treatment phase and the weight loss maintenance phase.

Energy Balance

In order to produce weight loss, it is necessary to modify energy balance by eating less and/or exercising more. A substantial body of research suggests that the combination of diet plus exercise is most effective for long-term maintenance of weight loss.

Diet

Within the context of diet, weight loss researchers have focused primarily on the level of caloric restriction and the degree of structure in the diet. Typically, behavioral weight loss programs recommend a low-calorie, low-fat diet. Participants are instructed to eat 1000–1500 kcals/day (low-calorie diet), depending on their initial body weight, and to reduce dietary fat to 20–25% of calories. There are no specific foods that are required or prohibited, but consumption of complex carbohydrates and guidelines based on the Food Guide Pyramid are stressed. Participants are instructed to self-monitor the calories and fat grams in all foods they consume. Self-monitoring is recommended daily for the first 6 months, and 1 week per month thereafter. Adherence to self-monitoring has been shown to be one of the best predictors of maintenance of weight loss.

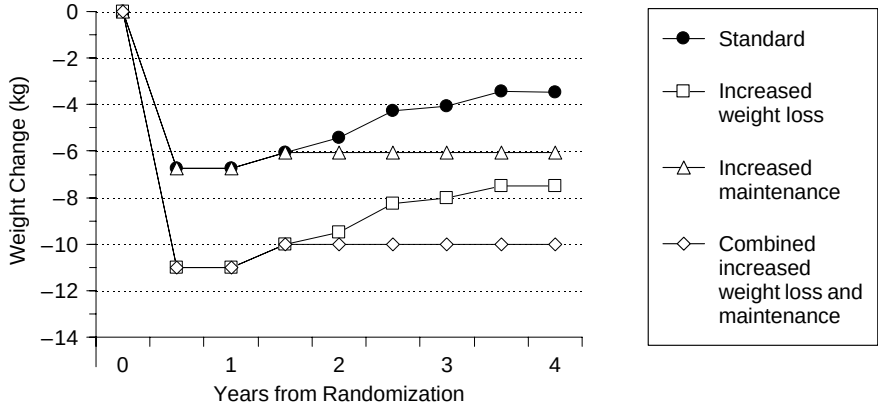


Figure 2 Ways to increase average long-term weight loss maintenance achieved in experimental research.

Very low-calorie diets Very low-calorie diets (VLCDs) are dietary regimens that provide approximately 400–600 kcal per day usually as a liquid formula. VLCDs have been shown to produce excellent initial weight losses (−20 kg at 12 weeks); this effect is due in part to the degree of caloric restriction and in part to decreased dietary variety and the use of portion-controlled foods in these regimens. Given the large initial weight loss produced by VLCDs, it was hoped that combining these diets with behavioral approaches would maximize long-term weight loss. Although VLCDs improve initial weight loss, they do not appear to produce better long-term weight loss than low calorie diets (LCDs). Difficulty with weight maintenance in programs with a VLCD appears to occur during the transition from the VLCD to a diet composed of conventional foods.

Since VLCDs have been very effective at decreasing intake, the effect of intermittent use of VLCDs (initiating weight loss with a VLCD, transitioning to conventional foods, and then returning to a VLCD) on long-term weight loss has also been investigated, but results have been less than promising. During a 50-week behavioral obesity intervention, in which a VLCD was prescribed for weeks 1 through 12 and 24 through 36, weight loss at week 50 was not significantly different between an intermittent VLCD and a LCD.

VLCDs with caloric levels between 400 and 800 kcals/day have been compared to examine if greater caloric restriction produces better weight loss. One study compared two outpatient groups with different caloric prescriptions, 420 and 800 kcals/day, and found that weight loss was not significantly different between the groups. This suggests that VLCDs may produce greater initial weight loss not only by restricting calories but also by increasing the structure of the diet.

Structured low-calorie diets Several studies have investigated different ways to increase the structure of LCDs. Structure in the diet can be strengthened by decreasing variety and/or food choices and by controlling portion sizes consumed. A study examined whether providing food to participants, which controls portion size and decreases food choice, improved long-term weight loss during a standard behavioral intervention using an LCD. Participants were provided all of the food they should eat for five breakfasts and dinners each week for 18 months. Participants receiving the food provisions had greater weight loss at 6 months (−10.1 vs. −7.7 kg), 12 months (−9.1 vs. −4.5 kg), and 18 months (−6.4 vs. −4.1 kg) than those participants receiving a standard intervention, even though both groups had identical calorie goals (1000–1500 kcals/day). However, even with the greater dietary structure, participants still regained weight during the maintenance phase.

Structure in the diet, by decreasing food choices, can also be increased by providing structured meal plans and detailed grocery lists. One investigation that provided meal plans and grocery lists along with a standard intervention showed greater weight loss than the standard intervention alone. The weight losses achieved with the meal plans were similar to those achieved with food provisions.

Using portion-controlled foods available in the marketplace, such as frozen entrees and meal replacement products such as Slim-FastR, also increases dietary structure. When an LCD composed of conventional foods was compared to an LCD using two Slim-FastR meal replacements, two Slim-FastR snack bars, and a healthy dinner, the diet using the Slim-FastR portion-controlled foods produced better weight loss at 3 months (−7.1 vs. −1.3 kg). For the next 24 months, both groups were instructed to consume one Slim-FastR meal replacement and snack bar per day. At 27 months, the Slim-FastR group still had better weight loss (−10.4 vs. −7.7 kg), and the greater weight loss was maintained at 4 years (−9.5 vs. −4.1 kg) in those participants available for follow-up.

Food provisions have also been used during a maintenance intervention as a rescue strategy. However, used in this manner, food provisions were not helpful in improving weight loss maintenance compared to a maintenance program without food provisions.

Consequently, increasing dietary structure by decreasing variety and food choices and/or using portion-controlled foods appears to improve long-term weight loss. These changes in the diet may increase adherence to an LCD, thereby producing greater weight loss, especially during the first 6 months of obesity treatment.

Physical Activity

Correlational studies suggest that physical activity is the single best predictor of long-term maintained weight loss. Physical activity is important because it increases energy expenditure, but it may also reduce hunger and improve mood. Physical activity is usually prescribed at a level of 1000 kcals/week or 150 minutes of moderate–intense activity per week; however, long-term weight loss has been found to be greater in participants who are active 200 minutes or more per week compared to those who are active 150 minutes or less per week during weight loss

interventions. Similarly, the NWCR data discussed earlier show that successful weight loss maintainers are very active, reporting more than 2500 kcals/week of activity. This suggests that an exercise prescription of at least 200 minutes per week may be needed to improve long-term weight loss. One study compared the effect of a standard activity recommendation (1000 kcals/week) versus a higher physical activity prescription (2500 kcals/week, equivalent to walking 75 minutes 5 days per week) in a standard behavioral weight loss intervention. The group with the higher physical activity prescription had greater long-term weight loss at 12 months (−8.5 vs. −6.1 kg) and 18 months (−6.7 vs. −4.1 kg). However, even with the higher exercise prescriptions, participants still regained weight during the maintenance phase.

Strategies for improving maintenance of physical activity One of the most challenging problems with physical activity in weight control programs is adherence to activity prescriptions. One way to increase activity adherence is to prescribe activity in short bouts (40 minutes/day in four 10-minute bouts) rather than in long bouts (40 minutes/day in one bout). Accumulating activity during the day may make it easier to achieve physical activity goals. Although short bouts of exercise improved initial adoption of exercise, they did not appear to increase physical activity adherence or weight loss at 12 and 18 months. Participants have also been provided with personal trainers, supervised walks, home exercise equipment, and financial incentives to improve physical activity adherence. Although personal trainers and financial incentives did increase attendance at exercise sessions, neither improved total exercise achieved or weight loss at 18 months. Providing participants with home exercise equipment has been shown to improve both adherence to physical activity and weight loss. Participants given home exercise equipment and encouraged to exercise in multiple short bouts had greater long-term weight loss (18 months) than those prescribed short bouts without equipment (−7.4 vs. −3.7 kg). The results of this study suggest that home exercise equipment and other approaches that make exercise more convenient may facilitate long-term adherence and consequently weight loss maintenance.

Another approach evaluated for increasing long-term physical activity adherence is focusing specifically on physical activity during a weight maintenance program. A 6-month exercise-focused maintenance program, which included supervised group walking sessions, individual and group incentives for exercise completion, and relapse prevention training aimed at maintaining physical activity, was compared to a weight-focused maintenance program, which focused on group problem solving of weight-related problems. No differences were found between the groups in terms of exercise participation and energy expenditure at the end of the maintenance program or at 6-month follow-up. The weight-focused group had better maintenance of weight losses over 6-month follow-up (3.1 vs. 5.2 kg). These results and findings from other studies suggest that placing too much emphasis on activity may detract from the dietary component of weight loss interventions and consequently decrease long-term weight loss success.

Taken together, it appears that physical activity at higher prescriptions, $\geq$200 minutes per week, aids long-term weight loss. However, adherence to this amount of activity may be difficult, and providing home exercise equipment with a physical activity prescription of multiple short bouts may be a promising approach.

Intensifying the Behavioral Component

Whereas the research described previously focused on ways to enhance negative energy balance through modifications in diet and physical activity, other investigations have examined ways to intensify behavioral components in weight loss or weight loss maintenance interventions. These strategies include extending professional contact, increasing social support, enhancing motivation using incentives, providing skills training, and combining some of these strategies into multicomponent maintenance programs.

Extending professional contact As noted previously, the maximum weight loss in a behavioral weight loss intervention is typically attained at 6 months, which also represents the end of the weekly phase of therapy and the start of the less intense maintenance phase. Weight regain is commonly assumed to be due to a failure to continue practicing effective behavioral techniques when treatment transitions. One way to sustain behavioral strategies is to lengthen treatment or to continue to provide some form of professional contact during the maintenance phase.

Lengthening the initial phase of treatment has been shown to increase initial weight loss. For example, when behavioral treatments of identical content, differing only in length of treatment (20 vs. 40 weeks), are compared, the two programs produce similar weight losses at 20 weeks (−9.5 kg), but the extended treatment produces greater weight

loss at 40 weeks (−13.6 vs. −6.4 kg). Based on this, several investigators tried to develop year-long programs with weekly meetings throughout. Weight losses at the end of the year were 10–14 kg, but attendance became quite poor toward the end of the program and the cost-effectiveness of such long-term weekly programs was questioned. Thus, investigators have considered how best to provide contact after the end of the 6-month weekly program.

One of the first methods employed to extend professional contact during the maintenance phase was the use of booster sessions. Booster sessions take place on a fairly infrequent basis after treatment, with an increasing interval of time between sessions to fade professional contact (e.g., meeting at months 1, 3, 6, and 12). Booster contacts have yielded inconsistent results. This finding and the fact that better maintenance of weight loss occurs when participants continue to be seen biweekly suggest that patients need a fairly high level of contact during maintenance. Studies using biweekly maintenance programs have found better weight loss maintenance at 6-month (120% (continued weight loss) vs. 83% (weight regain)) and 18-month (87 vs. 33%) follow-ups compared to a control intervention receiving no maintenance component. The specific content of the maintenance sessions appears less critical than the frequency of ongoing contact, the regular weighing of patients, and the emphasis on continued self-monitoring. Although face-to-face contact appears most effective, it may also be possible to provide extended contact by phone, mail, or e-mail.

Social support Another approach to provide long-term support is to involve friends and family of participants in the treatment program. Spouses have been included in treatment, but the effects have been mixed. A meta-analysis of the spouse support literature showed a small positive effect through 2 or 3 months of follow-up. One study examined the effectiveness of natural social support (participants were recruited with three other friends and family members who were all losing weight in the same program) and experimentally created social support (through the use of intragroup activities and intergroup competitions) during a standard behavioral weight loss intervention. Sixty-six percent of participants recruited with a friend and given the social support intervention retained their weight loss in full from month 4 to month 10 compared to 24% of individuals recruited alone and given the standard behavioral intervention without any social support intervention.

Peer support can also be developed among group members in the same weight loss intervention. Support from other members of the group may explain the finding that group treatment tends to be more successful than individual therapy. One investigation conducted a 7-month weight loss maintenance program involving peer support following a behavioral obesity treatment. Participants formed peer self-help groups, which met biweekly and used group problem-solving skills to handle difficulties with weight loss. The peer support group maintained a greater weight loss at 1-year follow-up than the control group that received no maintenance program (−6.5 vs. −3.1 kg). These studies suggest that social support is helpful in long-term weight loss and weight loss maintenance.

Incentives for weight loss and weight loss maintenance Behavioral interventions used in obesity treatments focus on changing antecedents and consequences of behaviors that influence energy balance. Behaviors that produce negative energy balance, and consequential weight loss, can be reinforced, thereby increasing the likelihood that these behaviors will continue. The effects of contracting for healthy behaviors and weight loss have been inconsistent. Contracting with participants to attend supervised exercise sessions doubled the number of walks attended but had no effect on overall activity level or weight loss. Likewise, providing financial incentives (a substantial cash payment was given each week to participants when the weekly weight loss goal was met) during a standard behavioral weight loss program had no effect on long-term weight loss. Procedures in which patients deposit money at the start of the program and then earn portions back each week for meeting specific weight loss goals or self-reported caloric intake goals appear more effective. In one study showing positive results, the financial deposits were returned based on the average weight loss of the whole group.

Skills training Another approach to improve maintenance of weight loss is to provide participants with training in specific skills. These specific skills can provide participants with the ability to cope with high-risk situations that increase the likelihood of a relapse of problematic eating and activity behaviors.

Two types of skill-based maintenance programs, provided after the completion of a standard behavioral weight loss intervention, have been investigated. One approach focuses on relapse prevention, in which participants are taught a

variety of methods to anticipate and cope with the problem of relapse in weight loss maintenance. In the second approach, participants are taught to use the steps of problem-solving to manage difficulties during weight loss maintenance. After 1 year of a maintenance program, participants in the problem-solving intervention had better weight loss than those who had received no maintenance program following treatment (−10.8 vs. −4.1 kg). There was no difference in weight loss between participants in the relapse prevention program and those participants who received no maintenance program (−5.9 vs. −4.1 kg). These results suggest that strengthening problem-solving skills during weight maintenance improves weight loss maintenance.

Multicomponent programs Since long-term weight loss maintenance is believed to be difficult for many reasons, an approach that combines several different strategies used after initial weight loss treatment may produce better weight loss maintenance. These multicomponent maintenance programs have used different combinations of extending professional contact, increasing peer support, providing incentives, and increasing physical activity. All programs using a multicomponent maintenance program show better weight loss maintenance at 18-month follow-up compared to no maintenance program, but the multicomponent programs do not produce greater weight loss maintenance than simple programs that just extend professional contact.

Conclusion

Successful weight loss maintenance can be challenging. However, information obtained from the NWCR, a registry of long-term successful weight loss maintainers, and from experimental studies that have examined different approaches to improve weight loss maintenance indicates that there are several strategies that may assist with long-term weight loss maintenance. These strategies are described in **Table 2**. Most notably, it is vital to recognize that for successful weight loss maintenance, individuals must continue to consume less calories and to engage in a greater level of physical activity than they did prior to weight loss; otherwise, they will return to a state of positive energy balance, in which weight (re)gain occurs.

Information from the NWCR suggests that to sustain weight loss, a low-calorie, low-fat diet is needed. Experimental studies show that at least during the weight loss phase, a structured low-calorie diet improves long-term weight loss. The structure of the diet can be increased by using food provisions, structured meal plans, and/or meal replacements. In addition, maintaining structure in the diet may help sustain a lower calorie diet, thus helping with weight maintenance.

Being physically active, for at least 200 minutes per week, also seems to aid in successful weight loss maintenance. Both experimental studies and self-reported activity information from registry participants support this recommendation. Ways to assist with achieving this level of activity include having

Table 2 Helpful strategies for successful weight loss maintenance

Area	*Strategy*
Diet	Consume a low-calorie (~1500 calories per day) diet
	Consume a low-fat (<30% calories from fat) diet
	To assist with consuming a low-calorie diet, increase structure in the diet using
	Food provision
	Meal plans
	Meal replacements
Physical activity	Be active at a moderate–intense level for at least 200 minutes per week
	To assist with achieving at least 200 minutes of activity per week,
	Have exercise equipment available at home
	Allow activity to accumulate during the day (in at least 10-minute bouts)
Self-monitoring	Record body weight at least once per week and not more than once per day
Professional contact	Extend professional contact beyond initial treatment
	Should occur on at least a biweekly basis
	Can occur in person or by phone, possibly Internet
Social support	Include support from other individuals also working toward weight loss or weight loss maintenance
	Can be from family, friends, or other members of treatment group
Skills training	Include problem-solving focused on weight loss maintenance during maintenance programs

exercise equipment available at home and accumulating activity in short (at least 10 minutes) bouts during the day.

In formalized behavioral programs, other strategies that may improve weight loss maintenance include extending professional contact beyond the initial 6 months of weekly treatments. To be most helpful, this contact should occur on a biweekly basis. Extending contact using technologies such as the Internet is being investigated and shows promise. Including peer support during both the weight loss and the weight maintenance phases of treatment can also improve long-term weight loss. In addition, during the maintenance phase, the focus of the intervention should be problem-solving difficulties related to maintaining weight loss. To date, studies suggest that effective long-term treatment of obesity may require several different strategies, implemented over an extended period of time. Although these tactics improve long-term weight loss, many patients still regain weight after the initial 6-month treatment. Consequently, further research on improving successful weight loss maintenance is needed.

See also: **Appetite**: Psychobiological and Behavioral Aspects. **Energy**: Balance. **Exercise**: Beneficial Effects; Diet and Exercise. **Obesity**: Treatment. **Weight Management**: Approaches.

Further Reading

Jeffery RW, Drewnowski A, Epstein LH *et al.* (2000) Long-term maintenance of weight loss: Current status. *Health Psychology* **19**: 5–16.

Jeffery RW, Wing RR, Sherwood NE, and Tate DF (2003) Physical activity and weight loss: Does prescribing higher physical activity goals improve outcome? *American Journal of Clinical Nutrition* **78**: 684–689.

Jeffery RW, Wing RR, Thorson C *et al.* (1993) Strengthening behavioral interventions for weight loss: A randomized trail of food provision and monetary incentives. *Journal of Consulting and Clinical Psychology* **61**: 1038–1045.

Perri MG (2002) Improving maintenance in behavioral treatment. In: Fairburn CG and Brownell KD (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*, 2nd edn., pp. 593–598. New York: Guilford Press.

Perri MG and Corsica JA (2002) Improving the maintenance of weight lost in behavioral treatment of obesity. In: Wadden TA and Stunkard AJ (eds.) *Handbook of Obesity Treatment*, pp. 357–370. New York: Guilford Press.

Peri MG, Nezu AM, McKelvey WF *et al.* (2001) Relapse prevention training and problem-solving therapy in the long-term management of obesity. *Journal of Consulting and Clinical Psychology* **69**: 722–726.

Renjilian DA, Perri MG, Nezu AM *et al.* (2001) Individual versus group therapy for obesity: Effects of matching participants to their treatment preferences. *Journal of Consulting and Clinical Psychology* **69**: 717–721.

Wilson GT and Brownell KD (2002) Behavioral treatment for obesity. In: Fairburn CG and Brownell KD (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*, 2nd edn., pp. 524–528. New York: Guilford Press.

Wing RR (2002) Behavioral weight control. In: Wadden TA and Stunkard AJ (eds.) *Handbook of Obesity Treatment*, pp. 301–316. New York: Guilford Press.

Wing RR and Hill JO (2001) Successful weight loss maintenance. *Annual Review of Nutrition* **21**: 323–341.

Wing RR and Klem M (2002) Characteristics of successful weight maintainers. In: Fairburn CG and Brownell KD (eds.) *Eating Disorders and Obesity: A Comprehensive Handbook*, 2nd edn., pp. 588–592. New York: Guilford Press.

Weight Cycling

L Lissner, Sahlgrenska Academy at Göteborg University, Göteborg, Sweden

> Is it better to have lost and regained than never to have lost at all?
>
> —Weight cycling refrain loosely adapted from Alfred Lord Tennyson

Weight Cycling—A Health Risk?

The term 'weight cycling' is used in the fields of nutrition and obesity research to refer to losses and subsequent regains of body weight typically occurring in association with weight loss dieting. Interest in this phenomenon was initially based on the observation that conventional weight loss programs are often unsuccessful in the long term. Dieting recidivism thus sets the stage for weight cycling, popularly referred to as yo-yo dieting, whereby dieters undergo multiple cycles of weight loss and regain in pursuit of their ideal body weights.

There seems to be little disagreement that weight cycling is one of the most difficult therapeutic aspects in the management of obesity. The majority of obese subjects seeking treatment have previously experienced cycles of weight loss and regain. Examples of weight cycling in two case studies are shown in **Figure 1**. These examples illustrate several phenomena that are important when considering weight cycling in humans: Weight losses and subsequent regains frequently occur in conjunction with intentional weight loss dieting; however, weight changes vary in magnitude and important fluctuations may be missed if weight measurements are taken at infrequent time intervals. Moreover, these two cases

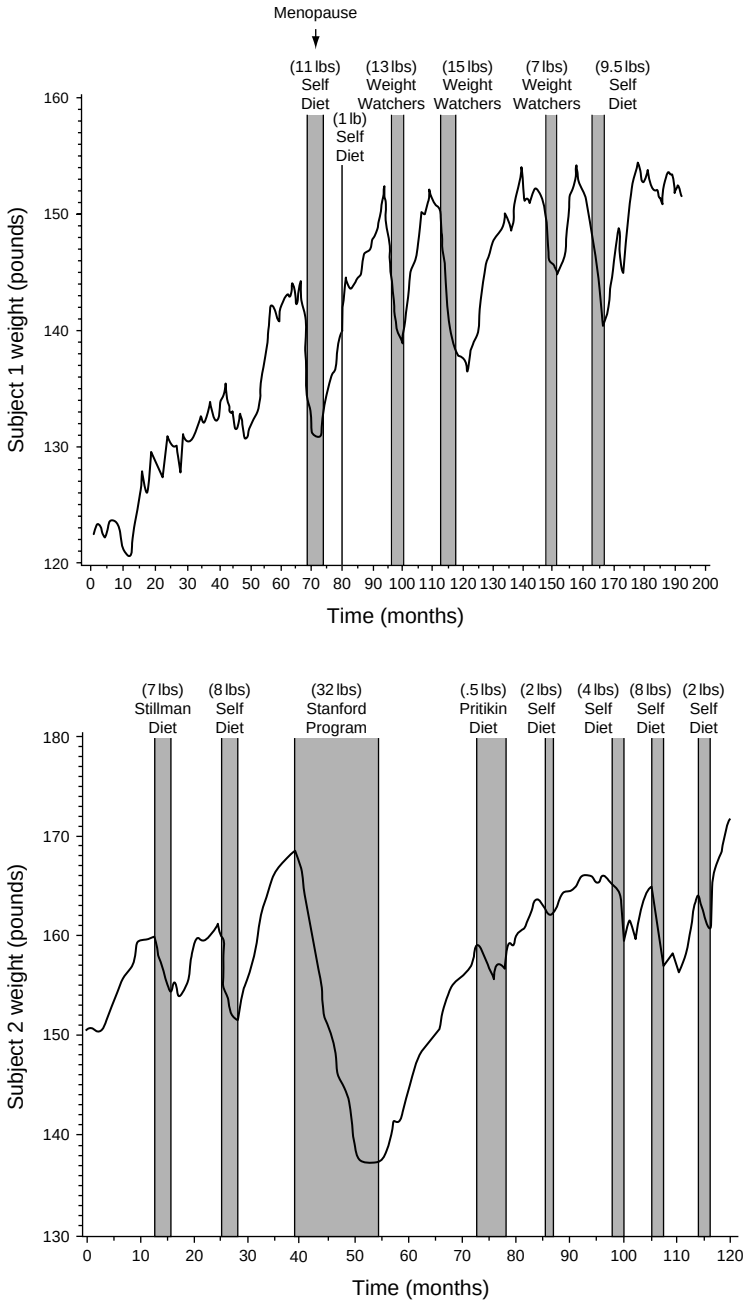


Figure 1 Monthly body weights (average of daily weights) of two female subjects, self-monitored over time. Reproduced with permission from Black DR, Pack DJ, and Hovell MF (1991) A time-series analysis of longitudinal weight changes in two adult women. *International Journal of Obesity* **15**: 623–633.

illustrate the common observation that intentional weight losses are frequently followed by regains in excess of the original body weight, and that true weight stability may be difficult to achieve.

Dieting to control body weight is not confined to overweight individuals but has been widely reported even among men and women who have never been overweight. As described by Jeffery in 1984, 72.5 and 43.7% of surveyed women and men, respectively, had dieted to lose weight; even among women who had never been overweight, the majority reported having been on weight loss diets. Although it is traditionally assumed that adherence to weight reduction diets is beneficial to health, the high rates of dieting and weight loss recidivism, among the nonobese as well as the obese, have naturally created concern regarding potential negative health consequences. However, weight cycling can be intentional or unintentional, and the most recent generation of research on weight cycling has

focused on intentional weight loss as the risk factor of relevance to the weight cycling debate.

In this article, some of the main points of this debate are highlighted. The first epidemiological studies suggesting health implications of weight cycling were reviewed in 1992, at which time the majority of available observational evidence indicated adverse consequences. Subsequently, this topic became a source of considerable controversy, and a number of investigators continued to examine this issue, focusing on possible effects of weight cycling on metabolism, chronic disease, and mental health. This article also provides an overview of knowledge in this area, together with some methodological controversies surrounding existing research on weight cycling.

The Metabolic Hypothesis

Given the fact that most people who lose weight are unable to sustain their losses, a metabolic hypothesis was formulated. It was proposed that if weight loss dieting caused permanent decreases in metabolic rate, the weight would be easily regained and every subsequent weight loss attempt would be more difficult. In the 1990s, the National Task Force on the Prevention and Treatment of Obesity in the United States reviewed the evidence and reported an overall lack of support for the hypothesis that weight cycling promoted obesity, increased body fat, or had permanent effects on metabolism. This report also concluded that the majority of available data in animals did not independently link weight cycling to any parameter of energy balance (food intake, body composition, or energy expenditure). This conclusion was supported by studies in humans, using a variety of designs, that failed to document irreversible effects of weight loss on metabolic rate, body composition, or adipose tissue distribution after regain. Also of interest in this context is the observation that, despite their regains in lost weight, weight cyclers tend to gain less (not more) weight over time than their weight-stable peers.

Weight Cycling and Mortality

Although most studies have not borne out the original idea that weight cycling alters metabolic rate, the possibility that weight fluctuation predicts chronic disease and death has been more difficult to discount. A number of prospective epidemiological studies have shown that an individual's variations in body weight over time, a proxy for weight cycling, can be used as statistical predictors for subsequent mortality and disease end points. Positive associations have been reported between body weight fluctuation and all-cause mortality in several but not all such studies.

These findings are often expressed in terms of relative risk estimates, which represent the mortality rates in a weight-fluctuating group compared to the rates in a weight-stable group. The relative risk estimates for all-cause mortality have been reported to be as high as 2, indicating a double excess mortality risk in the weight fluctuators. Some investigators have reported that significant associations are restricted to certain types of individual (i.e., nonobese or nonsmoking subgroups). It has been reported that the excess risk of mortality in weight-unstable men may be to a large extent explained by preexisting disease. Moreover, an analysis of data from the Multiple Risk Factors Intervention Trial concluded that any adverse effects of weight fluctuation were occurring in relatively normal weight subjects.

As an example, results are shown from a reanalysis of a longitudinal population study of Swedish women started in 1968 when subjects were 38–60 years old. Women were weighed on three occasions during a 12-year observation period, based on which four subcategories of weight change could be created: stable, weight gain, weight loss, and weight cycle. Specifically, these groups were defined as (i) women whose weights remained stable within ± 3 kg; those who (ii) gained or (iii) lost at least 3 kg between the first and last observation; and (iv) those who had lost and then regained at least 3 kg, or gained then lost 3 kg, without an overall change of more than 3 kg from the first to final observation. When these groups were followed for an additional 20 years, it was found that the weight loss group and weight cycling group both had approximately double the risk of mortality compared to the weight stable women (**Figure 2**). It may be argued that the weight losses and subsequent gains may not be voluntary but rather reflect preexisting diseases. However, after exclusion of women with prevalent or incident cancer, diabetes, or cardiovascular disease during the entire period of weight observation, the excess mortality in the fluctuating group was not attenuated, suggesting that morbidity from these conditions was not the underlying cause of the fluctuations or reason for the association. These findings and similar observations in a number of other populations have not been adequately explained by biologically plausible mechanisms.

Weight Cycling and Cardiovascular Disease

Most investigators have considered it more informative to focus on the association between weight

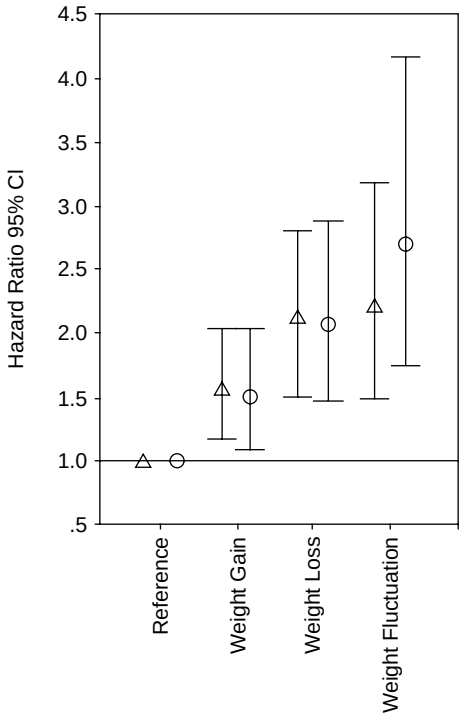


Figure 2 Relative risks with 95% confidence interval for 20-year all-cause mortality in relation to previous 12-year weight changes in the Prospective Population Study of Women in Göteborg. Triangular symbols refer to risks of total mortality, adjusted for age and final body mass index in 800 women. Circular symbols display results after excluding 99 prevalent cases of coronary heart disease, diabetes, or cancer during 12 years of weight observation.

fluctuation and specific diseases and causes of death and have frequently observed positive associations between weight fluctuation and cardiovascular disease end points. However, the results have not always been in agreement; data from the Framingham study showed excess cardiovascular disease among male and female weight fluctuators, whereas in men from the Baltimore Longitudinal Study on Aging there was no association between weight fluctuation and coronary heart disease. Two additional studies in male populations are illustrated in **Figure 3**. Both show a pattern of elevated risk of mortality from cardiovascular or coronary heart disease among weight cycling men and consistently lowest risk associated with stable body weight, based on which it has been concluded that stable weight over time is associated with best health.

It has been hypothesized that some of the observed associations between weight fluctuation and cardiovascular disease may be explained by changes in cardiovascular risk factors occurring during weight gain that are not fully reversible with weight loss. This possibility has been explored using longitudinal data on body weight and risk factors that are concurrently measured on multiple occasions. A systematic review of these studies by the National Task Force on the Prevention and Treatment of Obesity revealed no consistent associations between weight fluctuation and concomitant increases in traditional cardiovascular disease risk factors, such as blood pressure and serum cholesterol.

Hypertension has also been examined as an end point in a number of studies with somewhat mixed results. Based on data from the Nurse's Health Study II, it has been reported that intentional weight cycling is not associated with significant excess risk of development of hypertension. In contrast, a retrospective study indicated that a positive history of weight cycling among obese women as well as the

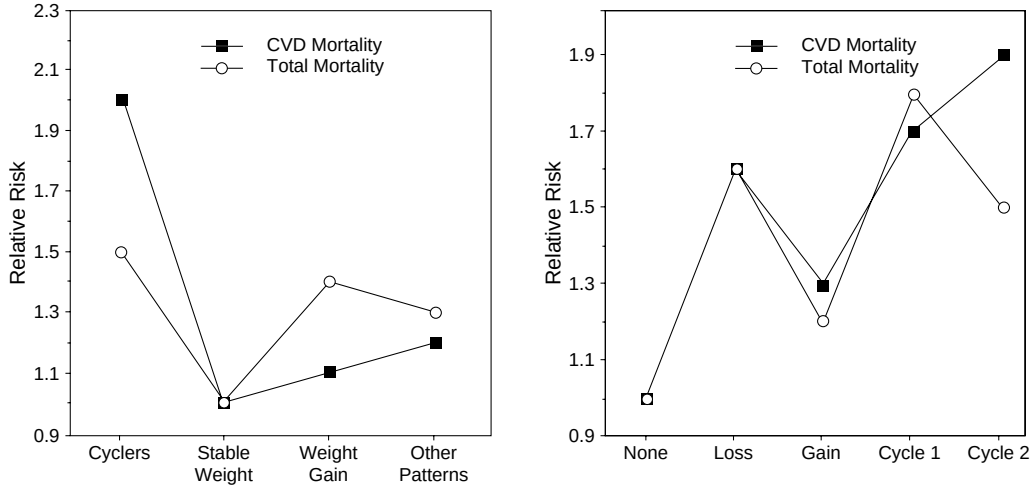


Figure 3 Total and cardiovascular disease (CVD) mortality in male subjects with different weight change patterns. (Left) Data from the Chicago Western Electric Study (Hamm *et al.*, 1989); (right) data from MRFIT (Blair *et al.*, 1991). Cycle 1 refers to cycle ending with weight loss, while cycle 2 ends with a weight gain. Reproduced with permission from Jeffrey RW (1996) Does weight cycling present a health risk? *American Journal of Clinical Nutrition* **63**(supplement): 452S–455S.

sum of weight regained increased the likelihood of being hypertensive. Interestingly, a possible beneficial effect of one weight cycle has been reported: In this blood pressure reduction trial, weight returned to baseline levels after 3 years, whereas blood pressure remained well below control levels. Control subjects in this study experienced a net weight gain, and one interpretation of this study is that a weight cycle did not predispose further hypertension but, rather, seemed to deter further weight gains at control levels, as also observed in the Nurses Health Study II.

Other Health Outcomes: Bone Status, Cancer, and Diabetes

Although a number of the original weight cycling studies also tested associations between weight cycling and cancer, cancer end points have typically not followed the same patterns as cardiovascular disease. It has also been observed that temporary weight cycling (weight loss followed by weight gain) is not associated with increased risk of postmenopausal breast cancer.

It has been proposed that weight cycling may affect bone density and fracture risk, but the evidence here is also mixed. For instance, one study observed that men and women with most weight variability had an increased incidence of hip fracture, whereas another observed no apparent relation between weight cycling and bone density.

Finally, a number of studies have examined associations between weight cycling and diabetes and have yielded little evidence of a relation. According to findings from the Nurses Health Study, no association was found between weight fluctuation and diabetes incidence. In another study, glucose tolerance and weight fluctuations were directly monitored in obese patients, and no deterioration was observed to be directly associated with weight cycling. Interestingly, the Diabetes Prevention Program Research Group found that the diabetes reduction achieved over 4 years with a lifestyle intervention was not diminished with the gradual regaining of more than half of the weight lost. This observation is an indication that a period of weight reduction may exert a net benefit for diabetes, even if weight is subsequently regained.

Psychological Consequences

It has often been assumed that the experience of dieting followed by involuntary regain of the weight lost must take a psychological toll, independent of any medical consequences of weight fluctuation. The possible psychological effects of weight cycling among obese people were the topic of a literature review that reported that weight cycling was not associated with depression or other psychopathology or depressogenic cognitive styles. It was observed, however, that weight cycling was associated with decreased perceptions of health and well-being, decreased eating self-efficacy, and weak increases in binge eating severity. Subsequently, it was concluded that an individual's perception of being a weight cycler may be more related to psychological problems than the actual number of pounds lost and regained over time. In 2000, the National Task Force on the Prevention and Treatment of Obesity concluded that concerns that dieting induces eating disorders or other psychological dysfunction in overweight and obese adults are generally not supported by empirical studies. This is in contrast to the wide belief that dieting is a necessary precursor for subsequent development of eating disorders.

Methodological Issues

Although a number of studies have shown that increased weight fluctuation is associated with subsequent occurrence of adverse health outcomes, a number of methodological problems make interpretation of these findings difficult. For example, weight gain, weight loss, and weight cycling are often considered separately, but it is almost impossible to determine their degree of overlap in observational studies. An individual who is observed to be systematically gaining weight at two points in time may experience a number of unmeasured fluctuations in the interim. The statistical complexities in defining weight cycling were reviewed in 1994 by the National Task Force on the Prevention and Treatment of Obesity.

A frequent critical observation surrounding this type of research is that different studies may be measuring quite different kinds of weight change—voluntary and involuntary. Involuntary changes may reflect serious underlying illness, depression, and other nondieting phenomena. However, intentional dieting may also occur by a variety of dietary methods, some of which are more detrimental to health than others. When focusing exclusively on voluntary losses, it has been suggested that preexisting medical conditions may profoundly influence associations with subsequent mortality risk. The issue of volition may shed light on the problem, but the specific impact of previous and current illness on epidemiological associations between weight cycling and longevity is still not fully understood. Other covarying factors besides intentionality of weight change and underlying illness may be producing artifactual associations in observational studies. These include

aging, smoking and other lifestyle choices, degree and regional pattern of adiposity, and psychological factors. In epidemiological analyses, various types of statistical corrections can be made for potential confounding factors of this type, although adjustment may be incomplete. Finally, biological plausibility is always an issue to consider when reviewing any epidemiological evidence and has been a particular concern when considering the observations of excess risk in association with weight fluctuations.

Conclusion

As summarized in **Table 1**, many but not all epidemiological studies indicate that men and women undergoing body weight fluctuations are at higher risk of mortality and/or cardiovascular disease than individuals experiencing less fluctuation, but the lack of biologically plausible mechanisms to explain these associations has limited the conclusions that can be drawn. Moreover, the evidence for effects on psychological and other health end points is inconclusive. Uncontrolled confounding from disease states resulting in a loss–gain or gain–loss pattern must be considered a plausible explanation for some of these findings, underscoring the difficulty of using observational data to study weight cycling. Moreover, some studies are emerging suggesting that even limited periods of weight reduction may be beneficial in the long term. Experimental data and intervention studies are required for confirmation of the weight cycling hypothesis. The published observational studies of subjects whose weight changes are known to be caused by dieting have been important contributions to the critical discussion of the weight cycling phenomenon, but additional studies are needed in which weight fluctuations are assessed in a more controlled manner.

Regarding the hypothesis that dieting exacerbates the problem of obesity and weight gain, most studies have failed to demonstrate that weight fluctuation per se depresses metabolic rate. Available research on the effects of weight cycling on both metabolism and disease thus provides little basis to discourage overweight patients from losing weight. Nevertheless, one of the conclusions of the 1994 report from the National Task Force on the Prevention and Treatment of Obesity was that individuals who are not obese and who have no risk factors for obesity-related illness should not undertake weight loss efforts. The only uncontroversial message of the weight cycling research is that overweight individuals need to be counselled in skills to maintain weight loss, and that relapse prevention should be a more central focus of weight loss programs.

In conclusion, although this research has drawn attention to the necessity of developing improved behavioral and nutritional strategies for sustaining weight reductions and thus preventing weight cycling, the evidence relating weight cycling to adverse health outcomes must be considered equivocal. Regarding the hypothesis that weight cycling exacerbates weight problems, a 1995 opinion survey of obesity researchers concluded that weight cycling was not considered a very important cause of obesity, and little convincing evidence has emerged in recent years to change that conclusion. With regard to the question in the 'weight cycling refrain' at the beginning of the article regarding possible consequences of weight cycling, the current knowledge suggests that it is probably not worse to have lost and regained than never to have lost at all. However, there remain some curious and persistent results in the experimental as well as epidemiological literature suggesting that we do not completely understand the phenomenon.

Table 1 Summary of hypothesized weight cycling effects and selected sources of evidence

Hypothesized adverse health consequences of weight cycling	*Comments (with suggested reading)*
Psychological consequences	Not supported by evidence (reviewed by National Task Force (2000))
Metabolic rate and body composition	Not supported by evidence (reviewed by National Task Force (1994))
All-cause mortality	Supported by most prospective studies, not by others (reviewed in Lissner and Brownell (1992) and by National Task Force (1994))
Cardiovascular disease	Supported by most prospective studies, not others (reviewed in Lissner and Brownell (1992) and National Task Force (1994))
Diabetes	Not supported by epidemiological studies (Field *et al.*, 2004); mixed evidence from clinical setting (Guagnano *et al.*, 2000)
Cancer	Not supported by epidemiological studies (Lissner *et al.*, 1991; Trentham-Dietz *et al.*, 2000)
Bone health	Evidence mixed and limited (Meyer *et al.*, 1998; Gallagher *et al.*, 2002)

See also: **Appetite**: Physiological and Neurobiological Aspects; Psychobiological and Behavioral Aspects. **Cancer**: Effects on Nutritional Status. **Coronary Heart Disease**: Prevention. **Diabetes Mellitus**: Dietary Management. **Hypertension**: Dietary Factors; Nutritional Management. **Meal Size and Frequency**. **Obesity**: Treatment. **Weight Management**: Approaches; Weight Maintenance.

Further Reading

Black DR, Pack DJ, and Hovell MF (1991) A time-series analysis of longitudinal weight changes in two adult women. *International Journal of Obesity* **15**: 623–633.

Bray GA and DeLany J (1995) Opinions of obesity experts on the causes and treatment of obesity—A new survey. *Obesity Research* **3**: 419S–423S.

Diabetes Prevention Program Research Group (2002) Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *New England Journal of Medicine* **346**: 393–403.

Field AE, Byers T, Hunter DJ *et al.* (1999) Weight cycling, weight gain, and risk of hypertension in women. *American Journal of Epidemiology* **150**: 573–579.

Field AE, Manson JE, Laird N *et al.* (2004) Weight cycling and risk of developing type 2 diabetes among adult women in the United States. *Obesity Research* **12**: 267–274.

Field AE, Wing RR, Manson JE, Spiegelman DL, and Willett WC (2001) Relationship of a large weight loss to long-term weight change among young and middle-aged US women. *International Journal of Obesity* **25**: 1113–1121.

Foster GD, Sarwer DB, and Wadden TA (1997) Psychological effects of weight cycling in obese persons: A review and research agenda. *Obesity Research* **5**: 474–488.

French SA, Jeffery RW, Folsom AR, Williamson DF, and Byers T (1995) History of intentional and unintentional weight loss in a population-based sample of women aged 55–69 years. *Obesity Research* **3**: 163–170.

Friedman MA, Schwartz MB, and Brownell K (1998) Differential relation of psychological functioning with the history and experience of weight cycling. *Journal of Consulting and Clinical Psychology* **66**: 646–650.

Gallagher KI, Jakicic JM, Kiel DP *et al.* (2002) Impact of weight-cycling history on bone density in obese women. *Obesity Research* **10**: 896–902.

Guagnano MT *et al.* (2000) Risk factors for hypertension in obese women. The role of weight cycling. *European Journal of Clinical Nutrition* **54**: 356–360.

Hamm PB, Shekelle RB, and Stamler J (1989) Large fluctuations in body weight during young adulthood and 25-year risk of coronary death in men. *American Journal of Epidemiology* **129**: 312–318.

Jeffery RW (1996) Does weight cycling present a health risk? *American Journal of Clinical Nutrition* **63**(supplement): 452S–455S.

Lissner L and Brownell K (1992) Weight cycling, mortality, and cardiovascular disease: A review of epidemiologic findings. In: Brodoff B and Björntorp P (eds.) *Obesity*. Philadelphia: JB Lippincott.

Lissner L, Odell P, D'Agostino R *et al.* (1991) Variability of body weight and health outcomes in the Framingham population. *New England Journal of Medicine* **324**: 1839–1844.

Meyer HE, Tverdal A, and Selmer R (1998) Weight variability, weight change and the incidence of hip fracture: A prospective study of 39,000 middle-aged Norwegians. *Osteoporosis International* **8**: 373–378.

National Task Force on the Prevention and Treatment of Obesity (1994) Weight cycling. *Journal of the American Medical Association* **272**: 1196–1202.

National Task Force on the Prevention and Treatment of Obesity (2000) Dieting and the developing of eating disorders in overweight and obese adults. *Archives of Internal Medicine* **160**: 2581–2589.

Podar T, Solntsev A, Vali M, Vinogradova T, and Podar I (1997) No deterioration of glucose tolerance in weight cycling obese. *International Journal of Obesity* **20**: 921–924.

Stevens VJ, Obarzanek E, Cook NR *et al.* (2001) Long-term weight loss and changes in blood pressure: Results of the trials of hypertension prevention, phase II. *Annals of Internal Medicine* **134**: 1–11.

Trentham-Dietz A *et al.* (2000) Weight change and risk of postmenopausal breast cancer (United States). *Cancer Causes & Control* **11**: 533–542.

Wannamethee SG, Shaper AG, and Walker M (2002) Weight change, weight fluctuation and mortality. *Archives of Internal Medicine* **162**: 2575–2580.

Williamson DF, Pamuk E, Thun M *et al.* (1995) Prospective study of intentional weight loss and mortality in never-smoking overweight US white women aged 40–64 years. *American Journal of Epidemiology* **141**: 1128–1141.

WHOLE GRAINS

R Lang, University of Teeside, Middlesbrough, UK
S A Jebb, MRC Human Nutrition Research, Cambridge, UK

Introduction

There is growing interest in the benefits of traditional dietary patterns with an emphasis on unrefined, plant-based foods. Whole-grain foods are rich in dietary fiber, antioxidants, and a range of other nutrients that may offer health benefits. This chapter will review the epidemiological analyses showing a reduced risk of premature death and decreases in the incidence of cardiovascular disease (CVD), type 2 diabetes, and cancer. However since whole-grain consumption is frequently linked to other positive dietary and life style behaviors (including increased consumption of fruit and vegetables and increased physical activity) more randomized controlled intervention studies are required to support these association studies. Moreover, further research is required to confirm putative mechanistic hypotheses, which include the positive effects of dietary fiber on lipid metabolism, improved glucose homeostasis as a consequence of a reduced glycemic response, or the

antioxidant properties of these foods, which may have beneficial effects on vascular reactivity and inflammation.

What are Whole Grains?

Cereal grains are the seeds of the plant and they house the embryo and the necessary food reserves required for germination. Forming the dietary staples in many countries, the major grains in the human diet are wheat, rice, and corn (maize). Consumption of oats, millet, barley, sorghum, and rye are more limited.

The basic structure of the grain (regardless of plant type) is shown in **Figure 1**. There are essentially three layers, the endosperm, bran, and germ layers, each of which have a unique role within germination, but which also contain essential nutrients and phytochemicals important within the human diet and linked to health benefits.

The Bran Layer

The bran layer is the outer thick-walled structure of the grain. It is rich in B vitamins and phytonutrients such as flavonoids and indoles plus a small amount of protein. It also contains antioxidant compounds including phytoestrogens such as lignans and isoflavones. These hormonally active compounds, similar to estrogen, may influence sex hormone metabolism and may impact on hormone-related disease. The bran also contains factors that may decrease bioavailability of nutrients such as phytic acid, tannins, and enzyme inhibitors. It is also where the bulk of insoluble fiber is found. The insoluble fiber contained within the bran layer has long been recognized to play an important role in intestinal health, by optimizing bowel transit time and increasing fecal weight. But some of the health benefits associated with a high-fiber diet may come from other components, and not just from the fiber itself. For example, the oligosaccharides found within the starchy endosperm layer behave in a similar manner to soluble nonstarch polysaccharides (NSP) and may therefore be useful in controlling blood lipid profiles and blood glucose. In addition, oligosaccharides are natural prebiotics, which encourage the proliferation of healthy microflora within the gut. Colonization of specific bacteria within the colon, such as bifidobacteria, has been implicated in benefits to the immune system, cholesterol lowering, and reducing the risk of colon cancer through the fermentation of these carbohydrates into short-chain fatty acids.

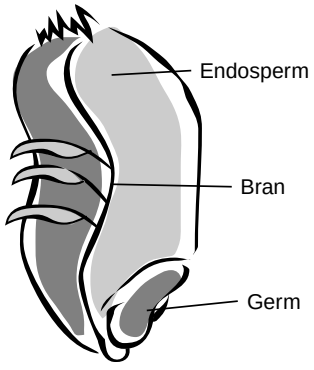


Figure 1 A grain of wheat. (Reproduced with kind permission from the British Nutrition Foundation: http://www.nutrition.org.uk)

The Germ Layer

This is the plant embryo and it contains a concentrated source of minerals such as iron and zinc plus vitamin E. These and other antioxidants provide defense systems against reactive oxygen species not only for the plant, but also for those who consume the grain. Indeed, the pH conditions of the stomach have been shown to cause a dramatic increase in the activity of these antioxidants. Whole grains contain a greater concentration of antioxidants than many fruits and vegetables.

The Endosperm

This makes up about 80% of the grain and is the starchy component comprising mainly carbohydrates including resistant starch and oligosaccharides such as fructans, inulin, and oligofructose. These behave in a similar manner to soluble NSP within the gut. The endosperm also contains B vitamins, in particular riboflavin and pantothenic acid, and some protein.

Definition of Whole Grain

Whole grains are defined as those that are used in their entirety in the food production process, so that all three layers are present within the product. This distinguishes them from refined grains. Regardless of how much of the grain is used, the milling process determines particle size and hence has an impact on handling within the body. Many whole and refined grains are processed in some way to enhance flavor, texture, color, and shelf life. As long as the whole of the grain is used in the process, the food is described as whole grain. However, there is some debate as to whether the disruption of the intact grain modulates the health impact. Certainly, it tends to increase the glycemic index of the food, which may have

implications for the risk of metabolic disease, especially type 2 diabetes.

The vast majority of grains consumed within Western countries are refined, and the outer germ and bran layers are removed to leave only the starchy endosperm. The refining process may reduce some nutrients such as zinc, selenium, and vitamin E by as much as 90%. In some cases vitamins and minerals are reintroduced through fortification and restoration, but the bioavailability and relative health effects of these nutrients when consumed in their natural state compared to artificial methods of fortification is not known.

At present there is no uniform definition of a whole-grain food, but for the purposes of health claims the proportion of whole-grain cereal used within a food product is critical. In 1999 the Food and Drug Administration in the US approved a health claim for use on packing to help consumers choose foods that contained a significant amount of the whole grain. Products must contain at least 51% whole grain by weight, i.e., must be the largest component of the product, to be entitled to carry the health claim. This allows a clear distinction between the refined and partially or nonrefined products available to the consumer. For example, a product made of 100% whole wheat could be labeled as whole grain, but a multigrain loaf containing 75% white flour and 25% wholemeal flour could not. In the UK, the Joint Health Claims Initiative has adopted a similar definition.

Consumption of Whole Grains

Recent studies have shown low levels of consumption of whole-grain foods in the general population of the most affluent countries. In the US data was collected from over 9000 US citizens aged 20 years or over who participated in the 1994–96 USDA's continuing survey of food intakes. Using an interview technique and serving sizes defined by the Food Guide Pyramid, food consumption over 2 nonconsecutive days was examined. It was found that 29% of the sample were nonconsumers of whole grains and the average number of daily servings was less than 1 per day. The proportion of the sample reaching the recommended 3 servings per day was 8%.

Intakes are similarly low in the UK. Using data from two nationally representative surveys whole-grain consumption was assessed: the Diet and Nutritional Survey of British Adults 1986–87 included over 2000 adults aged between 16 and 64 years, and the National Diet and Nutrition Survey of people aged 65 years and over included over 1000 free-living adults (from a total of over 2000 participants which included free-living and institutionalized individuals) during 1994–95. Dietary data was collected using a 7-day diary recording weighed food in 1986–87, and a 4-day weighed food diary in 1994–95. Whole-grain foods were identified as those having at least 51% whole-grain ingredients by weight and a serving was defined as each occasion the food appeared within the recording period. Both surveys showed that approximately 30% of individuals did not consume any whole-grain foods during the survey period (29% in 1986–87 and 33.5% in 1994–95), and over 97% of adults did not meet the US recommendation of 3 servings per day. Median consumption was less than 1 serving per day.

In comparison to the US and the UK, whole-grain foods are consumed in greater quantities in Scandinavian countries. In Finland, rye bread has always been a staple and consequently whole-grain intakes have always been high. In Norway, food disappearance data suggest that consumption of whole-grain foods is four times that seen in the US, but is lower than that seen in Finland. In the Scandinavian studies, however, estimation of consumption of whole grain was not based on number of servings but utilized measures such as a bread score based on number of slices of bread consumed multiplied by the proportion of whole-grain flour; 24-h recall techniques of whole-grain foods consumed; and simply the number of slices of bread consumed. These differing techniques make it difficult to directly compare findings in Scandinavia with the UK and US.

However, within populations the consumption of whole grains is influenced by a number of social and demographic factors. Most surveys have found an increase in whole grain consumption with age. For example, in the UK, there was a median of 1 serving per week in the 16–24-year-olds rising to 3 servings per week in the 35–64-year-olds (1986–87 survey).

Consumption in the survey of people aged 65 years and over showed higher intakes of 5 servings per week (1994–95 survey) but it is not clear whether this reflects a secular trend or a continuing effect of age. In general, men consume more whole-grain foods than women. This may reflect a greater food intake overall, rather than a specific preference for whole-grain varieties. In the US, white adults consume more whole-grain foods than black Americans, and Mexican Americans consume the least whole grains.

In the US and UK, income and level of education are also positively associated with whole-grain consumption, but in Finland, the highest intakes of rye bread are observed in the lower socioeconomic groups. In the US and UK, whole-grain

consumers are less likely to smoke, tend to be regular exercisers and consume more fruits and vegetables. These findings suggest an association of whole-grain consumption with other positive life style traits.

Whole Grains and Health

Epidemiological evidence suggests an inverse relationship between the consumption of whole-grain foods and the relative risk of a number of chronic diseases. Studies have found that habitual consumption of whole-grain foods is associated with reductions in premature mortality, risk of coronary heart disease, ischemic stroke, and type 2 diabetes.

All Cause Mortality

Three large prospective epidemiological studies have considered the relationship between whole-grain consumption and all cause mortality (**Table 1**). Using a variety of measures to assess whole-grain consumption, all three studies concluded that the more whole-grain foods were consumed, the lower the risk of death from a number of chronic diseases. Women in the Iowa Women's Study were followed over a 9-year period. The population was divided into quintiles of whole grain intake using a food frequency questionnaire (FFQ). Intakes varied considerably from 1.5 servings per week in the lowest quintile and 22.5 servings per week in the highest, yet an inverse association between whole-grain intake and the risk of death was observed across quintiles, with a 40% reduction in total mortality ($P < 0.0001$) in those consuming at least one serving of whole-grain foods per day. Even when adjusted for confounders such as age, energy intake, hypertension, heart disease, diabetes, cancer, body mass index (BMI), waist–hip ratio (WHR), physical activity, alcohol intake, smoking, positive dietary habits such as fruit and vegetable consumption, fat intake, and red meat and fish consumption, the significant inverse relationship remained ($P = 0.005$). In contrast, refined grain intake was associated with increased mortality for those in the quintile of highest refined grain intake compared to those in the quintile of lowest intake (**Figure 2**).

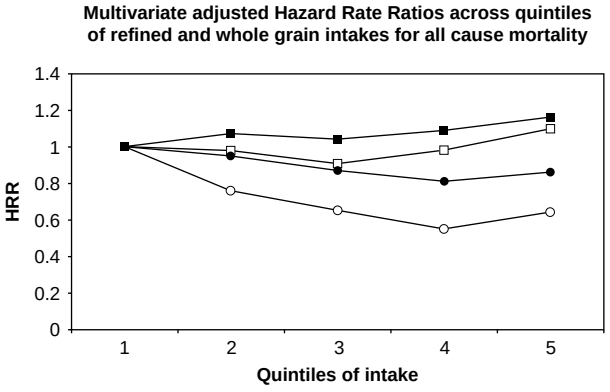


Figure 2 ○ (WG) and □ (refined) adjusted HRR for age and total energy intake. ● (WG) and □ (refined) adjusted for age, energy intake, marital status, education, high blood pressure, diabetes, heart disease, cancer, BMI, WHR, physical activity, smoking, alcohol intake, use of vitamin supplements, HRT, total fat, saturated fat, intake of fruits and vegetables, intake of meat and intake of fish and seafood.

Similarly in over 47 000 men and women in Norway aged 35–56 years studied over 9 years, hazard rate ratios (HRR) for total mortality were inverse and graded across whole-grain bread score categories in men and women between the highest and lowest bread score after adjustment for a range of dietary and lifestyle factors. The bread score was calculated using number of slices consumed per day and the proportion of whole-grain flour used. The analysis found that both components of the whole grain scoring system contributed to these inverse trends (% deaths between highest and lowest

Table 1 Effect of whole-grain consumption on all cause mortality

Cohort	*Measure of WG consumption*	*Reported association*	*Reference*
Iowa Women	FFQ	40% ↓ with at least 1 serving per day Refined grains ↑ mortality with ↑ consumption	Jacobs *et al.* (1999) Is whole grain intake associated with reduced total and cause-specific death rates in older women? The Iowa Women's Health Study. *American Journal of Public Health* **89**: 322–329.
Norwegian County Study	Bread score (slices × %WG)	HRR graded and inverse with ↑ consumption	Jacobs *et al.* (2001) Reduced mortality among whole grain bread eaters in men and women in the Norwegian County Study. *European Journal of Clinical Nutrition* **55**: 137–143.
Physician's Health Study	Breakfast cereal FFQ (WG classified as 25% w/w)	HRR ↓ with ↑ consumption	Liu *et al.* (2003) Is intake of breakfast cereals related to total and cause-specific mortality in men? *American Journal of Clinical Nutrition* **77**: 594–599.

scores categories: 7% versus 10% for men and 2.7% versus 4.6% for women).

In the Physicians' Health Study, breakfast cereal consumption was used as an indicator of whole-grain intake in over 86 000 US male physicians aged 40–84 years studied over 5.5 years. Breakfast cereal consumption was assessed using a semiquantitative FFQ where men had to report the amount, frequency, brand, and type of cereal consumed over the previous year. Whole-grain cereals were classified as those with >25% whole grain or bran by weight; all others were considered to be refined grains. Whole-grain breakfast cereal consumption was inversely associated with total mortality independent of a range of dietary and lifestyle considerations.

The use of bread and cereal intakes as a measure of total whole-grain consumption is of some concern, as the extent to which they correlate with overall whole-grain consumption is uncertain. Indeed, such studies also fail to distinguish whether it is in fact something within the whole-grain package that is of benefit, or something else entirely.

Cardiovascular Disease

Cardiovascular diseases are responsible for over a third of all deaths and are the biggest contributor to the global burden of disease. There are a number of studies to suggest that individuals who consume a diet rich in whole-grain foods have a lower incidence of heart disease, although the mechanism is still unclear (see **Table 2**).

Increases in the consumption of whole grains have been shown to decrease CHD deaths and the risk of stroke and heart disease in some, but not all, epidemiological analyses. In the study of postmenopausal Iowan women there was a reduction in relative risk (RR) of ischemic heart disease of about a third in those consuming at least 1 serving of whole-grain foods per day. This relationship was attributable to differences in the consumption of dark breads and whole-grain breakfast cereals while less common whole-grain foods such as popcorn, brown rice, and oatmeal showed no relationship with CVD. A significant inverse relationship between increasing whole-grain intake and risk was also observed for CHD and total CVD, but not stroke alone (**Figure 3**).

Similar results were obtained in the Nurses' Health Study of 75 000 women aged 38–63 years who were free from existing diabetes, angina, myocardial infarction, stroke, or other CVDs at baseline. Here, a significant inverse relationship was observed between CHD and whole-grain consumption even after multivariate adjustment for known confounders such as age, smoking, BMI, alcohol, and other dietary and lifestyle factors. For each additional serving of whole-grain food per day, the authors found a relative risk of 0.91 (95% confidence interval (CI) 0.85, 0.97) for CHD risk.

In this study there was also a significant inverse relationship between whole-grain intake and risk of ischemic stroke. After adjustment for smoking and other known CVD risk factors, the relationship was

Table 2 Summary of the evidence relating a reduced risk of CVD to increased whole-grain consumption

Evidence for a reduced risk of:	*Cohort*	*Reported association*	*Reference*
CHD	Californian Seventh Day Adventists	Lower RR for preference of whole grain bread	Fraser *et al.* (1999) Associations between diet and cancer, ischemic heart disease, and all-cause mortality in non-Hispanic white California Seventh-day Adventists. *American Journal of Clinical Nutrition* **70**: 532S–538S.
IHD	Iowa Women's Health Study	Lower RR for increasing whole grain consumption	Jacobs *et al.* (1998a) Whole grain intake may reduce the risk of ischaemic heart disease in postmenopausal women: The Iowa Women's Health Study. American *Journal of Clinical Nutrition* **68**: 248–257.
CHD and CVD	Iowa Women's Health Study	Lower RR for increasing whole grain consumption (except for stroke after adjustment)	Jacobs *et al.* (1999) Is whole grain intake associated with reduced total and cause-specific death rates in older women? The Iowa Women's Health Study. *American Journal of Public Health* **89**: 322–329.
CHD	Nurse's Health Study	Lower RR for increasing whole grain consumption	Liu *et al.* (1999) Whole grain consumption and risk of coronary heart disease: results from the Nurses' Health Study. *American Journal of Clinical Nutrition*. **70**: 412–419.
Ischemic stroke	Nurse's Health Study	Lower RR for increasing whole grain consumption (total stroke cases)	Liu *et al.* (2000) Whole grain consumption and risk of ischemic stroke in women: A prospective study. *JAMA* **284**: 1534–1540.

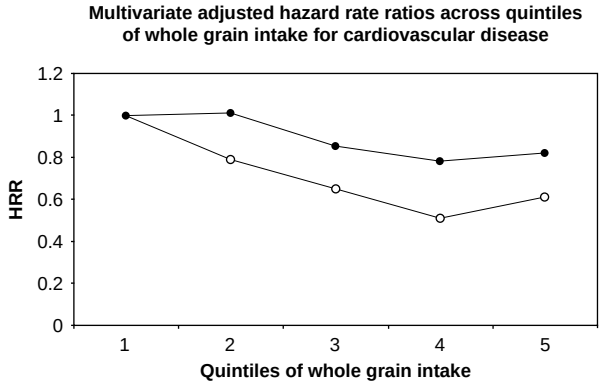


Figure 3 ○ adjusted HRR for age and total energy intake. ● adjusted for age, energy intake, marital status, education, high blood pressure, diabetes, heart disease, cancer, BMI, WHR, physical activity, smoking, alcohol intake, use of vitamin supplements, HRT, total fat, saturated fat, intake of fruits and vegetables, intake of meat and intake of fish and seafood.

attenuated but remained significant. However, after further adjustment for assorted dietary variables (folate, vitamin E, fiber, magnesium, and potassium), the effect was no longer significant. Unlike previous studies, the authors defined the different categories of stroke and found that although risk of hemorrhagic stroke or incident fatal strokes did not appear to be influenced by whole-grain consumption, total stroke risk was inversely related to consumption of whole-grain foods.

It is notable that in many studies subjects with the highest intake of whole-grain foods also had the healthiest lifestyles and the relationship with whole-grain foods is attenuated after adjustment for other diet and lifestyle variables. The exact mechanisms of protection are unclear. Diets rich in whole-grain foods tend to reduce serum LDL-cholesterol and TAG levels whilst increasing HDL-cholesterol concentrations and blood pressure is lower. This may be due in part to the dietary fiber, but the effect usually persists after adjustment for fiber intake. Whole grains also contain a number of specific components that may have heart health benefits, including antioxidants (vitamin E and selenium), B vitamins, flavonoids, and indoles. These may reduce oxidative stress and homocysteine levels, and the isoflavone content of these grains may positively influence vascular reactivity and the inflammatory state.

Type 2 Diabetes

The prevalence of type 2 diabetes has reached epidemic proportions with over 150 million cases diagnosed worldwide; this number is expected to double by 2025. The concurrent rise in obesity has been directly linked to insulin resistance and compensatory hyperinsulinemia and eventual type 2 diabetes, with over 80% of diagnosed type 2 diabetes being the result of excess body fat. Public health recommendations to reduce fat intake, especially saturated fat, have led to a rise in the proportion of carbohydrates (particularly refined carbohydrates) in the diet with consequences for postprandial glucose and insulin metabolism. The source of carbohydrate is also important. Whole-grain foods commonly have a low glycemic index because whole-grain foods with an intact bran and germ layer have a much smaller impact on blood glucose than refined carbohydrate foods because of their larger particle size, slowing the rate of enzymic attack. The level of soluble fiber within whole grains has also been identified as a possible protector and the higher amylose content is also thought to be beneficial. Slower rates of digestion are observed when foods have more compact granules, contain high levels of viscous soluble fiber, and have a higher amylose to amylopectin ratio.

The relationship between whole grains and diabetes has been studied in five large cohorts as highlighted in **Table 3**. All of the studies have found an inverse relationship between consumption of whole grains or cereal fiber and disease reduction despite slight variations in methodology.

As a proxy measure of whole-grain consumption, the relationship between the intake of total and specific sources of dietary fiber, dietary glycemic index, and glycemic load in the Nurses' Health Study and the Health Professional's Study was examined. Among the 65 173 women who participated during 1986–1992, women in the highest quintile of cereal fiber intake had a 28% lower risk of diabetes than those in the lowest quintile of intake (RR 0.72; 95% CI 0.58, 0.90; $P = 0.001$), a significant reduction that was not observed with fruit or vegetable fiber intakes. In men there was an inverse relationship between cereal fiber intake and risk of type 2 diabetes: a reduction in risk of 30% following adjustment for confounders. Again, no significant relationship of fruit or vegetable fiber to diabetes risk was observed.

The fiber content of whole grains has been suggested as a possible explanation for the inverse relationship between total and whole-grain intakes and risk of type 2 diabetes observed in a 10-year follow-up of Finnish men ($n = 2286$) and women ($n = 2030$). When the highest and lowest quartiles of whole-grain consumption were compared there was an over 30% reduction in risk following adjustment for age, sex, geographic area, and energy intake. Cereal fiber, but not that from fruits and vegetables, was inversely related to risk of type 2 diabetes even after adjustment for a number of

Table 3 Summary of the evidence relating a reduced risk of type 2 diabetes to increased whole grain consumption, including studies where cereal or dietary fiber intake is taken as a surrogate marker for whole-grain intakes

Evidence for a reduced risk of:	*Cohort*	*Reported Association*	*Reference*
Epidemiological			
Type 2 diabetes	Nurse's Health Study	Lower RR with increased dietary fiber	Salmeron *et al.* (1997a) Dietary fiber, glycemic load, and risk of non-insulin-dependent diabetes mellitus in women. *JAMA* **277**: 472–477.
Type 2 diabetes	Health Professionals Follow-up Study	Lower RR with increased dietary fiber	Salmeron *et al.* (1997b) Dietary fiber, glycemic load, and risk of NIDDM in men. *Diabetes Care* **20**: 545–550.
Type 2 diabetes	Finnish Mobile Clinic Health Examination Survey	Lower RR with increased whole grains	Montonen *et al.* (2003) Whole-grain and fiber intake and the incidence of type 2 diabetes *American Journal of Clinical Nutrition* **77**: 622–629.
Type 2 diabetes	Health Professionals Follow-up Study	Lower RR with increased whole grains	Fung *et al.* (2003) Whole-grain intake and the risk of type 2 diabetes: a prospective study in men. *American Journal of Clinical Nutrition* **76**: 535–540.
Type 2 diabetes	Nurse's Health Study	Lower RR with increased whole grains	Liu *et al.* (2000) A prospective study of whole-grain intake and risk of type 2 diabetes mellitus in US women. *American Journal of Public Health* **90**: 1409–1415.
Risk factors for type 2 diabetes and CVD	Framingham Offspring Study	Reduction in fasting insulin with increasing whole-grain intake	McKeown *et al.* (2002) Whole grain intake is favourably associated with metabolic risk factors for type 2 diabetes and cardiovascular disease in the Framingham Offspring Study. *American Journal of Clinical Nutrition* **76**: 390–398.
Intervention			
Insulin sensitivity	11 hyperinsulinemic overweight patients	Reduction in fasting insulin following diet rich in whole grains	Pereira *et al.* (2002) Effect of whole grains on insulin sensitivity in overweight hyperinsulinaemic adults. *American Journal of Clinical Nutrition* **75**: 848–855.

confounders. Adjustment for cereal fiber considerably weakened the association between whole-grain consumption and risk of type 2 diabetes, suggesting that this may be a significant component of the whole-grain package.

The effect of whole-grain consumption specifically, rather than fiber intakes, on incidence of type 2 diabetes was examined in the Health Professional's Follow-up Study. Over a 12-year follow-up period, intakes of whole and refined grains were analyzed using a validated semiquantitative FFQ. Despite no baseline history of diabetes or CVD, 1197 cases of incident type 2 diabetes were identified in this male cohort. Following adjustment for dietary and life style confounders including age, smoking, physical activity, and fruit and vegetable intake, there was a reduced risk of type 2 diabetes of almost 40% in those with the highest quintile compared with the lowest quintile of whole-grain intakes. The results were attenuated after adjustment for BMI, although the relationship remained significant. In those with a BMI $>30\,kg\,m^{-2}$ the association between whole grain and type 2 diabetes was weak, whereas in those men with a BMI $<30\,kg\,m^{-2}$ a 50% risk reduction was observed in those who consumed the most whole grains. However, after adjusting for components of the whole-grain package such as cereal fiber, magnesium, and glycemic load, the statistical significance was lost (**Figure 4**).

These findings in men were similar to those observed by Liu *et al.* (2000) when they looked specifically at whole and refined grain intakes in the women participating in the Nurses' Health

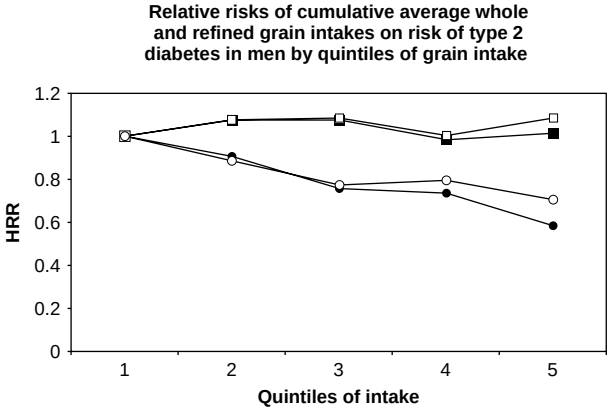


Figure 4 ○ adjusted for age, period, physical activity, energy intake, missing FFQ data, smoking, family history of diabetes, alcohol intake, fruit intake and vegetable intake. ● additionally adjusted for BMI $<30kg/m^2$ and $>30kg/m^2$.

Study. During the 10-year follow-up, 1879 cases of incident type 2 diabetes were confirmed. Although the women with the highest intake of whole grain had other beneficial dietary and lifestyle factors, whole-grain intake was inversely related to risk. There was a significant inverse association between the highest and lowest quintiles of whole-grain intake after adjustment for age and energy intake. Although attenuated after adjustment for BMI and other lifestyle factors, the relationship remained significant. Again BMI appeared to be the strongest confounding factor. Women in the lowest quintile of intake ratio (those with low whole grain or large refined grain intakes) had a 57% greater risk of type 2 diabetes than women in the highest quintile.

In a cross-sectional assessment of 2941 subjects in the Framingham Offspring Study the effect of whole-grain intake on metabolic risk factors for type 2 diabetes and CVD was examined. Dietary intake was assessed using a semiquantitative FFQ in the participants who were free from diabetes or high cholesterol. Breakfast cereal type was used to quantify whole-grain intakes based on a whole-grain content of over 25%. Other foods identified as whole grain were dark breads, popcorn, and oatmeal. Whole-grain intakes were similar between men and women (mean 8.3 and 8.8 servings per week, respectively) but refined grain intakes were much higher (22.0 and 18.5 servings per week, respectively). Similar to other studies, those in the highest quintile of whole-grain intakes (20.5 servings per week) had lower BMI, were less likely to smoke or drink, and dietary habits were better. Following adjustment for a host of confounding factors, whole-grain consumption in the highest quintile was associated with a significant reduction in fasting insulin in comparison to those in the lowest quintile of intake. Even after further adjustment for BMI and dietary factors such as vegetable and fat intakes, this relationship remained significant, but was no longer significant after further adjustment for magnesium, and insoluble and soluble fibers. The association between whole grain and fasting insulin was most striking in those with a BMI $>30\,\mathrm{kg\,m^{-2}}$ with the highest fasting insulin levels being observed in those with the highest BMI and the lowest intake of whole-grain foods.

Prospective epidemiological studies are generally stronger than cross-sectional associations. In the CARDIA study by Pereira and coworkers, a significant inverse relationship was observed between whole-grain foods and fasting insulin levels among over 3500 black and white young Americans aged 18–30 years. A dietary history was collected at baseline and 7 years later, while insulin measurements were collected at 10 years follow-up. After adjustment for a number of dietary and lifestyle factors an inverse and graded response was observed between whole-grain intake at 7 years and the insulin measurements collected at 10 years follow-up, although the relationship was not significant in black women.

There is only one small intervention study that examines the impact of increasing wholegrain consumption. In this study, it was found that after 6 weeks there was a 10% reduction in fasting insulin compared to results observed following the refined grain diet ($141 \pm 3.9\,\mathrm{pmol\,l^{-1}}$ versus $156 \pm 3.9\,\mathrm{pmol\,l^{-1}}$; $P < 0.01$). This relationship remained even after adjustment for body weight changes (nonsignificant change of $-0.7\,\mathrm{kg}$ on whole-grain diet) and physical activity.

As for other diseases the mechanism of the effect of wholegrain on insulin sensitivity is not entirely clear and may in part be mediated through effects on body weight. Cereal fiber and possibly certain micronutrients such as magnesium may also be important since the wholegrain effect is attenuated after adjustment for these variables.

Cancer

Dietary factors are thought to account for about 35% of all cancers but the role of any specific dietary factor or dietary regime has only been established for certain types of cancer. Only a few studies have looked at the links between wholegrain intake and cancer. In the Iowa Women's Study there was a 30% reduction in cancer deaths when comparing those with the highest quintile of whole-grain intake to those in the lowest quintile, after adjustment for age and energy intake. However, once other dietary and lifestyle factors were included within the multivariate analysis this relationship was attenuated and lost its statistical significance. Similar findings were observed in the Norwegian County Study with a 28% reduction in cancer deaths from the highest to lowest quintile of wholegrain when adjusted for age and energy intake. However, the effect was no longer significant after further adjustment for other dietary and lifestyle factors.

A number of studies have used case–control designs to investigate the relationship between whole-grain consumption and cancer incidence, although these suffer from the inherent flaws of such study designs, especially those involving recall of past dietary habits. In an analysis of 40 case–control studies, 90% of the studies included had an odds ratio (OR) <1, of which 55% reached statistical significance in favor of a benefit of wholegrain. The pooled OR for high versus low intakes of whole-grain foods was 0.66 (95% CI

0.60, 0.72). Most data pertain to the link with cancers of the digestive tract. The majority of pooled-odds ratios for specific cancers were between 0.5 and 0.8, except for breast and prostate cancers, which were 0.86 and 0.90, respectively. The differing types of dietary data collection impacted on the findings. Where only whole-grain frequency was recorded the pooled OR was 0.82 in those who ate whole grains infrequently and 0.59 among habitual consumers ($P < 0.0001$ for trend). Similar results were found in those studies reporting intake by tertiles (OR 0.81 and 0.62 for the second and third tertiles, respectively; $P = 0.0001$) and those reporting actual quantities of intake found a downward trend in ORs as the dose increases, although the trend is not strong ($P = 0.18$) suggesting that the dose-response relationship between whole grain and cancer types may only be modest.

The whole-grain package contains a number of components that have been identified as having beneficial effects on cancer risk, including antioxidants and flavonoids, isoflavones, fermentable carbohydrates, and resistant starch. Although many studies have been conducted looking at the effect of individual components that can be found within the whole-grain package on cancer risk, few studies have looked specifically at whole grains *per se*. Although a relationship has been observed between whole-grain consumption and cancer deaths, studies of cancer incidence and whole-grain consumption are not significant after adjustment for potential confounders. This is a good example of how consumption of whole-grain foods appears to be a marker of dietary habits associated with a reduction in risk, and they are not necessarily specifically important in their own right.

Dietary Recommendations

The epidemiological data suggest that health benefits can be obtained at relatively low levels of whole-grain consumption, typically 1–3 servings per day. In most studies, there was no clear dose–response relationship and a suggestion of a threshold effect as benefits were seen at the third quintile of whole-grain intakes with no further reduction in risk as intakes increased. However, assessment of intakes in different countries show that this threshold level of intake is not being achieved.

International dietary Guidelines recommend increased grain consumption. At present, the USA is the only nation to specify exact quantities of whole grain foods, and it is only within the last few years that whole grains have been considered seperately from total grain foods.

The recommendations for grains have evolved over time to reflect changes in research and to simplify and clarify consumer messages. The latest American Dietary Guidelines (2005) now state a recommendation of 3 or more ounce-equivalent portions of whole-grain foods daily, with a further recommendation that at least half of grain consumption should be whole-grain.

This recommendation of three servings per day was also specifically incorporated in the Department of Health and Health Services nutrition objectives for 2010 (US Department of Health and Human Services, 2000).

In the UK, the Food Standards Agency explicitly encourages consumers to select whole-grain varieties in their healthy eating advice, although no exact quantities are given. Many other European countries also tend to place emphasis on only cereals and fiber, without necessarily specifically highlighting whole grains.

Using data compiled from focus groups and consumer interviews in the US, a number of reasons as to why consumption of whole-grain foods may be low have been indentified. Consumers report difficulties in identifying whole-grain foods and express limited knowledge about the preparation and cooking of whole-grain foods. Adolescents in particular reported that whole-grain foods were bland and have a dry taste. While breakfast cereals appear to be well received in this age group, whole-grain breads were described as dry and bitter. Furthermore, whole-grain varieties of bread, pasta, and rice tend to be more expensive and this may deter those in low income and vulnerable groups.

In a UK intervention study to increase whole-grain consumption, 25–40-year-olds were encouraged to increase whole-grain food consumption over a period of 2 weeks by gradually increasing servings from 1 to 5 servings per day. Volunteers were given positive health messages about eating more whole-grain foods, were helped in identifying such food products, and were also given advice on how to incorporate them easily into their existing diet. In post study focus groups, participants were positive about the changes made and were happy to continue consuming whole-grain foods but at lower levels than that prescribed during the study period. Similar to findings in the US, breakfast was found to be a good meal to change habit and breakfast cereals and bread type was deemed the easiest way of incorporating whole-grain foods.

Consumer research suggests that few people are aware of the health benefits of whole-grain foods. Although other food groups, such as fruits and vegetables, have been identified as possessing health benefits, the association between whole-grain foods and

a reduced risk of a number of chronic diseases is not recognized among the general population or indeed health professionals. Recent health claims in the US and UK may help to address this knowledge gap. In addition, consumer initiatives such as the 'Whole Grain for Health' campaign in the UK, and 'Go Grains' in Australia provide ongoing education relating to the benefits of including whole-grain products within the diet.

See also: **Cancer**: Epidemiology and Associations Between Diet and Cancer; Epidemiology of Gastrointestinal Cancers Other Than Colorectal Cancers. **Cereal Grains**. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Dietary Fiber**: Physiological Effects and Effects on Absorption; Potential Role in Etiology of Disease; Role in Nutritional Management of Disease.

Further Reading

Anderson JW (2002) Whole-grains intake and risk for coronary heart disease. In: Marquart L, Slavin JL, and Fulcher RG (eds.) *Whole-grain Foods in Health and Disease*, pp. 155–185. Minnesota: American Association of Cereal Chemists, Inc.

Fung TT, Hu FB, Pereira MA, Liu S, Stampfer MJ, Colditz GA, and Willett WC (2002) Whole grain intake and the risk of type 2 diabetes in men: a prospective study in men. *American Journal of Clinical Nutrition* **76**: 535–540.

Jacobs DR Jr, Marquart L, Slavin J, and Kushi LH (1998b) Whole-grain intake and cancer: an expanded review and meta-analysis. *Nutrition and Cancer* **30**: 85–96.

Jacobs DR Jr, Meyer KA, Kushi LH, and Folsom AR (1998a) Whole grain intake may reduce the risk of ischaemic heart disease death in postmenopausal women: the Iowa Women's Health Study. *American Journal of Clinical Nutrition* **68**: 248–257.

Jacobs DR Jr, Meyer KA, Kushi LH, and Folsom AR (1999) Is whole grain intake associated with reduced total and cause-specific death rates in older women? The Iowa Women's Health Study. *American Journal of Public Health* **89**: 322–329.

Jacobs DR Jr, Meyer HE, and Solvoll K (2001) Reduced mortality among whole grain bread eaters in men and women in the Norwegian County Study. *European Journal of Clinical Nutrition* **55**: 137–143.

Lang R and Jebb SA (2003) Who consumes whole grains, and how much? *Proceedings of the Nutrition Society* **62**: 123–127.

Lang R, Thane CW, Bolton-Smith C, and Jebb SA (2003) Whole-grain food consumption by British adults from two national dietary surveys. *Public Health Nutrition* **6**: 479–484.

Liu S (2002) Dietary carbohydrates, whole grains, and the risk of type 2 diabetes. In: Marquart L, Slavin JL, and Fulcher RG (eds.) *Whole-grain foods in health and disease*, pp. 155–185. Minnesota: American Association of Cereal Chemists, Inc.

Liu S, Manson JE, Stampfer HJ, Hu FB, Giovannucci E, Colditz GA, Manson JE, Hennekens CH, and Willett WC (2000b). A prospective study of whole grain intake and risk of type 2 diabetes mellitus in US women. *American Journal of Public Health* **90**: 1409–15.

Liu S, Manson JE, Stampfer MJ, Rexrode KM, Hu FB, Rimm EB, and Willett WC (2000a) Whole grain consumption and risk of ischaemic stroke in women: a prospective study. *Journal of the American Medical Association* **284**: 1534–1540.

Liu S, Stampfer HJ, Hu FB, Giovannucci E, Rimm E, Manson JE, Hennekens CH, and Willett WC (1999) Whole grain consumption and risk of coronary heart disease: results from the Nurses' Health Study. *American Journal of Clinical Nutrition* **70**: 412–419.

McIntosh GH and Jacobs DR (2002) Cereal-grain foods, fibers and cancer prevention. In: Marquart L, Slavin JL, and Fulcher RG (eds.) *Whole-grain Foods in Health and Disease*, pp. 155–185. Minnesota: American Association of Cereal Chemists, Inc.

McKeown NM, Meigs JB, Liu S, Wilson PWF, and Jacques PF (2002) Whole grain intake is favourably associated with metabolic risk factors for type 2 diabetes and cardiovascular disease in the Framingham Offspring Study. *American Journal of Clinical Nutrition* **76**: 390–398.

Pereira MA, Jacobs DR, Pins JJ, Raatz SK, Gross MD, Slavin JL, and Seaquist ER (2002) Effect of whole grains on insulin sensitivity in overweight hyperinsulinaemic adults. *American Journal of Clinical Nutrition* **75**: 848–855.

Smith AT, Kuznesof S, Richardson DP, and Seal CJ (2003) Behavioural, attitudinal and dietary responses to the consumption of wholegrain foods. *Proceedings of the Nutrition Society* **62**: 1–13.

Wilson's Disease *see* **Copper**

Wine *see* **Alcohol**: Absorption, Metabolism and Physiological Effects; Disease Risk and Beneficial Effects; Effects of Consumption on Diet and Nutritional Status

WORLD HEALTH ORGANIZATION

J Akré, World Health Organization, Geneva, Switzerland

The World Health Organization (WHO) is the intergovernmental organization within the United Nations (UN) system that acts as the directing and coordinating authority on international health work. It performs its functions through three principal bodies—the World Health Assembly, the Executive Board, and the Secretariat. The objective of WHO, which has 191 member states, is the attainment by all peoples of the highest possible level of health. While its headquarters is located in Geneva, Switzerland, the organization is decentralized in six regions, each with its own regional committee and regional office. WHO's regular budget—US$842 654 000 for the biennium 1998–1999—is augmented by voluntary contributions, which are roughly equivalent to regular budget levels. WHO works closely with and through others, including other agencies of the UN system, nongovernmental organizations, and collaborating centres around the world in numerous disciplines. In restructuring its programs in the mid-1990s in the face of resource constraints, WHO decided to place emphasis on meeting the most pressing health needs. In seeking to prevent and overcome malnutrition, WHO promotes the tailoring of approaches to fit circumstances. While the rapidly increasing threat of noncommunicable diseases accounts for at least 40% of all deaths in developing countries and 75% in industrialized countries, many millions still cannot meet basic needs for energy and protein, are deficient in essential micronutrients, or are severely malnourished. Thus, coordinated action is called for on both fronts. Consistent with the unique normative, scientific, and advisory role that WHO has played for the last half century, the organization strives to support all its member states in developing food and nutrition policies that will make healthy choices the easy choices for their populations.

Birth of the World Health Organization

For many thousands of years, people have exchanged remedies and diseases without really thinking of ways to work together to promote health that could go beyond purely parochial concerns. Early attempts at international cooperation in health were limited to small groups of countries, which discussed a few obviously contagious diseases such as cholera and smallpox, and strategies such as quarantine to keep them at bay. Although 11 international sanitary conferences were held in Europe between 1851 and 1903, it was not until 1907 that a worldwide international institution—the Office International d'Hygiène Publique—was founded to prepare and administer international sanitary conventions and to provide national health administrations with an opportunity for regular contacts and discussion.

The terrible epidemics which raged through Europe at the end of World War I, coupled with the mass movement of liberated prisoners of war, constituted a menace to Europe of such magnitude as to require coordinated international effort. The League of Red Cross Societies, created in 1919, attempted the task but quickly realized that intergovernmental action was essential to cope with a problem of such magnitude. As the charter of the Office International d'Hygiène Publique did not give enough power for action in individual countries, provision for necessary measures had to be made by the League of Nations, then in the process of creation.

The Geneva-based League, during its short and unhappy history between the two world wars, had been the first to invoke international health cooperation to deal with many kinds of health problems. In the same period, the Pan American Sanitary Organization, originally established in 1902, continued to work in its own geographical sphere. (In 1958 it became the Pan American Health Organization (PAHO), which serves as WHO regional office for the Americas (see below).)

In 1945, the UN Conference on International Organization, meeting in San Francisco, unanimously approved a proposal by Brazil and China to establish an autonomous international health organization within the UN system. The following year, an international conference held in New York set up an interim commission and approved the *Constitution of the World Health Organization*. This came into force on 7 April 1948, when the 26th government, out of a total of 61 signatories, formally ratified it in its national parliament. Since then, 7 April is celebrated every year as World Health Day, when attention around the globe is focused on a theme of major international public health importance. Today, the health agency based in Geneva, Switzerland, is owned and operated by the governments of 191 countries, representing almost the entire population of the world, as reflected in its emblem (**Figure 1**).

Figure 1 WHO emblem.

The Task Entrusted to WHO

WHO is defined by its constitution as the directing and coordinating authority on international health work. Its aim is 'the attainment by all peoples of the highest possible level of health,' which is 'one of the fundamental rights of every human being.' The constitution lists specifically a number of responsibilities. These include the following:

- to stimulate the eradication of epidemic, endemic, and other diseases;
- to promote improved nutrition, housing, sanitation, working conditions, and other aspects of environmental hygiene;
- to propose international conventions and agreements in health matters;
- to promote and conduct research in the field of health;
- to develop international standards for food, biological, and pharmaceutical products;
- to assist in developing an informed public opinion among all peoples on matters of health.

The agreed policy of WHO is a determined and structured effort by all countries to bring health within the reach of everyone. 'Health' is defined by WHO's constitution as a 'state of complete physical, mental, and social well-being and not merely the absence of disease and infirmity.' It is seen as a shared responsibility, calling for a high degree of self-reliance from the individual, the family, the community, and, of course, the nation as a whole. Because the determinants of health are so broad, the efforts of the health sector must be supported and augmented by those of many other related sectors, including agriculture, water and sanitation, finance, industry, planning, communication, and education.

WHO provided the first truly global framework for setting international standards to promote and protect health. In keeping with postwar faith in the power of technology, WHO initially operated as a technical organization, fuelled by advances in bio-medical research and the belief that new medical discoveries would bring spectacular improvements in health. One of its first tasks was to develop mechanisms, still in effect today, for identifying urgent research needs and then linking the world's leading specialists and research institutes in a concerted attack on the problem. Tangible results came in the form of new diagnostic tests, therapeutic drugs, and vaccines. WHO also standardised the classification of diseases, terminology, nomenclature, reporting systems, research protocols, and quality and safety specifications for foods, drinking water, and pharmaceutical products. Consumers the world over benefit from these standards, which are continually revised in the light of new knowledge.

As research advances produced the means for conquering one disease after another, WHO shifted its emphasis to problems of logistics. Research took on social and ethical dimensions as the organization sought ways to extend the benefits of modern medicine to the world's populations. The early promise that sophisticated technology would bring spectacular improvements in health paled, however, against the reality of the millions of people who had no access to basic medical services. Given its constitutionally defined universal mandate and humanitarian mission, WHO began advocating changes that would eventually revolutionize the way public health was perceived. In a world that remained diseaseridden and suffering despite unprecedented technical advances, it became a matter of equity and social justice to make health progress available to all people through new approaches, new strategies, and better management of resources.

High technology as an end in itself was replaced by the concept of appropriate technology, affordable and culturally acceptable to the people who would use it. To come to terms with the soaring costs of medical care in affluent countries as well as the lack of funds in developing countries, WHO placed preventive—as opposed to curative—medicine in the forefront. The concept of primary health care, with its emphasis on individual responsibility for health and its conviction that the best help is self-help, began to take shape.

The age of technical paternalism came to a formal close in 1979 when the member states of WHO unanimously adopted the goal of 'Health for All by the Year 2000,' founded on the principles of primary health care

that had been elaborated in 1978 during the International Conference on Primary Health Care at Alma-Ata (Kazakstan). Commitment to this time-limited goal guided much of the organization's work over the next decade, though the promotion of research, particularly on disease prevention and control, continues in full force as part of WHO's global plan to push the world forward through the protection and promotion of health (see below).

Structure

WHO is a specialized agency of the UN, as provided for in the *Charter of the United Nations*. A goal-oriented organization with policies, program, and budget defined through well-developed mechanisms, WHO consists of three constituent bodies:

- The *World Health Assembly*, which is the highest decision-making body, is held in May each year and is attended by delegations from WHO's 191 member states and two associate members. Its main tasks are to decide on major policy matters and to approve the biennial program budget.
- The *Executive Board* consists of 32 persons, acting in their personal capacity, highly qualified in the field of health and designated for a 3-year term by as many member states, which are chosen by the Health Assembly on the basis of equitable geographical distribution. The board, which normally meets twice a year, gives effect to the decisions and policies of the Assembly, while advising it and preparing its agenda.
- The *Secretariat* serves to carry out the decisions of the World Health Assembly and the Executive Board; it is the entire staff of WHO headed by the director-general, who is appointed as its chief technical and administrative officer for a 5-year term by the World Health Assembly on the nomination of the Executive Board.

In general, all technical activities that are of universal applicability—such as biological and epidemiological standardization, the overall assessment of the efficacy of methods and materials, and promoting the control of disease—are the responsibility of the headquarters in Geneva (**Figure 2**). WHO's highly decentralized structure enables it to respond directly to the needs of its membership, upon request, through its six regions, each consisting of a regional committee and a regional office. The regional offices, with their own directors, are responsible for formulating policies of a regional character and for monitoring regional activities. In many countries there is a resident WHO representative who is the main intermediary for support of WHO and who participates with the government in

Figure 2 The headquarters of WHO in Geneva, Switzerland. Photograph by T. Farkas.

planning and managing national health programs. The location of the six regional offices and the member states covered are shown in **Figure 3**.

Some 40% of WHO's 4300 staff members, including PAHO, work in countries all over the world, either in field programs or as WHO representatives; 30% are in the six regional offices and 30% at headquarters in Geneva.

WHO's normative, i.e., standard-setting, functions also include preparation and updating of the *International Classification of Diseases*, assignment of generic names for pharmaceuticals, and, since 1957, evaluating the safety for human consumption of selected food additives and contaminants in food and establishing acceptable daily intakes for these substances through the Joint FAO (Food and Agriculture Organization)/WHO Expert Committee on Food Additives.

The committee's reports, as well as those of a similar FAO/WHO group responsible for evaluating the safety of pesticide residues, are used in the formulation of national food legislation intended to protect consumers from hazardous additives or contaminants and by the Codex Alimentarius Commission—another joint FAO/WHO body—in establishing international food standards. (Food legislation is one of the many topics regularly covered by one of WHO's half-dozen specialized international periodicals, the quarterly *International Digest of Health Legislation*.) Codex originated at a time—the early 1960s—when international efforts were being made to increase world trade by reducing tariff barriers, as well as nontariff barriers resulting from differing food regulations. Consistent with a dynamic system that is still changing to deal with ever-changing circumstances, the international community has decided to use health-related Codex standards, guidelines, and recommendations as a reference in implementing relevant aspects of the trade agreements administered by the World Trade Organization (WTO) since 1995.

While the member making the largest contribution to the WHO regular budget is assessed at a maximum 25%, members making the smallest each pay 0.01%. Apart from its regular budget—US$842 654 000 for the biennium 1998–1999—WHO receives voluntary contributions from both governmental and nongovernmental sources. In recent years the total amount of these contributions has been roughly equivalent to regular budget levels. They include contributions for fostering research in tropical diseases and human reproduction, improving community water supply, expanding immunization, preventing and controlling diarrheal diseases, leprosy, malaria, and yaws, and preparing a credible emergency health response to disasters and natural catastrophes.

Working with Others

From its beginning, WHO set out to work not through its small staff alone but with and through others. Many thousands of individual researchers

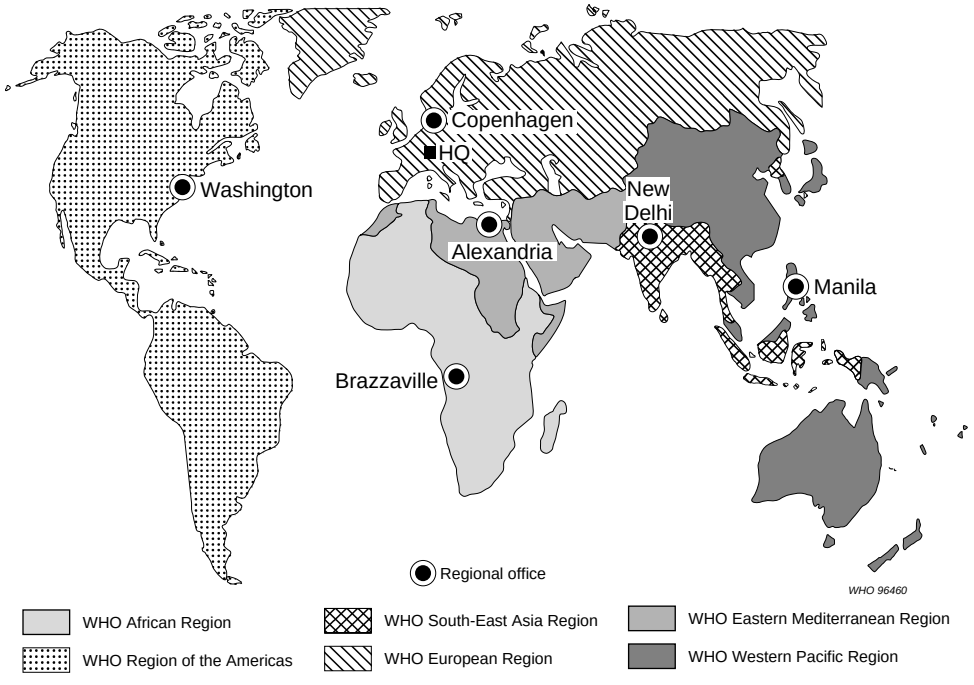


Figure 3 WHO regional offices and the areas they serve.

and scientists, including Nobel laureates, have put their talents at the disposal of WHO—and their number continues to grow. The same is true of WHO collaborating centers, which have grown steadily in number and breadth of disciplines. Rather than duplicate efforts, WHO entrusts critical technical functions to established laboratories and research institutes.

From the outset, WHO was also mandated to work with other agencies within the newly created UN family of organizations. Food and nutrition work, for example, quite naturally came to involve the FAO, as did work against animal diseases and, as mentioned above, in the area of food additives, food standards, pesticide residues, and contaminants. Later, WHO joined forces with the International Labor Organization (ILO) and the United Nations Environment Program (UNEP) in establishing the International Program on Chemical Safety. Recognition of the increasing seriousness of the global burden of malnutrition led WHO and FAO to convene jointly, in Rome in December 1992, the International Conference on Nutrition, which was attended by over 1000 representatives of 159 member states and the European Union. The resulting World Declaration and Plan of Action for Nutrition pledged to eliminate or substantially reduce the major forms of malnutrition and their contributing factors before the end of the decade. The declaration's nine goals for the year 2000 and the strategy and actions of the Plan of Action serve as the platform for WHO's support to countries, especially those most in need, in five priority areas:

- assessment, prevention, and management of protein–energy malnutrition;
- overcoming micronutrient malnutrition (chiefly iodine deficiency disorders, and vitamin A and iron deficiencies);
- improvements in infant and young child feeding (breast feeding and complementary feeding);
- nutrition emergencies, particularly training in preparedness and management;
- prevention of diet-related noncommunicable diseases (including obesity, cardiovascular diseases, and some cancers) and food-related communicable diseases (including diarrhea and parasites).

The views of the United Nations Educational, Scientific and Cultural Organization (UNESCO) are regularly sought, for example, on questions relating to bioethics and the health of schoolchildren. Occupational health is likewise a shared activity with the ILO, while drug dependence and abuse call for collaboration with the United Nations International Drug Control Program. After nearly a decade of providing direct financial support and technical guidance for AIDS (acquired immunodeficiency syndrome) activities in more than 150 developing countries, in January 1996 WHO became one of the cosponsors, together with the United Nations Children's Fund (UNICEF), the United Nations Development Program (UNDP), the United Nations Population Fund (UNFPA), UNESCO, and the World Bank, of the Joint United Nations Program on HIV/AIDS (UNAIDS). Building on the relationship already established with the General Agreement on Tariffs and Trade (GATT), WHO works with the GATT's successor, the WTO, in connection with health-related Codex standards, guidelines, and recommendations.

One of WHO's closest partners has been UNICEF, with which, for example, the early yaws and malaria campaigns were carried out. More recently, joint activities include support to countries in preventing and controlling micronutrient malnutrition and improving infant and young child feeding practices by promoting breast feeding and appropriate complementary feeding practices with emphasis on using locally available foods. To guide their concerted efforts in all fields, since 1948 the two agencies regularly confer on matters of joint health policy.

Collaboration was also initiated with professional, charitable, and other nongovernmental organizations (NGOs) pursuing aims consonant with those of WHO. By the end of the first decade, WHO had established official relations with no fewer than 40 such bodies, ranging from the International Council of Nurses to the International Commission on Radiation Units and Measurements, and from the World Federation for Mental Health to the International Union of Nutritional Sciences. Work of vital importance for WHO's technical programmes of support for its membership has been made possible through the enthusiasm and resources of these and other valuable organizations, which have in turn benefited from the moral support and the technical information provided by WHO.

Collaboration continues unabated; examples include joint work on cancer pain relief with the International Association for the Study of Pain; efforts to ensure safe blood and blood products undertaken by WHO's Blood Safety unit together with the blood programs of the International Federation of Red Cross and Red Crescent Societies, the International Society of Blood Transfusion, and the World Federation of Hemophilia; support for polio eradication provided through Rotary International; reinforcement of technical support to member states for the prevention and control of iodine deficiency disorders with the help of the

International Council for Control of Iodine Deficiency Disorders; and implementation of the Joint WHO/UNICEF Baby-Friendly Hospital Initiative—which strives to ensure the world over a health care environment for new-borns where breast-feeding is the norm—with the help of La Leche League International and the International Lactation Consultant Association. The success of these joint ventures is most strikingly illustrated by the ever-lengthening list of NGOs admitted into official relations with WHO, which now numbers more than 190.

Gearing Up for the Twenty-First Century

Using pragmatic tools in pursuit of its visionary goal, WHO strives to respond to its constitutional mandate through its evolving program of activities. These are based on the needs and priorities of activities determined by member states themselves through the organization's governing bodies, the World Health Assembly and the Executive Board. Changes in program emphasis occur in response to altered political, social, economic, and environmental realities.

For example, respecting the new issues and priorities of the times for WHO includes meeting the challenge implicit in redirecting the human and financial resources that, logically, should be released by the breakup of old Cold War alliances; forging new partnerships to mobilize social, political, and therefore financial support for health development and international health cooperation; and opening itself up still further to all sectors of society, including NGOs and the private sector. It means grappling with the multiple consequences for health of accelerating social and demographic changes, including population growth and encroachment on forest areas and other ecological zones, aging population structure, migration—including the mass movement of disaster-affected populations—and urbanization. It calls for dealing creatively with the impact of the indebtedness, incurred over time, that has led many countries to reduce public spending in health services, which are often regarded as being only an expenditure rather than an investment in human potential. Ironically, this last challenge comes at the very moment when WHO itself is operating within severe financial constraints. After working with a zero-growth budget policy since 1984–1985, WHO's 1996–1997 budget suffered a drastic reduction estimated at 14% in real terms, while the number of staff worldwide dropped from just under 5400 in 1990 to 4300 in 1998. In contrast, the extraordinary changes in the political landscape during the same period saw the number of WHO member states increase from 166 to 191.

WHO must face up to threats to human health that respect no national boundaries, such as uncontrolled dumping of toxic wastes and pollution of land, water, and air, and prodigious consumption and mismanagement of natural resources. The solution of these problems hinges, to a large extent, on the degree to which WHO is able to harness the interdependence of nations and peoples as a positive force for conciliating between competing present and future needs.

In the face of changing demands, WHO is doing its utmost to preserve its inclusive approach to health and to emphasize the continuity between prevention, care, rehabilitation, and health promotion for all people through the different stages of their lives. In restructuring its programs and activities in the light of resource constraints, WHO is placing emphasis on meeting the most pressing health needs based on the following priorities:

- those which present a health emergency;
- those which affect the poorest countries and the most vulnerable groups;
- those which produce the heaviest burden of death, suffering, and disability;
- those which represent a major impediment to social and economic development.

In 1995 the Executive Board and the World Health Assembly, as part of an overall global reform process, identified the need to review the basic primary health care principles of Alma-Ata, recommit member states to those principles, and renew aspects of the strategy to achieve *Health for All* in the light of changing global circumstances. Since 1978 many countries not present at Alma-Ata had attained statehood; a generation of health workers had graduated; and several key determinants of health—social, political, economic, demographic, and epidemiological—had profoundly affected the health profile of populations and the level of inequalities between various subgroups. Moreover, the opportunity for improving health through multisectoral approaches, application of appropriate technology, and greater emphasis on participatory approaches all require that countries, regions, and the international community look afresh at how international health policy for this century can truly improve the long-term health status of the world's poorest countries and communities.

Food, Nutrition, and World Health—Meeting the Global Challenge

Freedom from hunger and malnutrition, essential to the enjoyment of the highest attainable standard of

health, is among the fundamental rights of every human being. What is more, there can be no sound social and economic development without adequate food and nutrition. This is not the same as saying that people in rich countries necessarily have a better chance of being properly nourished than do people in poor countries. There is much more to achieving healthy nutritional status than can be conveyed by such simplistic labels as 'developed' and 'developing' where countries—and most of all people—are concerned.

Even if there is some truth in the axiom 'we are what we eat,' it is clear that nutritional status—a characteristic common to all that can be measured and monitored—depends on considerably more than diet. For individuals, it is best understood as the result of the complex interaction between health at any given moment, the food that is eaten, and the surrounding physical, social, and economic environment. Nutritional status not only reflects the quantity of available food but also its quality, including safety, while showing to what extent the body can transform food into nutrients that will protect and promote health and permit people to function to the best advantage. Because the environment and its impact vary so greatly from individual to individual, there can be no 'standard' answer to the problem of malnutrition. Moreover, no single strategy to combat it will produce the same results in every case. The approach to preventing and overcoming malnutrition thus has to be tailored to fit the circumstances.

Nutrition and the Sweep of History

Looking back in history, one can see that a major shift is taking place in the impact on humanity of the main factors—poor diet and ill health—that have traditionally accounted for most malnutrition. Over the centuries the human species survived hand-to-mouth on whatever it could manage to hunt, gather, harvest, or hoard. The diet which fuelled most of human evolution was low in fat and very low in sugar but high in fiber and other complex carbohydrates.

In the overall context of human development, it is only recently that people in high-income countries could stop worrying about the threat of occasional hunger and increasingly indulge in preferred foods owing to radical improvements in methods of food production, processing, storage, and distribution. Only in recent decades, as people benefited from greater control of infectious disease and better access to safe food, have research findings confirmed the well-founded suspicion that dietary preferences may influence the onset of several major chronic diseases, including coronary heart disease, stroke, various cancers, diabetes mellitus, gastrointestinal disorders, and various bone and joint diseases. Although many dietary factors have been investigated, those most frequently linked to such diseases figure prominently in a pattern of eating typified by the high consumption of energy-dense foods of animal origin and of foods processed or prepared with added fat, sugar, and salt.

As a result, grade 2 overweight, which warrants close attention, is relatively common in industrialized and many developing countries—up to 20% of Europeans and of Whites in the US 60 years of age are affected. The figure increases to 40% for women in eastern European and Mediterranean countries and Black women in the USA. Even higher prevalences are observed among American Indians, Hispanic Americans, and Pacific islanders. While the prevalence of grade 2 overweight is much lower in some African and Asian countries—the range is about 3% to 17%—in South America and the Caribbean it is close to that in many European countries, about 25%. WHO estimates that major noncommunicable diseases (NCDs) are responsible for at least 40% of all deaths in developing countries and 75% in industrialized countries, where cardiovascular diseases are the first cause of mortality and cancer is the third. By the year 2020, NCDs will account for about three-quarters of all deaths in the developing world.

Meanwhile, over 800 million people still cannot meet basic needs for energy and protein, more than 2000 million people are deficient in essential micronutrients, and an estimated 174 million children under 5 years of age in developing countries are malnourished, as indicated by low weight for age, while 230 million are stunted. Malnutrition results in poor physical and cognitive development as well as lower resistance to illness. Nearly half of the estimated 11 million deaths occurring annually among children under 5—or about 49% of young child mortality in developing countries—are associated with malnutrition.

Simultaneous Action on Two Overlapping Fronts

For these reasons, to promote healthy nutrition for all people it is necessary to take simultaneous action on two distinct, if overlapping, fronts. On the one hand, many less-favored nations remain handicapped by a formidable array of development constraints, including rapidly

increasing population, unproductive agriculture, environmental degradation, limited health service coverage, and war and civil strife. Among the most visible—and tragic—consequences are the many millions of wasted and stunted children who do not have enough protein and energy in their diets, who suffer from cretinism and other permanent brain damage because their diets and those of their parents are deficient in iodine, or who go blind or even die for lack of vitamin A. It is in just such environments that diarrheal diseases resulting from contaminated food and water, frequently compounded by seasonal or chronic food shortages, take their heaviest toll in terms of malnutrition, ill health, and premature death.

On the other hand, there has been a significant drop in recent years in the prevalence of infectious disease, while food availability and the quality of diets have improved for populations the world over. The result in many countries has been a sharp reduction in infant and child mortality and longer adult life expectancy. These and related factors have paved the way for a dramatic expansion of a different type of nutrition crisis: Diet-related chronic disorders are now flourishing in environments where, not so long ago, infectious diseases were the greatest menace to health. Too often, one type of malnutrition is being exchanged, or—worse still—being superimposed upon another, with no net gain for human health in the process.

Labels such as 'rich' or 'poor' and 'developed' or 'developing,' by themselves, provide little insight into what causes malnutrition and how it can be overcome. The fact is, wherever people reside, it is the *way they live, what they eat*, and *how they interact with their environment* that determine their nutritional status. Healthy nutritional status is by no means the monopoly of rich countries, any more than malnutrition is somehow the prerogative of poor ones. Whatever the cultural influences at work in a given milieu, all governments are challenged to develop food and nutrition policies that will make healthy choices the easy choices for their populations. Consistent with the unique normative, scientific, and advisory role that WHO has played for the last half century, the organization strives to support all its 191 member states in doing just this.

See also: **Aging**. **Anemia**: Iron-Deficiency Anemia; Megaloblastic Anemia. **Antioxidants**: Intervention Studies. **Appetite**: Physiological and Neurobiological Aspects; Psychobiological and Behavioral Aspects. **Bioavailability**. **Body Composition**. **Breast Feeding**. **Cancer**: Epidemiology and Associations Between Diet and Cancer; Epidemiology of Gastrointestinal Cancers Other Than Colorectal Cancers; Epidemiology of Lung Cancer; Effects on Nutritional Status. **Carotenoids**: Chemistry, Sources and Physiology; Epidemiology of Health Effects. **Coronary Heart Disease**: Hemostatic Factors; Lipid Theory; Prevention. **Diabetes Mellitus**: Etiology and Epidemiology; Classification and Chemical Pathology; Dietary Management. **Diarrheal Diseases**. **Dietary Guidelines, International Perspectives**. **Energy**: Balance. **Energy Expenditure**: Indirect Colorimetry. **Food Fortification**: Developed Countries; Developing Countries. **Food Safety**: Mycotoxins; Pesticides. **Growth and Development, Physiological Aspects**. **Infants**: Nutritional Requirements; Feeding Problems. **Infection**: Nutritional Interactions. **Iodine**: Physiology, Dietary Sources and Requirements; Deficiency Disorders. **Iron**. **Lactation**: Physiology; Dietary Requirements. **Low Birthweight and Preterm Infants**: Nutritional Management. **Malnutrition**: Primary, Causes Epidemiology and Prevention; Secondary, Diagnosis and Management. **Nutrient Requirements, International Perspectives**. **Nutrition Policies In Developing and Developed Countries**. **Nutritional Assessment**: Anthropometry; Biochemical Indices; Clinical Examination. **Nutritional Surveillance**: Developed Countries; Developing Countries. **Obesity**: Definition, Etiology and Assessment; Fat Distribution; Childhood Obesity; Prevention; Treatment. **Protein**: Synthesis and Turnover; Requirements and Role in Diet; Digestion and Bioavailability; Quality and Sources; Deficiency. **United Nations Children's Fund**. **Vitamin A**: Physiology.

Further Reading

WHO (1987) *Principles for the Safety Assessment of Food Additives and Contaminants in Food*.

WHO (1988a) *Four Decades of Achievement: Highlights of the Work of WHO*.

WHO (1988b) *A Guide to Nutritional Assessment*.

WHO (1990a) *Diet, Nutrition and the Prevention of Chronic Diseases*, World Health Organization Technical Report Series 797.

WHO (1990b) *Principles for the Toxicological Assessment of Pesticide Residues in Food*.

WHO (1991a) *Food Additives, Food Contaminants and Veterinary Drug Residues in Food*, A complete guide to reports issued by the Joint FAO/WHO Expert Committee on Food Additives.

WHO (1991b) *Joint FAO/WHO Expert Committee on Food Additives: Index of Substances Evaluated from 1st to 37th Meeting* (1957–1991).

WHO (1991c) *Strategies for Assessing the Safety of Foods Produced by Biotechnology*.

WHO (1992) *Hazard Analysis Critical Control Point Evaluations. A Guide to Identifying Hazards and Assessing Risks Associated with Food Preparation and Storage*.

WHO (1994a) *An Evaluation of Infant Growth*.

WHO (1994b) *Indicators for Assessing Iodine Deficiency Disorders and Their Control through Salt Iodization*.

WHO (1994c) *Safety and Nutritional Adequacy of Irradiated Food*.

WHO (1995a) *Physical Status: The Use and Interpretation of Anthropometry*, World Health Organization Technical Report Series 854.
WHO (1995b) *Vitamin A Deficiency and Its Consequences. A Field Guide to Detection and Control.*
WHO (1995c) *Global Prevalence of Vitamin A Deficiency.*
WHO (1996a) *The World Health Report 1996. Fighting Disease, Fostering Development.*
WHO (1996b) *Trace Elements in Human Nutrition and Health.*
WHO (1996c) *Indicators for Assessing Vitamin A Deficiency and Their Application in Monitoring and Evaluating Intervention Programmes.*
WHO (1996d) *The International Code of Marketing of Breast-Milk Substitutes.* A common review and evaluation framework.
WHO (1997a) *WHO Global Database on Child Growth and Malnutrition.*
WHO (1998a) *Obesity: Preventing and Managing the Global Epidemic.*
WHO (1998b) *Safe Vitamin A Dosage during Pregnancy and Lactation.*
WHO (1998c) *Preparation and Use of Food-Based Dietary Guidelines.*

Y

Yogurt *see* **Dairy Products**. **Functional Foods**: Health Effects and Clinical Applications; **Microbiota of the Intestine**: Probiotics; Prebiotics

Z

ZINC

Contents
Physiology
Deficiency in Developing Countries, Intervention Studies

Physiology

H C Freake, University of Connecticut, Storrs, CT, USA

Introduction

Zinc is only moderately abundant in nature, ranking 23rd of the elements. Of the trace elements in the body, it is second only to iron, but, in contrast to iron, it has a single redox state. Together with its size and charge characteristics, this has led to its widespread use in proteins of the body. The number of zinc proteins is unknown but growing, and they include numerous enzymes and many more nuclear proteins that regulate gene expression. Further sets of proteins are responsible for zinc homeostasis. The binding sites and functions of zinc within some of these proteins are well understood, but for others these are less clear. In particular, the links between the biochemical roles of zinc within proteins and its physiological functions are often obscure. The range of physiological functions of zinc is broad and can be observed in all tissues of the body. In general, zinc is required for DNA synthesis, cell division and growth, for protein synthesis and macronutrient metabolism, and for the development and appropriate function of most body systems. The lack of an appropriate assessment tool makes it difficult to estimate the prevalence of zinc deficiency, but undiagnosed marginal zinc deficiency may be a concern.

The History of Zinc as a Nutrient

The essentiality of zinc for bacterial growth has been known for almost 150 years. Later, it was shown to be required by plants and then, in 1934, by rats. In the succeeding years, the essentiality of zinc was demonstrated for other species. The fact that zinc is used widely by plants and animals and is therefore reasonably widespread in the food supply led to the position that human zinc deficiency was unlikely. It was not until the early 1960s that Prasad and others in Iran described a syndrome of dwarfism and lack of sexual development in teenage boys and young adults. The young Iranian men consumed a diet based on unleavened bread with very little animal protein and also ate large amounts of clay (geophagia). They were anemic and responded to treatment with ferrous sulfate coupled with a more balanced diet including animal protein. The other symptoms also resolved, but it seemed unlikely that lack of iron itself was responsible. Prasad then moved to Egypt, where he encountered a similar syndrome. His Egyptian patients were not geophagic, but they ate mostly bread and beans and were infested with *Schistosoma* and hookworm. Zinc deficiency was documented in these individuals, and treatment with zinc was more effective at increasing growth rates than either iron supplementation or a diet including animal protein. Thus, dietary zinc deficiency was

demonstrated, presumably due to impaired absorption because of the high fiber and phytate contents of the diet. While severe zinc deficiency is not a frequent problem in developed countries, the prevalence of milder symptoms is unknown. In the USA, mild symptoms of zinc deficiency have occasionally been reported (see below under Human Zinc Deficiency).

Chemistry of Zinc

The conjunction of the chemical properties of zinc underlies its biological significance. It is a relatively small ion (with an atomic number of 30) and carries a positive charge of two. It attracts electrons as a strong Lewis acid, and this property can be important in its catalytic functions. It has relatively flexible coordination geometry and, while binding its ligands with high affinity, exhibits rapid rates of exchange, which can facilitate chemical reactions and biological processes. All this is coupled with its single redox state, in contrast to the multiple redox states of iron and copper, which eliminates the danger of oxidative damage. While other trace elements may share some of these properties, none share them all. This is what makes zinc so valuable for protein structure and function.

Zinc in Foods

Zinc is associated with proteins in the body and is found associated with proteins in food. Thus, protein-rich foods tend to be good sources (**Table 1**). However, there is great variability, from egg whites, which have almost no zinc, to oysters, at $750\,mg\,kg^{-1}$. The physiological function of these high concentrations in oysters is unknown, though the zinc is concentrated in cells thought to serve a phagocytic/host defense function. In addition, the bioavailability of zinc may be quite variable, owing to other food components eaten at the same time. Grains and legumes may be relatively rich sources, but bioavailability is limited owing to their phytate content. On the other hand, animal proteins appear to enhance zinc absorption.

Table 1 Dietary sources of zinc

Food	*Zinc content ($mg\,kg^{-1}$ raw weight)*
Oysters	750
Beef, lean	59
Pork	26
Chicken breast	8
Chicken leg	18
Salmon	4
Egg, whole	11
Egg white	0.3
Milk, whole	4
Cheese, cheddar	31
Wheat, whole flour	29
Wheat, white flour	7
Rice, brown	20
Rice, polished	12
Kidney beans	27
Lentils	36
Potatoes	3
Broccoli	4
Apples	0.4

Control of Zinc Homeostasis

The size and charge characteristics of zinc mandate the use of carriers to traverse biological membranes. Two families of transporters have been described and partially characterized. The ZIP family (ZRT (zinc-regulated transporter)- and IRT (iron-regulated transporter)-related proteins, named after homologous transport proteins in yeast and plants) appears to move zinc into the cytoplasm of the cell, either from outside the cell or from subcellular compartments. The second group of transporters, the CDF (cation diffusion factor) family, is responsible for zinc egress from the cytoplasm. This latter family includes ZnT-1, which has been localized to plasma membranes and functions as a cellular efflux protein, and ZnT-2, which transports zinc into storage vesicles under conditions of high cellular zinc. Collectively, the ZIP and CDF proteins are likely to underlie the homeostatic control of zinc distribution around the body.

Zinc Absorption

The absorption, distribution, and excretion of zinc are shown in **Figure 1**. Overall, about 20–40% of consumed zinc is absorbed, depending on the bioavailability within the particular food source. Zinc is absorbed by both saturable and non-saturable processes, with the greatest rates of absorption occurring in the jejunum. Absorption is adjusted to meet needs, being proportionately increased in deficiency states and reduced when intake is high. Zinc status is reflected by the intestinal concentration of the zinc-binding protein metallothionein (MT). MT may trap zinc within the epithelial cells, causing it to be lost in feces as the cells are sloughed off. This may be part of the explanation of how zinc absorption is adjusted to meet needs. Acrodermatitis enteropathica is an autosomal recessive condition of zinc malabsorption, which can lead to severe deficiency. The gene alteration that leads to this condition has recently been identified in *ZIP4*, which encodes one of the zinc transporters. This protein has been localized to the apical membrane of intestinal epithelial cells and, given the severity of the symptoms associated with its inactivation,

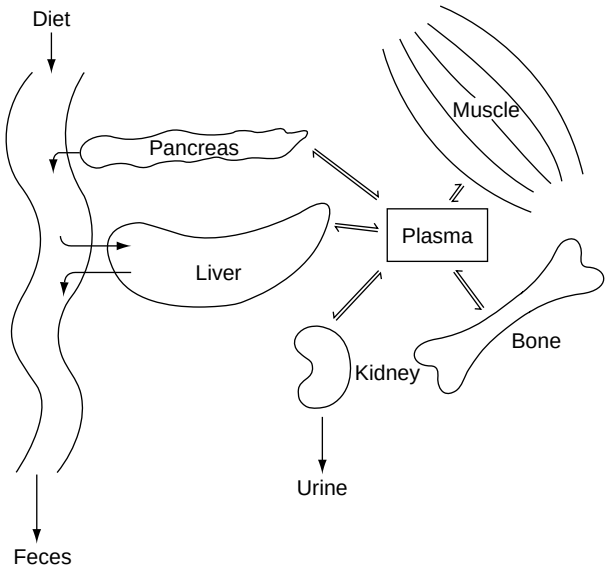


Figure 1 Whole-body zinc homeostasis. Zinc in the intestine comes from the diet and from endogenous secretions. A portion is absorbed, but much is lost in the feces, which are the major route of excretion. Absorbed zinc passes through the liver and then to the general circulation. Zinc is distributed throughout the body, with muscle and bone constituting the largest pools. A minor but controlled amount of zinc is lost in the urine.

appears to be necessary for normal zinc absorption. The zinc efflux protein ZnT-1 is found at the basolateral membrane and probably promotes the passage of zinc out of the intestine. Acrodermatitis enteropathica can be treated with large doses of zinc, supporting the existence of paracellular transport at high intake levels. A large amount of zinc is secreted into the gut from the pancreas and intestine (**Figure 1**). Malabsorption syndromes can lead to a failure to reabsorb these endogenous secretions and, hence, to rapid loss of body zinc.

Transport and Distribution

The zinc pool in plasma is relatively small, representing only about 0.1% of total body zinc. It circulates bound to albumin and α-2-macroglobulin, and about 3% is complexed with amino-acids. About fivefold greater amounts of zinc are found in whole blood, with erythrocytes accounting for about 75% of the total. However, about 85% of erythrocyte zinc is complexed within carbonic anhydrase and therefore does not exchange easily. The egress of zinc from the circulation across endothelial cells and into tissues of the body is not well understood. Uptake in association with albumin has been suggested, but members of the ZIP family of transporters are likely to play a role here. The tissue distribution of zinc is relatively uniform. All cells require the mineral, and no cell stores it. The concentration of zinc in the adult human is about $0.5\,\mu mol\,g^{-1}$, giving a total body content of about 2 g. More than half is found in skeletal muscle, and about 30% is found in bone. The bone pool appears to be more labile than the muscle pool and has been used as an index of zinc status in experimental animals. The liver represents another labile pool. It receives dietary zinc from the portal circulation and contains about 5% of body zinc.

Excretion

Zinc is lost from the body primarily through the feces (**Figure 1**). Feces contain unabsorbed dietary zinc, zinc contained within intestinal epithelial cells that have been sloughed off, and endogenous zinc secretions into the gut from the pancreas, the gall bladder, and the cells lining the gastrointestinal tract. The endogenous secretions and the extent to which they are reabsorbed can be controlled and constitute an important homeostatic mechanism for regulating zinc status. Zinc losses in urine are relatively minor but do respond to extremes of intake to help maintain homeostasis. Shed skin cells, sweat, hair, menstrual blood, and semen are additional routes of loss.

Zinc Biochemistry

Zinc homeostasis and action involve an intimate association of the mineral with proteins. These proteins include membrane transporters responsible for the absorption of zinc in the gut and its passage into and out of cells and subcellular organelles, transport and delivery proteins (both in the circulation and within cells), sensing proteins that will adjust homeostasis and function according to zinc availability, and a large range of proteins to which zinc is ultimately delivered. Two major classes of these latter proteins are the enzymes and transcription factors. In addition to its association with proteins, zinc within cells is also found associated with membrane lipids and both DNA and RNA. The functions of these pools of zinc are not clear.

Homeostasis

The interaction of zinc with its transporters has not been well characterized, though transmembrane domains have been identified that are thought to be responsible for the transport function. Free concentrations of zinc within the cell appear to be extremely low and may not constitute a sufficient pool for the supply of zinc to its protein ligands. This suggests the existence of delivery proteins, and this role has been suggested for MT, which has been shown to transfer zinc to apoenzymes *in vitro*. MT

was originally discovered as a cytoplasmic heavy-metal-binding protein, which was thought to prevent metal toxicity within cells. Additional more significant roles were suggested by the realization that there are multiple MT genes, which have been conserved through evolution. It is a small protein that is unusually rich in cysteine and can bind seven atoms of zinc. MT may influence the subcellular distribution and availability of zinc, since its own distribution varies. For example, the nuclear content of MT varies with the cell cycle. MT expression is regulated not only by heavy metals but also by a range of other signals including glucocorticoids, interleukins, and cyclic adenosine monophosphate. In addition, its zinc-binding activity is influenced by the cellular redox state. For example, an increase in the glutathione disulfide–glutathione ratio results in the release of zinc from MT and thus an increase in its availability for other proteins. However, deletion of individual MT genes in mice has not resulted in major pathologies, questioning the significance of these proteins.

Investigation of the mechanism whereby zinc regulates the expression of MT led to the discovery of the single protein known to act as a zinc sensor within mammalian cells, MTF-1 (metal response element (MRE)-binding transcription factor-1). MTF-1 binds to MREs in the promoter region of MT and other genes and regulates their expression. The ability of MTF-1 to localize to the nucleus and bind to its target genes depends on its zinc content. Thus, an increase in cellular zinc levels results in greater MTF-1 activity and, consequently, increased expression of its target genes. In addition to MT, which will bind more zinc, these include ZnT-1, which will transport zinc out of the cell. These mechanisms underlying cellular zinc homeostasis are illustrated in **Figure 2**.

Zinc Enzymes

The three-dimensional structures of more than 200 zinc-containing enzymes have now been characterized, and many more enzymes have been identified. All six International Union of Biochemistry classes are represented. Zinc enzymes can be divided into three groups according to the role played by zinc within the protein. In the catalytic group (e.g., carbonic anhydrase), zinc is a direct participant in the catalytic function of the enzyme. The zinc atom is coordinated by three amino-acids from the enzyme and a molecule of water at the active site. In enzymes with structural zinc sites (e.g., protein kinase C), the metal binds four amino-acids within the protein and ensures appropriate folding for bioactivity. Enzymes in which zinc serves a co-catalytic function (e.g., superoxide dismutase) contain two or three zinc atoms, two

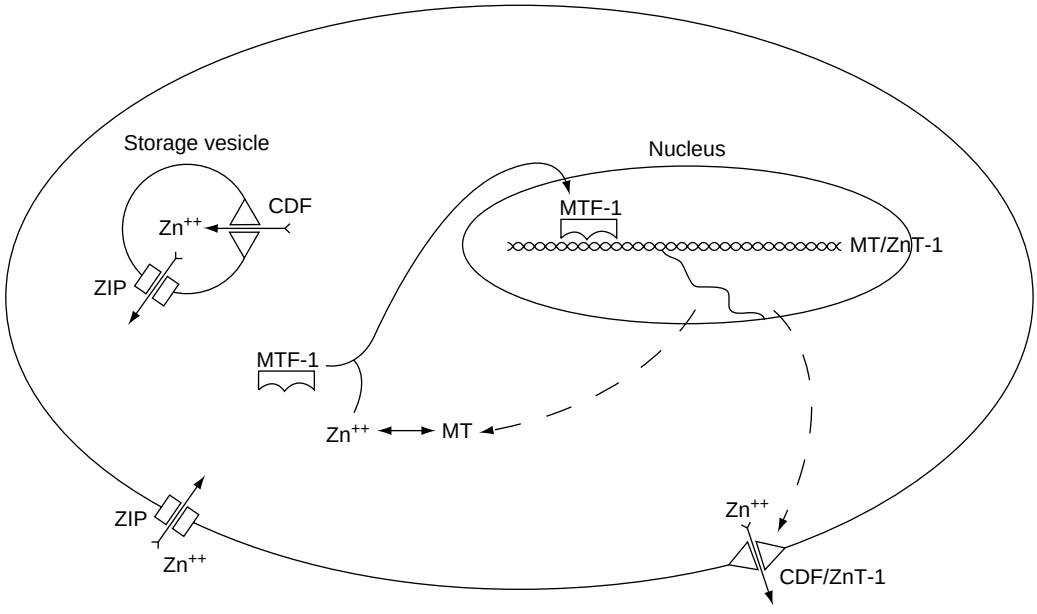


Figure 2 Cellular zinc homeostasis. Zinc is delivered to the cytoplasm from either the extracellular space or vesicles within the cell by members of the ZIP family of transporters. A rise in cellular zinc results in activation and nuclear translocation of MTF-1. In the nucleus, MTF-1 regulates the transcription of a set of target genes, including MT and ZnT-1. MT will bind zinc, and ZnT-1 will transport zinc out across the plasma membrane. MT may govern the delivery of zinc to other proteins within the cell. Other members of the CDF family transport zinc into vesicles.

Table 2 Examples of mammalian zinc-dependent enzymes

Enzyme	*Function*
RNA polymerase	Transcription and synthesis of mRNA
Carboxypeptidase A	Protein digestion in the intestine
Protein kinase C	Signal transduction
Carbonic anhydrase	Respiration, buffering, and hydration of carbon dioxide
Cytochrome *c* oxidase	Respiration and electron transport chain
Alcohol dehydrogenase	Ethanol metabolism
Superoxide dismutase	Inactivation of free radicals
Nitric oxide synthase	Signalling and vasodilation
Angiotensin converting enzyme	Blood-pressure regulation and activation of angiotensinogen

of which are coordinated by a shared amino-acid residue. Coordination sites are widely spaced on the protein, and the zinc may be used for both structural and catalytic functions. In addition to these three groups, zinc has also been found to serve a bridging function between two separate polypeptides to stabilize a biologically active larger complex (e.g., nitric oxide synthase). A selection of zinc enzymes is included in **Table 2**, which helps to illustrate the wide variety of metabolic functions requiring zinc.

Zinc Transcription Factors

There are many zinc enzymes, but there appear to be even more transcription factors that use zinc. These sites have all been identified in the last 20 years and have been less thoroughly investigated than the zinc enzymes. Variable numbers of zinc atoms are each coordinated by four cysteine/histidine residues to stabilize a DNA binding structure. A search of the human genome has revealed over 1000 genes (about 3% of those identified) containing these characteristic zinc finger domains. An important class of zinc finger transcription factors is the steroid/thyroid receptor superfamily, which is responsible for mediating the biological response to a wide range of hormonal and metabolic signals, including retinoic acid and vitamin D. These factors all have nine conserved cysteine residues in the DNA binding region, eight of which are coordinated by two atoms of zinc. Loss of zinc from these sites would interrupt biological function, but it is not clear that this ever happens in a physiological context. Recently, the new array technologies have been used to assess the genome-wide response to changing zinc availability in different tissues, including intestine, liver, and cells of the immune system. The gene products that have been identified as zinc sensitive by these approaches amount to about 5% of the expressed genes within a tissue. They do not necessarily encode zinc proteins themselves but rather proteins whose transcription is altered by zinc. MTF-1 is likely to mediate some but not all of these changes, and other transcription factors whose activity is dependent on zinc may soon be found.

Zinc Physiology

The enormous range of biochemical roles for zinc predicts a large number of physiological functions. The physiological roles of zinc may be further extended to include secondary effects mediated by altered food intake and effects on the functions of other nutrients. While the physiological roles for zinc are well described, it is important to note that the connections between the biochemistry and physiology of zinc remain unclear. Thus, in zinc deficiency the specific zinc-sensitive biochemical step leading to altered physiology is usually unknown. This disconnection will become apparent as the physiological roles of zinc are considered. The broad distribution of zinc through the body at the organ, cellular, and even protein levels suggests that the functions of most systems are dependent upon zinc. Its physiological roles become manifest in cases of deficiency, and that framework will be used to discuss the principal functions here.

Growth

The requirement of zinc for the growth of numerous organisms, ranging from bacteria to humans, is well established. Growth failure is a relatively early consequence of zinc deficiency in experimental animals. Given the lack of a zinc store, in the absence of a sufficient dietary supply zinc will be immediately unavailable for new tissue. Numerous processes seem to contribute to the growth failure. Experiments in animals have shown that zinc deficiency leads to a drop in food intake, though the use of control animals pair-fed an identical amount of a zinc-sufficient diet demonstrates a clear role for a lack of zinc beyond its effects on feeding behavior. The endocrine system is involved with multiple effects of zinc deficiency on the somatotrophic axis, notably a reduction in circulating concentrations of insulin-like growth factor 1 (IGF-1). Again, this appears to be only part of the story since force-feeding a zinc-deficient diet and administering exogenous IGF-1 both fail to correct the growth failure caused by zinc deficiency. Growth of cultured cells is

dependent on media zinc. DNA synthesis is interrupted. Production of thymidine kinase mRNA is diminished by the removal of zinc, but again this appears to be only a partial explanation. The IGF-1 signalling pathway within cells also seems to be affected. Zinc is also required for wound healing, presumably owing to related processes.

Immune Function

The immune system appears to be particularly sensitive to zinc deficiency, in comparison with the rest of the body. Lymphopenia and thymic atrophy are observed, and both cell-mediated and antibody-mediated responses are reduced. As with growth, multiple mechanisms appear to be at play. In addition to its generalized effects on DNA synthesis, zinc deficiency appears to induce apoptosis, resulting in a loss of B-cell and T-cell precursors within the bone marrow. Thymulin is a zinc-dependent enzyme that stimulates the development of T cells within the thymus. The production of cytokines by mononuclear cells is also reduced by zinc deficiency. It appears likely that these effects can be of clinical significance. Infections occur more frequently in individuals with acrodermatitis enteropathica, and reduced immune function is accompanied by zinc deficiency in several other conditions, including sickle-cell anemia and various gastrointestinal disorders. In the USA, zinc lozenges have become popular as a treatment for the common cold. Results from controlled trials of this treatment have been variable, but a shortening of cold duration may occur. It would appear reasonable to suppose that treatment effectiveness would depend on initial zinc status, with greater success being seen in individuals with marginal undetected zinc deficiency.

Reproduction

The original description of zinc deficiency in humans included lack of pubertal development. Spermatogenesis is a zinc-dependent process. Seminal fluid is particularly rich in zinc, and the sperm appear to accumulate zinc from this source prior to ejaculation. Zinc is also crucial for normal fetal development, and deficiency leads to abnormalities in humans and animals. Maternal zinc deficiency has been linked with pregnancy-associated morbidity, including pre-term delivery.

Nervous System

The brain is one of the sites that has been shown to be particularly sensitive to zinc deficiency during fetal development, with neural-tube defects and other disorders being found. While this work was performed in animals, a similar relationship appears likely in humans. Zinc is distributed throughout the brain, but greater concentrations are found within the hippocampus. Here, a brain-specific transporter, ZnT-3, concentrates zinc in vesicles within glutamatergic neurones. It is co-secreted with the neurotransmitter and appears to serve as a modulator of neurotransmission. Very high concentrations of zinc ($>100\,\mu M$) are found within the synaptic cleft during this process. In addition, brain injury resulting from ischemia or trauma causes the release of massive amounts of zinc, which is thought to be responsible for the resultant cell death.

Antioxidant Defense System

Although zinc is not itself an antioxidant, there are several ways in which it participates in the antioxidant defense system of the body, with important implications for health. It can bind to thiol groups in proteins, making them less susceptible to oxidation. By displacing redox-reactive metals such as iron and copper from both proteins and lipids it can reduce the metal-induced formation of hydroxyl radicals and thus protect the macromolecules. Its role in inducing MT has already been mentioned, and this protein scavenges hydroxyl radicals. Increased oxidative stress results in the release of zinc from MT, presumably making it more available for other proteins. Copper/zinc superoxide dismutase is an important zinc-containing antioxidant enzyme whose activity is impaired in the deficient state. In general, animal studies have revealed an association between zinc deficiency and increased oxidative stress. The likelihood of increased oxidative stress under conditions of zinc deficiency suggests a potential anticarcinogenic role for this mineral. This connection is further supported by the finding that the tumor suppressor gene p53, which is frequently mutated in human cancers, is a zinc-containing transcription factor whose expression is also dependent on zinc.

Macronutrient Metabolism

Many of the enzymes of intermediary metabolism contain zinc, and deficiency affects all macronutrients. Protein synthesis and DNA and RNA synthesis require zinc. Insulin is secreted from the pancreas and circulates in association with zinc. This secretion is diminished under conditions of zinc deficiency, leading to impaired glucose metabolism. Lipid metabolism is also affected, with zinc deficiency being associated with reductions in circulating high-density lipoproteins.

Human Zinc Deficiency

In addition to dietary inadequacy, there are several routes that lead to zinc deficiency. Acrodermatitis enteropathica, the genetic disorder of zinc malabsorption, has already been mentioned. Other, more generalized, malabsorption syndromes (e.g., coeliac disease) can also lead to zinc deficiency. Deficiency has also resulted from inappropriate intravenous feeding and the use of chelation therapy. Children are likely to be particularly at risk of zinc deficiency, because of its involvement in growth.

Mild

Given the difficulty of assessing marginal impairments in zinc status, the effects of deficiency can often be verified only by a response to treatment. Growth provides a good example of this. Children in Denver, Colorado, who were of low height for their age increased their growth rates in response to zinc supplementation, whereas zinc had no effect in children of normal height. In addition to improved growth, improvements in immune function, taste and smell acuity, and reproductive function have been noted with zinc supplementation.

Severe

Severe human zinc deficiency has been well characterized by the original descriptions in the Middle East and in patients with acrodermatitis enteropathica. The symptoms of mild deficiency are continued and exaggerated. Thus, stunting can be extreme and is accompanied by delayed sexual maturation and impotence. Characteristic skin lesions are found, originating around the mouth and nose but becoming widespread as deficiency develops. Diarrhea is also present. Deficits in taste and smell are accompanied by anorexia and other behavioral changes, including increased irritability and impaired cognitive function. Eye pathologies similar to those seen in vitamin A deficiency are observed.

Zinc Toxicity

Toxicity of zinc from food sources has not been reported and seems unlikely since absorption is homeostatically regulated. Acute gastrointestinal symptoms and headaches have been reported after ingestion of amounts about 10–20-fold higher than the recommended intakes. Chronic ingestion of these large amounts has been shown to impair immune response and lipoprotein metabolism. However, the key danger of excessive zinc intake is reduced copper status. This is probably due to a zinc-induced blockage of copper absorption and in fact is clinically useful in individuals with Wilson's disease, a condition of copper toxicity. In the USA, an upper limit of $40\,mg\,day^{-1}$ has been set for adults, because of the threat to copper status. The popularity of zinc lozenges for treatment of the common cold could lead to this intake being exceeded. Thus, the use of these treatments should be of limited duration.

Assessment

The prevalence of marginal zinc deficiency in human populations is unknown because of the lack of a good means of assessing zinc status. Measurement of plasma zinc is straightforward, but it does not serve as a reliable indicator of zinc status. Plasma zinc is a quantitatively minor pool that can be easily influenced by minor shifts in tissue zinc. Plasma concentrations do not fall with decreasing dietary intake, except at very low intakes. Plasma zinc can also be affected by factors unrelated to zinc status (e.g., time of day, stress, and infection). Cellular components of blood can be assayed, but erythrocyte concentrations of zinc are maintained in deficient states and variable results have been found with leucocytes. Hair zinc concentrations may reflect available zinc but will also depend on the rate of hair growth.

Several different zinc-dependent enzymes have been investigated as potential markers of zinc status, but none have proved reliable. MT in blood cells has been suggested as a useful indicator of zinc status, assayed at either the protein or the mRNA level. MT expression is likely to be regulated by factors other than zinc and therefore may lack the specificity required of a good indicator. The gene-array approaches that have recently been used to determine the global effects of zinc deficiency within a tissue would appear to offer hope for the identification of an appropriate functional marker of zinc status.

Recommended Intakes

In the absence of a reliable index of zinc status, both the US Food and Nutrition Board and the Food and Agriculture Organization (FAO)/World Health Organization (WHO) Expert Committee used the factorial approach to estimate human zinc requirements. As shown in **Table 3**, the FAO/WHO give three sets of recommendations, depending on the zinc bioavailability of the diet. The US Food and Nutrition Board figures fall between those given for moderate- and low-availability diets. Both groups also set upper limits for intake, based largely on the risk of impairing copper status. These values

Table 3 Recommended intakes of zinc

Age group		*US–Canadian recommended dietary allowance*	*FAO/WHO reference nutrient intake Bioavailability*		
			High	*Moderate*	*Low*
Children (1–3 years old)		3	2.4	4.1	8.3
Adolescents (14–18 years old)	Female	9	4.3	7.2	14.4
	Male	11	5.1	8.6	17.1
Adults (>19 years old)	Female	8	3.0	4.9	9.8
	Male	11	4.2	7.0	14.0
Pregnant women	Third trimester	11	6.0	10.0	20.0
Lactating women	0–3 months	12	5.8	9.5	19.0

are similar (40 mg for the US Food and Nutrition Board, 45 mg for FAO/WHO, for adults).

See also: **Antioxidants**: Diet and Antioxidant Defense; Observational Studies; Intervention Studies. **Bioavailability**. **Children**: Nutritional Requirements. **Cofactors**: Inorganic. **Copper**. **Cytokines**. **Immunity**: Effects of Iron and Zinc. **Inborn Errors of Metabolism**: Classification and Biochemical Aspects. **Nutrient–Gene Interactions**: Molecular Aspects; Health Implications.

Further Reading

Andrews GK (2001) Cellular zinc sensors: MTF-1 regulation of gene expression. *Biometals* **14**: 223–237.

Cousins RJ, Blanchard RK, Moore JB *et al.* (2003) Regulation of zinc metabolism and genomic outcomes. *Journal of Nutrition* **133**(5S-1): 1521S–1526S.

FAO/WHO (2002) *Human Vitamin and Mineral Requirements*. Report of a joint FAO/WHO expert consultation, Bangkok, Thailand, pp. 257–270. Rome: Food and Nutrition Division of the Food and Agriculture Organization.

Gaither LA and Eide DJ (2001) Eukaryotic zinc transporters and their regulation. *Biometals* **14**: 251–270.

Institute of Medicine (2001) Dietary reference intakes for vitamin A, vitamin K, arsenic, boron, chromium, copper, iodine, iron, manganese, molybdenum, nickel, silicon, vanadium and zinc. pp. 442–501. Washington, DC: National Academy Press.

MacDonald RS (2000) The role of zinc in growth and cell proliferation. *Journal of Nutrition* **130**(5S): 1500S–1508S.

Maret W (2001) Zinc biochemistry, physiology, and homeostasis – recent insights and current trends. *Biometals* **14**: 187–190.

Mills CF (ed.) (1989) *Zinc in Human Biology*. London: Springer-Verlag.

Prasad AS (1991) Discovery of human zinc deficiency and studies in an experimental human model. *American Journal of Clinical Nutrition* **53**: 403–412.

Vallee BL and Falchuk KH (1993) The biochemical basis of zinc physiology. *Physiological Reviews* **73**: 79–118.

Deficiency in Developing Countries, Intervention Studies

C Hotz, National Institute of Public Health, Morelos, Mexico

Introduction

Knowledge of the occurrence of zinc deficiency and its importance to human health has increased greatly in recent years. Available evidence indicates that zinc deficiency is an important contributing factor to impaired growth and development, morbidity, and mortality among children in underprivileged settings. Presently, there are few estimates of the prevalence of zinc deficiency in developing countries based on dietary intake or biochemical indices. However, national level estimates of the adequacy of zinc in the food supply and the prevalence of childhood growth stunting can be used to inform on the relative risk of zinc deficiency among countries. National programs to improve zinc status through either supplementation or food fortification are just being initiated.

Recognition of Zinc Deficiency in Developing Countries

The recognition of zinc deficiency as an important contributor to the high rates of morbidity, mortality, and delayed growth and development among

children is relatively recent in contrast to the earlier recognition of the importance and widespread occurrence of deficiencies of iodine, vitamin A, and iron. Coordinated efforts to address vitamin A deficiency in less developed countries were formally initiated by the establishment of the International Vitamin A Consultative Group (IVACG) in 1975. In the mid-1980s, similar groups were founded for the control of iodine deficiency disorders (International Council for the Control of Iodine Deficiency Disorders; ICC/IDD) and iron deficiency (International Nutritional Anemias Consultative Group; INACG). It was not until the year 2000 that a similar group emerged, the International Zinc Nutrition Consultative Group (IZiNCG), to promote the control of zinc deficiency in more vulnerable populations.

The detection of zinc deficiency in populations and the recognition of its association with health outcomes have been somewhat more challenging for zinc than for other nutrients, contributing to the delay in efforts to control it. The ability to diagnose zinc deficiency in individuals using biochemical measures is somewhat limited. For example, the concentration of zinc in serum or plasma may not diminish until the depletion of zinc is more advanced, making it less useful for diagnosing mild to moderate zinc deficiency states in individuals. Other possible biochemical indicators of zinc status have not been consistently demonstrated to reflect change in zinc status. These limitations may have subsequently dampened enthusiasm for evaluating zinc status at the population level. Furthermore, the health conditions that are clearly associated with zinc deficiency (e.g., childhood growth stunting, common childhood infections, and mortality; described in further detail below) are general in nature and have multiple causes. This is in contrast to the strong iconic association of iodine deficiency with goiter and cretinism, vitamin A deficiency with eye disorders and blindness, and iron deficiency with easily diagnosable anemia. The nonspecific nature of health outcomes associated with zinc deficiency is in concordance with the role of zinc in a wide variety of biological functions, covering all human physiological systems. Thus, the very nature of zinc metabolism and the ubiquity of zinc in biological functions at the molecular, cellular, and physiological levels has likely contributed to the difficulties and delays in recognizing the important contribution of zinc deficiency to impaired health and development. A brief history of the knowledge of zinc deficiency in developing countries is presented in **Table 1**.

Table 1 History of knowledge of zinc deficiency in developing countries

Year	*Event or publication*
1963	Relationship between zinc deficiency and hypogonadal dwarfism noted in Egypt
1972	The role of zinc deficiency in hypogonadal dwarfism described in Iran
1974	Zinc demonstrated to increase linear growth, weight, and bone age in Iranian pubertal boys
1982	Supplemental zinc demonstrated to increase linear growth in Chinese preschool children
1993	Supplemental zinc during pregnancy demonstrated to increase birth weight and gestational age of infants in India
1996	Supplemental zinc demonstrated to decrease the prevalence of diarrhea and pneumonia among malnourished children in Vietnam
1999	Pooled analysis of randomized, controlled, zinc supplementation trials indicates a significant positive effect of zinc on reducing the incidence of diarrhea and pneumonia
2000	Establishment of the International Zinc Nutrition Consultative Group (IZiNCG)
2001	Mortality reduced by zinc supplementation among low-birth-weight infants in India
	Zinc supplementation recommended as adjunctive therapy for childhood diarrhea
2002	Meta-analysis of randomized, controlled zinc supplementation trials indicates a modest but significant overall improvement in growth

Causes of Zinc Deficiency in Developing Countries

Although the etiology of zinc deficiency in developing countries has not been thoroughly studied, the main contributing factor is believed to be inadequate intake of zinc in bioavailable (i.e., available for absorption across the intestine) forms.

Inadequate Dietary Intake of Zinc

In general, the risk of inadequate intake of dietary zinc within a population may be associated with the nature of the food supply, and its content and relative bioavailability of zinc. Animal source foods, in particular shellfish, small whole fish, beef, and organ meats such as liver and kidney, are rich sources of zinc. Furthermore, the zinc contained in animal source foods is more highly bioavailable than from plant source foods; the presence of certain amino acids (e.g., histidine, methioinine), or perhaps other unidentified factors, may facilitate the intestinal absorption of zinc from animal flesh foods. Plant source foods, such as most fruits and vegetables including green leaves, and starchy roots and tubers, have relatively low zinc content. While whole grains and legumes have moderate to high

zinc content, these foods also contain large quantities of phytate (phytic acid or myo-inositol hexaphosphate), the most potent identified dietary inhibitor of zinc absorption. The zinc and phytate content, and the phytate:zinc molar ratio of some foods are shown in **Table 2**.

Plants synthesize phytate, which occurs in highest concentration in seeds and to a lesser extent in vegetative plant parts. Phytate forms chelates with zinc and other minerals; as this compound is largely undigested and is not absorbed, it carries the chelated portion of dietary zinc out of the intestine, thus reducing the amount of zinc available for absorption. The phytate:zinc molar ratio of the diet can be used to estimate the bioavailability of zinc. Populations with a heavy dietary reliance on unrefined cereals or legumes, complemented with only small amounts of zinc-rich animal source foods, will have lower intakes of bioavailable zinc. Although milling cereal grains removes large amounts of phytate, it also removes large amounts of zinc. Thus, among populations with a heavy dietary reliance on refined cereals (e.g., rice) or starchy roots and tubers (e.g., potatoes, cassava) with small amounts of zinc-rich animal foods, the total intake of dietary zinc will be low. In either case, low food intakes due to food insecurity will exacerbate the risk of not meeting daily physiological requirements for absorbed zinc.

Table 2 The content of zinc and phytate, and the phytate:zinc molar ratio in uncooked foods

Food	*Zinc (mg/100 g)*	*Phytate (mg/100 g)*	*Phytate:zinc molar ratio*
Cereals			
Corn	1.8	800	44
Pasta	0.7	282	40
Rice (milled)	1.1	352	32
Wheat or whole-wheat bread	2.9	845	29
White bread	0.9	30	3
Nuts and legumes			
Lentils/mung beans	1.3	358	27
Peanuts	3.3	1760	53
Peas	2.9	1154	39
Red beans	2.9	1629	56
Roots and tubers			
Cassava	0.3	54	18
Potato	0.3	81	27
Sweet potato	0.5	50	10
Vegetables			
Cabbage	0.1	0	–
Green leaves	0.2	42	21
Onion	0.2	0	–
Tomato	0.1	6	6
Fruits			
Banana	0.2	0	–
Coconut	1.1	324	29
Orange	0.1	0	–
Mango	0.0	20	–
Animal source foods			
Beef	3.0	0	–
Chicken	1.3	0	–
Eggs	1.1	0	–
Fish	0.5	0	–
Milk	0.4	0	–
Pork	1.9	0	–

Other Causes of Zinc Deficiency

There are a few other commonly occurring conditions in developing country settings that may contribute to zinc deficiency. Diarrhea may not only lead to a reduced absorption of dietary zinc during the episode due to increased intestinal transit time, but may also cause an increase in the loss of body zinc. Under normal physiological conditions, zinc is secreted into the intestine in large quantities together with digestive juices but is largely reabsorbed again; diarrhea may interfere with the reabsorption of this zinc. Given the important role of the intestine in regulating dietary zinc absorption, and the secretion and reabsorption of body zinc during digestion, conditions that affect the health or integrity of the intestine, such as tropical enteropathy, could interfere with the adequate maintenance of zinc balance. The contribution of these conditions to zinc deficiency in developing countries requires investigation.

Prevalence of Zinc Deficiency in Developing Countries: Available Evidence

Relatively little information on population zinc status has been collected at the national or subnational level in developing countries. Thus, only very limited estimates of the prevalence of zinc deficiency are available that are based on the proportion of the population with low concentrations of serum zinc or inadequate dietary zinc intakes. Estimates of the magnitude of risk of zinc deficiency in a population have therefore been derived from more indirect indicators, such as the:

- adequacy of zinc in the national food supply;
- national prevalence of childhood growth stunting; and
- occurrence of a positive response of health conditions to supplemental zinc as determined by randomized, controlled zinc supplementation trials.

Adequacy of Zinc in the National Food Supply

As described above, the nature of the food supply will provide some information on the likelihood of risk of

Table 3 Adequacy of dietary zinc in the food supply in major developing country regions, as compared to North America

	Population (millions)	*Zinc (mg/caput/day)*	*Phytate:zinc molar ratio*	*Zinc from animal source foods (%)*	*Estimated population at risk of inadequate zinc intake (%)*
North America	305	12.5	11	61	10
China	1256	12.4	16	37	14
Latin America and Caribbean	498	10.3	20	42	25
South Asia	1297	10.8	26	11	27
Southeast Asia	504	9.2	24	21	33
Sub-Saharan Africa	581	9.4	26	15	28

Adapted with permission from Food and Nutrition Bulletin (2004) (suppl 2) **25**: S135.
International Zinc Nutrition Consultative Group (Brown KH, Rivera JA, Bhutta Z, Gibson RS, King JC, Ruel M, Sandström B, Wasantwisut E, Hotz C, Lönnerdal B, Lopez de Romaña D, and Peerson J) (2004) Assessment of the risk of zinc deficiency in populations and options for its control. *Food and Nutrition Bulletin* **25**: S91–S202.

inadequate dietary zinc within a population. Information compiled by the United Nation's Food and Agriculture Organization has been used to estimate the potential risk of inadequate zinc in the food supply for a large number of countries. This estimate uses country level data on the per capita amounts of 95 different food commodities available for human consumption, and estimates of the zinc content and phytate:zinc molar ratio of these foods, to calculate the per capita amount of bioavailable zinc in the food supply. The per capita amount of bioavailable zinc is compared to the physiological requirement for absorbed zinc weighted for the demographic distribution of the population. The theoretical proportion of the population at risk of inadequate dietary zinc is used to estimate the relative risk of zinc deficiency at the national level. For example, countries with 25% or more of the population at risk of inadequate dietary zinc are considered to be at elevated risk. This information is limited in that it represents the national average situation and cannot identify subnational populations that may be at elevated risk. In the absence of more direct measures of zinc status, such estimates will justify the need to conduct population surveys that measure risk of zinc deficiency more directly.

Estimates of the proportion of the population at risk of inadequate dietary zinc based on food supply data have been calculated for 176 countries; a summary of the tabulations by developing country region is given in **Table 3**, and compared to those from North America. Overall, these estimates suggest that about 20% of the world's population is at risk of inadequate dietary zinc intake.

National Prevalence of Childhood Growth Stunting

Zinc deficiency is a common limiting factor to adequate child growth in developing country settings. A meta-analysis of 25 studies using a randomized, placebo-controlled design, which measured change in linear growth of children following zinc supplementation for at least 2 months, indicated that supplemental zinc had an overall, positive effect on linear growth. This meta-analysis also demonstrated that a low group mean index of child height-for-age (i.e., 1.58 SD below the reference median for height-for-age) predicts an improvement in linear growth in response to supplemental zinc. Therefore, a high prevalence of childhood growth stunting in a population represents an elevated risk of zinc deficiency. The World Health Organization suggests that when the prevalence of children with height-for-age of 2 SD below the reference median is 20% or higher, childhood growth stunting should be considered a problem of public health concern; this prevalence may likewise be indicative of an elevated risk of zinc deficiency. The World Health Organization maintains a global database on the prevalence of low height-for-age at the national and subnational level for a large number of countries.

Occurrence of a Positive Response of Health Conditions to Supplemental Zinc

Suggestive evidence for the widespread occurrence of zinc deficiency in developing regions is derived from the large number of countries from a wide geographical range where positive health changes were observed in response to supplemental zinc. The health conditions that have been positively affected by supplemental zinc, as demonstrated through randomized, controlled, community-based zinc supplementation trials and the locations of these studies are described in detail in the following section.

Consequences of Zinc Deficiency in Developing Countries: Evidence Derived from Zinc Supplementation Trials

In the context of developing country settings, present knowledge on the health consequences of zinc

deficiency has been almost entirely derived from community-based trials of zinc supplementation among populations at possible risk of zinc deficiency. In these trials, individuals in the study population are randomly allocated to receive either a zinc supplement, usually in the form of tablets or syrups, or the same supplement format without zinc (i.e., placebo). The condition under study is then monitored for a given period (typically for 2 months to one year), and the occurrence of or change in the condition is compared between the zinc-supplemented group and the corresponding control group. Given that several other nutritional and environmental factors can influence the health conditions hypothesized to occur with zinc deficiency, such studies have been essential in demonstrating unequivocally the causal role of zinc deficiency in these conditions among human populations. The following section provides an overview of the population groups at elevated risk of zinc deficiency, and the health consequences associated with zinc deficiency, as concluded from these studies.

Groups at Elevated Risk of Zinc Deficiency

In accordance with age and physiological status, some population groups have increased daily physiological requirements for absorbed zinc. The incorporation of zinc in new tissues being synthesized such as occurs during growth and pregnancy or the secretion of zinc in breast milk during lactation require that relatively larger amounts of zinc are absorbed daily. These increased needs for zinc increase the challenge of acquiring sufficient amounts of absorbable zinc from the food supply. Those groups with higher zinc requirements and who are thus at elevated risk of zinc deficiency include:

- infants (particularly those born prematurely);
- young children;
- children recovering from severe malnutrition;
- adolescents; and
- pregnant and lactating women.

At least some evidence exists for the occurrence of zinc deficiency among each of these groups in developing country settings. The elderly may also be at elevated risk of zinc deficiency, due to a decline in adequacy of zinc intakes and possibly a reduction in the absorption of dietary zinc. However, evidence for zinc deficiency among the elderly has thus far only been derived from industrialized countries; elderly populations have not been the subject of study of zinc deficiency in developing countries.

Growth and Development of Children

Many children in developing country settings experience poor growth, in comparison to relatively healthy children from more developed countries. The prevalence of low height-for-age and weight-for-age indices among children under 5 years of age are used as indicators of poor living conditions, to which poor diet, poor environmental and social conditions, and higher exposure to infectious diseases contribute. Similar conditions can result in impaired neurobehavioral development and cognitive function, putting children in developing countries at further disadvantage. Evidence exists for a specific role of zinc in both of these aspects of child development. **Table 4** provides a summary of countries in which improved growth or development in response to supplemental zinc has been clearly demonstrated.

Growth Zinc plays an important role in child growth. Several mechanisms may be involved, including the role of zinc in the transcription and translation of genetic material and, perhaps more importantly, the regulatory role of zinc in the primary endocrine system, which controls growth (i.e., the growth hormone-somatomedin axis). Specifically, zinc status is associated with the concentration of circulating insulin-like growth factor-1, the principal growth factor that controls early childhood growth. Among populations where growth retardation occurs, both height and weight gain have improved following supplemental zinc. Stimulation of linear growth appears to be the primary response, while the increase in body weight likely reflects the synthesis of lean tissue such as bone, cartilage, and muscle associated with linear growth. This is evident because, in general, weight does not increase independently of increased height in response to supplemental zinc.

The magnitude of improvement in linear growth in response to supplemental zinc is, not surprisingly, greater among children experiencing growth retardation (or 'stunting'; >2 SD below the median height-for-age of international reference data). Zinc deficiency has been demonstrated to be an important limiting factor to growth of children across a wide range of geographical settings in developing regions (**Table 4**). It should be noted that not all studies have demonstrated a significant, positive effect of zinc on growth. Possible explanations for this include: the prevalence or severity of growth stunting in the study communities was low; zinc status was adequate; or deficiencies of other growth-limiting nutrients coexisted thus preventing a positive effect of zinc on growth. The latter

Table 4 Countries from developing regions with documented evidence of improved growth or development in response to supplemental zinc

Region	*Country*	*Population group*	*Development outcome improved*
Eastern Mediterranean	Iran	Pubertal boys	Height, weight, bone age
Latin America and Caribbean	Belize	Preschool children	Height
	Brazil	Low-birth-weight infants	Weight
	Chile	Low-birth-weight infants	Length
		Severely malnourished infants	Length gain
		Preschool children (boys only)	Height
		Preadolescent and adolescent children (boys only)	Height
	Guatemala	Infants (growth stunted)	Length, lean body mass, physical activity
		Preadolescent children	Mid upper arm circumference
	Jamaica	Severely malnourished infants and preschool children	Lean tissue synthesis
South and Southeast Asia	Bangladesh	Infants (low serum zinc concentration)	Weight
		Severely malnourished infants and preschool children	Weight gain
	China	Infants	Length, weight
		Preschool children	Height, weight
		Preadolescent children	Heel-to-knee height[a]
			Neuropsychological performance
	India	Preschool children	Physical activity level
	Japan	Preadolescent children	Height
	Vietnam	Preschool children (growth stunted)	Height, weight
Sub-Saharan Africa	Ethiopia	Infants (growth-stunted)	Length
	Uganda	Preschool and school-aged children	Mid upper arm circumference

[a]An improvement with supplemental zinc was observed only when administered simultaneously with other micronutrients.

situation may also explain the observation in some studies of a transient effect of zinc on growth.

Low-birth-weight infants (<2.5 kg) may have additional needs for zinc, presumably to facilitate their rapid postnatal catch-up growth. Some benefits of supplemental zinc to growth have been observed among low-birth-weight infants in the first 6 months of life.

Severely malnourished infants and children have exhibited improved rates of weight gain, height gain, or synthesis of lean tissue when supplemental zinc has been included in their usual rehabilitation treatment regimen. In these recovering children, zinc has been shown to augment the deposition of lean tissue by increasing protein synthesis.

Cognitive function and behavior There are a few possible mechanisms by which zinc may be speculated to affect neurobehavioral function; these include neurotransmission in the synapses or development of the central nervous system via the synthesis of genetic material, proteins, and cell replication. Adequate zinc status appears to be important for certain aspects of neurobehavioral development among infants and children, although these associations are not conclusive. Higher levels of activity, specifically more frequent engagement in walking or playing as opposed to sitting or watching, have been observed among infants in developing country settings in response to supplemental zinc. Nonetheless, other studies have failed to demonstrate an effect of zinc on motor scores as assessed by Bayley Scales or Griffiths' Developmental Assessment and, in one case, a negative effect on the mental development index was observed.

Evidence for improved cognitive function among school-aged children has been derived from studies of urban and rural children in China. In the rural population of children, the positive effect of zinc on cognitive function was dependent on the provision of other supplemental micronutrients, while in the urban group, supplemental zinc had a positive effect that was independent of the provision of other micronutrients. It is possible that some of the inconsistencies in the studies of neurobehavioral development occur due to concurrent deficiencies of other nutrients that also play a role in cognition (e.g., iodine, iron).

Infectious Diseases Among Children

Zinc has many roles in the immune system, contributing both to specific and nonspecific immune functions. Indeed, there is ample information indicating that zinc deficiency makes an important

Table 5 Countries from developing regions with documented evidence of a reduced prevalence of infectious disease in response to supplemental zinc for prevention

Region	*Country*	*Population studied*	*Health condition*
Latin America and Caribbean	Mexico	Preschool children	Diarrhea
	Guatemala	Infants	Diarrhea
	Peru	Preschool children	Diarrhea
South and Southeast Asia	India	Infants and preschool children	Diarrhea
	India	Infants and preschool children (recovered from acute diarrhea)	Pneumonia
		Infants (term, small-for-gestational-age)	Mortality
	Vietnam	Infants and preschool children (growth stunted and underweight)	Diarrhea Pneumonia
Sub-Saharan Africa	Burkina Faso	Infants and preschool children	Diarrhea
	Ethiopia	Infants	Diarrhea
Western Pacific	Papua New Guinea	Infants and preschool children	Malaria

contribution to some of the most common childhood infections that occur in developing countries, as summarized in **Table 5**.

Diarrhea Zinc has an important role in both the prevention and treatment of diarrhea, which may be mediated both through functions in immune competence and maintenance of the integrity of the intestine. Studies in various settings indicate that provision of supplemental zinc on a nearly daily basis reduces the incidence of childhood diarrhea by nearly 20%, and reduces the prevalence of diarrhea by about 25%. The magnitude of this decrease is similar to that expected from programs to improve water quality and sanitation. There is no strong evidence to suggest that greater benefits of zinc in diarrhea prevention would occur among children who are growth stunted. Rather, all children living under poor conditions with exposure to diarrheal pathogens may potentially benefit from improved zinc intakes.

Zinc also has therapeutic benefits for recovery from diarrheal infections. Overall, supplemental zinc provided to children during recovery from either acute or persistent diarrhea leads to a reduction in the duration and severity of the episode. It has been recommended that zinc be used in the management of acute diarrhea, in conjunction with oral rehydration therapy. The current recommendation is to provide 10–20 mg of zinc once daily for 10–14 days.

Lower respiratory tract infections Zinc deficiency appears to be associated with an increased incidence of pneumonia. Evidence thus far indicates that supplemental zinc reduces the incidence of pneumonia in children by about 40%.

Malaria Only a few studies to date have considered the possible importance of zinc in protection against malaria. Nonetheless, while it is unlikely that improved zinc status could prevent infection with malarial parasites, it does appear that zinc may reduce the severity of the infection or the symptoms of morbidity associated with the infection. Evidence for this is suggested by a reduction in the number of visits to health facilities due to malaria, but not in the number of cases of malaria as determined during daily surveillance at the child's home, when children in malaria endemic areas were provided with supplemental zinc.

Mortality Given the contribution of zinc deficiency to three of the most common causes of death among children in developing countries (i.e., diarrhea, pneumonia, and malaria) it can be expected that zinc deficiency also contributes substantially to childhood mortality among these populations. Although still limited, available information does suggest that supplemental zinc leads to sizeable reductions in mortality among vulnerable groups of children. In Bangladesh, evaluation of a program that provided supplemental zinc for 14 days as treatment for diarrhea demonstrated a 68% reduction in mortality among infants and preschool children. Mortality was also reduced by two-thirds following supplemental zinc among low-birth-weight infants in India. A nearly 60% reduction in child mortality was observed among children in Burkina Faso, although this was not statistically significant. Further large-scale studies are required to better quantify the impact of zinc on child mortality.

Pregnancy: Maternal, Fetal and Infant Health

Few firm conclusions can be made as to the consequences of zinc deficiency during pregnancy on

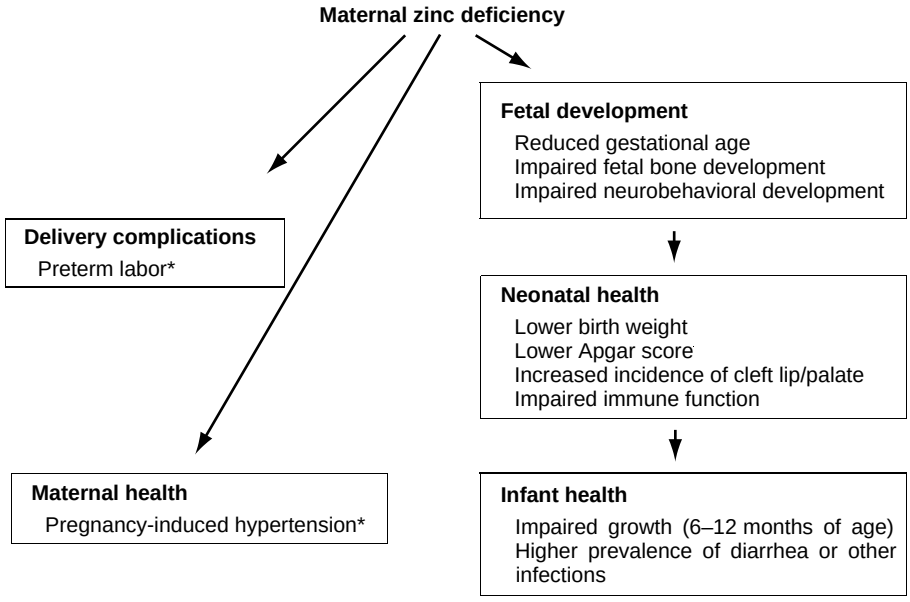


Figure 1 Several consequences of maternal zinc deficiency during pregnancy on maternal health, fetal development, and infant health have been observed in developing and more industrialized countries. These consequences have been confirmed by randomized, placebo-controlled trials of maternal zinc supplementation. Not all of the consequences have been observed in all studies, and the reasons for inconsistent results among studies are not well understood. *Determined from studies in industrialized countries only.

maternal, fetal, and infant health. Results from zinc supplementation trials have been inconsistent and therefore difficult to interpret. This may be partly attributed to inadequate study design or failure to consider the zinc status of the women studied. Most earlier studies focused on the evaluation of gestational age and birth weight as primary outcomes. However, zinc deficiency may also manifest itself in more specific qualities of health and development of the fetus and infant, as summarized in **Figure 1**. While there is some evidence from industrialized countries that zinc deficiency contributes to complications during pregnancy, delivery, and postpartum, these outcomes have not been adequately studied in developing countries.

Control of Zinc Deficiency in Developing Countries

Efforts to control zinc deficiency in national programs are only just being initiated. The following information describes the current state of development of zinc nutrition programs, and some direction for the future.

Zinc-Containing Pharmacological Supplements

As exemplified by the results of controlled trials, zinc supplementation is an efficacious (i.e., effective under controlled study conditions) strategy to prevent zinc deficiency. However, the effectiveness of this strategy under realistic conditions will depend on the success of in-country programs to distribute zinc supplements to vulnerable populations and on their use by the intended recipients. At present, few such zinc supplementation programs are in place. It may be more feasible to add zinc to iron supplements, use of which is widely advocated for the prevention of iron deficiency anemia in young children and women of childbearing age. However, some evidence indicates that when these two minerals are combined, their ability to improve either zinc or iron status diminishes. The competitive interaction between iron and zinc at the level of intestinal absorption or post-absorption may explain this observation. Research is required to determine optimal supplementation schemes for the prevention of iron and zinc deficiencies simultaneously.

Given the recent recommendation for the use of zinc in the management of acute diarrhea and the resultant reduction of childhood mortality observed in one study to date, it is expected that diarrheal treatment programs including supplemental zinc will ensue.

Enrichment (Fortification) of Foods with Zinc

A few countries from developing regions have implemented a policy for the fortification of staple foods with zinc. Mexico established a program whereby wheat and corn (maize) flour producers could

voluntarily add zinc to their products (20 mg/kg flour). Indonesia has also implemented a national program for the fortification of wheat flour, which includes addition of zinc. The fortification of condiments, such as fish sauce or seasoning powders in Asia, may serve as an additional vehicle for zinc fortification in the future. Several countries are adding zinc (and other micronutrients) to foods that are distributed in programs targeted to specific, vulnerable population groups. For example, in Chile and Argentina milk powder for use by young children is fortified with zinc, while in Mexico a milk powder-based supplement with added zinc is directed towards young children as well as pregnant and lactating women. As yet, there is an absence of information on the effectiveness of these programs to improve population zinc status.

Modification of Foods and Diets

Several strategies apart from the use of pharmacological supplements and food fortification have been suggested for the improvement of dietary zinc intake in developing country settings. Cereal crops, such as wheat, corn, and rice, are being bred to contain higher concentrations of zinc in the grain portion. Cereals that have a reduced content of phytate have also been produced but still require further testing of their agricultural viability and effect on improving zinc status when used in the context of a usual diet. Promotion of the production and use of zinc-rich foods through community level education and provision of starter materials could also be used. The efficacy of most of these alternative strategies to improve population zinc status has not yet been tested.

See also: **Anemia**: Iron-Deficiency Anemia. **Diarrheal Diseases**. **Growth and Development, Physiological Aspects**. **Pregnancy**: Nutrient Requirements. **Supplementation**: Dietary Supplements. **Vitamin A**: Biochemistry and Physiological Role; Deficiency and Interventions. **Zinc**: Physiology.

Further Reading

Black RE (2003) Zinc deficiency, infectious disease and mortality in the developing world. *Journal of Nutrition* **133**: 1485S–1489S.

Brown KH, Peerson JM, Rivera J, and Allen LH (2002) Effect of supplemental zinc on the growth and serum zinc concentrations of prepubertal children: a meta-analysis of randomized controlled trials. *American Journal of Clinical Nutrition* **75**: 1062–1071.

Brown KH and Wuehler SE (2000) *Zinc and Human Health: Results of Recent Trials and Implications for Program Interventions and Research*. Ottawa: The Micronutrient Initiative/ International Development Research Center.

Brown KH, Wuehler SE, and Peerson JM (2001) The importance of zinc in human nutrition and estimation of the global prevalence of zinc deficiency. *Food and Nutrition Bulletin* **22**: 113–125.

Gibson RS (1994) Zinc nutrition in developing countries. *Nutrition Research Reviews* **7**: 151–173.

Gibson RS and Ferguson EL (1998) Nutrition intervention strategies to combat zinc deficiency in developing countries. *Nutrition Research Reviews* **11**: 115–131.

International Zinc Nutrition Consultative Group (Brown KH, Rivera JA, Bhutta Z, Gibson RS, King JC, Ruel M, Sandström B, Wasantwisut E, Hotz C, Lönnerdal B, Lopez de Romaña D, and Peerson J) (2004) Assessment of the risk of zinc deficiency in populations and options for its control. *Food and Nutrition Bulletin* **25**: S91–S202.

Osendarp SJM, West CE, and Black RE on behalf of the Maternal Zinc Supplementation Study Group (2003) The need for maternal zinc supplementation in developing countries: an unresolved issue. *Journal of Nutrition* **133**: 817S–827S.

Wood RJ (2000) Assessment of marginal zinc status in humans. *Journal of Nutrition* **130**: 1350S–1354S.

Zinc Investigators' Collaborative Group (Bhutta ZA, Black RE, Brown KH, Meeks Gardner J, Gore S, Hidayat A, Khatun F, Martorell R, Ninh NX, Penny ME, Rosado JL, Roy SK, Ruel M, Sazawal S, and Shankar A) (1999) Prevention of diarrhea and pneumonia by zinc supplementation in children in developing countries: Pooled analysis of randomized controlled trials. *Journal of Pediatrics* **135**: 689–697.